

Constructed Gastric Pouch Size and Weight Loss After
Roux-en-Y Gastric Bypass:
A Systematic Review and Meta-Analysis

Hjalti

Thordur Bjarnason

Hannes Jon Larusson

September 1, 2026

Abstract

Roux-en-Y gastric bypass is one of the most studied operations in medicine, and the gastric pouch is its central constructed artifact: a stomach deliberately rebuilt to a chosen size. That choice is made hundreds of thousands of times a year, yet it rests on habit rather than evidence — in a 2026 international survey barely half of surgeons (53.5%) measured the pouch in any way, with a median target length of 6 cm — no ASMBS or IFSO guideline prescribes any size, and the literature has disagreed for forty years about whether the choice matters for the operation’s primary goal, durable weight loss. The largest cohort ever assembled (14,168 patients, from a Scandinavian registry) found that pouch size stopped predicting weight loss six weeks after surgery. The randomized trials disagree with that registry and with each other. The single systematic review could not pool its 14 studies. No meta-analysis exists.

This dissertation answers the question quantitatively for the first time, inside a full scholarly account of its subject: how obesity became a surgical disease and how the incretin-drug era reframes what surgery is for; how bariatric technique evolved by convention rather than trial; what the pouch is anatomically and physiologically; how it is measured, and how badly; and how the field’s own outcome metrics obscure its results. Against that foundation it conducts a PRISMA systematic review (526 candidate records cataloged) and attempts the missing meta-analysis: random-effects pooling of study-defined small-versus-large contrasts on percent total weight loss, with dose-response and pre-specified subgroup analyses. If heterogeneity defeats pooling, the dissertation demonstrates rigorously why — which is itself the finding.

Its signature contribution is the measurement-artifact hypothesis. Pouch size reaches the literature through four incompatible modalities — construction proxies (stapler loads, bougie size), contrast radiography, CT volumetry, and swallow MRI — with no reference standard and validation resting on handfuls of patients. If the pouch-size effect appears only under weak proxies and vanishes under direct volumetry, the suspect is not noise — measurement error attenuates; it cannot manufacture an effect — but whatever travels with the proxy: confounding by surgeon, era, and technique, or selective reporting. If the effect survives measurement quality, surgical habit needs revisiting. Either outcome is new knowledge. A measurement-fidelity simulation completes the argument: calibrated to the documented error of the four modalities and the realized study designs alone, it derives the detectability floor of the pouch-size question — what effect sizes the field’s instruments could ever have detected — and the measurement precision a decisive future study would require. Modern physiology — the pouch as signaling organ rather than container — frames the test, a safety analysis in both directions bounds it, and a one-anastomosis-bypass subgroup widens the anatomical variance the analysis can exploit.

Contents

I	The Field	6
1	Introduction	7
1.1	An operation built around a number	7
1.2	The doctrine: born with the operation	8
1.3	Physiology moved; the doctrine did not	8
1.4	The forty-year stalemate	9
1.5	The measurement-artifact hypothesis	10
1.6	Aim and research questions	11
1.7	Original contributions	12
1.8	How this dissertation is organized	13
2	Obesity and its treatments: the stakes	14
2.1	The scale of the problem	14
2.2	The physiology of body-weight defense	15
2.2.1	The energy-homeostasis system	15
2.2.2	What happens when weight falls	16
2.3	The treatment ladder	17
2.3.1	Lifestyle intervention	17
2.3.2	Pharmacotherapy before the incretins	18
2.3.3	Endoscopic therapy	18
2.3.4	Surgery's position on the ladder	18
2.4	The incretin era	19
2.4.1	Surgery-adjacent weight loss from a syringe	19
2.4.2	Chronic therapy for a chronic disease	20
2.4.3	Pharmacology borrowed from surgery	20
2.4.4	Why the drugs sharpen the dissertation's question	21
2.5	Where surgery still wins, and what it costs	21
2.5.1	Durability	21
2.5.2	Diabetes remission	22
2.5.3	Mortality	22
2.5.4	The price	23
3	Bariatric surgery: how technique congealed	24
3.1	Reading guide: the vocabulary of reconstructed anatomy	24
3.2	The malabsorptive prologue: jejunoileal bypass	25
3.3	Mason's loop and the birth of the small pouch	26
3.4	Gastroplasty and the vertical banded decade	28
3.5	The laparoscopic revolution: standardization by stapler	29
3.6	Sleeve gastrectomy: a staging maneuver becomes the default	30
3.7	The loop returns: one-anastomosis gastric bypass	32
3.8	Consolidation, and the state of the convention	32

3.9	The genealogy of a number	34
4	The gastric pouch: anatomy, construction, physiology	36
4.1	Surgical anatomy of the proximal stomach	36
4.1.1	Blood supply and the tension between vascularity and size	37
4.1.2	Innervation	37
4.1.3	The mucosa that ends up in the pouch	37
4.2	Construction, step by step	38
4.3	What varies between surgeons, and what the instruments constrain	42
4.4	Emptying dynamics, and the steelman for size dependence	44
4.4.1	What the pouch actually does to nutrient flow	44
4.4.2	Two models, both predicting size dependence, with opposite signs	45
4.4.3	What the emptying data actually show	46
4.4.4	Where the physics is being abused	46
4.4.5	The mechanistic expectations entering Part II	48
4.5	Distension, compliance, and what “size” names	48
4.6	Ghrelin territory	49
4.7	Complications, with anatomical logic	50
4.7.1	Marginal ulcer	50
4.7.2	Stricture	52
4.7.3	Dumping	52
4.8	What the pouch is	53
5	Mechanisms of weight loss after RYGB	54
5.1	The legacy model: restriction and malabsorption	56
5.1.1	The model as engineered design	56
5.1.2	Malabsorption, measured	56
5.1.3	Restriction, measured	57
5.2	The neuroendocrine model	58
5.2.1	Ghrelin: the excluded fundus, and a complicated thread	58
5.2.2	GLP-1 and PYY: the hindgut surge	59
5.2.3	Bile acids, FXR, and TGR5: a contested thread	60
5.2.4	The gut microbiota	61
5.2.5	Vagal and central rewiring: appetite, reward, and food preference	62
5.2.6	Energy expenditure: clear in rodents, mixed in humans	62
5.2.7	The threads assembled	63
5.3	What each mechanism predicts about pouch size	63
5.3.1	The restrictive prediction: volume as dose	63
5.3.2	The pure neuroendocrine prediction: volume as packaging	63
5.3.3	The emptying-dynamics steelman: geometry as filter	65
5.3.4	The safety gradients: harms in both directions	66
5.4	From mechanism to hypotheses	66
6	Measuring the pouch	68
6.1	What quantity is being measured?	68
6.2	Construction proxies: bougie and stapler	71
6.3	Contrast radiography	72
6.4	Computed tomography volumetry	73
6.5	Swallow magnetic resonance imaging	74
6.6	Endoscopy and functional measurement	74
6.7	Dilation over time: the moving target	74
6.8	What the measurement problem does to the evidence	76

6.9	The modality-quality hierarchy	76
7	Measuring the outcome	80
7.1	The metric family	80
7.1.1	One signal, many normalizations	80
7.1.2	The %EWL problem: a reference correlated with the population	81
7.1.3	%TWL’s rise, and what the older metrics still contaminate	82
7.1.4	Dichotomous “success” and its ceiling	82
7.2	Timepoints: nadir versus fixed	83
7.3	Weight regain: a definition for every study	84
7.4	The ASMBS 2015 standard and its uptake	84
7.5	Attrition and the survivorship problem	85
7.5.1	Magnitude	85
7.5.2	Non-randomness	85
7.5.3	Why attrition biases anatomical correlations specifically	86
7.6	Harmonization rules for the meta-analysis	86
II	The Study	89
8	Methods	90
8.1	Study design and protocol	90
8.1.1	Pre-specification and the preliminary scan	91
8.2	Eligibility criteria	91
8.2.1	Population	91
8.2.2	Exposure and comparator	91
8.2.3	Outcomes	92
8.2.4	Study designs	92
8.3	Information sources and search strategy	93
8.3.1	Electronic searches	93
8.3.2	Search validation	93
8.3.3	Other sources	94
8.4	Study selection	94
8.5	Data collection: LLM-assisted, span-verified extraction	95
8.5.1	The span-verification rule	95
8.5.2	Well-formedness gating	95
8.5.3	Model selection and adversarial checking	95
8.5.4	Human review	96
8.5.5	Data items	96
8.6	Exposure classification: the measurement-modality hierarchy	96
8.7	Outcome harmonization	97
8.8	Risk of bias and certainty of evidence	97
8.9	Statistical synthesis	98
8.9.1	Primary analysis	98
8.9.2	Heterogeneity, subgroups, and dose-response	99
8.9.3	Defensibility criteria and the SWiM fallback	99
8.10	Software, provenance, and reproducibility	100
8.11	Ethics	101
8.12	Protocol deviations	101

9	The evidence base: corpus and study characteristics	102
9.1	From search to catalog	102
9.2	Screening and the PRISMA flow	103
9.3	The corpus in time	103
9.4	The corpus in print and in place	104
9.5	A census of record types	106
9.6	Anchor studies: the fixed points visible before screening	107
9.7	The anatomy of heterogeneity, pre-stated	109
10	Primary analysis: does pouch size predict weight loss?	110
10.1	The estimand	110
10.2	The model, and how to read the forest plot	112
10.3	Defensibility criteria for pooling, stated in advance	113
10.4	The pooled estimate at one year	114
10.5	Weight loss at three years and beyond	115
10.6	Dose-response: gradient or threshold? (descriptive)	116
10.7	Heterogeneity, quantified and dissected	117
10.8	If pooling fails: synthesis without meta-analysis	117
10.9	What carries forward	118
11	The measurement-artifact test and the evidence conflict	119
11.1	The modality-stratification design	119
11.1.1	What measurement error can and cannot manufacture	120
11.2	A framework for dissecting evidence conflicts	122
11.3	The registry versus the trials	122
11.4	Trial versus trial	124
11.5	Adjudication, pre-committed	125
12	The detectability floor: a measurement-fidelity simulation	127
12.1	Motivation: a question the existing data can already answer	127
12.2	Error models	127
12.3	Design	128
12.4	Results	130
12.5	What a simulation can and cannot show	130
13	Boundaries: safety trade-off and the OAGB contrast	134
13.1	The safety counterweight: marginal ulceration, and the harms at the other end	134
13.1.1	The harm, anatomically	134
13.1.2	The registry signal, verified	135
13.1.3	Corroboration, and the limits of it	135
13.1.4	The safety trade-off, in both directions	136
13.2	The OAGB contrast	138
13.2.1	A longer lever arm	138
13.2.2	What the contrast can and cannot show	139
13.2.3	The safety gradient, reversed	139
13.3	Risk of bias and certainty of evidence	140
13.4	Exploratory corpus-mined patterns	140

III	Synthesis	142
14	Discussion	143
14.1	Interpretation committed before data	143
14.2	Lens I: no effect under sound measurement	144
14.3	Lens II: the effect survives measurement quality	145
14.4	Lens III: undecidable — and exactly what would decide it	146
14.5	This review against the existing synthesis	147
14.6	Scoring the mechanistic predictions	147
14.7	The incretin-era stakes, revisited	149
14.8	Clinical implications, drawn conservatively	150
14.9	Limitations	150
15	Conclusions and future directions	152
15.1	The seven questions, answered	152
15.1.1	Q1 — Does a smaller constructed pouch produce greater weight loss? . .	152
15.1.2	Q2 — Is there a gradient, or is any effect threshold-like?	153
15.1.3	Q3 — Does the answer depend on how pouch size was measured?	153
15.1.4	Q4 — Why do the randomized trials disagree with the largest registry, and with each other?	153
15.1.5	Q5 — How does construction size trade off against harm in both directions? .	153
15.1.6	Q6 — Does the relationship differ in one-anastomosis gastric bypass? . .	154
15.1.7	Q7 — Which model’s predictions do the findings match?	154
15.2	Original contributions	154
15.3	The registry recommendation	155
15.4	The definitive trial	155
15.5	Closing	156
A	Search strings and hit counts	172
B	Extraction schema and gold cases	175
C	Per-study evidence tables	176
D	Excluded studies	177
E	Figure provenance	178
F	Pipeline documentation	180

Part I

The Field

Chapter 1

Introduction

1.1 An operation built around a number

Obesity is now the dominant form of malnutrition in most of the world. In the largest pooled analysis of population-representative anthropometry, covering 222 million participants in 200 countries, the combined prevalence of underweight and obesity rose between 1990 and 2022 in 162 countries for women and 140 for men, with the increase driven almost everywhere by obesity rather than undernutrition [1]. Against that burden, metabolic and bariatric surgery — surgery that reduces body weight by reconstructing the gastrointestinal tract — remains the most effective durable treatment, and it is performed at industrial scale: the international federation of bariatric societies (IFSO) documented 598,834 surgical and endoluminal procedures worldwide in 2021 alone, a figure drawn from only 57 of its 74 national societies and therefore an undercount [2].

The Roux-en-Y gastric bypass (RYGB) is the reference procedure of this field — the operation against which alternatives are benchmarked, and together with sleeve gastrectomy one of the two most commonly performed [2]. Its anatomy can be stated in engineering terms. The stomach is divided so that a small chamber at its top — the *gastric pouch*, typically a few centimeters long and a few tens of milliliters in capacity — remains in the food stream, while the remainder of the stomach — the great majority of the organ, in a proportion that varies with patient anatomy and construction — is sealed off and left in place, excluded. The small intestine is then divided and rearranged into a Y-shaped circuit (the eponymous configuration described by César Roux): one limb is brought up and joined to the pouch at a new outlet called the *gastrojejunostomy* or *stoma*, so that food passes from the esophagus into the pouch, through the stoma, and directly into the jejunum, bypassing the excluded stomach and duodenum entirely.

Every element of that circuit is a design parameter chosen by the surgeon: pouch size, stoma diameter, and the lengths of the intestinal limbs. This dissertation is about the first of these. The pouch is the operation’s central constructed artifact — a stomach deliberately rebuilt to a chosen size — and the choice of that size is made hundreds of thousands of times a year. The founding rationale for making it small was mechanical: a small reservoir restricts how much can be eaten, the bypass reduces how much is absorbed, and weight loss follows from the arithmetic of intake and absorption. Under that model, pouch size is the operation’s primary control knob, and smaller is better. The premise of this dissertation is that this model has quietly collapsed while the parameter it justified lives on, that the empirical literature on whether the parameter matters has disagreed for forty years, and that no pooled quantitative synthesis of that literature has been identified. The disagreement itself, this dissertation will argue, may be a measurement problem as much as a biological one.

1.2 The doctrine: born with the operation

The pouch-size doctrine is as old as the operation. Mason and Ito reported the gastric bypass for obesity in 1967 [3], and by 1975 Mason’s group, reviewing 442 patients over nine years, had already codified the recipe: the best weight loss required three components — “Bypass of stomach and duodenum,” “a small fundic segment,” and “a small (12 mm diameter) gastroenterostomy stoma” — and a stoma found too large at reoperation could be revised smaller to induce further loss [4]. The same report logged the doctrine’s shadow cost — stomal ulceration, the mucosal ulcer that forms at the new gastrointestinal junction, at that time one per 140 patient-years [4]. By 1982 the doctrine had a comparative test: Kirkpatrick and Siegel, noting that “[m]ost investigators have emphasized that both a small reservoir and a narrow stoma are critical,” prospectively varied the stoma while holding the reservoir at 30–45 mL and concluded that stoma size mattered little *provided* a tiny reservoir was consistently created — the reservoir, not the outlet, was the critical determinant of success [5]. The claim that this dissertation examines was thus stated within fifteen years of the operation’s invention, and it was stated as settled.

What followed was not a trial but a propagation. The small pouch hardened into training, instruments, and habit. When the American Society for Bariatric Surgery surveyed its practicing surgeons in 2008, the average reported pouch was 25 cm³, and only about half of respondents (49%) measured pouch size even indirectly, by distance from the gastroesophageal junction [6]. Nearly two decades later the picture is unchanged in kind: in a 2026 worldwide survey of 245 IFSO surgeons across 47 countries, only 53.5% routinely measured pouch length at all; those who did reported a median of 6 cm (interquartile range 5–7 cm); 67.3% characterized their pouch only by the subjective label “short or small”; and the calibrating bougie — the rubber sizing tube passed down the esophagus as a template for the staple line — ranged from 28 to 46 French between practices [7]. The survey’s own conclusion is that pouch descriptors “lack consistent anatomical definitions” and that the field cannot yet say which of its technical variations matter [7].

No guideline resolves this. The current joint statement of the American and international societies (ASMBS/IFSO, 2022) — the document that governs modern practice — specifies *who* should be offered metabolic surgery and contains no recommendation on how large the pouch should be built [8]; nor is any such recommendation identified in the 2026 IFSO survey just cited, whose conclusion is instead that the field lacks the consistent anatomical definitions a recommendation would presuppose [7]. Pouch size is therefore a parameter set hundreds of thousands of times a year by convention: by where a surgeon trained, which stapler loads the hospital stocks, and a number written into a textbook before the question was ever measured. Chapter 3 traces this congealing of technique in full; the point here is only that the number’s provenance is authority, not evidence.

1.3 Physiology moved; the doctrine did not

While the doctrine propagated, the mechanistic understanding of the operation was rebuilt underneath it. Three decades of physiological work — reviewed in depth in Chapter 5 — relocated the operation’s effect from the staple line to the gut–brain axis. Direct measurement dismantled both halves of the founding model: RYGB and sleeve gastrectomy “engender weight loss and metabolic improvements by mechanisms other than restriction and malabsorption,” in the summary of Pucci and Batterham’s mechanistic review — reduced hunger and altered food preference driven by an orchestra of gut-derived signals: the hunger hormone ghrelin (produced largely in the excluded stomach), the satiety hormones GLP-1 and PYY released when nutrients reach the lower gut, bile acids, and the gut microbiota [9]. In this model the pouch is not a container whose capacity limits intake; it is a low-impedance antechamber. Scintigraphic studies

show that after RYGB the pouch empties *faster* than an intact stomach — “the bulk of foods pass without hindrance into the small intestine” — and that faster pouch emptying is associated with stronger GLP-1 and PYY responses [10]. On this evidence the organ built to restrict flow behaves less like a reservoir than like a nozzle — an analogy, not a derived mechanism — whose geometry may shape how quickly nutrients arrive at the hormone-secreting jejunum.

The incretin-drug era sharpens the stakes of this reframing rather than retiring them. GLP-1 receptor agonism with semaglutide produces a mean 14.9% body-weight reduction at 68 weeks [11]; dual GIP/GLP-1 agonism with tirzepatide reaches 20.9% at 72 weeks at the highest dose [12] — surgery-adjacent numbers achieved by pharmacologically mimicking the hormonal consequences of the operation. If drugs can replicate the operation’s endocrine output, then what remains distinctively surgical is exactly the mechanical configuration — the constructed anatomy that no molecule can copy — and the effect size of each mechanical parameter becomes more worth knowing, not less (Chapter 2 develops this argument).

The mechanistic reframing does not, however, simply predict that pouch size is irrelevant — and this is the steelman the analysis must respect. Both models predict *some* pouch-size dependence, with opposite signs of optimism about small pouches. Under the legacy restrictive model, smaller is better, monotonically. Under emptying-dynamics logic, what matters is the pouch’s geometry as a metering device: a longer, narrower pouch might dilate less and empty differently than a short wide one of equal volume, making the effect shape-dependent rather than volume-dependent — the explicit rationale, grounded in the physics of wall tension and flow (Laplace and Poiseuille), offered by the one trial group that randomized patients to a *longer* pouch and reported better maintenance of weight loss [13]. Chapter 4 builds this mechanistic bridge in anatomical detail; Chapter 5 derives from it the competing predictions that the empirical analysis will adjudicate.

1.4 The forty-year stalemate

The empirical literature that should decide between these models is instead a stalemate, and its flagship results contradict each other at every level of the evidence hierarchy. Table 1.1 states the conflict in the studies’ own numbers.

The largest evidence mass points to no effect. The Scandinavian registry cohort of 14,168 gastric bypasses — the largest study ever conducted on this question, exploiting Sweden’s highly standardized technique — used the total length of stapler cartridge fired to construct the pouch as a size proxy, and found that it predicted excess-BMI loss at six weeks but *not* at one year [14]. The same cohort quantified the safety counterweight: each additional centimeter of staple line raised the relative risk of marginal ulcer — the ulcer at the gastrojejunal junction that is modern surgery’s descendant of Mason’s stomal ulcer — by 14% (95% CI 9–20%) [14]. On this evidence a smaller pouch is safer but no more effective, and size stops predicting weight loss six weeks after the operation.

The comparative literature refuses to agree with the registry or with itself. The only systematic review of the question, by Mahawar and colleagues in 2020, found fourteen studies of pouch size — two of them randomized — of which nine reported no significant association with weight loss and five found larger pouches worse; no study found larger pouches better; and the authors judged the quality and heterogeneity of the data such that a meta-analysis could not be performed [15]. No pooled estimate of the pouch-size effect has been identified anywhere in the literature: forty years of data, and the field’s central technical question has never been answered with a number. The randomized evidence, meanwhile, is thin, and its two trials test different exposure constructs in different directions. The Dutch trial that randomized 132 patients to an extended 10-cm pouch versus a standard 5-cm pouch — a *length* contrast — found *better* three-year total weight loss with the longer pouch (31% versus 27%, $p = 0.023$), driven by less weight regain [13], though at five-to-nine-year follow-up — with only 47% of the cohort reachable —

the difference was no longer significant and the authors declared the comparison underpowered [16]. The other randomized study, of 76 Chinese patients with type 2 diabetes and BMI 30–35, contrasted estimated pouch *volumes* (10–20 versus 25–35 mL) and found greater excess-weight loss with the smaller pouch at twelve months (68.9% versus 59.1%, $p = 0.033$) [17]. In the same direction, a six-year retrospective cohort of 95 Chinese patients with type 2 diabetes and BMI below 35 found the *small*-pouch group achieved greater BMI reduction and more diabetes remission, with more marginal ulcers in the large-pouch group [18]. A registry that says size stops mattering at six weeks; a randomized length trial in which the longer pouch wins; a randomized volume trial and a cohort in which the smaller pouch wins; a review that could not pool any of it: that is the state of the evidence on the most frequently chosen parameter of the world’s benchmark metabolic operation.

Table 1.1: The headline conflict: flagship studies of constructed pouch size and weight loss after RYGB. All values as reported by the cited studies. %TWL, percent total weight loss; %EBMIL, percent excess-BMI loss; RR, relative risk; CI, confidence interval.

Study	Design	N	Exposure definition	Finding
Edholm 2016 [14]	Registry cohort (SOREg)	14,168	Stapler length fired (proxy; mean 145 mm)	Predicted %EBMIL at 6 weeks, not at 1 year; marginal-ulcer RR +14% per cm (95% CI 9–20%)
Mahawar 2020 [15]	Systematic review	14 studies	Heterogeneous	9 studies null, 5 favor smaller, 0 favor larger; pooling precluded; no meta-analysis exists
Boerboom 2020 [13]	RCT	132	Pouch length: 10 vs. 5 cm	Extended (longer) pouch: %TWL 31% vs. 27% at 3 years ($p = 0.023$), less weight regain
Ren 2015 [17]	RCT	76	Estimated volume: 10–20 vs. 25–35 mL	Smaller pouch: %EWL 68.9% vs. 59.1% at 12 months ($p = 0.033$)
Harker 2026 [16]	RCT follow-up	62	Pouch length: 10 vs. 5 cm	%TWL 26.1% vs. 24.1% at 5–9 years, not significant; 47% follow-up, underpowered
Gao 2022 [18]	Retrospective cohort	95	Intraoperative small vs. large	Small pouch: greater BMI reduction and diabetes remission at 6 years; large pouch: more marginal ulcers

1.5 The measurement-artifact hypothesis

This dissertation’s organizing hypothesis is that the stalemate is, in substantial part, a measurement problem — that the field has spent forty years regressing an outcome on an exposure it has never reliably measured. Pouch size reaches the literature through four mutually incompatible modalities: construction proxies recorded in the operating room (stapler cartridges fired, bougie caliber, length in centimeters along a curved and compliant organ); upper-gastrointestinal contrast radiography, a two-dimensional shadow of a three-dimensional chamber; computed-tomography volumetry, the nearest thing to a reference standard — though any postoperative image measures the pouch as it exists at the time of imaging, not as it was constructed; and swal-

low magnetic-resonance imaging. No reference standard is agreed, and the validation literature connecting these modalities is strikingly thin — the agreement between swallow MRI and 3D-CT volumetry, for instance, rests on a series of twelve patients [19]. The 2026 global survey draws the same conclusion from the practice side: because pouch size is reported through “heterogeneous proxies, including subjective descriptors, stapler-based estimates, anatomic landmarks, and non-uniform volumetric measures,” the literature’s associations “remain hard to compare and translate” [7].

In statistical terms this is a classical errors-in-variables problem. Nondifferential noise in an exposure attenuates its estimated association with any outcome — regression dilution — so a true effect measured through a poor proxy can shrink to null; conversely, a proxy correlated with other features of surgical technique can manufacture an association where none exists. A staple line’s length is not a volume; a bougie’s caliber is not a pouch; and a literature in which most studies measure the exposure worst is a literature in which both null and positive findings are individually uninformative. The remedy is not another narrative review but a stratified quantitative test: order the modalities by measurement quality *before* seeing the data — the hierarchy is defined in Chapter 6 — and ask whether the pouch-size effect depends on how the pouch was measured. The logic of that test has to be stated carefully, because classical nondifferential measurement error can only attenuate a true association; it cannot create one. An effect that survives under direct volumetry despite the noisier strata below it is evidence that surgical habit needs revisiting and the mechanistic model needs an amendment. An association that lives *only* in the weak proxies, by contrast, cannot be dismissed as noise: it points instead toward proxy-correlated confounding — a stapler count or bougie caliber tracking other features of surgical technique — or toward selective reporting in the proxy literature. And because measurement modality is entangled with era, study design, and study size, the stratified comparison is an exploratory effect-modification analysis that weighs competing explanations; it cannot by itself deliver a verdict (Chapter 11 states the competing explanations explicitly). Either pattern is new knowledge about where the forty-year disagreement comes from. A final scope decision follows from the same measurement logic: pouches dilate substantially over the years after surgery, so any late assessment conflates the constructed size with its subsequent remodeling; this dissertation therefore studies *constructed* size only — postoperative dilation, a downstream and time-varying process that is part mediator of construction and part consequence of the weight trajectory, is excluded as an exposure and discussed as context (Chapter 6).

1.6 Aim and research questions

The aim of this dissertation is to answer, quantitatively and for the first time by meta-analysis, whether the constructed size of the gastric pouch in Roux-en-Y gastric bypass relates to weight loss — and to do so inside a full scholarly account of where the pouch-size doctrine came from, what the pouch actually is and does, and how both the exposure and the outcome are measured. The vehicle is a systematic review conducted per Cochrane Handbook guidance and reported to PRISMA 2020 standards, over a corpus of 526 candidate records already cataloged, with random-effects pooling of the mean difference in percent total weight loss at one year across study-defined small-versus-large construction contrasts, a separate pre-specified slope estimand for continuous-exposure studies, pre-specified subgroups, and an exploratory dose-response analysis; the full protocol is given in Chapter 8. Because the evidence base mixes randomized contrasts with confounded observational comparisons, the pooled quantity is an association; causal interpretation is reserved for the randomized evidence, which is weighed separately. Seven questions, fixed in the project’s governing specification before screening began — and refined once, still before any screening or extraction, after an adversarial methods review whose sequence the repository’s version history documents — structure the work:

1. **Primary.** In adults undergoing primary RYGB, does a smaller constructed pouch produce

greater weight loss (percent total weight loss at 1 year and at ≥ 3 years)?

2. **Dose-response.** Is there a gradient, or is any effect threshold-like?
3. **Measurement.** Does the answer depend on how pouch size was measured (imaged volume vs. length vs. stapler/bougie proxies)? If the association only appears under weak proxies, the field’s 40-year disagreement traces to what travels with those proxies — proxy-correlated confounding or selective reporting — rather than to the anatomy; weighed as exploratory effect modification.
4. **Evidence-level conflict.** Why do the randomized trials disagree with the largest registry, and with each other?
5. **Safety trade-off.** How does construction size trade off against harm in *both* directions: the marginal-ulcer gradient that penalizes larger construction (more staple line, more ulcers), and the harms plausibly attached to very small construction (stricture, anastomotic tension, leak, food intolerance)?
6. **OAGB subgroup.** Does the relationship differ in one-anastomosis gastric bypass — the single-connection variant of the operation in which the constructed pouch is deliberately long and varies far more widely?
7. **Mechanistic framing.** Which model’s predictions do the findings match? The legacy restrictive model predicts a monotonic benefit of smaller construction. The neuroendocrine model does *not* predict a strict null — emptying dynamics permit pouch effects — but it predicts that any effect tracks emptying-relevant geometry rather than capacity per se. The two models thus make distinct, testable predictions, not a simple effect-versus-no-effect contrast.

1.7 Original contributions

Three intended contributions of this dissertation address gaps for which no published counterpart has been identified — a claim held as “none identified,” pending a reproducible search maintained through submission, not as a guarantee:

1. **The first pooled estimate.** No meta-analysis of constructed pouch size against weight loss has been identified; the single systematic review explicitly could not pool its fourteen studies [15]. This dissertation attempts that pooling under a pre-specified, versioned protocol — and if heterogeneity defeats it under the pre-stated defensibility criteria, delivers instead a rigorous, study-by-study demonstration of *why* pooling fails, which is itself a finding the field does not have.
2. **The measurement-modality test.** No identified study has stratified the pouch-size effect by measurement modality under a quality hierarchy fixed before analysis. Because modality is entangled with era, design, and study size, this is an exploratory effect-modification analysis: it can shift the weight among competing explanations of the field’s central contradiction, not adjudicate it outright — and it may prove non-estimable at the achievable study counts, an outcome that would itself be reported.
3. **The dose-response analysis.** On the subset of studies with volumetric exposure data, the dissertation explores — within the ecological limits of aggregate data and a small expected study count — whether any effect is graded or threshold-like — the question that would tell a surgeon not just *whether* size matters but *which* sizes differ.

Beyond these, the structured per-study evidence tables of Chapter 9 are intended to stand alone as a reference resource on this literature, and any patterns mined from the assembled corpus beyond the pre-specified analyses are reported strictly as labeled exploratory, hypothesis-generating findings.

1.8 How this dissertation is organized

The monograph is built in three parts, and the order is the argument.

Part I (The Field, Chapters 2–7) assembles everything the analysis needs before a single datum is pooled. Chapter 2 establishes the stakes: obesity as a disease of energy regulation, the treatment ladder, and the incretin-drug era that makes surgery’s mechanical parameters more, not less, worth quantifying. Chapter 3 gives the history of bariatric technique as a history of decisions made without trials — where the pouch-size number came from. Chapter 4 treats the pouch itself: its surgical anatomy, its construction step by step, and the physiology by which its size *could* matter under either mechanistic model. Chapter 5 lays out the full mechanistic account of weight loss after RYGB and derives from it the predictions the results must adjudicate. Chapters 6 and 7 are the measurement-science core: Chapter 6 dissects each pouch-measurement modality and fixes the quality hierarchy used later; Chapter 7 does the same for the outcome side — the competing weight-loss metrics, regain definitions, and attrition patterns — and derives the harmonization rules the meta-analysis applies.

Part II (The Study, Chapters 8–13) is the systematic review and meta-analysis. Chapter 8 gives the methods per the versioned protocol. Chapter 9 maps the evidence base from 526 cataloged records to the included set, with structured evidence profiles of every included study. Chapter 10 delivers the primary pooled analysis and dose-response (Questions 1–2). Chapter 11 is the dissertation’s pivot: the measurement-modality analysis and the design-level dissection of the field’s flagship contradictions (Questions 3–4). Chapter 12 turns the measurement problem into a quantitative instrument: a simulation study that computes, from the documented error of each measurement tier and the realized study designs, the detectability floor — the smallest true pouch-size effect the field’s literature could ever have detected — and the measurement fidelity a decisive future study would need. Chapter 13 bounds the result with the two-sided safety trade-off and the one-anastomosis subgroup (Questions 5–6), closing with risk-of-bias and certainty-of-evidence summaries.

Part III (Synthesis, Chapters 14–15) interprets the results through a set of interpretive lenses committed to in advance — among them no effect under good measurement, an effect surviving measurement quality, direction reversal, discordance between efficacy and safety, and undecidable heterogeneity — and states the answer to each of the seven questions at its earned confidence, the limitations without cushioning, and the forward program: what bariatric registries and the decisive future trial would have to measure for the field’s oldest technical question to become answerable at scale.

Chapter 2

Obesity and its treatments: the stakes

The question this dissertation answers — whether the constructed size of the gastric pouch determines how much weight a patient loses after Roux-en-Y gastric bypass — is a technical question inside a much larger one: what obesity is, why it resists treatment, and where surgery sits in the hierarchy of things that work. This chapter supplies that frame. It argues three things. First, obesity is a disease of energy regulation, not of willpower: the body defends its fat mass with a feedback architecture that actively reverses voluntary weight loss, which is why the treatment ladder is a ladder at all — each rung exists because the rung below it fails against that regulation. Second, the arrival of potent incretin-based drugs has, for the first time, produced pharmacological weight loss in the neighborhood of surgical weight loss, and it did so by mimicking the hormonal consequences of the operations — a lineage that matters for this dissertation, because it makes the *mechanical* parameters of surgery, pouch size chief among them, exactly the part of the operation that drugs cannot reproduce and whose independent contribution therefore needs to be known. Third, surgery retains distinct claims — durability without continued dosing, diabetes remission, and mortality reduction measured over decades — and those claims are what make its technical execution worth interrogating with the rigor this dissertation applies.

2.1 The scale of the problem

Obesity is conventionally defined by body-mass index (BMI), body weight in kilograms divided by the square of height in meters, with $\text{BMI} \geq 30 \text{ kg/m}^2$ labeled obese and $\geq 35 \text{ kg/m}^2$ commonly labeled severe obesity. BMI is a crude proxy — it measures load, not adiposity, and says nothing about fat distribution or organ function — and its limitations as a diagnostic instrument have become a live controversy, taken up below. But as a population-level index it is measured consistently enough, across enough countries and decades, to support one robust conclusion: the distribution has shifted, nearly everywhere, in one direction.

The NCD Risk Factor Collaboration’s pooled analysis of 3,663 population-representative studies covering 222 million people — the largest harmonized anthropometric dataset ever assembled — found that between 1990 and 2022 the combined prevalence of underweight and obesity rose in roughly four fifths of countries for women (162 of 200) and seventy percent for men (140 of 200), with the rise driven almost everywhere by obesity; by 2022 obesity was more prevalent than underweight in 89% of countries for women and 73% for men [1]. The historical problem of too little food has been overtaken, in nearly every national population on earth, by the problem of energy surplus. The Global Burden of Disease analysis put absolute numbers on the health cost: in 2015, 603.7 million adults and 107.7 million children were obese, and high BMI accounted for four million deaths that year — nearly forty percent of them in people

whose BMI was below the obesity threshold, and more than two thirds through cardiovascular disease [20]. In the United States, where the bariatric-surgery literature is disproportionately generated, self-report-corrected projections from more than six million measurements indicate that by 2030 nearly one in two adults will have obesity and nearly one in four will have severe obesity — with severe obesity becoming the single most common BMI category among women, non-Hispanic Black adults, and low-income adults [21]. The population from which bypass candidates are drawn is not shrinking; it is becoming the mode of the distribution.

Two features of this epidemiology matter for everything that follows. The first is that the burden concentrates in exactly the range where bariatric surgery operates. Excess mortality and diabetes risk climb steeply with BMI, and the severe-obesity stratum — the surgical population — is the fastest-growing one [21]. The second is the recent redefinition effort: a 2025 Lancet Commission of 58 experts proposed distinguishing *clinical obesity* — excess adiposity that has already produced organ dysfunction — from *preclinical obesity*, precisely because BMI alone both over- and under-diagnoses the disease [22]. The direction of this move is instructive: the field is converging on obesity as a chronic systemic illness diagnosed by physiological consequence, not a lifestyle descriptor diagnosed by a threshold on a ratio. That framing — disease, not choice — is not rhetoric. It is the empirically forced conclusion of the physiology reviewed next.

Scale is also the multiplier on this dissertation’s question. Gastric bypass and its siblings are performed roughly 280,000 times a year in the United States alone [23], each one embedding a surgeon’s choice of pouch dimensions into a patient’s anatomy permanently. At that volume, even a modest true effect of pouch size on weight loss — one or two percentage points of total weight — aggregates into population-level consequence, and so does a modest true effect on ulcer risk in the opposite direction. Parameters chosen hundreds of thousands of times a year deserve effect sizes, not traditions.

FIGURE PLACEHOLDER: Epidemiology panel (data-rendered per spec:011, never LLM-drawn). Panel A: global adult obesity prevalence trend 1990–2022 by sex, from NCD-RisC published country-level estimates (source data to be downloaded from NCD-RisC repository and re-plotted). Panel B: projected US prevalence of obesity and severe obesity to 2030 from Ward et al. published estimates. Rendered with the repository plotting toolchain from tabulated source data; provenance recorded in Appendix E.

Figure 2.1: The epidemiological backdrop: obesity prevalence trends worldwide (1990–2022) and projected US prevalence to 2030. Data-rendered figure; sources [1, 21].

2.2 The physiology of body-weight defense

Why does a condition of energy surplus not correct itself when the person decides to eat less? The engineering answer is that body weight is not an open-loop accumulator; it is the state variable of a negative feedback control system, and dieting is an attempt to command a new setpoint from outside the loop without changing the controller.

2.2.1 The energy-homeostasis system

The modern account begins with the 1994 positional cloning of the mouse *obese* gene and its product leptin, a hormone secreted by adipose tissue in proportion to fat mass; mutation of the gene produces profound obesity as part of a syndrome resembling human morbid obesity [24]. Leptin closed a loop that physiologists had postulated for decades: fat mass reports its own magnitude to the hypothalamus (the brain region that integrates hunger and energy signals), and the brain adjusts intake and expenditure to hold that magnitude near a defended level. In

signal-processing terms, leptin is the feedback path of a lipostat — and the system’s failure mode in common obesity is not absent feedback but something closer to receiver desensitization at a raised operating point. The Endocrine Society’s scientific statement on obesity pathogenesis synthesizes the evidence bluntly: obesity is a disorder of the energy-homeostasis system, in which the defended level of fat mass is shifted upward and then *defended there*, rather than a passive accumulation of excess intake; individuals completing behavioral weight-loss programs eventually regain most of the lost weight because the homeostatic system actively opposes the new, lower mass [25]. The distinction is the whole argument: a passive store can be drained at will; a regulated variable fights back.

2.2.2 What happens when weight falls

The counterattack has been measured, on both the expenditure side and the intake side, and its two signatures are worth stating precisely because bariatric surgery must ultimately be judged against them.

On the expenditure side, Leibel, Rosenbaum, and Hirsch performed the canonical perturbation experiment: hold subjects at their usual weight, then drive weight 10–20% down (or 10% up) by controlled under- or overfeeding, and measure total energy expenditure by calorimetry at each plateau. Maintaining a body weight 10% or more below baseline reduced total energy expenditure by 6 ± 3 kcal per kilogram of fat-free mass per day in never-obese subjects and 8 ± 5 in obese subjects — a decline in burn rate over and above what the loss of tissue itself predicts, opposing further loss; maintaining an elevated weight produced the mirror-image increase [26]. This surplus decline is *adaptive thermogenesis*: the controller reduces the plant’s idle power draw when the regulated variable falls below its reference. It is not a transient. In fourteen contestants of the television weight-loss competition “The Biggest Loser,” resting metabolic rate was about 610 kcal/day below baseline at the end of the 30-week intervention — and 704 ± 427 kcal/day below baseline six years later, by which time 41 of the original 58 kg mean loss had been regained; the metabolic adaptation component, the residual after adjusting for body composition and age, was still roughly -500 kcal/day at six years [27]. The subjects who kept the most weight off were running the largest expenditure deficit — a pattern consistent with a persistent, displacement-scaled opposition (the integral-control language is engineering shorthand for that pattern, not a measured controller), and not with waning discipline.

On the intake side, Sumithran and colleagues followed the circulating appetite-regulating hormones of fifty patients through a ten-week very-low-energy diet (mean loss 13.5 kg) and for a year afterward. Weight loss reduced leptin, peptide YY, cholecystokinin, insulin, and amylin — the satiety-side signals — and raised ghrelin, the stomach-derived hormone that stimulates hunger; subjective appetite rose in parallel. At 62 weeks, one full year after the loss, every one of those deviations was still present and hunger remained elevated [28]. The authors’ conclusion is the one that matters for the treatment ladder: the hormonal configuration that encourages regain had not reverted one full year after the loss — the longest window measured, with nothing in the data suggesting it was on its way back. The lay narrative — diets fail because people stop trying — has the causality inverted. Diets fail because they succeed: the achieved loss itself generates the orexigenic (appetite-driving) signal that reverses it. Dieting against this system is impedance-mismatched by construction: the harder the drive, the stronger the reflected wave.

Three properties of this control system deserve explicit statement, because each one reappears as an interpretive rule later in the dissertation. The first is asymmetry. The system defends against depletion far more vigorously than against surplus — an intelligible bias for a regulator shaped under famine selection pressure, though that reading is evolutionary plausibility rather than evidence — so the same person who drifted upward over twenty years without effective opposition will meet a coordinated counterattack within ten weeks of dieting downward [25, 28]. Obesity is therefore easy to enter and hard to leave in a way that no symmetric-penalty model predicts. The second is that the defense is proportional and sustained, not transient:

the expenditure deficit scales with the displacement from the defended weight and persists for at least six years [27], which means treatments are not fighting an adjustment period but a standing force. Any therapy evaluated at one year and pronounced durable has been measured against an opponent that does not tire at one year. The third is that the defended level itself is plastic in one direction: the system that resists loss so effectively evidently permitted — or tracked — the original gain, an “upward setting” of the reference over years of positive energy balance [25]. The practical corollary is the central one for surgery: an intervention that merely subtracts calories leaves the reference where it was and fights the error signal forever, while an intervention that changes what the gut reports to the brain has at least the possibility of moving the reference itself. Which of these gastric bypass does is, in compressed form, the mechanistic question behind this dissertation’s clinical one.

One caution belongs on the record here. What these experiments measure is the compensatory response, not the controller: whether the architecture is best described as a defended set point, a settling point emerging from competing influences, or a dynamic equilibrium remains contested, and the interpretive rules above rest on the measured responses, not on the winner of that dispute.

This is the physiological stage on which every obesity treatment performs. An intervention succeeds durably only if it either overrides the defense continuously (pharmacotherapy, for as long as it is taken) or alters the signaling architecture itself — which, as Chapter 5 develops in detail, is the modern account of what gastric bypass actually does. The dissertation’s subject sits precisely at the junction of the two readings of the operation: if bypass works by re-tuning gut-to-brain signals, what work is left for the size of the reservoir the surgeon builds?

2.3 The treatment ladder

Clinical practice arranges obesity treatments as an escalation ladder: lifestyle intervention, then medication added to lifestyle, then endoscopic procedures, then surgery. The ladder is a pedagogical simplification, not a care pathway: contemporary practice runs the rungs concurrently rather than strictly in sequence, and no rung is a mandatory gate for the next — current surgical eligibility is defined by BMI and comorbidity, not by documented failure of the rungs below (the criteria close this section). The ladder’s ordering is by invasiveness and risk, not by mechanism; what actually ascends from rung to rung is the degree to which the intervention counters, rather than merely exhorts, the defense physiology of Section 2.2.

2.3.1 Lifestyle intervention

Intensive lifestyle intervention is the ladder’s traditional first rung and the best-studied. Its definitive test is Look AHEAD, which randomized 5,145 overweight or obese adults with type 2 diabetes to an intensive program of caloric restriction and physical activity versus support and education, powered for cardiovascular events over more than a decade. The intervention achieved 8.6% weight loss at one year against 0.7% in controls — but the gap decayed to 6.0% versus 3.5% by study end, and the trial was stopped for futility at a median 9.6 years with no reduction in cardiovascular events [29]. Read against Section 2.2, the shape of that curve looks less like a compliance artifact than like the step response of a regulated system returning toward its defended value. Lifestyle change remains the substrate on which every other treatment is layered, but as a monotherapy for established severe obesity it is, on randomized evidence, insufficient — and the physiology explains why effort alone was never going to be the missing ingredient.

2.3.2 Pharmacotherapy before the incretins

The pre-incretin pharmacopoeia was modest. A systematic review of long-term trials of the agents approved in the United States found placebo-subtracted losses at one year ranging from approximately 3% of initial weight (orlistat, lorcaserin) to 9% for top-dose phentermine–topiramate, with no agent shown to reduce cardiovascular morbidity or mortality [30]. Several drugs of that generation were subsequently withdrawn for safety, and prescribing remained dominated by noradrenergic appetite suppressants approved only for short-term use [30]. Persistence is poor even now: in a health-system cohort of 1,911 adults starting anti-obesity medication between 2015 and 2022, only 44% were still on drug at three months and 19% at one year [31]. A rung that most patients step off within a year, delivering single-digit efficacy while they stand on it, left a wide gap between medical and surgical treatment. That gap is where the incretin drugs landed.

2.3.3 Endoscopic therapy

Between drugs and surgery sits a newer intermediate rung: endoscopic bariatric therapy, of which the most consequential is endoscopic sleeve gastroplasty (ESG) — a suturing procedure performed from inside the stomach through a flexible endoscope, reducing gastric volume without incisions or resection. In the MERIT randomized trial, ESG plus lifestyle modification produced a mean excess-weight loss of 49.2% versus 3.2% for lifestyle alone at 52 weeks (total body-weight loss 13.6% versus 0.8%), with serious device-related adverse events in 2% of participants and no mortality [32]. Excess-weight loss (%EWL) — the fraction lost of the weight above ideal — and total-weight loss (%TWL) are the field’s two competing outcome currencies; their non-interchangeability is a methodological problem this dissertation treats as a first-class subject in Chapter 7. For present purposes the placement is what matters: ESG’s roughly 14% total-weight loss at one year is genuine but intermediate, and its long-term durability record is short.

2.3.4 Surgery’s position on the ladder

Metabolic and bariatric surgery is the top rung and the standard against which the others are measured. The 2022 joint statement of the American and international bariatric-surgery societies (ASMBS and IFSO) — updating eligibility criteria that had stood since 1991 — recommends surgery for BMI ≥ 35 regardless of comorbidity, advises considering it from BMI 30 with metabolic disease, and lowers the thresholds for Asian populations, on the grounds that long-term results consistently demonstrate safety and efficacy [8] — criteria defined by BMI and comorbidity, with no requirement of documented failure of lifestyle therapy first. Yet the rung is barely used relative to the population standing below it: the ASMBS estimates approximately 280,000 procedures in the United States in 2022 — about one percent of the qualifying population [23]. Sleeve gastrectomy (removal of most of the stomach along its greater curvature, leaving a narrow tube) is now the most common operation, with Roux-en-Y gastric bypass — the operation whose construction this dissertation examines — stable as the reference procedure [23]. Two randomized head-to-head comparisons have reported five-year follow-up. In SLEEVEPASS, estimated mean excess-weight loss was 49% after sleeve versus 57% after bypass — a difference of 8.2 percentage units (95% CI 3.2–13.2) favoring bypass whose confidence interval nevertheless crossed the trial’s prespecified ± 9 -unit equivalence margin, so the procedures were not shown equivalent; a failed equivalence test, not a demonstration of clinically decisive superiority [33]. In SM-BOSS, the Swiss multicenter trial of 217 patients, excess-BMI loss at five years did not differ significantly between the procedures (61.1% after sleeve versus 68.3% after bypass) [34]. How this particular operation came to be the benchmark, and how its technical parameters — pouch size among them — were fixed by convention rather than trial, is the subject of Chapter 3.

Table 2.1: The treatment ladder in numbers: representative randomized (or, where noted, controlled observational) evidence for each rung. Outcomes are as reported by each source; metrics are not directly comparable across rows (%TWL vs. %EWL vs. placebo-subtracted loss; see Chapter 7).

Rung	Source	Weight outcome	Timepoint
Intensive lifestyle	Look AHEAD RCT [29]	−8.6% vs. −0.7%; −6.0% vs. −3.5% at end	1 y; ~10 y
Pre-incretin drugs	systematic review [30]	≈ 3–9% placebo-subtracted	1 y
Semaglutide 2.4 mg	STEP 1 RCT [11]	−14.9% vs. −2.4% placebo	68 wk
Tirzepatide 15 mg	SURMOUNT-1 RCT [12]	−20.9% vs. −3.1% placebo	72 wk
Retatrutide 12 mg	phase 2 RCT [35]	−24.2% vs. −2.1% placebo	48 wk
Endoscopic sleeve	MERIT RCT [32]	−13.6% vs. −0.8% %TWL	52 wk
RYGB	SOS controlled cohort [36, 37]	−32% peak; −25% maintained	1–2 y; 10 y
RYGB	Adams et al. cohort [38]	−26.9%	12 y

2.4 The incretin era

2.4.1 Surgery-adjacent weight loss from a syringe

Incretins are gut hormones secreted in response to nutrient arrival in the intestine that amplify insulin secretion — the two principal ones being glucagon-like peptide-1 (GLP-1) and glucose-dependent insulintropic polypeptide (GIP). GLP-1 also acts centrally to suppress appetite and slows gastric emptying, which is what made its receptor agonists candidates for obesity treatment once their duration of action was engineered from minutes (the native peptide is degraded almost immediately) to a week. The efficacy step change arrived in three trials in rapid succession. STEP 1 randomized 1,961 non-diabetic adults with obesity to weekly semaglutide 2.4 mg or placebo: −14.9% versus −2.4% body weight at 68 weeks, with 86% of treated participants losing at least 5% [11]. SURMOUNT-1 did the same for tirzepatide, a dual GIP/GLP-1 receptor agonist, in 2,539 adults: −20.9% at the 15 mg dose at 72 weeks, against −3.1% for placebo [12]. And the phase 2 trial of retatrutide, a triple agonist adding glucagon-receptor activity, reached −24.2% at 48 weeks at the highest dose — with the trajectory not yet at plateau [35]. For scale: the last figure approaches the 10-year maintained weight loss of gastric bypass in the Swedish Obese Subjects cohort [37], and the drug class has also crossed the hard-outcome threshold that pre-incretin pharmacotherapy never reached — in the SELECT trial of 17,604 patients with cardiovascular disease and overweight or obesity but no diabetes, semaglutide reduced major adverse cardiovascular events by 20% (hazard ratio 0.80, 95% CI 0.72–0.90) over a mean 40 months [39]. It is no longer defensible to describe pharmacotherapy as categorically inferior to surgery on the axis of achievable weight loss. Whatever the treatment ladder looks like in ten years, its rungs have moved.

Two caveats keep the comparison honest. The first is population: the incretin trials enrolled cohorts with mean BMI around 38 [12], whereas the classical surgical cohorts sit substantially higher — median BMI 46.5 in the LABS consortium [40], mean 41 in SOS [37] — and percentage weight loss is not transportable across that gap without argument. The second is response heterogeneity: mean trajectories conceal a nonresponder tail (in STEP 1, roughly one in seven treated participants failed to lose even 5% [11]), and no randomized head-to-head trial of an incretin agonist against bariatric surgery in a surgical-eligible population has reported. The claim the trials support is therefore precise but bounded: modern pharmacotherapy occupies weight-loss territory formerly exclusive to surgery, in populations somewhat leaner than the surgical mainstream, for as long as the drug continues.

2.4.2 Chronic therapy for a chronic disease

What the drugs have not done is repeal Section 2.2; they negotiate with it, continuously, for as long as they are administered. The withdrawal experiments are unambiguous. In the STEP 1 extension, participants who had lost 17.3% on semaglutide regained 11.6 percentage points — two thirds of the loss — within one year of stopping drug and lifestyle support, with cardiometabolic risk factors reverting in parallel [41]. SURMOUNT-4 made the same measurement by randomized design: after a 36-week tirzepatide lead-in producing 20.9% loss, participants re-randomized to placebo regained 14% of body weight over the following year while those kept on drug lost a further 5.5% [42]. On this evidence the drugs do not durably reset the defended level: they oppose it for as long as they are administered — an inference from the regain curves, not a direct measurement of the controller — and withdrawal releases the system to defend the old weight. This would be unremarkable — antihypertensives behave the same way — except that real-world persistence with anti-obesity medication was the 19%-at-one-year figure cited above, from a 2015–2022 cohort dominated by pre-incretin agents and an earlier access environment, so current incretin persistence may differ [31], and the economics of lifelong weekly biologics for a tenth of the adult population are unresolved. The impasse has numbers. At estimated United States net prices of \$6,830 per year for semaglutide and \$7,973 for tirzepatide, the 2026 ICER appraisal judges both drugs cost-effective at a \$100,000 per quality-adjusted life-year threshold — and simultaneously estimates that less than 1% of the eligible population could be treated without exceeding its budget threshold [43]; a pooled analysis of published economic evaluations reaches the opposite verdict on the first question, finding GLP-1 receptor agonists not cost-effective for obesity without diabetes against lifestyle intervention from the payer’s perspective in high-income countries at current prices [44]. A therapy can apparently be fairly priced per life-year and still unaffordable in aggregate. The clinical reality of the incretin era is therefore not drug-replaces-surgery but a chronic disease acquiring its first adequate chronic medical therapy, with adherence, supply, and cost determining who actually receives it.

2.4.3 Pharmacology borrowed from surgery

The intellectual lineage of these drugs runs through the operating room, and the point deserves emphasis because it inverts the usual framing of drugs-versus-surgery. The reason anyone believed profound weight loss could be hormonally mediated is that gastric bypass demonstrated it. Cummings and colleagues showed in 2002 that ghrelin — the stomach-derived hunger hormone that rises with every diet-induced kilogram lost, part of the siege described above — was not merely un-elevated after bypass but suppressed 72–77% below controls, its meal-related pulsatility abolished, despite 36% weight loss [45]. le Roux and colleagues then showed that bypass patients mount exaggerated postprandial responses of the satiety hormones peptide YY and GLP-1 — an anorectic gut-hormone profile that purely restrictive banding does not produce [46]; the postprandial GLP-1 excursion after bypass is roughly tenfold normal [47]. Mechanistic reviews of the past decade converge on the conclusion that the weight loss of modern bariatric surgery is driven principally by these altered gut–brain signals — appetite, food preference, and energy handling — rather than by mechanical restriction of intake or malabsorption of nutrients [9]. Chapter 5 examines this evidence, and its genuinely contested threads (whether GLP-1 is cause or epiphenomenon of surgical diabetes control is itself debated [47]), in the depth the dissertation’s hypotheses require. Here the point is genealogical: the incretin drugs are, functionally, an attempt to package the endocrine consequences of gastric bypass in a syringe. The genealogy is not mediation evidence — that pharmacological activation of one pathway reproduces much of the weight loss does not show that this pathway carries most of surgery’s own effect, and the newer agents engage receptor combinations (GIP, glucagon) beyond the measured postoperative profile — but the lineage stands: surgery was the proof of concept for its own competition.

2.4.4 Why the drugs sharpen the dissertation’s question

It is tempting to conclude that the incretin era makes the fine points of surgical technique obsolete — that if hormones do the work, the dimensions of a surgically constructed reservoir are a historical curiosity. The opposite is true, and the argument has three steps.

First: what distinguishes surgery, now that its weight-loss magnitude has company, is precisely the set of things a drug cannot do — create a permanent anatomical rearrangement that re-tunes gut–brain signaling without further dosing. That durability is not free — lifelong supplementation, surveillance, and revision risk are priced in Section 2.5 — but it does not depend on continued adherence to, or continued payment for, a weekly biologic. The durable component of surgery is its *construction*. If construction is what surgery uniquely sells, then the effect size of each construction parameter is exactly what the field needs to know, and pouch size is the parameter the surgeon controls most directly when building a bypass.

Second: the mechanical parameters are plausibly not hormonally inert. The pouch is the antechamber through which nutrients reach the jejunum, and its geometry may modulate the rate at which they arrive — a rate that drives the very hindgut hormone responses — GLP-1 and PYY chief among them — that the drugs mimic. A volumetric study with objective emptying measurement found that small pouches (≤ 40 mL) empty faster, and that faster emptying — not volume per se — was associated with better maintained excess-weight loss and food tolerance at a median of four years [48]. In other words, even under a fully neuroendocrine model of the operation, pouch construction could matter — as a modulator of signal timing rather than as a flow restrictor. The two models make distinct — and in places opposite — predictions about *whether and which way* size should matter, a tension Chapter 4 develops as the steelman for the analysis.

Third: the field cannot currently answer the question. The evidence on the construction’s central parameter is simultaneously null, positive, and inverted depending on where one looks — a registry of 14,168 patients in which pouch size stops predicting weight loss six weeks after surgery, a randomized trial won by the *longer* pouch, a retrospective cohort won by the smaller, and a systematic review that could not pool any of it [14, 13, 18, 15]. That forty-year stalemate is stated in the studies’ own numbers in Section 1.4 (Table 1.1) and is not restated here. Forty years into the operation, with the drugs forcing surgery to justify itself by exactly its construction, that is the gap this dissertation was designed to close, or to explain.

2.5 Where surgery still wins, and what it costs

The claims that surgery retains against surgery-adjacent pharmacotherapy are three, and each rests on evidence of a kind the drugs do not yet have: decades.

2.5.1 Durability

The Swedish Obese Subjects (SOS) study — a prospective, controlled (non-randomized) comparison of 2,010 surgical patients against 2,037 conventionally treated matched controls — remains the longest controlled record. Gastric-bypass patients reached 32% weight loss at one to two years and maintained 25% at ten, while controls’ weight changed by less than $\pm 2\%$ throughout [36]; at the ten-year laboratory follow-up the surgical cohort as a whole held -16.1% against $+1.6\%$ in controls [37]. The Utah cohort of Adams and colleagues, with follow-up exceeding 90% at twelve years — a retention figure almost unheard of in this literature, and a standing rebuke to it — found bypass patients at -26.9% of baseline weight at year twelve, versus -2.0% and -0.9% in two nonsurgical comparison groups [38]. No pharmacological dataset extends remotely this far, and the withdrawal trials above give strong reason to expect that off drug, no pharmacological cohort ever will. Surgery’s weight loss, once established, is a property of the patient’s new anatomy, with attenuation over years — some of it attributed, rightly or wrongly,

to dilation of the constructed pouch, a confound between construction and measurement that Chapter 6 disentangles.

2.5.2 Diabetes remission

Type 2 diabetes — chronically elevated blood glucose from insulin resistance and progressive pancreatic beta-cell failure — was long managed as strictly incurable. Metabolic surgery falsified that. *Remission* here means glycated hemoglobin (HbA1c, the standard integrated measure of blood glucose over the preceding months) held in the nondiabetic range without glucose-lowering medication. In the STAMPEDE randomized trial, 29% of gastric-bypass patients versus 5% on intensive medical therapy met the glycemic primary endpoint at five years — but that endpoint was $\text{HbA1c} \leq 6.0\%$ *with or without* glucose-lowering medication, a target-attainment measure, not the medication-free remission just defined, and the two must not be read interchangeably [49]. The endpoint came alongside a weight advantage (-23% for bypass versus -5% for medical therapy) and superiority on lipids, insulin use, and quality of life [49]. In the Italian randomized trial with the longest follow-up, 37.5% of surgical patients maintained continuous remission for the full ten years (25% for bypass specifically, against 5.5% for medical therapy); the honest counterweight in the same data is that 58.8% of patients in remission at two years relapsed during follow-up — though even relapsers retained adequate glycemic control (mean HbA1c 6.7% at ten years), and diabetes-related complications were fifteenfold less frequent in the surgical arms [50]. The pooled ARMMS-T2D follow-up of four US randomized trials found an HbA1c advantage of 1.4 percentage points at seven years and remission in 18.2% versus 6.2% at seven years and 12.7% versus 0% at twelve [51]. Remission erodes with time — the honest reading of all three trials — but a durable, medication-free minority and a large glycemic advantage for the rest, achieved by a single intervention, remains beyond any drug regimen’s demonstrated reach. In the Utah cohort of Adams and colleagues, gastric bypass also cut the twelve-year *incidence* of type 2 diabetes among participants free of diabetes at baseline — an odds ratio of 0.08 (95% CI 0.03–0.24) against the matched nonsurgical comparison group, an observational estimate but a striking one — prevention, not just remission [38].

2.5.3 Mortality

The hardest endpoint is survival — though here the evidence, for all its consistency, is observational, and settles less than its effect sizes suggest. SOS reported an unadjusted all-cause mortality hazard ratio of 0.76 for surgery over a mean 10.9 years [36], extending by its 24-year analysis to a hazard ratio of 0.77 and an adjusted median life expectancy 3.0 years longer than conventional care — though still 5.5 years shorter than the general population [52]. The matched Utah cohort found 40% lower adjusted all-cause mortality over a mean 7.1 years after bypass, driven by deaths from coronary disease (-56%), diabetes (-92%), and cancer (-60%) — while deaths from accidents and suicide were 58% *higher* in the surgical group, a signal the field has still not fully explained [53]. A one-stage meta-analysis reconstructing patient-level survival from sixteen matched cohort studies and one prospective trial — 174,772 participants, 1.2 million patient-years — estimated a 49.2% reduction in hazard of death and a median 6.1 years of life expectancy gained, rising to 9.3 years in patients with diabetes; the ten-year number needed to treat to prevent one death was 8.4 with diabetes and 29.8 without [54]. These are large associations with total mortality, replicated across continents and decades — but every one of them comes from designs (a nonrandomized trial, matched cohorts, and pseudo-individual-level data reconstructed from published cohort curves) that cannot exclude healthy-candidate, access, socioeconomic, and treatment-selection effects. They are read here as consistent observational associations, not as hierarchy-settling causal effects.

2.5.4 The price

The costs are real and belong in the same ledger. Perioperatively, modern bariatric surgery is safer than its reputation: in the prospective multicenter LABS consortium, 30-day mortality across Roux-en-Y bypass and adjustable banding combined was 0.3%, with 4.3% experiencing a major adverse outcome — composite figures that pool bypass with the lower-risk banding procedure and therefore flatter bypass somewhat [40]; SOS recorded 90-day postoperative mortality of 0.2% [52]. For calibration, the LABS composite risk concentrated in identifiable strata — prior venous thromboembolism, obstructive sleep apnea, impaired functional status, and extremes of BMI — rather than being uniform across candidates [40]. The chronic costs are larger. At five years in the randomized SLEEVEPASS comparison, overall morbidity was 26% after bypass and 19% after sleeve gastrectomy [33] — a between-operation difference that is part of why the technically simpler sleeve overtook bypass in volume [23], and a reminder that bypass buys its outcomes with complexity. In SOS, 2.9% of surgical patients underwent repeat surgery — a figure the source reports alongside 90-day mortality and best read as early reoperation, not lifetime revision burden [52]; in ARMMS-T2D, anemia, fractures, and gastrointestinal adverse events were all more frequent in the surgical arms over twelve years [51]; and the accident-and-suicide excess in the Utah data [53] warns that rearranging the gut rearranges more than metabolism. Bypass anatomy specifically carries lifelong obligations — micronutrient supplementation and surveillance for deficiency states. The standard of care is codified in the ASMBS nutritional guideline, whose 2016 micronutrient update issues 92 graded recommendations for screening, preventive supplementation, and repletion spanning thiamine, vitamin B12, folate, iron, the fat-soluble vitamins, calcium, vitamin D, copper, and zinc [55] — and the obligation is not precautionary: twelve years after RYGB in a Norwegian multicenter cross-sectional study of 490 patients, suboptimal vitamin D persisted in 52% and suboptimal vitamin B12 in 16% of patients *adherent* to their recommended supplements, with markedly higher deficiency rates among the non-adherent [56]. And among the operation’s specific complications, one is directly coupled to this dissertation’s exposure variable: marginal ulceration at the gastrojejunal junction, whose risk in the largest registry rises monotonically with the length of staple line used to build the pouch [14] — a gradient that penalizes *larger* construction, and therefore points the same way as any small-pouch benefit rather than against it. The safety ledger is nonetheless two-sided, because very small construction carries its own plausible harms — stricture, anastomotic tension, leak, and food intolerance — and Chapter 13 prices both directions of the trade explicitly.

The stakes, restated

The chapter’s argument compresses to this. Obesity is a regulated-system disease whose scale is still growing and whose defense physiology defeats intention reliably enough that the effective treatments are the ones that change the system rather than argue with it. For four decades the only such treatment was surgical, and the world used it at one percent of its indicated capacity. The incretin drugs have now demonstrated that a hormonal channel can, by itself, carry weight loss of near-surgical magnitude — which does not prove that it carries most of surgery’s own effect, but does convert the question of surgery’s mechanical parameters from a technical detail into the heart of surgery’s remaining case. If the pouch’s size does real work, surgeons should know its dose–response and both sides of its safety ledger; if it does not, then a parameter the profession has calibrated, taught, and fought over since 1982 is a legacy of a disproven model, and precision effort belongs elsewhere. Either answer changes practice. The next chapter turns to how the parameter got its value in the first place — a history of numbers chosen without trials — because the doctrine under test here cannot be understood, or fairly tested, without knowing where it came from.

Chapter 3

Bariatric surgery: how technique congealed

The operations that this dissertation studies were not designed; they accreted. Every number that today defines a Roux-en-Y gastric bypass — the volume of the gastric pouch, the diameter of its outlet, the lengths of the intestinal limbs — entered the canon at a particular moment, for a particular reason, and almost never because a comparative trial had identified it as optimal. This chapter reconstructs that history as a history of *decisions made without dose-finding evidence*. Trials were run — procedures were randomized against each other from the 1970s on, and a pouch-length trial eventually appeared in the modern era — but the parameters themselves were set before any trial addressed them: each technical parameter was set by anatomy, by the instruments available in the operating room, by the need to avoid a complication that had just injured a patient, or by the authority of the surgeon who published first — and was then propagated by training until it hardened into convention. The question the dissertation poses statistically in Part II — does the constructed size of the gastric pouch predict weight loss? — is posed here historically: *where did the number come from?*

3.1 Reading guide: the vocabulary of reconstructed anatomy

Because the operations are easiest to follow as plumbing, a few terms are worth fixing at the outset. An *anastomosis* is a surgically created connection between two hollow structures; its opening is a *stoma* (Greek for mouth). A *bypass* routes food around an organ or intestinal segment without removing it; the excluded tissue remains in place, with its blood and nerve supply, but no longer sees nutrients. A *Roux-en-Y* reconstruction — named for the Swiss surgeon César Roux, and shaped like the letter Y — divides the small intestine and brings one end up to drain the upper gastric segment (in bariatric use, the constructed pouch; the bypassed remnant continues to empty through its own pylorus) while the digestive juices from the bypassed segment rejoin the food stream downstream; its practical virtue is that far less bile washes back into the drained segment than with a simple loop — reduced, not abolished. The *gastric pouch* is the small upper-stomach reservoir that the surgeon constructs by stapling; the rest of the stomach (the *remnant*) is left in the abdomen, disconnected from the food stream. A *marginal ulcer* is a peptic ulcer arising at or near the gastrojejunal stoma; acid exposure of the intestinal mucosa is one contributor among several — local ischemia, tobacco smoking, NSAID use, foreign suture material — and the lesion is multifactorial, a point Chapter 13 returns to [57, 58]. Outcomes in this literature are reported as percent excess weight loss (%EWL, loss expressed as a fraction of the weight above an arbitrary “ideal”), percent total weight loss (%TWL), or change in body-mass index (BMI); the metric problem these conventions create is the subject of Chapter 7.

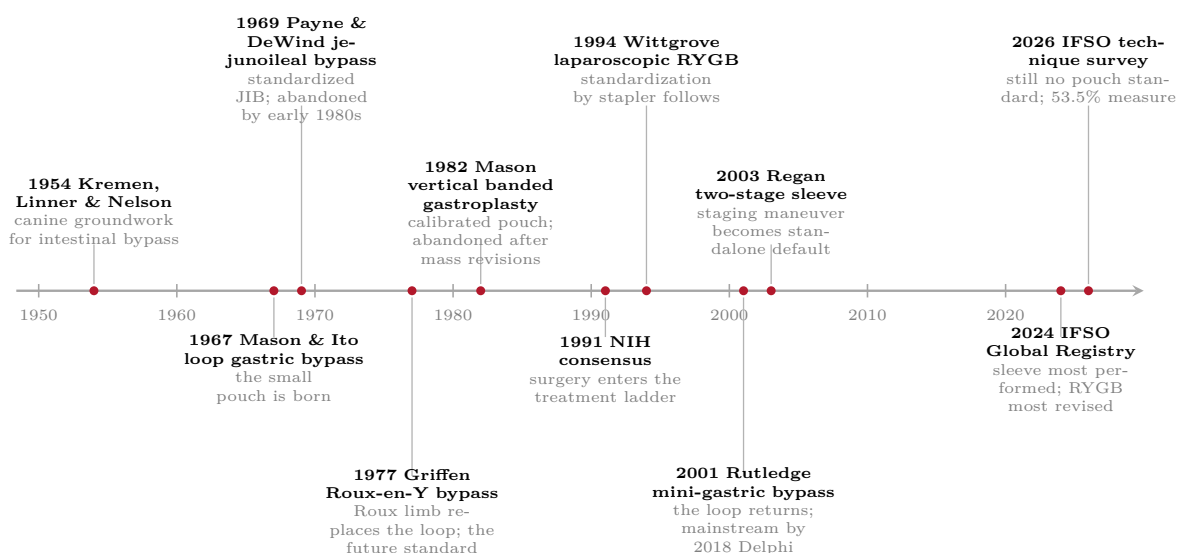


Figure 3.1: Timeline of bariatric operations, 1954–2026. Each milestone carries the year, the report or event, and a one-line fate; dates and fates are those cited in this chapter, and the horizontal axis is linear in time. Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

3.2 The malabsorptive prologue: jejunoileal bypass

Surgery for obesity began not at the stomach but in the small intestine, and its first era supplies the template for everything that follows: a plausible mechanism, a fixed set of construction numbers chosen without dose-finding, wide adoption, and a reckoning delivered years later by complications rather than by trials. The experimental foundation was laid in 1954, when Kremen, Linner, and Nelson published a canine evaluation of the relative nutritional importance of proximal versus distal small intestine — the experimental groundwork for the idea that excluding absorptive surface could be a deliberate therapeutic act [59]. The clinical operation that grew from this work, the jejunoileal bypass (JIB), joined a short segment of proximal jejunum directly to the terminal ileum, excluding most of the roughly six meters of absorptive small bowel from the nutrient stream. Payne and DeWind’s 1969 report standardized the operation for a generation of surgeons [60]; in Alden’s series the configuration was named by its numbers — a “14–4” jejunoileostomy, fourteen inches of jejunum to four of ileum [61] — and in the randomized comparison by Griffen and colleagues it was thirty centimeters of jejunum to twenty-five of terminal ileum [62]. Note what these numbers are: round figures in customary units, transparently chosen for memorability and reproducibility, not the output of any titration experiment. The first congealed parameters in bariatric surgery were lengths picked to be easy to teach.

The operation worked, in the narrow sense that mattered least. In Griffen’s randomized trial of 59 patients, JIB patients lost weight at least as well as gastric bypass patients; in Buckwalter’s four-year clinical trial of 38 patients, the jejunoileal group lost more [62, 63]. But the mechanism was indiscriminate: an intestine that cannot absorb calories also cannot absorb bile salts, fat-soluble vitamins, or electrolytes; in patients with fat malabsorption and an intact colon, oxalate is hyperabsorbed — the colon was shown to be both the site of and the requirement for the enteric hyperoxaluria behind the era’s kidney stones [64]; and the defunctionalized loop became a bacterial fermenter with systemic consequences. Griffen’s trial recorded significant diarrhea in 56% of JIB patients and chronic medication dependence in 74%, plus kidney stones and gallstones not seen after gastric bypass; most damningly, the fatty liver present in every patient at baseline *worsened* at one year in three-quarters of the JIB group while improving or

remaining stable in every gastric-bypass patient [62]. Alden, comparing one hundred consecutive patients of each type operated by the same surgeon, reached the same verdict: equal safety on the table, but long-term complications of the jejunoileal bypass “greater in frequency, duration, and seriousness” [61]. Buckwalter concluded flatly that gastric bypass was the better operation even while conceding it produced less weight loss [63]. By the early 1980s JIB was abandoned; its survivors still surface in surgical clinics decades later, and the long-term management of bypass-type complications remains a live clinical topic [65].

Two lessons from this prologue frame the rest of the chapter. First, the field learned that weight loss alone cannot validate an operation — the two-sided benefit–harm accounting that Chapter 13 formalizes, outcome by outcome, was discovered here, empirically and expensively. Second, pure malabsorption did not die; it was domesticated. Scopinaro’s biliopancreatic diversion, first reported in humans in 1979, retained controlled malabsorption but combined it with a gastric reduction and avoided the blind loop, producing 33.7% body-weight loss at one year in the initial cohort without the hepatic deterioration that had condemned JIB [66]. The lineage continued through the duodenal switch — Hess’s series of 440 patients reported 80% excess-weight loss at two years, holding near 70% at eight [67] — and thereby, as Section 3.6 recounts, accidentally created the sleeve gastrectomy. Malabsorptive dose — limb length — remained a live technical variable ever after, which is why limb lengths are treated as confounders in the analysis of Part II.

3.3 Mason’s loop and the birth of the small pouch

The gastric alternative originated in Iowa City. Edward Mason and Chikashi Ito introduced the gastric bypass in 1967 [3]; in the form Mason’s group reported through the operation’s first decade, the stomach and duodenum were bypassed in favor of a small upper gastric segment emptying through a narrow gastroenterostomy stoma [4], carried on a loop of jejunum — the “standard loop gastric bypass” of the era’s clinical reports [68]. In the origin story the field has retold ever since, the insight was borrowed from ulcer surgery: patients who lost part of their stomach to peptic-ulcer resection lost weight and kept it off. The operation thus began life with a mechanical theory attached — a small reservoir fills quickly, and a full reservoir terminates a meal — and the reservoir’s size was, from the first publication, a parameter the surgeon chose.

What did the founders choose, and why? The clearest early answer is Mason’s own 1975 report on 442 patients accumulated over nine years, titled, with disarming honesty, “Optimizing results of gastric bypass” [4]. The paper is worth pausing over, because it is the founding document of the doctrine this dissertation audits, and its texture is diagnostic of how the era reasoned. Early postoperative deaths had occurred in 3% of patients, most of them, in Mason’s judgment, preventable by better operative technique and earlier recognition of leaks; the expected trajectory for “the average patient of 142 kg” was a weight of about 107 kg at one year and 103 kg at five — point estimates from a single institution’s experience, offered without comparison arms, because there were none to offer. Within that frame, Mason reported that the best weight loss required adherence to three components: bypass of the stomach and duodenum, “a small fundic segment,” and a small gastroenterostomy stoma of 12 mm diameter; inadequate weight loss could be rescued by revising a large stoma down to 9 mm [4]. The epistemic form is that of engineering practice, not clinical science: iterate on the design, attribute outcome shifts to the most recent change, publish the current best configuration. It is how one tunes a process; it is also how one manufactures a convention. But the paper’s second preoccupation is the more revealing one: stomal ulcers occurred in 1.8% of patients — once per 140 patient-years at risk — and Mason expected the “present emphasis on a smaller fundic pouch” to lower that incidence [4]. The small pouch, at its origin, was therefore doing double duty, and only one of its two justifications was about weight. A smaller pouch meant fewer acid-secreting parietal cells left in contact with the vulnerable stoma — a physiological premise later supported when Smith

and colleagues measured basal and stimulated acid output from small cardia-based pouches and found it virtually absent [69]. Support for the premise is not proof of the clinical claim: marginal ulceration is multifactorial (Section 3.1), and whether smaller construction actually lowers ulcer incidence is a comparative-outcome question, not a physiological one. Still, ulcer avoidance, not weight-loss dose-finding, supplied much of the original pressure toward smallness. The marginal-ulcer gradient that Chapter 13 quantifies — in the largest registry, ulcer risk rising with each additional centimeter of staple line, penalizing *larger* construction — points the same way as Mason’s oldest argument; the harms that weigh against very small construction are different ones (stricture, anastomotic tension, intolerance), and Chapter 13 treats the two directions separately.

Through the late 1970s the loop bypass was progressively re-engineered. Alden simplified it by stapling the stomach in continuity, bypassing “90% of the stomach” [61]; Griffen’s group performed it as a 90% gastric exclusion and replaced the loop with a Roux-en-Y reconstruction [62] — the variant that would eventually own the field. The Y-limb’s central virtue is the one named in Section 3.1: it reroutes bile and pancreatic juice downstream, away from the pouch — a liability of the loop that became directly measurable when the loop returned decades later, bile in the pouch being a common scintigraphic and endoscopic finding after one-anastomosis bypass [70] and conversion to Roux-en-Y its surgical remedy [71]. Halverson’s 1981 assessment distilled the American consensus of the moment: a “50–60 cc pouch and a small (1–1.2 cm) gastrojejunostomy” constituted “the operation of choice for morbid obesity,” with 90% of patients losing more than half their excess weight within two years [72]. The numbers were now smaller and more confident, but their provenance had not changed character: they were round figures descended from Mason’s practice, adjusted downward each time a series with large pouches disappointed.

And large pouches did disappoint, repeatedly — this is the empirical kernel inside the doctrine, and it deserves to be stated fairly. Buckwalter attributed the poor weight loss of his early gastric-bypass patients to “an excessively large proximal gastric pouch” and reported that the smaller-pouch revision of the operation matched jejunoileal bypass [63]. Knecht, following 171 Mason-type bypasses for eight years, found poor long-term weight loss in his first 26 patients with large pouches and durable loss — 42% of total weight, 72% of excess weight — in the 145 subsequent patients given smaller pouches [73]. These are before–after comparisons across a surgeon’s learning curve, confounded by everything else that improves with experience; but they were the available evidence, and they all pointed the same way. By 1982 the doctrine was explicit in a title: Kirkpatrick and Siegel’s “Critical determinants of a successful gastric bypass: reservoir versus stoma” reported that if a “tiny (less than 45 ml) reservoir” was consistently created with a Roux-en-Y, the stoma could be left wide — 3.2 cm on average, the native jejunal lumen — with no penalty in weight loss and with post-bypass vomiting virtually eliminated [5]. The reservoir had been promoted to *the* critical determinant of success; the pouch-size doctrine whose empirical basis this dissertation audits was now fully formed, resting on one prospective single-arm study, several learning-curve series, and the mechanical model of restriction that Chapter 5 shows was later dismantled.

One further 1983 refinement completed the anatomical template. Torres, Oca, and Garrison moved the pouch off the fundus entirely, partitioning from the angle of His toward the *lesser* curvature to create a 25–35 ml pouch drained through a stapled lesser-curve gastrojejunostomy — explicitly motivated by the vascular reliability of the lesser-curve arcade and the mechanical reliability of stapled stomas [74]. The vertical lesser-curve pouch became the dominant template of the modern RYGB — contemporary construction still varies in length, width, and first-firing position (Section 4.3) — and its anatomy is dissected in Section 4.2. Note again the driver: blood supply and instrument reliability, not a weight-loss experiment.

3.4 Gastroplasty and the vertical banded decade

If a small reservoir with a narrow outlet is the active ingredient, a natural inference follows: skip the bypass altogether, and simply partition the stomach. That inference — pure restriction, the mechanical model taken at face value — was tested at scale for two decades, and its failure is among the most instructive episodes in the field’s history. Gastroplasty’s conventional starting point is Printen and Mason’s 1973 report on gastric surgery for morbid obesity [75], from the same Iowa group; Mason himself later described the vertical banded gastroplasty as having “evolved from gastric bypass and earlier forms of gastroplasty” [76], and by decade’s end the concept had proliferated into a family of “gastric partitioning” operations. Pace and colleagues stapled across the stomach below the gastroesophageal junction, leaving a 50–60 cc food reservoir emptying through a 1 cm gap left in the middle of the staple line — an operation whose stated appeal was that jejunoileal bypass was unacceptably morbid and gastric bypass “technically difficult” [77]. Their mean operative time was 48 minutes; the operation could be taught quickly and done safely, and in a field still counting perioperative deaths, that was a powerful argument. Simplicity, again, was a design input — and note the reservoir number: 50–60 cc, the same figure Halverson would codify for gastric bypass two years later [77, 72]. The number was circulating between procedures as inherited furniture, adopted for each new anatomy rather than re-derived from it.

The gastroplasty era also produced something the gastric-bypass literature of the time largely lacked: measurement. Andersen and Pedersen, in a 1984 prospective study, radiologically measured pouch volume and stoma diameter serially after Gomez-type gastroplasty and correlated them with weight loss. Early pouch volumes ranged from 51 to 244 ml — a near-fivefold spread in a supposedly standardized operation — and within that range *neither pouch volume nor stoma diameter was significantly correlated with weight loss*; “other factors,” they concluded, “seem to be of major importance” [78]. Four decades before this dissertation’s question was posed as a meta-analytic problem, its answer-in-miniature was already in print: measured size failed to predict outcome even for the purely restrictive operation whose entire theory was size.

Mason’s own response to the failures of horizontal partition was characteristically constructive: the vertical banded gastroplasty (VBG) of 1982, the most deliberately *calibrated* operation of the era. (Two landmarks recur from here on: the *angle of His*, the acute notch where the esophagus meets the gastric fundus, marking the upper limit of any vertical staple line; and the *crow’s foot*, the fan-shaped terminal branching of the anterior vagus nerve on the lesser curvature, marking a reproducible lower boundary. Their surgical anatomy is treated in Section 4.2.) A stapled window through both stomach walls above the crow’s foot allowed a vertical staple line up to the angle of His, creating a lesser-curve pouch of less than 50 ml — measured intraoperatively at a standard distending pressure of 70 cm H₂O — with the outlet encircled by a polypropylene mesh collar of fixed circumference, 5.0 cm for patients between 170 and 225% of ideal weight and 4.5 cm for heavier patients [79, 76]. Here, for once, is a genuine attempt at metrological discipline: a defined filling pressure, a measured volume, a banded outlet that cannot stretch. “Weight control,” Mason wrote, “depends on a small pouch and stoma. Bypass is not necessary” [76]. The claim was testable, and it was tested.

It failed on both arms. In the randomized comparison of Näslund and colleagues — 57 patients randomized to bypass or gastroplasty — gastric bypass produced significantly greater weight loss at one year than gastroplasty, with significantly fewer failures, and four gastroplasties were converted to bypass for failure within two years [80]. Buried in that trial is the single most elegant early datum on this dissertation’s question: peroperatively measured pouch volume correlated with weight loss after *gastroplasty* but *not* after *bypass* — pouch volume and stoma diameter together explained roughly 40% of outcome variance for the purely restrictive operation and a negligible fraction for the bypass [80]. The randomization was to operation, not to pouch size, so these are within-arm observational correlations in small arms — open to confounding by construction difficulty, surgeon, and measurement error, and no formal test of

the between-operation difference in association is reported — but the pattern is suggestive. In signal-processing terms: when restriction is the only channel, reservoir geometry carries information about outcome; add whatever else bypass does — the neuroendocrine channels of Chapter 5 — and the geometric signal drops below the noise floor. Within a randomized trial, though not by randomization, the pouch-size association appeared confined to the operation the field was about to abandon. The long-term VBG record completed the verdict: in a Dutch cohort of 392 primary VBGs, 58% of patients developed long-term complaints and 39% ultimately required revisional surgery, for weight regain, vomiting, outlet stenosis, or pouch and outlet dilation [81]. Pure restriction dilates, obstructs, or is out-eaten; the operation disappeared from practice. What did *not* disappear was its central habit of thought — that the constructed reservoir is the operation’s dose — which migrated intact into the bypass literature even though the best early comparative datum — the within-arm correlations of Näslund’s trial — had found the association confined to the operation that died. The migration is visible even in the VBG’s afterlife: the same Dutch group reporting the operation’s failure noted that interest in it had been renewed by the “banded gastric bypass, with a similar mechanical outlet restriction and control of the pouch size” [81] — the calibration idea returning to the bypass a generation later, still untested against the bypass’s own outcome data. (The 2026 worldwide survey finds ring augmentation in routine use by 20.4% of RYGB surgeons, with strong regional clustering and no consensus [7].)

3.5 The laparoscopic revolution: standardization by stapler

Two events in the first half of the 1990s converted gastric bypass from a contested specialty operation into an industry. The first was institutional: the 1991 NIH Consensus Development Conference concluded that gastric restrictive or bypass procedures “could be considered for well-informed and motivated patients” with acceptable operative risk, embedding surgery in the legitimate treatment ladder for severe obesity [82]. The second was technical: in 1994 Wittgrove, Clark, and Tremblay reported five laparoscopic Roux-en-Y gastric bypasses — the operation performed through ports rather than an open incision — claiming, in the authors’ own preliminary assessment of a five-patient feasibility series, comparable weight loss and reduced morbidity and disability [83]. Within a decade the laparoscopic approach had been elaborated into teachable, reproducible technique [84, 85], and by 2008, in an online survey of American Society for Bariatric Surgery surgeons, 98% of respondents performed the operation laparoscopically [6].

For the history of the pouch-size parameter, the laparoscopic revolution matters less as an access route than as a *standardization mechanism* — and the standard it fixed was set by the instrument, not by evidence. A laparoscopic pouch is built by sequential firings of a linear stapler around a calibrating tube (a *bougie*) passed down the esophagus; its size is therefore quantized by cartridge lengths and bougie diameters. What the surgeon actually controls is where the first horizontal firing enters the lesser curvature and how many vertical firings follow; volume is a downstream consequence, estimated rather than measured. Sundbom and Gustavsson’s hand-assisted series specified the pouch as “4 × 3 cm” [85] — pouch size expressed as stapler geometry. The 2008 survey found a reported average pouch of 25 cm³, with roughly half of surgeons “measuring” pouch size as a distance from the gastroesophageal junction [6] — a length standing in for a volume, the first-rank example of the proxy problem that Chapter 6 elevates to the dissertation’s central methodological theme. The Scandinavian registry cohort that anchors the evidence conflict of Chapter 11 is the logical terminus of this development: in 14,168 standardized Swedish operations, pouch size was operationalized as the total length of stapler cartridge fired — mean 145 mm — because in a stapler-built operation that is the construction variable that exists in the record [14].

The surveys also allow the drift of convention itself to be glimpsed — with an important caveat: they are cross-sectional snapshots of *different* populations, a 2008 online survey of Amer-

ican ASBS members and a 2025–26 worldwide IFSO sample spanning 47 countries, with different sampling frames, memberships, and response biases, so any difference between them mixes temporal drift with geography and composition. In 2008, American surgeons reported an average biliopancreatic limb of 48 cm and Roux limb of 114 cm, split their gastrojejunostomy technique almost evenly between circular stapler (43%) and linear stapler (41%) with a hand-sewn minority, and overwhelmingly (95%) declined to band the pouch [6]. By 2025–26, the worldwide medians were a 100 cm biliopancreatic limb and 120 cm alimentary limb — rounder numbers, longer biliopancreatic limbs — and linear stapling had consolidated to 81.2% of gastrojejunostomies [7]. On these snapshots practice appears to have moved, and moved coherently; but what moved it was the same mixture of instrument economics, complication experience, and circulating opinion that set the original numbers. Nothing in either survey records a technique change driven by a randomized comparison of the parameter changed. Convention is not static — it drifts like any uncontrolled process — and a drifting convention is even harder to audit retrospectively than a fixed one, because “standard RYGB” denotes different anatomies in different decades. Chapter 9 returns to this as the era-subgroup problem. The laparoscopic era did not resolve the pouch-size question; it changed what “pouch size” meant, from a measured volume (Mason’s 70 cm H₂O calibration) to an instrument count. That semantic drift — volume, to length, to stapler firings — is part of what precluded pooling in the field’s one systematic review, whose authors attribute the decision to the “quality and heterogeneity of data” [15]; reading the drift itself as a product of workflow rather than science is this chapter’s interpretation, not the review’s.

3.6 Sleeve gastrectomy: a staging maneuver becomes the default

No episode illustrates evidence-free consolidation better than the rise of the sleeve gastrectomy, today the most performed bariatric operation on earth — a status it reached without ever being designed as a primary procedure. The sleeve began as a component: the gastric reduction of the biliopancreatic diversion with duodenal switch, in which the stomach is converted to a narrow tube along the lesser curvature [67]. In 2003, Regan, Inabnet, Gagner, and Pomp reported using the laparoscopic sleeve alone as the *first stage* of a planned two-stage bypass in super-super-obese patients (BMI > 60 kg/m²), the intention being to return after initial weight loss had lowered operative risk and complete the Roux-en-Y [86]. In Regan’s own seven-patient series the second stage was in fact completed; but as the approach diffused, first-stage weight loss proved substantial and the planned completion often never happened — in one long-term series of 41 two-stage patients, only a quarter ever underwent the second-stage bypass, while 44% still met criteria for it at a mean of 3.6 years [87] — and a staging maneuver drifted into a standalone operation. Two decades later the Eighth IFSO Global Registry report, aggregating 502,150 procedures from 24 national and 2 regional registries, records sleeve gastrectomy as the most performed primary bariatric procedure worldwide — with RYGB the most performed *revisional* procedure [88], a status often read as conversion of failed sleeves to bypass, though the registry’s aggregate data do not themselves link index to revisional operations or record revision indications. The same registry synthesis shows how little “bariatric surgery” names a single practice: Asian chapters operate on patients at substantially lower BMI and with higher rates of diabetes than Western ones [88], so a technical convention exported across chapters lands on different physiology than the one it congealed around. The sleeve’s dominance is itself convention-assisted — contemporary surveys attribute it to its ease as well as its outcomes [7] — while RYGB retains the role of the standard of comparison, the operation “developed in the 1960s and refined in the 1990s” against which the newer procedures are benchmarked [7].

The sleeve also replicated, in compressed time, the parameter history this chapter traces for the pouch. Its geometry is set by the bougie against which it is stapled, and bougie caliber

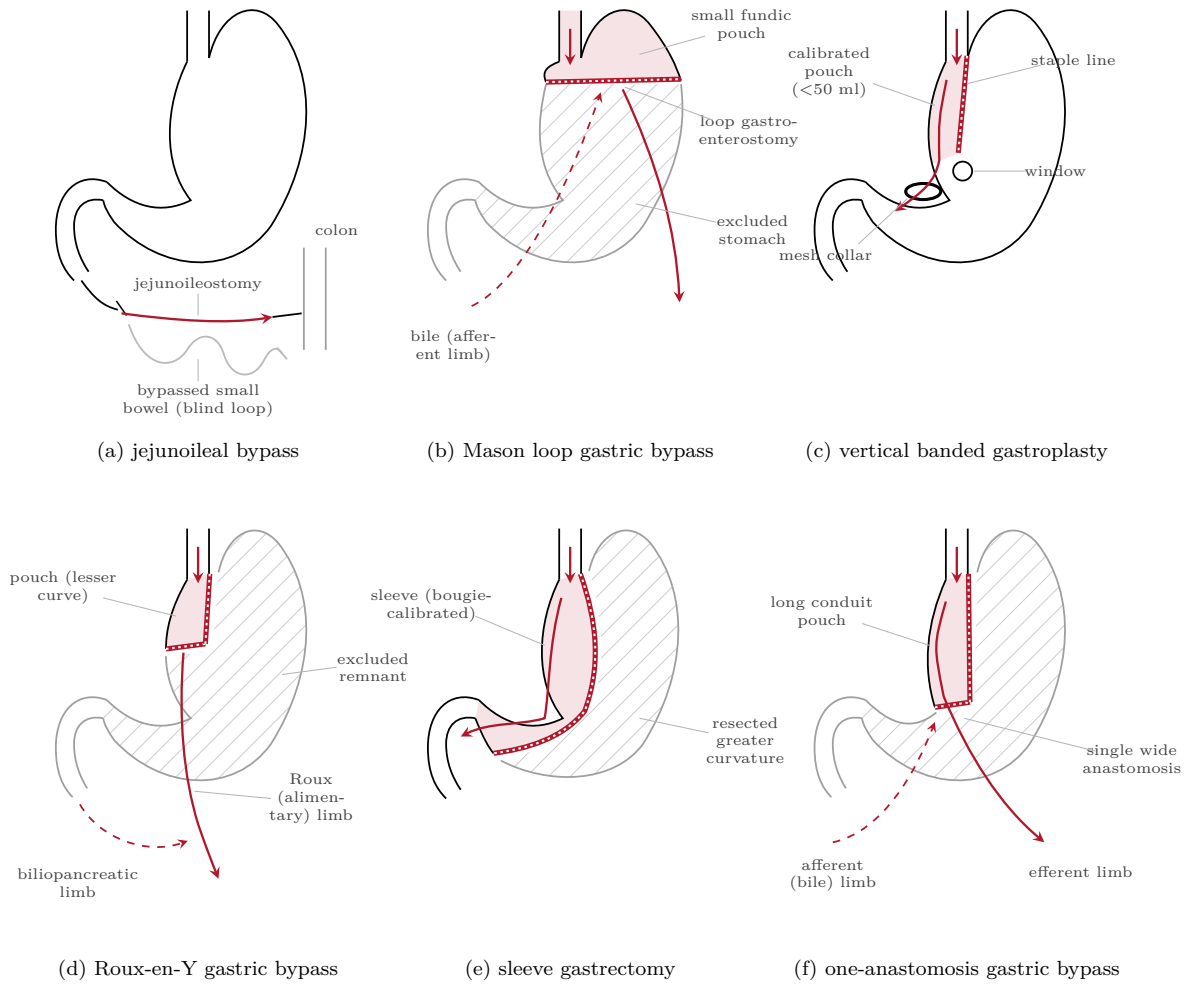


Figure 3.2: The principal bariatric operations, drawn schematically with one visual grammar: the constructed reservoir (pouch, sleeve, or conduit) is shaded; staple lines are the heavy segmented lines; excluded or resected tissue is hatched; the food stream is a solid arrow and the biliopancreatic stream a dashed arrow. Anatomy is stylized and not to scale. Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

became the sleeve’s pouch-size debate: reported diameters span under 32 to 50 French with no consensus on the optimum. The meta-analytic record is unsettled rather than settled: one recent pooled analysis found the smallest calibers (32–36 Fr) delivered a 6.7 percentage-point excess-weight-loss advantage over the largest at 12 months, attenuating to 3.2 points by 36 months, with stricture and staple-line leak more frequent in the smallest group — but earlier meta-analyses found no meaningful weight-loss advantage at all, a network meta-analysis put the *lowest* leak risk in the 33–36 Fr band, and the evidence base is largely observational with inconsistent cut points and definitions [89]. Expert opinion and the large majority of practice settle on the 36–40 Fr middle of the range [89]. A newer operation, the same epistemic structure: a continuous construction parameter, weak and conflicting dose–response evidence, and a convention stabilized by consensus rather than by trial.

3.7 The loop returns: one-anastomosis gastric bypass

The most consequential recent divergence in bypass technique is, from this dissertation’s vantage point, a natural experiment in pouch construction. Rutledge’s “mini-gastric bypass,” first reported in 2001 from 1,274 cases, resurrected Mason’s loop — one anastomosis instead of two — but hung it on a radically different pouch: not a small cardia reservoir but a long, narrow conduit beginning *below* the crow’s foot on the lesser curvature and stapled upward to the left of the angle of His [90, 91]. The originators are explicit that the geometry is the point: a long non-obstructive channel with a wide anastomosis, deliberately unlike the small high-pressure pouch of the RYGB tradition [91]. The operation’s appeal was the same appeal that had sold gastric partitioning in 1979: it is fast and simple. Rutledge’s initial series reported a mean operating time of 36.9 minutes — the shortest 19 — a 1.5-day hospital stay, a 5.2% overall complication rate, and excess-weight loss of 68% at one year and 77% at two [90]. Whatever else may be said of those uncontrolled figures, they explain the diffusion curve. Carbajo’s one-anastomosis variant likewise runs the pouch from the esophagogastric junction down to the crow’s foot, parallel to the lesser curvature, with an estimated volume of some tens of milliliters and an antecolic loop anastomosed 200 cm beyond the ligament of Treitz [92]. After years on the margins — the loop’s bile-reflux liability was exactly what Griffen’s Roux-en-Y had been adopted to avoid — the procedure was declared an “acceptable mainstream surgical option” by a unanimous 2018 Delphi consensus of 101 experts from 39 countries, with 96% agreeing it could no longer be called experimental [93]. Note the instrument of legitimation: a structured opinion poll. The one-anastomosis gastric bypass (OAGB) entered the mainstream the same way every operation in this chapter did — by accumulated practice and expert agreement — but in doing so it stretched the anatomical variance of “the pouch” far beyond anything the RYGB literature contains, from a few centimeters of cardia to a conduit spanning most of the lesser curvature. That widened variance is why spec:000 pre-specifies OAGB as a separate, descriptively analyzed subgroup (Q6): within-OAGB variation in pouch length offers the widest exposure range in the bypass literature. The caveat is that OAGB differs from RYGB in far more than the pouch — anastomotic configuration, bile exposure, biliopancreatic limb length, and likely patient selection — so the between-procedure contrast isolates nothing about pouch geometry; only within-OAGB comparisons can speak to it, and the subgroup runs as a descriptive parallel to the RYGB analysis, not as its validation.

3.8 Consolidation, and the state of the convention

By the 2010s the Roux-en-Y gastric bypass had settled into the role it occupies in this dissertation: the reference procedure — the operation against which alternatives are trialed, the most common revisional destination [88], and the anatomy on which the field’s mechanistic research program was built (Chapter 5). Its technical canon — vertical lesser-curve pouch, stapled gas-

trojejunostomy, Roux limb measured in round hundreds of centimeters — is the sediment of every episode above: Mason’s small-fundic-segment doctrine, ulcer avoidance, Torres’s vascular argument, the stapler’s quantization, and the survivorship of whatever did not cause visible complications. What the canon conspicuously lacks is a prescribed pouch size. The 2022 joint ASMBS–IFSO indications statement — the document that formally replaced the 1991 NIH criteria — specifies who should be offered surgery, not how the operation should be built; it is silent on pouch dimensions [8].

How, then, do surgeons choose? The most direct evidence is the 2026 worldwide survey of IFSO members — 245 RYGB surgeons across 47 countries — conducted precisely to map technical variation [7]. Its findings are a portrait of convention operating without a standard. Only 53.5% of respondents routinely measure the size of the gastric pouch at all; the median stated pouch length among those who do is 6 cm (interquartile range 5–7); and most describe their pouch with subjective labels the authors note lack consistent anatomical definitions [7]. The full distribution — landmarks, bougie calibers, descriptors — is tabulated in Section 4.3 (Table 4.1); the historical point is the pattern: in the survey’s matched comparison of 60 low-volume and 60 high-volume surgeons, no significant volume-based differences were detected in pouch measurement, stated length, or termination level, and the authors conclude that experience alone does not ensure procedural standardization [7]. (Absence of a detected difference in samples of this size is not proof of equivalence; but nothing in the data suggests variability is an apprentice phenomenon that experience cures.) The survey’s own conclusion is the one this chapter has been building toward: pouch size is “reported using heterogeneous proxies, including subjective descriptors, stapler-based estimates, anatomic landmarks, and non-uniform volumetric measures,” and registries should begin recording measured pouch length and calibration details so that technique–outcome relationships can be evaluated at all [7]. Contemporary technical reviews land in the same place: a pouch below roughly 40 ml “has generally been suggested as an acceptable target,” on evidence the reviewers themselves grade as heterogeneous, with reproducibility of construction likely mattering more than any numeric goal [89].

Meanwhile the evidence that should discipline the convention points in no single direction — the stalemate stated in Section 1.4 and dissected in Part II. The largest cohort ever assembled, 14,168 patients from the Scandinavian registry, found that stapler length (its pouch-size proxy) predicted excess-BMI loss at six weeks but not at one year, while each additional centimeter of stapler raised the relative risk (the registry’s own term) of symptomatic marginal ulcer at one year — a 0.9% absolute incidence — by 14% (95% CI 9–20%): a safety gradient without a detected efficacy gradient [14]. The temporal pattern invites a specific reading: pouch size behaving like a transient, governing the early response and washing out of the steady state — what a restriction effect *should* look like if restriction dominates only while other adaptations — hormonal, behavioral, anatomical — are still equilibrating, and not what a durable dose–response looks like. The reading must be held loosely: a significant association at six weeks beside a nonsignificant one at one year is not itself evidence that the two effects differ — no formal time-by-size interaction is reported, and differential precision, attrition, or outcome variance could produce the same pattern. Whether the apparent wash-out is physiology, statistical artifact, or the noise floor of a weak proxy is exactly the measurement question Q3 poses and Chapter 11 takes up. The rest of the stalemate — the Dutch extended-pouch trial won by the *longer* pouch, the low-BMI diabetic cohort won by the smaller, and the sole systematic review unable to pool any of it — is stated in the studies’ own numbers in Section 1.4 [13, 18, 15].

Why did dose-finding trials come so late, and so rarely? Part of the answer is structural. Surgical parameters are hard to randomize: the assessment of surgery is challenged by factors that depend on operator, team, and setting — learning curves and quality variations, in the surgical-evaluation literature’s own accounting [94] — so the intervention is entangled with the operator and technique drifts over any trial’s horizon; that no sponsor stands to profit from knowing the optimal pouch volume the way one profits from a drug dose is this chapter’s

conjecture, offered as conjecture. Part of the answer is historical, and it is this chapter's: by the time trial infrastructure existed, the parameter no longer looked like a question. A surgeon trained on a small pouch has seen thousands of them work; the doctrine is confirmed daily by outcomes that — as the mechanistic literature of Chapter 5 argues — may owe little to the pouch's dimensions at all. Convention does not merely fill the space evidence leaves empty; it removes the sensation that evidence is missing. The audit of that convention begins in Chapter 8.

3.9 The genealogy of a number

It is worth naming the machinery by which these decisions congealed, because the machinery, not any single decision, is the chapter's finding. Four mechanisms recur. *Anatomy*: the lesser-curvature blood supply and the angle of His define where a reliable pouch can be built, so the feasible design space was never continuous [74]. *Instruments*: stapler cartridge lengths, bougie calibers, and circular-stapler diameters quantize the space further, and what the instrument makes easy becomes what the profession does [6, 7]. *Complication avoidance*: nearly every downward revision of the pouch traces to ulcers, leaks, vomiting, or failed weight loss in a recent series — feedback control on the last observed error, with all the stability properties of a controller driven by noisy, one-sided measurements [4, 63, 73]. *Authority and training*: the configurations of high-volume pioneers were reproduced by their trainees and ratified, when ratification came at all, by consensus statements and Delphi polls rather than trials [72, 93]. None of these mechanisms is irrational; each is how a craft transmits hard-won safety knowledge. But their joint output is a parameter distribution shaped by history rather than by outcome gradients — and a literature in which the exposure variable of interest was recorded, when recorded at all, in whatever unit the era's machinery happened to emit.

Table 3.1 compresses this chapter into its through-line: fifty years of stated pouch parameters, each with the rationale or evidentiary context its source supplies. Read down the final column. Ulcer avoidance; technical simplicity; vascular reliability; stapler geometry; complication management; teachability. Define dose-finding strictly — a prospective comparison in which the pouch parameter itself is varied by design against a weight outcome — and it appears exactly once in the genealogy, in the modern era: the pouch-length trial whose result contradicts the doctrine [13]. Everything earlier is observational approximation — Mason's iterative revisions, the learning-curve series, Andersen's measured null [78], and Näslund's within-arm correlations, present after gastropasty and absent after bypass [80]. The pouch number was born as an ulcer-avoidance measure with a mechanical story attached, shrank each time a large-pouch series disappointed, was frozen by the stapler and the training pipeline, and now persists as a convention that half the world's surgeons do not measure and no guideline specifies. The doctrine, in short, is a fossil of the restrictive model — and whether the fossil records a real organism is a question that construction history cannot answer. Answering it requires knowing what the pouch actually is anatomically and physiologically (Chapter 4), how its size can be measured and how badly it usually is (Chapter 6), and what the assembled evidence says when it is finally pooled (Part II). The history having shown where the number came from, the dissertation now asks whether it matters.

Table 3.1: Stated gastric pouch (reservoir) parameters, each with the stated rationale or evidentiary context its source supplies (some entries are study designs or reported practice rather than contemporaneous reasons for the parameter choice). Values are as published; modalities range from intraoperative volumetric calibration to subjective description — the measurement heterogeneity formalized in Section 6.9.

Year	Report	Stated pouch parameter	Rationale / context
1975	Mason et al. [4]	“small fundic segment”; 12 mm stoma	weight loss; stomal ulcer
1977	Alden [61]	90% of stomach bypassed	safer alternative to JIB
1977	Griffen et al. [62]	90% exclusion; Roux-en-Y drainage	randomized against JIB
1979	Pace et al. [77]	50–60 cc reservoir; 1 cm gap	technical simplicity
1981	Halverson et al. [72]	50–60 cc; 1–1.2 cm stoma	“operation of choice”
1982	Kirkpatrick & Siegel [5]	<45 ml reservoir; stoma size free	reservoir doctrine
1982	Mason (VBG) [79]	<50 ml at 70 cm H ₂ O; collar	measured calibration
1983	Torres et al. [74]	25–35 ml, lesser curve	blood supply; stapled stoma
1984	Andersen & Pedersen [78]	measured 51–244 ml (gastroplasty)	no outcome correlation
2000	Sundbom & Gustavsson [85]	4 × 3 cm	stapler geometry
2008	ASBS survey [6]	mean 25 cm ³ ; 49% measure by length	reported practice
2016	SOReg cohort [14]	stapler length, mean 145 mm	registry proxy
2020	Boerboom RCT [13]	5 vs. 10 cm randomized	regain hypothesis
2026	IFSO survey [7]	median 6 cm (IQR 5–7); 53.5% measure	convention, unstandardized

Chapter 4

The gastric pouch: anatomy, construction, physiology

Chapter 3 traced where the pouch’s number came from; this chapter turns to the thing the number describes. Every other chapter of this dissertation treats the gastric pouch as a variable. This one treats it as an object. Before pouch size can be regressed against weight loss, three prior questions must be answered honestly: what tissue is actually in the pouch, how the pouch comes to have the size it has, and through what physiology a change in that size could plausibly propagate to body weight a year later. The answers turn out to constrain the empirical question sharply. The pouch is not a uniform object cut to a specification; it is a compliant, secretory, innervated remnant whose “size” is a family of loosely coupled quantities, produced by a construction sequence that is coarsened by instrument geometry and calibrated against landmarks that surgeons locate differently. And the physiology admits not one but two mechanistic hypotheses by which size could matter — hypotheses that make opposite predictions about which direction is better. That double prediction, developed in Section 4.4, is the steelman the empirical chapters of Part II must respect: it is not the case that modern mechanistic understanding predicts pouch size is irrelevant, and any analysis that assumes so is arguing against a weaker opponent than the one actually on the field.

4.1 Surgical anatomy of the proximal stomach

The stomach is conventionally divided into cardia, fundus, body (corpus), antrum, and pylorus. The distinctions are simultaneously topographic and histological, and for present purposes the histology matters more. The *cardia* is the short collar of mucosa immediately distal to the gastroesophageal junction, lined by mucus-secreting cardiac glands and containing few or no acid-producing cells. The *fundus* is the dome that rises above and to the left of the gastroesophageal junction under the left hemidiaphragm; it and the *corpus* share the same oxyntic (acid-producing) mucosa, populated by parietal cells that secrete hydrochloric acid, chief cells that secrete pepsinogen, and a set of hormone-producing enteroendocrine cells. The *antrum* is the distal, thicker-walled milling chamber whose mucosa contains the gastrin-producing G cells that drive acid secretion in the proximal stomach by an endocrine loop. The pouch of a Roux-en-Y gastric bypass (RYGB) is intended to be carved from the cardia and the proximal, lesser-curvature portion of the corpus; the fundus is deliberately excluded — though, as the next subsection shows, incompletely mobilized fundus can remain attached to the pouch unrecorded — and the antrum, with its G-cell mass, stays behind in the excluded remnant, still perfused, still secreting, still in communication with the duodenum, but no longer in the food stream.

Three anatomical facts govern how the pouch is constructed and what it subsequently does.

4.1.1 Blood supply and the tension between vascularity and size

The proximal stomach receives arterial inflow chiefly from the left gastric artery, whose ascending branches run along the lesser curvature and supply the cardia and the upper lesser-curve wall, and from the short gastric arteries arising from the splenic artery, which supply the fundus. Constructing a lesser-curvature pouch requires dividing the gastrohepatic (lesser omental) attachments and the vessels entering the stomach at the level chosen for the horizontal staple fire, then dividing the fundic attachments — including short gastric branches and the posterior fundic attachments to the diaphragm — so that the fundus can be excluded rather than left in continuity with the pouch. This dissection is the single most under-described step in the operation. Kassir and colleagues, in an anatomical analysis addressed explicitly to the question of why gastric pouches expand, noted that the upper fundus lies above the omental bursa and is therefore effectively retroperitoneal, and that surgeons who mobilize only until the left diaphragmatic crus becomes visible commonly leave a portion of fundus still adherent to the diaphragm — and therefore still attached to the pouch — while believing the release to be complete.[95] The consequence is a pouch that contains a tongue of fundus that neither the operative note nor the stapler count records.

The vascular constraint plausibly runs in the opposite direction from the size constraint. Whatever lesser-curve vessels are divided at the level of the horizontal fire no longer supply the pouch, so the perfusion of the distal pouch and the gastrojejunal anastomosis depends on which ascending branches remain intact, on staple-line geometry, and on anastomotic tension — an anatomical argument, it should be said, rather than a measured perfusion gradient; no direct measurement of pouch perfusion as a function of pouch length has been identified in this literature. Marginal ulceration — ulceration of the jejunal mucosa just distal to the anastomosis — has been attributed since the earliest series in part to relative mucosal ischemia at the staple line, alongside acid exposure and staple-line dehiscence.[96] Pouch construction therefore sits on a genuine trade-off surface rather than a monotone “smaller is better” gradient, a point Section 4.7 develops with the quantitative safety data.

4.1.2 Innervation

The anterior and posterior vagal trunks descend on the esophagus and divide at and below the gastroesophageal junction; the hepatic and celiac branches leave early, and the anterior and posterior nerves of Latarjet run distally along the lesser curvature within the lesser omentum, terminating in the fan-shaped terminal branches at the incisura angularis known as the *crow’s foot* (*patte d’oie*). Two observations follow. First, the crow’s foot is a visible operative landmark near the incisura, used as such by surgeons constructing long pouches and one-anastomosis bypasses — a surgical waypoint, not a validated histological boundary of the corpus. Second, it is plausible that a pouch extending far distally along the lesser curvature divides more of the Latarjet branches supplying the pouch wall than a pouch constructed high on the lesser curvature — though the realized innervation of any pouch depends on how the lesser omentum is taken down, and no study has measured pouch innervation as a function of pouch length; the claim is anatomical inference, not demonstrated dose-response. Since vagal afferents carry the gastric distension signal centrally and vagal efferents mediate receptive relaxation, the innervation of the pouch is not incidental to its function as a satiation sensor (Section 4.5).

4.1.3 The mucosa that ends up in the pouch

Although the pouch is intended to be cardia and proximal corpus, its mucosa at the anastomotic margin is predominantly oxyntic, and this is not a theoretical inference. Siilin and colleagues retrieved the gastric suture ring from the circular-stapled gastrojejunostomy in 23 consecutive RYGB patients and examined it histologically: every ring from a standard 4×3 cm pouch

consisted of corpus-fundus mucosa densely populated with parietal cells, histologically indistinguishable from reference biopsies taken from the gastric corpus. Deliberately reducing the pouch to 2×3 cm in a subsequent ten patients did not eliminate oxyntic mucosa; all ten modified rings still contained corpus-fundus mucosa, with cardiac mucosa present in only five and then only across a segment of about 6 mm.[97] The conclusion this small study (23 and 10 patients, two selected pouch geometries, sampling only the anastomotic margin) can actually carry is narrower than it is sometimes made to carry: acid-producing mucosa is commonly present at the distal margin of the pouch even when the pouch is deliberately made small. That the *total* oxyntic mucosal area retained in the food stream scales with pouch size is a geometric inference, plausible but not measured by this study — and no direct measurement of total pouch oxyntic area has been identified in this literature.

Total acid output from the pouch is nevertheless low in absolute terms. Smith and colleagues measured basal and pentagastrin-stimulated peak acid output from the proximal pouch after vertical Roux-en-Y gastric bypass in ten patients using a perfused orogastric tube: basal output was 0.01 ± 0.01 mEq/h versus 4.97 ± 0.66 mEq/h in matched controls, and peak stimulated output 0.08 ± 0.04 mEq/30 min versus 12.11 ± 1.34 ($p < 0.001$ for both).[69] Low is not zero, and the acid that is produced sits in a compartment with no duodenal bicarbonate buffering: Hedberg and colleagues passed a pH probe into the proximal pouch of six patients with endoscopically confirmed stomal ulcer and six asymptomatic controls and found the ulcer group’s pouches exposed to $\text{pH} < 4$ for a median 69% of the four-hour recording versus 20% in controls ($p < 0.01$).[98] That study compares six ulcer patients with six controls; low pH could contribute to ulceration, result from it, or merely accompany it, and the design cannot apportion attributable risk. What the two measurements together support is more modest: acid output from the pouch is quantitatively small, but the acid that is produced reaches jejunal mucosa — tissue that normally receives chyme already neutralized by pancreatic and biliary bicarbonate and lacks the gastric mucosa’s protective adaptations — making unbuffered acid exposure a plausible contributor to marginal ulceration alongside smoking, NSAIDs, ischemia, and foreign material, not a demonstrated decisive cause.

Finally, the pouch is depleted of one endocrine cell population that has dominated bariatric physiology for two decades. Ghrelin-producing cells are distributed throughout fundic and corpus mucosa, sitting basally in the glands near parietal cells as closed-type cells not in contact with the lumen;[99] quantitatively, ghrelin mRNA expression in the human stomach is graded along its long axis, with a ghrelin/actin ratio of 0.78 in the upper (fundus and upper body) portion, 0.20 in the middle, and 0.07 in the antrum and pylorus ($p = 0.001$).[100] Excluding the fundus therefore removes the region of highest ghrelin cell density from the food stream — though, as Section 4.6 shows, exclusion is not resection, and the empirical consequences are far less tidy than the anatomy suggests.

4.2 Construction, step by step

The modern laparoscopic RYGB pouch is built in a fixed sequence whose individual steps each admit variation. Describing the sequence precisely is a prerequisite for understanding why “pouch size” as reported in the literature is not one measurement but several.

Step 1: opening the lesser sac. The gastrohepatic ligament is opened on the lesser curvature at a chosen distance below the gastroesophageal junction. The choice of level is the single most consequential decision in the operation for pouch size, and it is made against soft landmarks. Contemporary practice descriptions in the corpus locate this entry point variously “at the crow’s foot position,” “1–2 cm below the incisura angularis (area of the crow’s feet),” or by counting vessels: between the first and the second (“middle”) gastric vein encountered ascending the lesser curvature.[7] In the OAGB variant, expert consensus places the entry lower still: 94% of a

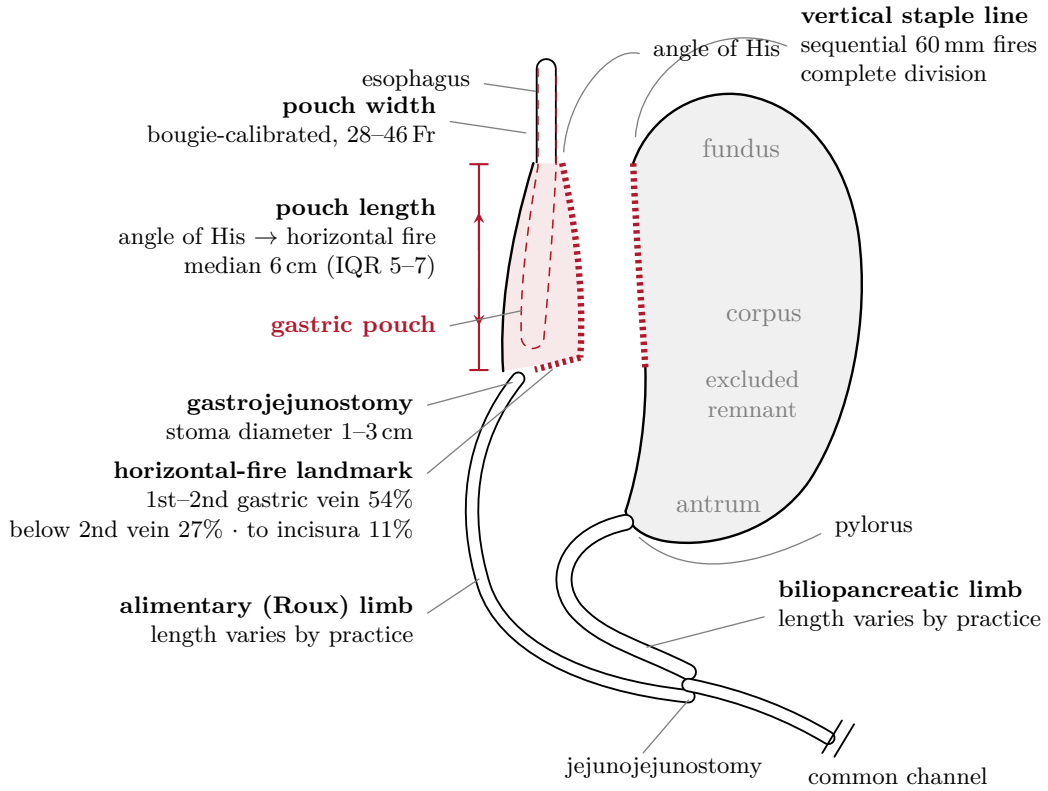


Figure 4.1: Roux-en-Y gastric bypass anatomy, drawn schematically, with the construction parameters that vary between surgeons: pouch length from the angle of His to the horizontal fire (median 6 cm, IQR 5–7, among surgeons who measure), pouch width set by bougie calibration (28–46 Fr in practice), the lesser-curve landmark at which the horizontal fire terminates (between the first and second gastric vein in 54% of surveyed surgeons, just below the second in 27%, down to the incisura in 11%), stoma diameter (1–3 cm), and the alimentary and biliopancreatic limb lengths. Parameter distributions are those of the 2026 IFSO worldwide survey [7]; the dissertation’s exposure variable is the pouch dimension set, while limb lengths and stoma diameter enter as confounders (Chapter 8). Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

101-expert modified-Delphi panel agreed that pouch construction should begin in the horizontal portion of the lesser curvature.[93]

Step 2: the horizontal fire. A linear stapler — typically a 45 mm cartridge — is fired perpendicular to the lesser curvature, creating the caudal boundary of the pouch. This fire sets pouch length.

Step 3: calibration. A bougie or orogastric tube is passed transorally and positioned along the lesser curvature; the subsequent vertical fires are laid against it. This sets pouch width, and by setting width it sets, via geometry, most of pouch volume for a given length. Calibration is not universal: in Sweden’s highly standardized practice the pouch is built without a calibrating tube and stapler length is recorded instead,[14] whereas in most other practices a tube of 28–46 Fr is used (Section 4.3).[7]

Step 4: fundic mobilization and the vertical fires. Successive linear fires — typically 60 mm cartridges — run cephalad along the bougie toward the angle of His, transecting the stomach completely and creating a vertical, lesser-curvature-based, divided pouch. The number of fires needed is the discrete quantity that registry proxies record. Complete division rather than stapling in continuity is standard contemporary practice — non-divided and partially divided variants persist mainly in historical and revisional material — and the lesson was learned the hard way: Capella and Capella reported gastrogastic fistula — recanalization between pouch and excluded remnant — in 49% of 189 patients whose stomachs were stapled in continuity or only partially divided, falling to 2.6% with transection and to zero when a jejunal limb was interposed between pouch and remnant.[101] Adequate release of the fundus is required for the last fire to reach the angle of His rather than leaving a fundic cuff at the top of the pouch;[95] the micropouch technique of Sapala and colleagues took this to its logical extreme, mobilizing the fundus completely — transecting the left phreno-esophageal and gastrophrenic ligaments — and limiting the transected pouch to the gastric cardia on the explicit reasoning that the fundus is elastic, aperistaltic, and dilates over time, and that the proximal lesser-curve wall carries a high parietal-cell concentration that potentiates marginal ulceration.[102]

The lesser-curvature orientation that all of this presupposes was itself a choice, and a recent one. Torres and colleagues described partitioning the stomach from the angle of His toward the lesser curvature to produce a 25–35 mL proximal pouch drained by a Roux limb, arguing the case on vascular integrity and ease of stoma construction rather than on weight-loss physiology.[74] The vertical lesser-curve pouch that is now standard was adopted because the lesser curvature is the least distensible part of the stomach and because the geometry suited the stapling instruments — not because a trial compared it with the alternatives.

Step 5: the gastrojejunostomy. The alimentary limb is brought up (antecolic or retrocolic) and anastomosed to the distal pouch by linear stapler, circular stapler, or hand suture. This step does not set pouch size, but it sets the pouch *outlet*, which Section 4.4 argues is plausibly the dominant resistance in the emptying circuit, and it sets much of the complication profile (Section 4.7).

Step 6: reinforcement, leak test, defect closure. Optional staple-line reinforcement with sutures or clips; intraoperative leak testing by methylene blue, air insufflation, or endoscopy; closure of Petersen’s space and the jejunojejunal mesenteric defect.

Two features of this sequence deserve emphasis because they propagate directly into the measurement problem of Chapter 6. First, the realized volume of the pouch is rarely measured directly at the time it is made. Some surgeons measure a length (Section 4.3), and some investigators have recorded intraoperative dimensions prospectively, but what construction fixes

is a set of proxies — a landmark, a measured or estimated length, a bougie caliber, a count of cartridges — and the mapping from proxy to realized volume passes through the patient’s gastric wall thickness, mesenteric tension, body habitus, and the surgeon’s traction on the stomach during firing. Second, the length dimension is coarse. Stapler cartridges come in fixed lengths, fires overlap and angle, and the first fire begins at a continuously varying entry point, so cartridge count maps only loosely onto final staple-line length — it is a coarse construction proxy, not a discrete measurement of pouch length. The randomized trials in this literature work at exactly this resolution: the standard arm of the Dutch extended-pouch trial fired its first stapler 5 cm below the angle of His and completed the pouch with two further 60 mm cartridges against a 40 Fr tube, and the extended arm fired at 10 cm and used three.[13, 16] The trial’s contrast is a construction package — first-fire position plus one additional cartridge — not a single dimension varied in isolation.

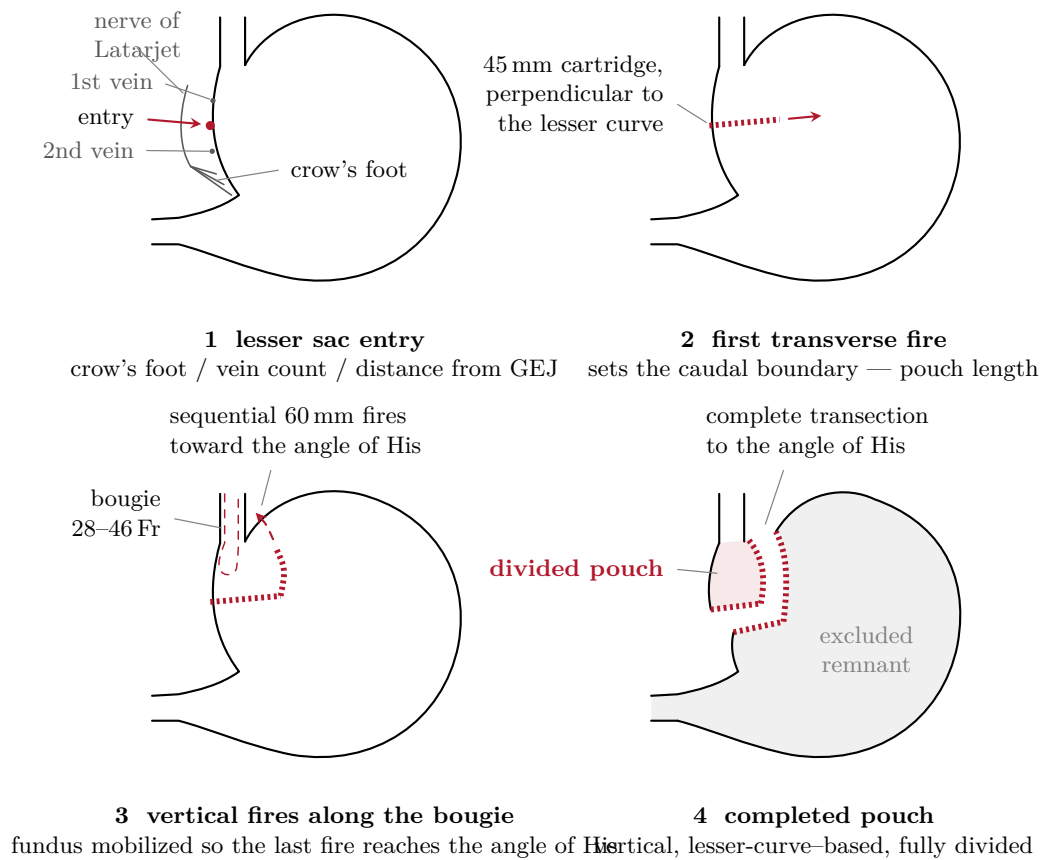


Figure 4.2: The pouch construction sequence in four steps, with the decision points at which inter-surgeon variation enters. (1) Lesser sac entry on the lesser curvature, located against soft landmarks — the crow’s foot of the nerves of Latarjet, a gastric-vein count, or a distance below the gastroesophageal junction. (2) The first transverse stapler fire, perpendicular to the lesser curvature, which sets pouch length. (3) Sequential vertical fires laid against the calibrating bougie toward the angle of His, with the fundus mobilized so the last fire reaches it. (4) The completed vertical, lesser-curve-based, fully divided pouch, separated from the excluded remnant. Technique descriptions per Section 4.2 [7, 13]. Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

4.3 What varies between surgeons, and what the instruments constrain

The dissertation’s exposure variable is a surgeon’s choice. It is therefore essential to know the empirical distribution of that choice, and — more sharply — how much of the observed between-study variance in reported pouch size is real anatomical variance versus variance in how the same anatomy is described.

The most recent and largest characterization comes from a worldwide survey of International Federation for the Surgery of Obesity and Metabolic Disorders (IFSO) members: 245 eligible responses from surgeons in 47 countries across all five IFSO regional chapters, with a median 14 years of metabolic and bariatric surgery experience and a median 50 primary RYGB procedures per year as lead surgeon.[7] Its findings on the pouch are summarized in Table 4.1, and three of them reframe the dissertation’s question.

First, **only 53.5% of surgeons report routinely measuring pouch length** — the survey item asks about routine length measurement, so the remainder may still calibrate dimensionally (bougie, landmarks) without measuring a length. Second, among those who do measure, the median reported pouch length is 6 cm (IQR 5–7), with a statistically significant difference across IFSO chapters ($p = 0.034$) — 5 cm (IQR 4–6) in Latin America against 7 cm (IQR 6–8) in Asia-Pacific. Third, and most corrosive for the literature, **the subjective descriptors do not partition the same way the measurements do**: 67.3% of respondents described their pouch as “short or small,” 30.6% as “long,” and 2.0% as a “micro pouch,” yet the underlying median lengths of the groups are close enough that a study reporting a “standard small pouch” and a study reporting a “long pouch” may be describing overlapping anatomy. The termination landmark varies correspondingly: 53.9% end the pouch between the first and second gastric vein, 27.3% just below the second gastric vein, 11.0% between that and the incisura, and 7.8% between the hiatus and the first gastric vein. Notably, none of these distributions differed significantly between low- and high-volume surgeons in a propensity-matched comparison. A cross-sectional survey cannot establish what does drive pouch practice, and failure to detect a difference is not proof of no experience effect; but the regional gradients, at least, are directly observed.

The 2008 survey of 215 practicing American Society for Bariatric Surgery members provides the historical comparison and shows how little the qualitative picture has changed in nearly two decades: mean stated pouch volume 25 cm³, with roughly half of surgeons (49%) sizing the pouch by distance from the gastroesophageal junction, and the gastrojejunostomy split roughly three ways between circular stapler (43%), linear stapler (41%), and hand suture (21%) — figures that sum to more than 100% as published, so they cannot be read as mutually exclusive practice shares.[6] Its authors concluded that technical variation existed and that research was needed to link the variations to outcomes. That research is what this dissertation attempts.

The most striking evidence that “pouch size” is partly a property of the observer rather than the patient comes from comparing two single-center cohorts that measured the same quantity by the same method. Roberts and colleagues measured pouch area on postoperative day 1 upper gastrointestinal contrast radiographs at maximum distension in 320 laparoscopic RYGB patients and reported a mean of 63.9 cm². [103] Campos and colleagues measured pouch area on postoperative day 1 contrast radiographs at maximum distension in 310 gastric bypass patients and reported a mean of 26 cm² (range 7–78). [104] Campos’s group, noting that Roberts’s mean was “more than double what we measured in our series,” attributed the gap both to measurement variability between the two studies and to genuine differences in how different surgeons in different centers create a pouch — and could not separate the two. A 2.5-fold difference in the central estimate of the exposure variable between two contemporaneous American academic centers using nominally the same imaging protocol is the single most important fact a meta-analyst in this field must hold in mind, and it is why Chapter 6 defines a modality-quality hierarchy before any data are pooled.

Table 4.1: Inter-surgeon variation in RYGB pouch construction. Source: worldwide survey of 245 IFSO-member surgeons across 47 countries [7], except the last block, from a survey of 215 American Society for Bariatric Surgery members [6]. Percentages are of respondents; lengths are median (IQR). The gastrojejunostomy-technique percentages in the last block sum to 105% as published in the source; the source does not state whether respondents could select more than one technique.

Domain	Practice	Value
Measurement	Routinely measures pouch length	53.5%
	Median measured length (cm)	6 (5–7)
	Chapter-level range of medians (cm)	5–7
Description	“Short or small” pouch	67.3%
	“Long” pouch	30.6%
	“Micro” pouch	2.0%
Landmark	Hiatus to first gastric vein	7.8%
	First to second gastric vein	53.9%
	Just below second gastric vein	27.3%
	Below that, to incisura	11.0%
Calibration	Bougie 36 Fr (modal)	41.2%
	Bougie 32 Fr	13.5%
	Bougie 40 Fr	10.6%
	Bougie range observed (Fr)	28–46
Reinforcement	Routinely reinforces pouch staple line	22.8%
Anastomosis	Linear-stapled gastrojejunostomy	81.2%
US survey (2008)	Mean stated pouch volume (cm ³)	25
	Sizes pouch by distance from GEJ	49%
	Circular / linear / hand-sewn GJ	43% / 41% / 21%

A complementary line of evidence shows how narrow the realized variation can be once a single surgeon standardizes. O'Connor and Carlin collected intraoperative pouch dimensions prospectively in 231 consecutive laparoscopic gastric bypasses, building a divided vertical lesser-curve pouch against an orogastric tube and recording pouch height as the number of staple fires; the calculated volumes spanned only 10–20 cm³, and %EWL at one year did not differ across that range.[105] A study of that design has severely attenuated power against any effect operating across the wider range surgeons actually produce: with almost no exposure variance, only a very large per-millilitre effect would be detectable, so its null is weak evidence — an instance of the restriction-of-range problem that recurs throughout this literature and that Chapter 9 quantifies study by study. Conversely, Madan and colleagues had three blinded experts grade postoperative contrast studies in 59 patients on separate ordinal scales for pouch size and for the amount of *fundus* visible in the pouch — an explicit acknowledgment that retained fundus is a distinct construction defect from a large pouch, and one that no numerical size measurement distinguishes.[106]

One further variation deserves separate mention because it dissolves the one-dimensional framing of the whole literature. Length and width are independent construction parameters — the first set by where the horizontal fire lands, the second by the bougie — and they are not interchangeable determinants of behavior. Mahawar argued explicitly that it is the width and not the size of the pouch that matters,[107] and the argument has a physical analogy behind it: by Laplace's relation, wall tension for a given internal pressure scales with radius, suggesting — though tissue remodeling, not instantaneous tension, is what actually governs dilation — that a long narrow pouch and a short wide pouch of identical volume may differ in their propensity to dilate. The Iranian long-versus-short comparison discussed in Section 4.4 built its “long” 80 cc pouch over a narrow 36 Fr bougie and found no penalty in weight loss at one year, which its authors read as direct support for the width hypothesis.[108] A literature that reports only “pouch size” has projected a two-dimensional exposure onto one axis, and the direction of projection differs between studies. The long-pouch Roux-en-Y gastric bypass developed by Ribeiro and colleagues makes the point institutionally: a deliberately long pouch combined with a 100 cm alimentary limb and a long biliopancreatic limb, adopted to capture the metabolic advantages associated with one-anastomosis bypass while retaining Roux-en-Y protection against bile reflux, and reporting 30.5% total weight loss at four years in 300 cases with 90% follow-up.[109] “Long pouch” here denotes a different operation, not merely a longer staple line — a confounding of pouch length with limb configuration that the inclusion criteria of Chapter 8 must handle explicitly.

4.4 Emptying dynamics, and the steelman for size dependence

Here is where the chapter earns its place in the argument. The conventional framing of this literature is that the restrictive model predicts a pouch-size effect and the modern neuroendocrine model predicts none, so that a null pooled result would simply confirm what mechanism already knows. That framing is wrong, and defeating it is necessary before the analysis can claim to have tested anything.

4.4.1 What the pouch actually does to nutrient flow

RYGB removes the pylorus from the food stream and replaces it with a gastrojejunal anastomosis of 1–3 cm diameter, held open by its own construction. The physiological consequence is dramatic and has been measured. Dirksen and colleagues studied 17 patients more than a year after RYGB and nine healthy controls with dual-isotope scintigraphy: at meal completion a smaller fraction of both liquid and solid radiolabeled marker was retained in the pouch of the patients than in the stomach of controls ($p = 0.002$ and $p < 0.001$), pouch emptying was faster

than gastric emptying in controls for both markers ($p < 0.001$ liquid, $p = 0.004$ solid), small intestinal transit of the solid marker was *slower* in patients ($p = 0.034$), and — the mechanistic hinge — GLP-1 and PYY₃₋₃₆ secretion was higher in patients, with fast pouch emptying of the liquid marker associated with high gut hormone secretion.[10] Stano and colleagues quantified the acceleration directly in a within-subject design: 21 patients studied before and one year after RYGB emptied the pouch 2.4 times faster after surgery, and meal texture strongly modulated the downstream endocrine response — a 600 kcal liquid meal produced a peak GLP-1 of 106.1 ± 67.2 pM against 45.3 ± 25.2 pM for the isocaloric solid meal ($p = 0.004$), with correspondingly more frequent hypoglycemia after the liquid meal.[110]

This is the mechanistic bridge the synopsis names. In the neuroendocrine account of RYGB — the account that has substantially displaced pure restriction and malabsorption as the presumed dominant mechanism, though RYGB physiology remains multifactorial and the hormonal changes have not been shown to account for the full durable weight effect[9] — a major driver of weight loss is a set of gut-derived satiety signals (GLP-1, PYY, and others) released when nutrients reach distal small bowel enteroendocrine cells rapidly and in high concentration. The pouch, in this account, is not primarily a container that restricts intake. It is the *antechamber that sets the arrival waveform*: the temporal profile with which a swallowed meal is delivered to the jejunum. And arrival waveform is exactly the kind of quantity that a pouch’s geometry could plausibly modulate. Under the neuroendocrine model, then, pouch size is not predicted to be irrelevant: emptying is one candidate channel through which it could matter — alongside the distension, innervation, and secretory channels this chapter treats separately.

4.4.2 Two models, both predicting size dependence, with opposite signs

The two live hypotheses can be stated as competing claims about the transfer function from pouch geometry to weight loss. Both are stated here as they appear in the surgical literature; neither of their geometry-to-emptying premises has been demonstrated, and each link in each chain is an assumption requiring direct mediation evidence (Section 4.4 reviews what little exists).

Model A — fast-emptying / hindgut amplification (hypothesis). A smaller pouch is posited to hold less, empty faster, and deliver each meal to the jejunum as a sharper bolus; a sharper bolus produces a larger incretin and PYY excursion; larger excursions produce stronger satiation and greater weight loss. None of these steps is automatic — emptying depends on outlet diameter, pressure, meal rheology, compliance, and motility at least as much as on reservoir capacity. Prediction, if the chain holds: *smaller is better*, and the mediator is emptying rate, not capacity.

Model B — slow-emptying / sustained signaling and dilation resistance (hypothesis). A longer, narrower pouch is posited to offer more resistance and empty more slowly — though a pouch is a compliant reservoir, not a uniform conduit, so added length does not automatically add resistance; slower emptying produces a longer, more gradual period of gut hormone secretion and a longer-lasting sensation of fullness; and by analogy with Laplace’s relation the narrower tube experiences lower wall tension for the same internal pressure and might therefore dilate less over time, preserving whatever effect it has. Prediction, if the chain holds: *longer and narrower is better*, with the benefit appearing late, as a difference in recurrent weight gain rather than in nadir weight loss. This is explicitly the reasoning of the Dutch extended-pouch program, which invoked Poiseuille’s and Laplace’s laws by name.[16]

Both hypotheses are size-dependent. They differ in sign, in the mediator, and in the time-point at which the effect should be visible. A pooled analysis that finds no overall association is therefore *not* automatically evidence against mechanism; it is equally consistent with a mixture of studies in which the two effects operate in opposite directions in different populations and cancel. This is the steelman, and Part II must engage it, not the strawman that “modern physiology says pouch size does not matter.”

4.4.3 What the emptying data actually show

The direct evidence on emptying and weight loss is small, heterogeneous, and contradictory in a way that maps neatly onto the two models (Table 4.2).

Against Model B and for Model A: Akkary and colleagues classified 304 patients by the presence or absence of contrast emptying from the pouch on routine postoperative day 1 upper gastrointestinal study, and found that the 188 patients with normal emptying lost more weight at one year than the 116 with very slow or absent emptying (50.6 versus 47.3 kg; BMI loss 18.1 versus 16.6 kg/m²; $p = 0.007$ adjusted for age, sex, and operative time) — a result they named “the paradox of the pouch” precisely because it inverted the restrictive expectation.[111] Riccioppo and colleagues combined three-dimensional pouch volumetry with four-hour scintigraphy in 67 patients at a median 47 months and found the two hypothesized links in series: pouches ≤ 40 mL emptied faster than pouches > 40 mL (1-hour retention 6% versus 44%), and patients with 1-hour retention $\leq 12\%$ had higher %EWL ($p = 0.033$) and less %EWL regain ($p = 0.036$) than those with retention $\geq 25\%$, as well as better food tolerance ($p = 0.003$).[48] Both results are hypothesis-generating rather than mediation evidence: day-1 contrast passage is dominated by edema, spasm, and perioperative conditions and is not a validated proxy for chronic emptying behavior, Akkary’s outcome is absolute weight lost (baseline-sensitive despite the listed adjustments), and Riccioppo’s late cross-sectional measurements are open to reverse causation — weight trajectory and eating behavior may enlarge the pouch rather than the reverse.

For Model B and against Model A: Deden and colleagues performed solid-meal pouch-emptying scintigraphy in the five women with the least and the five with the most weight loss at their center 2.5 years after RYGB. The poor-weight-loss group showed fast initial emptying with an exponential curve; the successful group showed slower initial emptying with a more linear curve. Intestinal content at meal completion was $42 \pm 18\%$ in the poor group versus $4.0 \pm 3.3\%$ in the successful group, and maximal pouch emptying rate 17 ± 4.7 versus $5.6 \pm 3.4\%/min$ — fast emptying with poor weight loss, the direction opposite to Riccioppo. The paper also reports a linear correlation between total weight loss and maximal emptying rate of Pearson $R = 0.82$ ($p = 0.004$, ten patients), a coefficient whose positive sign is inconsistent with the paper’s own group data and is best read as a magnitude; and because the ten patients were selected as the five best and five worst outcomes at the center, the correlation is mechanically inflated and cannot be treated as an ordinary population estimate.[112]

Neither: the two oldest studies, using scintigraphy in the open-surgery era, found nothing. Horowitz and colleagues studied 12 patients a year after gastric bypass against 11 controls and found solid emptying slower and liquid emptying faster after surgery, but no correlation between weight loss and emptying rate, stoma size, or pouch size.[113] Näslund and Beckman randomized 57 patients to gastric bypass or gastroplasty and measured emptying at 2 and 12 months; pouch emptying after bypass was slower at 2 months than at 12, stoma diameter correlated with emptying rate after gastroplasty but not after bypass, and no correlation between emptying rate and weight loss was found in either group.[114]

4.4.4 Where the physics is being abused

Both mechanistic arguments as stated in the surgical literature contain a physical error worth naming, because the mechanistic discussion in Chapter 14 depends on getting it right.

The invocation of Poiseuille’s law to argue that a longer pouch empties more slowly treats the pouch as the resistive element of the circuit.[16] For laminar flow in a rigid cylindrical tube, volumetric flow rate scales as r^4/L — inversely with length, but with the *fourth power* of radius. Applied as an intuition pump — and it can be nothing more, since a meal is a multiphase, non-Newtonian material moving through a compliant, irregular reservoir and anastomosis, and no pressure-flow measurements of the post-RYGB circuit exist — the scaling suggests where to look: at distension the pouch has a diameter of roughly 3–4 cm and the gastrojejunal stoma

Table 4.2: Studies relating pouch emptying to weight loss after gastric bypass. “Model” indicates which of the two mechanistic hypotheses in Section 4.4 the result is directionally consistent with. The studies differ in method, era, meal, timing, and endpoint, and the smallest (Deden) deliberately sampled outcome extremes, which mechanically exaggerates separation; the apparent sharpness of the small studies should be read in that light.

Study	Method	n	Finding	Model
Horowitz 1982[113]	Dual-isotope	12	No correlation with weight loss	—
Näslund 1987[114]	Scintigraphy	49	No correlation with weight loss	—
Akkary 2009[111]	POD-1 contrast	304	Prompt emptying $\rightarrow$ more loss	A
Deden 2017[112]	Solid scintigraphy	10	Fast emptying $\rightarrow$ less loss	B
Riccioppo 2018[48]	Scintigraphy + 3D volumetry	67	Small, fast $\rightarrow$ more loss	A

1–3 cm, and where the stoma is in fact substantially narrower than the pouch, the fourth-power dependence would make the stoma, not the pouch body, the dominant resistance. That is a hypothesis about geometry, not a calculation; the ranges overlap, stoma length and pouch shape enter the same formula, and none of the quantities have been measured together in vivo. What lends the intuition empirical weight is that the observational literature finds stoma diameter a more consistent correlate of outcome than pouch size: Heneghan and colleagues, endoscoping 380 post-RYGB patients, found that pouch length, pouch volume, and stoma diameter all correlated inversely with excess weight loss on univariate analysis, but that only *stoma diameter* survived multivariable analysis as an independent correlate of weight regain,[115] and the same group’s companion series found the most common abnormality in patients referred for weight regain to be an enlarged gastrojejunostomy (58.9%) rather than an enlarged pouch (28.8%).[116] Abu Dayyeh and colleagues reached the same conclusion from a different angle: in 63 patients evaluated for weight regain after nadir, stoma diameter predicted regain ($\beta = 0.39$, $p < 0.01$) while ghrelin level did not.[117] These are associations in cohorts largely referred *for* weight regain, where selection and reverse causation are live — stoma dilation may be a consequence or marker of eating behavior rather than its cause — and none of them measures flow resistance directly. The working hypothesis they support, stated with that caveat, is that the stoma rather than the pouch is the throttle of the reconstructed circuit.

The idea is not new; it is merely unheeded. Kirkpatrick and Siegel set out in 1982 to test the then-standard teaching that both a small reservoir and a narrow stoma were required for a successful gastric bypass. Holding reservoir volume fixed at 30–45 mL and allowing stoma diameter to be set by the jejunal lumen (mean 3.2 cm cross-sectional diameter), they found that stoma size had little effect on the success of the operation provided a reservoir under 45 mL was consistently created, while avoiding a narrow stoma virtually eliminated postoperative emesis.[5] Read forty years later, the series suggests that with the pouch held small, stoma size within the range produced by the jejunal lumen made little observable difference to weight loss. It is tempting to pair this with the modern regain cohorts as a quasi-factorial experiment showing that neither parameter binds in its usual range — but the pairing is not valid: the modern studies did not hold stoma size constant experimentally, and populations, techniques, outcomes, and eras all differ. The neither-parameter-binds reading is one consistent interpretation, not a demonstrated result.

Symmetrically, the Laplace argument, even taken on its own terms, addresses a different question than the one it is usually deployed for — and it too is an analogy, not a derivation. Laplace’s relation describes instantaneous wall tension under simplified geometry; whether a pouch *dilates* over years depends additionally on wall anisotropy, thickness, scarring, viscoelastic creep, remodeling, and pressure history, none of which the relation contains. What the analogy suggests is a conjecture about *dilation over time* — a narrower tube at lower wall tension

might enlarge less — not about emptying rate at construction. The conjecture would predict a late-appearing, width-dependent effect on durability, and the extended-pouch trial’s pattern (no difference through two years; a divergence at three years driven by regain in the standard arm[13]) is consistent with it — but that trial varied length and construction package, not width alone, so its result cannot validate a width-mediated mechanism; only serial width, pressure, and compliance measurements could. Chapter 6 takes up the dilation question as a measurement confound; here it is enough to note that Model B, stated carefully, is a hypothesis about *width* and *durability*, and that testing it with pouch *length* at *one year* — which is what most of the literature does — tests something else.

4.4.5 The mechanistic expectations entering Part II

Stated as predictions the analysis can confront with data:

1. If Model A holds, the pouch-size effect should be strongest where pouch *volume* is measured directly, should favor smaller pouches, should be visible at one year, and should be mediated by emptying rate rather than capacity.
2. If Model B holds, the effect should be a *width* effect, should favor narrow pouches, should be invisible at one year, and should appear at three years and beyond as a difference in recurrent weight gain.
3. If neither holds and the literature’s associations are creatures of measurement, the association should appear under construction proxies and fade under direct volumetry, and should show no coherent relationship with emptying data.

These three patterns motivate the pre-specified subgroup structure of Chapter 8, but a candid caveat belongs here rather than in the limitations: they are not cleanly distinguishable from study-level subgroup contrasts alone. Measurement artifact can coexist with either biological mechanism; length, width, stoma size, and follow-up time are confounded with era, design, and size across studies; and subgroup differences cannot establish mediation without direct mediator data and interaction tests. The subgroup patterns can therefore shift the weight of evidence among competing explanations — which is how Chapter 8 frames the modality analysis — but they cannot deliver a mechanistic verdict, and the analysis of Part II treats them accordingly — as exploratory effect modification, not adjudication.

It is worth recording that the only prior systematic review of this question did not organize its evidence this way, and could not have: it identified 14 studies of pouch size and weight loss, of which nine found no significant association and five found larger pouches associated with poorer weight loss, with no study finding larger pouches better, and judged the quality and heterogeneity of the data to preclude meta-analysis.[15] A vote count of nine null against five positive is compatible with all three of the patterns above, and distinguishes none of them. The studies whose nulls make up that majority are largely proxy- or radiograph-based and one-year: contrast-radiographic pouch area uncorrelated with %EWL at 3, 6, 12, and 24 months in 82 patients (Pearson r of 0.038, 0.110, 0.015, and 0.042 respectively),[118] and barium-assessed pouch volume uncorrelated with %EWL at a mean 35.7 months in 107 patients, including 59 with pouches over 50 mL who lost as much as those with normal-sized pouches (68 ± 3.6 versus 66 ± 3.6 %EWL).[119] Whether that null is the answer or the artifact is precisely Q3.

4.5 Distension, compliance, and what “size” names

A second physiological channel runs alongside emptying. Gastric distension is a satiation signal: vagal mechanosensitive afferents innervating the gastric wall encode wall tension and stretch and relay that information centrally, where it contributes to meal termination, and vagally dependent

gastric functions — motility, tone, compliance, and emptying rate — collectively regulate food intake and satiation.[120] A smaller pouch reaches a given wall tension at a smaller ingested volume; on the face of it, this is the restrictive model rehabilitated as neurophysiology rather than plumbing, and it predicts a size effect for reasons that have nothing to do with mechanical obstruction.

The difficulty is that the pouch is not rigid, and its compliance is of the same order as the construction differences under study. Felsenreich and colleagues measured pouch volume in the same 12 patients by both swallow magnetic resonance imaging and three-dimensional computed tomography. No significant between-modality differences were detected — mean pouch volume 40.4 ± 21.0 mL by MRI versus 43.5 ± 30.2 mL by 3D-CT ($p = 0.83$), anastomosis diameter 13.1 ± 3.6 versus 13.0 ± 4.0 mm ($p = 0.76$) — though with 12 patients and standard deviations of this size a failure to detect a mean difference is weak evidence of agreement, and no formal agreement analysis (paired limits of agreement or intraclass correlation between modalities) was reported. The more consequential observation is dynamic: the MRI series also captured the pouch *before* the liquid bolus — mean volume 9.5 ± 5.5 mL, short-axis diameter 15.9 ± 5.2 mm, rising to 40.4 ± 21.0 mL and 35.3 ± 5.9 mm at maximum distension.[19] On group means, the same pouches are a 10 mL organ at rest and a 40 mL organ during a swallow.

This is not a technical footnote; it is a definitional crisis for the exposure variable. When a study reports a “20 mL pouch,” the number is uninterpretable without knowing the distension state at which it was obtained, and the distension state is set by the imaging protocol — volume of contrast administered, use of effervescent granules, patient instruction to avoid burping, buscopan administration. A four-fold ratio of group-mean distended to resting volume — not a per-patient measurement of compliance, which would require simultaneous pressure and known infused volume — means that protocol differences between centers can generate apparent between-patient differences of the same magnitude as the deliberate surgical contrast in the randomized trials. It also means that the intraoperative proxies and the imaging measurements are estimating different physical quantities: the surgeon specifies a resting geometry, the radiologist measures a distended one, and the mapping between them depends on the compliance of that individual patient’s gastric wall.

The instrument-level consequences of this compliance — how CT volumetry, contrast radiography, and endoscopy disagree with one another, and what each protocol actually samples of the volume–pressure curve — are measurement science, treated in full in Chapter 6.

Three consequences follow for the analysis. First, any modality hierarchy must treat distension protocol, not merely modality, as the quality dimension — the point Chapter 6 develops. Second, the distension-signaling mechanism, whatever its physiological reality, is plausibly buffered by compliance: a pouch built 5 cm longer is not necessarily 5 cm’s worth of extra tension at a given meal volume, because the compliant wall redistributes. Third, compliance suggests — as a mechanistic extrapolation that would need pressure-volume curves and regional strain measurements to test, not as a demonstrated result — that a distension-mediated effect may depend more on pouch *wall properties* and *width* than on length — which converges, from an independent direction, on the same conjecture the Laplace analogy reached in Section 4.4.

4.6 Ghrelin territory

The strongest *a priori* argument that pouch size should matter hormonally is territorial: ghrelin, the only known circulating orexigenic gut hormone, is produced overwhelmingly in fundic and corpus mucosa,[100, 99] and the RYGB pouch excludes essentially all of the fundus. If the pouch’s size determines how much ghrelin-producing tissue remains in the food stream, and if food-stream contact governs ghrelin secretion, then pouch size should move ghrelin, and ghrelin should move weight. Each link in that chain has been tested, and past the first link the evidence does not support it.

The first link has real support. Lin and colleagues measured plasma ghrelin immediately before and after surgery in 34 patients undergoing RYGB with a completely divided 15 mL vertically oriented pouch, in eight patients undergoing gastric procedures without complete gastric division (four vertical banded gastroplasties, four antireflux operations), and in six lean antireflux controls. Ghrelin fell from 355 ± 20 to 246 ± 13 pg/mL in the divided-pouch group ($p < 0.001$) and did not change significantly in either non-divided group.[121] The signal is specific to *division*, that is, to disconnecting the fundus.

The second and third links fail, and the full mechanistic accounting belongs to Section 5.2.1; the summary suffices here. Exclusion is not resection: resecting the bypassed stomach deepens the one-year ghrelin fall by 36 percentage points relative to leaving it in place, yet buys no difference in weight loss or dietary intake — and the authors themselves concluded that the small gastric pouch probably matters more than the hormone.[122] On pooled evidence the suppression itself is transient — fasting ghrelin is lower only in the first three months and higher than baseline thereafter — and meta-regression finds gastric pouch volume, alimentary limb length, and biliopancreatic limb length all unassociated with the ghrelin change.[123] In diet-induced obese mice, RYGB with a large pouch, a small pouch, or no pouch at all improves body weight and glucose tolerance equally, with no ghrelin correlate of the metabolic improvement;[124] and in humans, at the outcome end of the chain, ghrelin level did not predict weight regain after nadir while stoma diameter did.[117]

The territorial argument therefore finds little support in the data — with the caveats that the acute studies are confounded by fasting, anesthesia, and energy balance, the resection and division comparisons are nonrandomized, the null meta-regression is study-level and may be underpowered, and mouse anatomy does not settle human physiology. Fundic exclusion produces an acute ghrelin drop that looks like a signature of the operation rather than a dose-dependent function of how much fundus was excluded; the ghrelin trajectory reverses within months; and neither pouch volume nor the resulting ghrelin level has been shown to predict weight outcome. Current evidence does not support ghrelin as a major mediator of pouch-size effects — a conclusion short of elimination, but one that sharpens rather than weakens the emptying-dynamics steelman, since it weakens the competing hormonal explanation and leaves the arrival-waveform hypothesis of Section 4.4 carrying most of the hormonal case. Chapter 5 takes this up in the full mechanistic account.

4.7 Complications, with anatomical logic

Three complications of the pouch bear directly on the dissertation’s question, because each has an anatomical rationale that links it to a construction parameter. The safety ledger they populate runs in both directions, and the directions must not be conflated. The evidence below ties marginal ulcer to *larger* construction — more staple line, more retained oxyntic mucosa — so it penalizes building bigger, not smaller. The harms plausibly attached to very small construction are different ones: anastomotic tension and ischemia on a short pouch, early stricture, leak, and food intolerance — harms for which the direct pouch-size evidence is thin and largely inferential. Stricture as observed is driven chiefly by stoma construction and anastomotic technique and is largely orthogonal to pouch size. The third complication, dumping, is not a counterweight at all but a physiological readout of the emptying hypothesis of Section 4.4.

4.7.1 Marginal ulcer

Marginal ulceration is ulceration of the jejunal mucosa at or just distal to the gastrojejunal anastomosis. Its anatomical logic is straightforward: jejunal mucosa evolved to receive chyme already neutralized by pancreatic and biliary bicarbonate. In the RYGB configuration it receives pouch effluent directly, and pouch effluent is acidic because the pouch retains oxyntic mucosa

(Section 4.1).[97] The dose is small in absolute terms[69] but the exposure is unbuffered,[98] and on this logic the ulcer risk would be expected to scale with the amount of acid-secreting mucosa left in the food stream — that is, with pouch size. In a systematic review of 41 studies comprising 16,987 patients, 787 (4.6%) developed marginal ulcer, with reported incidences ranging from 0.6% to 25%; the position and size of the pouch, smoking, and NSAID use were the associated factors.[125]

The dose-response evidence is, unusually for this literature, large and directionally consistent — though it is observational throughout and its exposures are construction proxies, not anatomical pouch dimensions. In the Scandinavian Obesity Surgery Registry, 14,168 laparoscopic RYGB patients with linear-stapled gastrojejunostomies were analyzed using total stapler length as a proxy for pouch size (mean 145 mm). Symptomatic marginal ulcers occurred in 0.9% at one year, and the relative risk of marginal ulcer rose 14% (95% CI 9–20%) for each additional centimeter of *stapler* used to construct the pouch — stapler length, affected by overlap, angulation, and pouch width, is not a centimeter of anatomical pouch length, and the gradient cannot be read as an acid-mucosal-area dose response. The same analysis found that stapler length predicted excess BMI loss at six weeks but *not* at one year — the central empirical conflict this dissertation exists to resolve — and its authors’ recommendation follows only from the safety side: additional stapling should be avoided.[14] A separate analysis of 37,301 Swedish patients found larger gastric pouch to be among the independent risk factors for acid-related complications at one year (OR 2.19, 95% CI 1.53–3.25), alongside smoking (OR 2.59) and diabetes treatment (OR 1.75), and found that prophylactic proton pump inhibition did not reduce the rate.[126] Endoscopic series concur on direction: in 103 patients with endoscopically documented marginal ulcers, gastric pouch length was associated with ulcer formation on univariate analysis (OR 1.2, $p = 0.02$) alongside diabetes and smoking, though only diabetes survived multivariable adjustment.[57] And the Chinese cohort comparing large (> 50 mL, antrum-based) with small (10–20 mL, lesser-curve) pouches found marginal ulcer in 23.7% of the large-pouch group versus 6.4% of the small-pouch group at six years ($p = 0.023$) — a comparison in which volume, pouch location, mucosal phenotype, and likely technique all change together, so it is a contrast of operative configurations, not a pouch-volume dose-response.[18]

Two mechanistically informative negatives complete the picture. Sapala and colleagues, using the near-total gastric bypass in which the transected pouch is limited to the cardia and the gastrojejunostomy is placed along the greater curvature, documented marginal ulcer in exactly one of 173 patients at three years (0.6%), and that patient had a dilated 60 cc reservoir.[96] Conversely, the Iranian comparison of 80 cc long pouches against 30 cc short pouches reported no marginal ulcers in either arm at twelve months — in a series of 219 patients, at a time-point at which the registry data would predict roughly 1–2 events.[108] Follow-up length and ascertainment method (routine endoscopy in all patients versus symptom-triggered endoscopy) plainly dominate the reported incidence, which is a caution the safety analysis of Chapter 13 must carry.

The technique of the anastomosis modifies the risk independently of pouch size. Schäfer and colleagues compared 126 circular-stapled with 115 linear-stapled gastrojejunostomies and found anastomotic ulcers significantly more frequent after circular stapling ($p = 0.0034$), with more refractory ulcers and more revisional surgery, concluding that a linear-stapled anastomosis with a long narrow pouch should be preferred.[58] Since 81.2% of IFSO respondents now use a linear-stapled gastrojejunostomy[7] whereas in 2008 only 41% of American surgeons did,[6] the era-dependence of marginal ulcer rates is partly a confounder of technique rather than a fact about pouches — which is why era is a pre-specified subgroup. The Michigan Bariatric Surgery Collaborative, linking a surgeon technique survey to registry outcomes for 9,904 laparoscopic gastric bypasses, found the same pattern for early morbidity: leak rate did not differ by anastomotic technique (0.6% circular, 0.3% linear, 0.6% hand-sewn; $p = 0.38$), but postoperative hemorrhage did (2.9%, 1.2%, 1.6%; $p < 0.0001$), as did wound infection, and staple-line reinforcement of

the pouch was associated with less hemorrhage (1.9% versus 2.7%, $p = 0.012$).[127] Technical parameters of pouch construction have demonstrable, separable effects on harm. Whether they have any on weight is the open question.

4.7.2 Stricture

Stricture at the gastrojejunostomy is two entities that the literature often pools. Late stenosis can follow marginal ulceration — ulceration heals by fibrosis and fibrosis narrows a lumen — and to that extent shares the ulcer’s substrate; but many early strictures arise without any documented ulcer, from technical narrowing, edema, ischemia, anastomotic tension, or stapler diameter, and are determined by the constructed stoma and the anastomotic technique, independent of pouch dimensions. Ahmad and colleagues followed 450 consecutive laparoscopic gastric bypass patients, all with a 15 mL pouch, and treated 14 (3.1%) for symptomatic stenosis with a total of 23 endoscopic balloon dilations at a mean 2.7 months postoperatively.[128] Khoraki and colleagues showed how sharply technique dominates: among 876 circular-stapled gastrojejunostomies, 76 (8.7%) developed stricture, but the rate ranged from 17% in the trans-gastric 21 mm group down to 1.6% in the trans-oral 25 mm group, with anvil size and insertion route as independent predictors.[129] A ten-fold range in stricture risk attributable to two binary technical choices, in a single center, is a useful calibration for how much of the between-study variance in this literature can be technique rather than pouch. It also underlines Section 4.4’s point: the stoma, not the pouch, is where the anatomy is tight enough to obstruct.

4.7.3 Dumping

Dumping syndrome is the direct symptomatic expression of the emptying physiology described above. The international consensus panel agreed on the pathophysiological relevance of rapid passage of nutrients to the small bowel, on the role of decreased gastric volume capacity, and on the role of GLP-1 release; it defined early dumping (vasomotor and gastrointestinal symptoms within about 30 minutes, diagnosed by a hematocrit rise $> 3\%$ or pulse rise > 10 bpm 30 minutes after a glucose load) and late dumping (postprandial hyperinsulinemic hypoglycemia, diagnosed by a nadir glucose < 50 mg/dL), and identified dietary adjustment as first-line therapy with acarbose effective for late dumping.[130]

Because dumping is caused by the same mechanism that Model A proposes as the route by which small pouches produce weight loss, it is the natural place to look for a dose-response that the weight-loss outcome may be too noisy to reveal. The evidence so far is null or weakly favorable to Model B. In the nine-year follow-up of the extended-pouch trial, 48.3% of standard-pouch patients versus 40.6% of extended-pouch patients scored ≥ 7 on the Sigstad questionnaire ($p = 0.61$), with identical mean Arts early- and late-dumping scores; the authors read the non-significant direction as a potential advantage of the longer, slower-emptying pouch.[16] The Iranian series found early dumping by Sigstad criteria in 18.72% overall with no difference between long and short pouches ($p = 0.476$),[108] and the Chinese six-year cohort likewise found no difference in dumping between large and small pouches.[18] Set against this, a limiting case is worth recording for illustration, not inference: a patient in whom endoscopy found no identifiable gastric pouch or gastric mucosa at all — jejunum encountered immediately distal to the esophagus — presented with debilitating dumping *and* persistent class III obesity and absent satiety.[131] A single uncontrolled case with unknown stoma, limb configuration, and behavioral history cannot falsify any model — Model A does not predict that dumping guarantees weight loss in every individual — but it illustrates that rapid transit and weight loss can dissociate completely. Consistent with the stoma-dominance hypothesis, though not proof of it, the therapeutic lever for refractory dumping is the outlet rather than the pouch: transoral outlet reduction produced complete response at two years in 66% of a small refractory series.[132]

4.8 What the pouch is

The object at the center of this dissertation can now be described — with its descriptions sorted by how well they are established, because this paragraph supplies premises for Part II and must not launder hypotheses into anatomy.

Established, or directly measured: the pouch is a compliant remnant of cardia and proximal corpus carrying oxyntic mucosa at least at its distal margin;[97] its realized volume is rarely measured at the time it is made, its length is coarsened by stapler cartridge geometry, and its width is set by an independent parameter that the literature usually does not report; its resting and distended group-mean volumes differ by a factor of four in the one dynamic-imaging series that measured both.[19] Slightly more than half of surveyed surgeons report routinely measuring pouch length, and those who do report a median of 6 cm with regional medians spanning 5 to 7 cm.[7] Two centers measuring the same quantity by the same imaging method in the same era reported means differing by a factor of 2.5.[103, 104]

Observed association, short of mechanism: stoma diameter correlates with weight regain more consistently than pouch size does in observational cohorts,[115, 117] supporting — but, given referral selection and reverse causation, not proving — the hypothesis that the outlet rather than the pouch is the dominant flow resistance. Marginal-ulcer risk rises with construction proxies for larger pouches — 14% more relative risk per additional centimeter of stapler in 14,168 registry patients[14] — an observational gradient on a proxy exposure, and the most consistent such gradient in this literature, penalizing larger construction.

Hypothesis, awaiting mediator data: the pouch matters, if it matters, as the antechamber that shapes nutrient arrival at the jejunum. Two mechanistic hypotheses make opposite predictions about which geometry produces the better arrival waveform, differ in mediator (emptying rate versus wall tension and dilation resistance), and differ in the timepoint at which their effects should appear. The extent of vagal denervation of the pouch is anatomical inference, not measurement. The ghrelin channel, the most intuitively appealing hormonal route from pouch size to weight, is unsupported by current evidence as a major mediator.[123, 124, 122] The distension channel is physiologically real but buffered by compliance and, like Model B, points — conjecturally — at width rather than length.

Two implications carry forward. First, the exposure variable is irreducibly multidimensional — length, width, volume, distension state, retained fundus — and the literature has collapsed it to a scalar in inconsistent ways; Chapter 6 formalizes this into the modality hierarchy the analysis will use. Second, mechanism does not license the expectation of a null result. It licenses the expectation of a possibly *sign-ambiguous*, *timepoint-dependent*, and *width-mediated* effect, which is a much more demanding thing to detect and a much more interesting thing to fail to detect. The analysis in Part II is designed against that harder standard — and the next chapter assembles the full mechanistic account of the operation from which these predictions, and their competitors, derive their standing.

Chapter 5

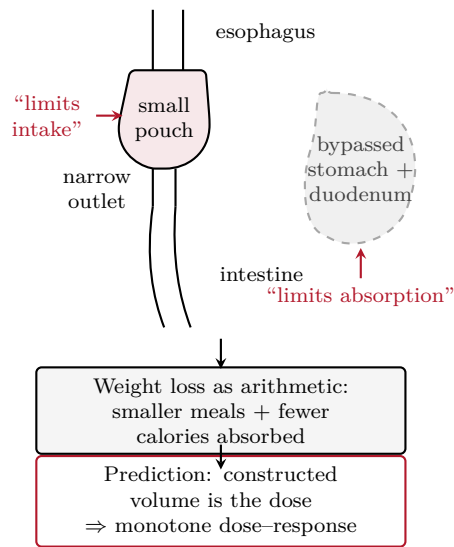
Mechanisms of weight loss after RYGB

This chapter is the referee of the dissertation’s empirical question. The chapters that follow will ask whether the constructed size of the gastric pouch predicts weight loss, and the answer will be a number with a confidence interval. But a number is only interpretable against a model of how the operation works, because the model determines what an effect — or its absence — would mean. If Roux-en-Y gastric bypass (RYGB) works by mechanically limiting how much food the stomach can hold, then pouch volume is the operation’s dose, and a dose–response relationship is almost obligatory: failing to find one would indict the evidence, not the hypothesis. If instead the operation works chiefly by rewiring the hormonal and neural signaling between gut and brain, constructed volume may be a weakly coupled parameter — packaging more than dose — and forty years of null and contradictory findings are what such a physiology would tolerate. These two positions are poles, not exhaustive alternatives: the defensible modern account is a multicomponent model in which signaling carries most of the demonstrated weight and mechanical geometry retains candidate channels — delivery dynamics, dumping, intolerance — whose relative contributions are uncertain. The mechanistic literature has moved decisively away from the first pole, and this chapter lays out that movement in full: the legacy restrictive/malabsorptive model and the failure of its strong form under direct measurement; the neuroendocrine model thread by thread — ghrelin, the hindgut hormones GLP-1 and PYY, bile acids, the gut microbiota, vagal and central rewiring, food preference; and the threads that remain genuinely contested. It closes by doing what a mechanism chapter in this dissertation must do: deriving from each mechanistic account a testable prediction about pouch-size dependence, so that the meta-analytic results of Part II can be scored against physiology rather than merely tabulated.

Two warnings frame everything that follows. First, much of the sharpest mechanistic evidence comes from rodent surgery, where anatomy can be manipulated and signaling pathways deleted in ways impossible in patients; rodent bypass differs from human bypass in scale, in limb proportions, and sometimes in the very direction of its effects, and this chapter flags species throughout. Second, mechanism papers and outcome papers largely do not cite each other in this field. The surgical outcome literature that Part II will pool was mostly designed under the legacy model; the physiology literature that dismantled that model rarely measured pouch size at all. The synthesis attempted here — reading each mechanism for its pouch-size prediction — is therefore itself part of the dissertation’s contribution, and its inferential status is labeled accordingly.

A Restrictive–malabsorptive model

legacy pole: the pouch as a container



B Multicomponent neuroendocrine model

modern pole: the pouch as a signaling conduit

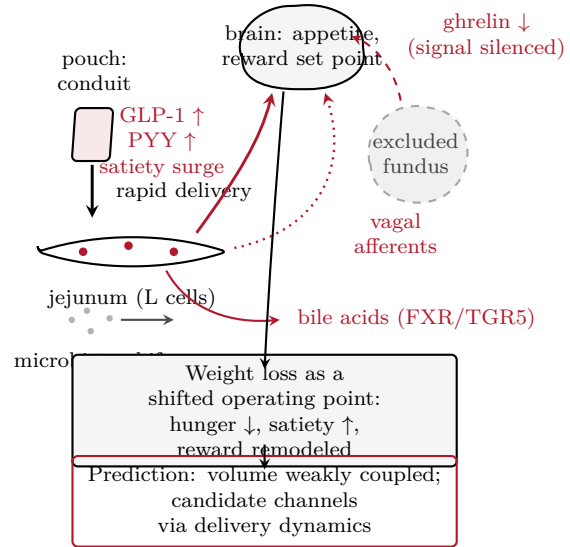


Figure 5.1: Two poles of the mechanistic debate over RYGB. (A) The legacy restrictive–malabsorptive model treats the pouch as a capacity-limiting container: a small reservoir limits intake, the bypassed foregut limits absorption, and weight loss follows arithmetically — making constructed volume the operation’s dose. (B) The multicomponent neuroendocrine model treats the pouch as a signaling conduit: rapid nutrient delivery to the jejunal L cells drives a GLP-1/PYY satiety surge, fundus exclusion silences the ghrelin signal, and bile-acid, microbiota, and vagal channels converge on a lowered defended operating point — leaving constructed volume weakly coupled. The panels are poles, not exhaustive alternatives: this chapter’s working model is multicomponent, with signaling the best-supported carrier of the effect and mechanical delivery dynamics a candidate channel (Section 5.3). Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

5.1 The legacy model: restriction and malabsorption

5.1.1 The model as engineered design

The restrictive/malabsorptive account was not an inference from data; it was the design specification of the operation. When Mason and Ito introduced gastric bypass in 1967, the procedure was conceived as a controlled reduction of gastric capacity with diversion of food past the duodenum, adapted from the weight loss observed after ulcer surgery [3]. The logic was that of a plumbing problem: a small reservoir limits how much can be eaten at a sitting, a bypassed foregut limits how much of what is eaten gets absorbed, and weight loss follows as an arithmetic consequence of the energy balance. Both terms of the model were mechanical, both were set at the operating table, and both therefore implied that surgical geometry — reservoir volume, outlet caliber, limb length — was the causal dose.

The pouch-size doctrine crystallized within fifteen years. By 1982, Kirkpatrick and Siegel could report a prospective series built explicitly to arbitrate between the two candidate restriction parameters, reservoir and stoma, and concluded that stoma size had little effect on success provided a tiny reservoir — less than 45 mL — was consistently created: the reservoir, not the outlet, was declared the critical determinant of a successful gastric bypass [5]. That conclusion, drawn from the restrictive premise and a small uncontrolled series, is the doctrine whose empirical standing this dissertation audits. It is worth pausing on what the doctrine assumes: that the pouch functions as a capacity-limiting vessel, that satiation arises from its distension, and that a smaller vessel produces earlier satiation, smaller meals, and more weight loss — a monotone, volume-mediated causal chain in which every link is checkable. Over the following decades the links were checked, and the chain did not survive in the strong form the doctrine requires: no link held as stated, though — as the measurements below show — several survive in weakened, partial forms.

5.1.2 Malabsorption, measured

The malabsorptive term was the first to be measured directly, because it is the easiest to measure: energy in the diet minus energy in the stool is a bomb-calorimetry problem, not a physiology problem. The measurement demoted it from mechanism to minor term. Odstrcil and colleagues performed exactly this measurement in nine severely obese patients before and after long-limb RYGB — the variant whose 150 cm Roux limb was specifically believed to induce malabsorption — computing absorption coefficients for fat, protein, carbohydrate, and total combustible energy [133]. Fat absorption did fall, from $92 \pm 1.3\%$ of intake before surgery to $72 \pm 5.5\%$ at five months and $68 \pm 8.7\%$ at fourteen months, with no significant change in protein or carbohydrate absorption. But the decomposition of the total energy deficit is the decisive result. At five months, malabsorption accounted for 124 ± 57 kcal/day of reduced energy absorption while reduced food intake accounted for 2062 ± 271 kcal/day; at fourteen months the figures were 172 ± 60 versus 1418 ± 171 kcal/day. Malabsorption thus contributed roughly 6% and 11% of the operation's total effect on absorbed energy at the two time points [133]. In signal-processing terms, the malabsorptive channel carries an order of magnitude less of the effect than the intake channel. The finding generalizes with stated caveats: nine patients, one long-limb variant, two timepoints within fourteen months, and macronutrient energy only — fat absorption genuinely fell, distal variants and micronutrient handling are separate ledgers, and later follow-up was not measured. But within its scope the decomposition is decisive against malabsorption as the *dominant* human weight-loss mechanism: whatever standard RYGB does, it does overwhelmingly by making patients eat less — and the question collapses into *why* they eat less.

5.1.3 Restriction, measured

The restrictive term is where the legacy model’s answer to that question lived, and it is where the model failed most instructively. The restrictive claim is really a claim about retention: the pouch limits intake because food sits in it, distending it, until the narrow outlet lets it through. The prediction is a slow-emptying reservoir. Direct measurement found the opposite.

The early radionuclide studies already contained the anomaly. Horowitz and colleagues, imaging twelve patients a year after gastric bypass with a double-isotope technique, found that liquids emptied *more rapidly* than in unoperated controls, and — the finding that should have been fatal to the doctrine on arrival — that neither pouch size nor stoma size correlated with emptying rate, nor emptying rate with weight loss [113]. Näslund and Beckman, in a randomized comparison of gastric bypass and gastroplasty with radiolabeled food, likewise found no correlation between emptying rate and weight loss in either group [114]. These were small studies with the imprecise instruments of their era, and the field absorbed them as curiosities rather than refutations. But modern scintigraphy reproduced and sharpened the result. Dirksen and colleagues studied seventeen patients more than a year after RYGB against nine healthy controls: at meal completion a *smaller* fraction of both liquid and solid marker remained in the pouch than in an intact stomach, pouch emptying was faster than gastric emptying for both phases, and — completing the inversion of the restrictive picture — small intestinal transit of solids was *slower* [10]. Stano and colleagues, using the acetaminophen absorption method around a standardized 600-kcal meal, measured pouch emptying 2.4 times faster after RYGB than the same patients’ gastric emptying before surgery [110]. For the liquid and marker meals these studies measured, the constructed pouch behaves less like a reservoir than like a low-impedance conduit: contents pass with little hindrance into the jejunum. The measurements do not exclude every mechanical channel — transient distension, outlet resistance to solid boluses, mechanosensitive signaling, and symptom-limited meal termination remain possible and largely unmeasured — but they are incompatible with the doctrine’s specific picture, a slow-emptying vessel whose capacity caps the meal. A model in which weight loss is caused by food being held back is difficult to sustain when the measured motility finding is rapid transit.

The animal models dismantled the remaining links. If capacity limitation drove weight loss, removing the reservoir entirely should represent the limiting case of maximal restriction, and varying reservoir size should produce graded outcomes. Neither holds. Stenström and colleagues followed rats for a year after “micro-gastric bypass” — esophagojejunal anastomosis, effectively a zero-volume pouch — and observed 30–40% weight reduction *without reduced food intake* [134]. Xu and colleagues constructed murine RYGB with three deliberately different pouch sizes — roughly 30% of the stomach, roughly 10%, and none (esophageal anastomosis) — and found that all three variants improved body weight and glucose tolerance equally; pouch size affected surgical complication rates (obstruction in the large-pouch animals) but not the metabolic phenotype [124]. And Hao and colleagues, in a carefully controlled murine series with sham-operated and weight-matched calorie-restricted comparison groups, found that energy intake after RYGB was suppressed only for the first ten days; thereafter the operated mice ate as much as sham controls — and roughly 35% *more* than the calorie-restricted mice held at the same body weight — while maintaining their weight loss through elevated energy expenditure [135]. A restriction mechanism cannot explain weight loss that persists at unrestricted intake.

Human eating behavior tells the same story in a quieter register. Patients after RYGB do eat small meals — but the proximate cause is not that larger meals are physically impossible. They report less hunger and earlier fullness; large or liquid meals provoke discomfort and dumping symptoms through rapid jejunal delivery, not through pouch distension against a full vessel [110]. And where the restrictive model predicts that patients should be hungrier than ever — carrying, as they do, a body defending a higher weight with only a 30-mL stomach to feed it — the observed phenotype after RYGB is the opposite: hunger is reduced, satiety enhanced, from the first postoperative weeks [9]. Something is signaling, not merely obstructing. The measure-

ment campaign that demoted the restrictive model thus simultaneously named its successor: much of the operation’s effect had to live in the signals, and the search for them defines the neuroendocrine model. The defensible summary of this section is deliberately narrower than the field’s obituaries for restriction: pure capacity limitation and calorie malabsorption are not the sole or dominant mechanisms of human weight loss after RYGB. That is what the measurements license — not that every mechanical contribution is zero.

Three caveats keep this section honest. First, the demotion of restriction as the *dominant* mechanism does not prove the mechanical geometry of the pouch is causally inert — the emptying dynamics of the conduit itself may modulate the signaling, a possibility developed as the steelman in Section 5.3. Second, the motility evidence is liquid- and marker-weighted; solid-food handling, mechanosensation, and symptom-mediated meal limitation are thinner threads, and geometry could act through any of them. Third, “restriction” survives in clinical language — and in patients’ subjective experience of early fullness — as a description of the phenotype; what fell is the claim that mechanical capacity limitation is the causal engine.

5.2 The neuroendocrine model

The neuroendocrine model holds that RYGB achieves weight loss chiefly by changing the gut-derived signals that govern appetite and energy homeostasis — reducing hunger and enhancing satiety so that, in the set-point vocabulary the model inherits (itself a contested construct, Section 2.2), the defended body weight falls and reduced intake follows from altered appetite rather than from a physical ceiling on meal size. The model’s authoritative modern statement is the review by Pucci and Batterham, which this dissertation treats as its mechanistic anchor: after fifty years of evidence it is “now accepted” that bariatric procedures produce weight loss and metabolic improvement “by mechanisms other than restriction and/or malabsorption,” with reduced energy intake — as a result of altered eating behavior — the main driver, produced by coordinated changes in gut hormones, bile acids, microbiota, and gut–brain neural signaling [9]. The contrast with dieting is the model’s sharpest exhibit: weight lost by caloric restriction triggers compensatory biology — rising hunger hormones, falling energy expenditure — that promotes regain, whereas RYGB produces the same weight loss with hunger reduced and food-cue reactivity blunted [9]. On this model the operation does not fight the body-weight regulatory system so much as shift its operating point. Two qualifications attach before the threads are examined. The model’s strong form is not established fact: the set-point account of the shift is an interpretive frame, and the blockade and comparison studies below quantify individual channels, not the whole. And “neuroendocrine versus mechanical” is a rhetorical simplification: mechanical geometry can alter intake without acting as a static container — through rapid delivery, pressure-related satiation, dumping, conditioned avoidance, altered meal texture, and intolerance — so the working model of this dissertation is multicomponent, with signaling the best-supported carrier of the effect and the mechanical channels’ contributions uncertain rather than zero. The threads are examined in turn, in roughly descending order of how directly they bear on the pouch.

5.2.1 Ghrelin: the excluded fundus, and a complicated thread

Ghrelin is the only known orexigenic (hunger-stimulating) gut hormone, produced predominantly by cells in the gastric fundus — the stomach’s upper dome, which RYGB excludes from the food stream and, in the standard lesser-curve-based construction, from the pouch itself (Chapter 4). The hormone rises with fasting and falls after meals, and rises chronically during diet-induced weight loss — part of the compensatory siege that defeats dieting (Section 2.2). The landmark observation belongs to Cummings and colleagues: a diet-induced weight loss of 17% of body weight raised 24-hour ghrelin profiles by 24%, whereas patients after RYGB, despite 36% weight

loss, had ghrelin profiles 77% *below* normal-weight controls, with loss of the normal meal-related oscillation [45]. Here was a hormonal mechanism tied directly to surgical anatomy: exclude the ghrelin-producing territory and the hunger signal falls silent, in stark opposition to the dieting response. The finding redirected the field’s attention from mechanics to endocrinology, and its anatomical logic is the one mechanistic thread that gives the pouch’s construction — though not its volume — a starring role. Early surgical series supported the anatomical reading: Lin and colleagues found that RYGB with a completely divided, 15-mL vertically oriented pouch produced an immediate fall in ghrelin (from 355 ± 20 to 246 ± 13 pg/mL) not seen after vertical banded gastroplasty or antireflux procedures that left the stomach in continuity [121], and Carrasco and colleagues found a greater one-year ghrelin decline when the bypassed stomach was resected than when it was left in place ($-25.3 \pm 22.5\%$ versus $+11.2 \pm 50.9\%$) [122].

The thread then complicated, and honesty requires saying so plainly. Pooling sixteen studies with 325 patients, Xu and colleagues found that fasting total ghrelin falls only transiently after RYGB — lower at three months or less (SMD -0.49), but *higher* than baseline beyond three months (SMD $+0.46$), with an overall upward tendency [123]. The same meta-analysis carries a finding of direct relevance here: in meta-regression, gastric pouch volume — along with alimentary and biliopancreatic limb lengths — was *not* associated with the ghrelin change [123]. Ghrelin levels do not track outcomes either: in patients evaluated endoscopically after RYGB, ghrelin bore no relation to weight regain, whereas gastrojejunal stoma diameter did [117]; in the Carrasco cohort the divergent ghrelin trajectories produced no difference in weight loss at twelve months, leading the authors themselves to conclude that the small pouch mattered more than the hormone [122]; and in the murine pouch-size experiment, plasma ghrelin did not correlate with any metabolic outcome [124]. Methodological variability — assay specificity for acylated versus total ghrelin, sample handling, timing — explains some of the discordance [9], but the sober reading is that ghrelin suppression is neither sustained nor necessary for the operation’s effect. For this dissertation the implication is directional rather than settled: the one mechanism that made pouch construction hormonally interesting predicts dependence on *which tissue is excluded* — fundus in or out of the food stream, stomach divided or merely partitioned — and not on the constructed volume of the pouch. The model therefore expects two fundus-excluding pouches of 15 and 50 mL to differ little in their ghrelin behavior. The null meta-regression is consistent with that expectation, but a study-level null cannot establish equivalence: retained ghrelin-producing tissue, vagal handling, and emptying were not measured against volume within patients, and the volume range represented is narrow.

5.2.2 GLP-1 and PYY: the hindgut surge

If ghrelin was the model’s opening argument, the hindgut hormones are its core. Glucagon-like peptide-1 (GLP-1) and peptide YY (PYY) are secreted by enteroendocrine L cells, whose density increases distally along the small intestine; both are released in proportion to nutrient contact with the gut epithelium, and both act centrally as satiety signals — GLP-1 doubling as an incretin, a hormone that amplifies glucose-stimulated insulin secretion. In an intact gastrointestinal tract, the stomach meters chyme into the duodenum at 1–4 kcal/min and the distal gut sees nutrients late and dilute. RYGB re-plumbs this entirely: the pouch delivers ingested nutrients rapidly, with minimal gastric processing and ahead of the biliopancreatic secretions that join at the jejunojejunostomy, to the mid-jejunum, and the L-cell response is transformed — a transformation not attributable to distal L-cell density alone, given how early in the meal the response appears. le Roux and colleagues showed that patients after RYGB have markedly increased postprandial PYY and GLP-1 with early, exaggerated insulin responses — and, critically, that none of these changes appear in patients losing equivalent weight by gastric banding, tying the hormonal signature to the anatomical rearrangement rather than to weight loss itself [46]. The postprandial GLP-1 response rises roughly ten-fold after RYGB while integrated GIP (the duodenal incretin) is unchanged — a spatial signature: the signal gain lives where the

anatomy changed [47]. The changes precede weight loss, appear within the first postoperative week, and persist for up to ten years [9].

The mechanistic coupling to delivery dynamics is direct and experimentally established. Dirksen and colleagues found that GLP-1 and PYY₃₋₃₆ secretion was highest in the patients with the fastest liquid pouch emptying [10] — the epithelium responds to the rate of nutrient arrival, like a detector responding to pulse slope rather than integrated charge. Stano and colleagues closed the loop within subjects: after RYGB a liquid meal, which traverses the pouch fastest, produced a peak GLP-1 of 106.1 ± 67.2 pM against 45.3 ± 25.2 pM for the identical meal in solid form, a five-fold versus two-fold amplification over the presurgical response, with peak GLP-1 correlating with pouch-emptying rate after liquid meals [110]. The same rapid delivery explains the operation’s signature adverse physiology — dumping symptoms and, in a subgroup, late postprandial hyperinsulinemic hypoglycemia, both provoked preferentially by large liquid carbohydrate loads [110, 47].

Is the hormonal surge causal for reduced intake, or an epiphenomenon? The blockade experiments say causal, with quantified caveats. Somatostatin-analogue suppression of the entire gut hormone response increases appetite, energy intake, and weight after RYGB; combined pharmacological blockade of GLP-1 action and PYY₃₋₃₆ formation increases food intake after RYGB; and patients with poor surgical weight loss show lower PYY and GLP-1 with higher ghrelin than good responders [9]. On the glycemic side, infusion of the GLP-1 receptor antagonist exendin-(9-39) shows that endogenous GLP-1’s contribution to meal-induced insulin secretion roughly doubles after RYGB [47]. Yet the same literature counsels against single-hormone triumphalism: GLP-1 receptor knockout mice lose weight normally after bypass, and antagonist studies in patients with remitted diabetes suggest the enhanced GLP-1 action, though real, may add little glycemic benefit beyond weight loss itself [47]. The model that survives is a redundant, multi-channel signaling system in which the hindgut surge is a major — but not sole — carrier of the satiety signal.

For the pouch question, this thread is the pivotal one, because it is the only major mechanism with a plausible *functional* (as opposed to anatomical) dependence on the pouch: the surge is driven by nutrient arrival dynamics, and the pouch and its outlet are the upstream elements that shape those dynamics. Whether that dependence runs through constructed volume, and with what sign, is taken up in Section 5.3.

5.2.3 Bile acids, FXR, and TGR5: a contested thread

Bile acids are detergent molecules synthesized from cholesterol, secreted into the duodenum to emulsify fat, and recycled from the distal ileum — but they are also signaling ligands, acting on the nuclear receptor FXR (regulating, through fibroblast growth factor 19, glucose and lipid handling) and the membrane receptor TGR5 (which stimulates L-cell GLP-1 secretion and, in muscle and brown adipose tissue, may raise energy expenditure). Circulating bile acid levels and composition change after RYGB, plausibly because accelerated nutrient flow reaches the distal gut and alters enterohepatic recirculation, and these changes have been proposed as contributors to both weight loss and glycemic improvement [9]. The strongest causal evidence is surgical-genetic and comes from the neighboring operation: in diet-induced obese mice, the benefits of vertical sleeve gastrectomy — explicitly shown in the same experiment not to result from mechanical restriction by the smaller stomach — are substantially lost in FXR knockout animals [136].

This dissertation labels the thread contested, for three reasons stated in the mechanistic literature itself. First, the causal demonstration is murine and sleeve-specific; whether FXR/FGF-19 signaling contributes to the human RYGB effect is, in Pucci and Batterham’s words, uncertain at present [9]. Second, the kinetics fit the early effect poorly: in the review’s account, substantial rises in circulating bile acids appear only around one year after surgery, whereas the PYY/GLP-1 surge and the metabolic improvement appear within a week [9]. The primary human kinetics

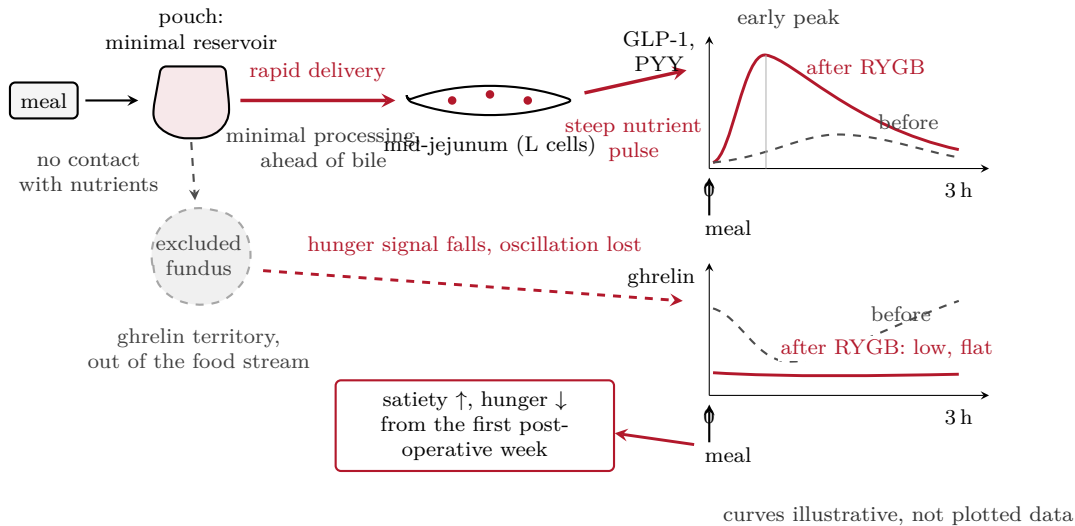


Figure 5.2: The hormonal response chain of RYGB. A meal traverses the constructed pouch with minimal storage or gastric processing and arrives in the mid-jejunum as a steep nutrient pulse, ahead of the biliopancreatic secretions; the L-cell hormones GLP-1 and PYY rise earlier and far higher than in the intact anatomy, while the excluded fundus removes the ghrelin-producing territory from nutrient contact, leaving the hunger signal low with its meal-related oscillation lost. The behavioral output — enhanced satiety, reduced hunger — appears from the first postoperative week. The time courses are illustrative schematics of the patterns reported in the studies cited in Sections 5.2.1 and 5.2.2, not plotted data. Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

support the lag while complicating its outline: in a longitudinal series of thirteen diabetic women measured before and at one month and two years after RYGB, fasting bile acids *fell* at one month before rising at two years, the exaggerated postprandial excursion appeared only at two years, and the compositional shift toward 12 α -hydroxylated species was likewise confined to the late timepoint [137] — a pattern that differs by bile-acid fraction and by fasting versus postprandial sampling, so the timing argument is suggestive rather than decisive. Third, the thread is entangled with the microbiota (bile acids shape microbial communities and microbes biotransform bile acids), so neither signal can be cleanly attributed [9]. For the pouch question the thread is nonetheless well-behaved: every proposed bile-acid mechanism runs through distal delivery and enterohepatic geometry — limb lengths, not pouch volume. No published account assigns bile-acid signaling a pouch-size dependence beyond whatever the pouch contributes to delivery rate.

5.2.4 The gut microbiota

RYGB restructures the intestinal microbial community rapidly and reproducibly: shifts detectable within a week in rodents, conserved across mice, rats, and humans — most prominently increased relative abundance of *Gammaproteobacteria* (*Escherichia*) and *Verrucomicrobia* (*Akkermansia*) — and, critically, independent of weight change and caloric restriction, which were controlled by sham-operated and calorie-restricted arms [138]. The causal experiment is the transfer: Liou and colleagues colonized germ-free mice with cecal contents from RYGB-operated donors and observed weight loss and reduced adiposity relative to recipients of sham-donor microbiota [138]. The altered community is thus *sufficient* to transmit part of the phenotype. Whether it is *necessary*, and how large its share of the human effect is, remain open: rodent and human microbiomes differ substantially, the mediating pathways (short-chain fatty acid production, bile acid biotransformation, energy harvest) are entangled, and the measured transfer

effect is modest beside the operation’s full effect [9]. As a candidate carrier of pouch-size dependence the microbiota is implausible on its face: the community shift is driven by the re-routed foregut, altered pH and oxygen exposure, and bile flow — consequences of the bypass topology, not of reservoir volume. No study linking constructed pouch size to microbial composition was identified: the mechanistic review’s microbiota account is silent on pouch size [9], and targeted searches of the corpus and of PubMed returned none.

5.2.5 Vagal and central rewiring: appetite, reward, and food preference

The hormonal threads terminate in the brain, and the brain is where the behavioral phenotype is assembled. Homeostatic circuits in the hypothalamus and hindbrain integrate the endocrine signals with neural traffic carried by the vagus nerve — the tenth cranial nerve, whose afferent fibers report gastric distension and intestinal nutrient sensing. RYGB leaves the vagal trunks along the pouch and excluded stomach largely intact when construction is careful, and vagal signaling appears to increase after the operation — a preserved, possibly amplified channel through which the re-routed gut reports to the brain [9]. Hedonic circuits change too, and here the evidence is human and quantitative. Scholtz and colleagues, comparing RYGB with banding patients by functional MRI, found lower brain-hedonic responses to food — particularly high-calorie food — in orbitofrontal cortex, amygdala, caudate, nucleus accumbens, and hippocampus after RYGB, together with lower rated palatability of energy-dense foods and lower fat intake; the differences tracked the RYGB group’s higher GLP-1, PYY, and bile acids rather than hunger or psychological traits [139]. Food, in short, becomes less rewarding, and specifically the energy-dense food that most efficiently drives positive energy balance.

The behavioral output is a measurable shift in food preference. In the randomized comparison of laparoscopic RYGB against vertical banded gastroplasty, RYGB patients selectively reduced the fat fraction of their diet — about 30% actively avoided fatty food — while the gastroplasty patients drifted toward sweet soft calories; the authors proposed the “steering away from fatty foods” as a mechanism of the bypass’s superior results [140]. The same steering appears in rodents as a shift toward lower-fat chow [135]. Two honest complications attach. First, the rat literature on motivated behavior toward fat and sugar after bypass is internally conflicting, and at least part of the disagreement traces to preoperative maintenance diet — animals raised on high-fat diet reduce their fat preference only gradually after surgery, where chow-raised animals do so immediately — a warning that preference shifts are conditioned as well as hormonal [141]. Second, the mediating pathways in humans remain unresolved: altered taste and olfaction, conditioned avoidance after aversive postingestive episodes, and hormone-driven reward remodeling are all plausible and poorly separated [9]. The conditioned-aversion pathway has direct rodent support — in the fat-preference experiments of le Roux and colleagues, a corn-oil gavage paired with saccharin induced a classical conditioned taste aversion in bypassed rats [142] — but the finding is not uniform: a second series found the oil-conditioned aversion after sleeve gastrectomy and *not* after RYGB [143], and no human study isolates the pathway from its hormonal companions. For the pouch question, note the geometry of the causal chain: reward remodeling sits downstream of the hormonal surge and the vagal traffic. It inherits whatever pouch dependence those upstream signals have, and adds none of its own — with the possible exception of dumping-conditioned avoidance, which rapid emptying provokes and which therefore shares the emptying-dynamics dependence discussed below.

5.2.6 Energy expenditure: clear in rodents, mixed in humans

Dieting lowers energy expenditure — adaptive thermogenesis — and this fall is part of why lost weight returns. Does RYGB, beyond escaping that fall, actively raise expenditure? In rodent models the answer is clearly yes, and the effect is load-bearing: in the murine model of Hao and colleagues, RYGB mice maintained an 18% weight reduction while eating roughly 35% more

than calorie-restricted mice held at the same weight, the balance paid by higher total and resting energy expenditure — up to 20–35% above weight-matched controls at thermoneutrality, concentrated in the dark (active) period [135]. The human evidence is mixed and methodologically fraught: expenditure estimates after major weight loss are sensitive to how body composition is adjusted for, to physical activity and the thermic effect of food, and to measurement timepoint. In the randomized RYGB-versus-gastroplasty trial with indirect calorimetry, basal metabolic rate after RYGB fell as predicted from the lost tissue mass, with no procedure-specific elevation [140], and reviews accordingly credit human RYGB’s advantage to the *absence* of the dieting-induced expenditure fall rather than to a positive thermogenic effect [9] — but one calorimetry trial cannot establish the absence of a human procedure-specific effect, and the escape from adaptive thermogenesis is itself not consistently demonstrated across studies. The species contrast is therefore a difference in the weight and quality of evidence, not a settled split; it still matters for reading the animal literature quoted throughout this chapter, because the rodent mechanism mix — expenditure-heavy — cannot be assumed to be the human mix. No account, in either species, ties expenditure to pouch size.

5.2.7 The threads assembled

Table 5.1 condenses the model: each thread, the strongest evidence for it, its standing, and — the column this dissertation adds — what it implies about constructed pouch size. The striking feature of the table is its final column. Of seven mechanistic threads, none predicts dependence on constructed pouch *volume* per se; one (ghrelin) predicts dependence on which anatomy is excluded, and one (the hindgut surge, with its downstream reward effects) predicts dependence on delivery dynamics that pouch geometry may shape. The neuroendocrine model, taken whole, is a theory under which the field’s forty-year pouch-size dispute is a dispute about a weakly coupled parameter.

5.3 What each mechanism predicts about pouch size

The models can now be made to disagree in public, which is this chapter’s purpose. Three positions generate distinct, testable predictions about the relationship Part II will estimate.

5.3.1 The restrictive prediction: volume as dose

Under the legacy model the prediction is unambiguous: constructed volume is the dose, smaller pouches produce earlier distension-mediated satiation, and the effect should be monotone, present at every follow-up horizon, and strongest exactly where measurement is best — true volumetry should show it more clearly than crude proxies, since proxies add noise and noise attenuates regression slopes toward zero (classical non-differential measurement error). A restrictive world is one where the modality hierarchy of Section 6.9 sharpens the effect as quality rises. The model has already absorbed direct hits — the emptying studies of Section 5.1 are incompatible with distension-retention mechanics — but the prediction remains the doctrine’s implicit content, and stating it precisely is what makes the doctrine falsifiable at last.

5.3.2 The pure neuroendocrine prediction: volume as packaging

Under the neuroendocrine model in its strong form, constructed pouch volume within the surgically practiced range is causally inert: the operative ingredients are fundus exclusion, gastric division, and the jejunal re-routing, all shared by every RYGB regardless of pouch size. The prediction for Part II is a pooled effect near zero on %TWL, *and* — the distinctive, risky part — a specific pattern across measurement modalities: any apparent association should concentrate in the weak-proxy stratum, where stapler counts and bougie sizes correlate with surgeon, era,

Table 5.1: Mechanistic threads of the neuroendocrine model of RYGB and their predicted dependence on constructed pouch size. Evidence classes: H = human experimental/observational, R = rodent, M = meta-analysis.

Thread	Key evidence	Standing	Predicted pouch-size dependence
Ghrelin exclusion	77% lower profiles vs. controls after RYGB (H)[45]; long-term rebound on pooling (M)[123]	Weakened; transient	On fundus exclusion and division, not volume; meta-regression: no pouch-volume association[123]
GLP-1/PYY surge	Anatomy-specific elevation (H)[46]; blockade raises intake (H)[9]; surge tracks emptying rate (H)[10, 110]	Robust; core thread	Indirect, via emptying dynamics; sign ambiguous (§5.3)
Bile acids (FXR/TGR5)	FXR-KO abolishes sleeve benefit (R)[136]	Contested: murine, sleeve-derived; kinetics heterogeneous, largely lagging the early effect [9]	None proposed beyond delivery rate
Microbiota	Conserved shift; transfer to germ-free mice transmits weight loss (R)[138]	Sufficient, share unquantified	None proposed
Vagal signaling	Preserved/increased after RYGB (H, indirect)[9]	Plausible, thin	On construction technique (nerve preservation), not volume
Central reward, food preference	Lower fMRI hedonic response (H)[139]; fat steering in RCT (H)[140]; rat findings conflict (R)[141]	Robust phenotype, contested mediation	Inherited from upstream signals; dumping-conditioned component shares emptying dependence
Energy expenditure	Expenditure-dominant in mouse (R)[135]; not seen in the one human RCT calorimetry study (H)[140]	Rodent-established; human evidence mixed	None proposed

technique, and patient factors, and should shrink toward zero in the imaged-volumetry stratum. This is the measurement-artifact hypothesis (Q3) in mechanistic dress. It is worth being precise about why the model tolerates the positive studies in the literature. Under proxy measurement, “pouch size” is a bundle: a surgeon who staples a longer pouch also, systematically, operated in a different era, with different limb lengths and different follow-up culture. The Scandinavian registry result — stapler length predicting weight loss at six weeks but not at one year across 14,168 patients [14] — reads naturally in this frame: whatever early, transient mechanical or edema-related effect exists washes out as the hormonal steady state takes over. And the one clearly positive long-term comparison, a retrospective cohort of small (10–20 mL) versus deliberately large (>50 mL, subtotal-stomach) pouches in low-BMI diabetic patients, showing better six-year BMI and remission outcomes with the small pouch [18], compares constructions so extreme that the large-pouch arm retains substantial acid-secreting and possibly ghrelin-producing mucosa in the food stream — a difference in *which stomach is excluded*, not merely in volume, exactly where the neuroendocrine model permits an effect.

5.3.3 The emptying-dynamics steelman: geometry as filter

The strongest mechanistic case that pouch construction *does* matter — the steelman this dissertation is obligated to respect, stated anatomically in Section 4.4 and restated here as a mechanism-level prediction — runs not through capacity but through dynamics. The hindgut surge is driven by the rate of nutrient arrival at the jejunum [10, 110]; the pouch and its outlet are the upstream hydraulic elements that set that rate; therefore pouch geometry could modulate the operation’s principal signal. In circuit terms the pouch is a matching network between the esophagus and the jejunal detector: its volume, compliance, and outlet impedance shape the pulse the L cells see. This argument makes pouch construction matter in a hormonal world — but it fractures, honestly followed, into two schools with opposite signs.

The *fast-emptying school* holds that smaller pouches empty faster, faster delivery means a sharper hormonal pulse, and a sharper pulse means stronger satiety: small is better, but through signaling rather than restriction. Its evidence is real. Riccioppo and colleagues, with 3D volumetry and scintigraphy at a median of 47 months, found that pouches of 40 mL or less emptied faster, and that faster emptying — not smaller volume directly — predicted higher %EWL, less regain, and better food tolerance [48]. Akkary and colleagues found that patients whose day-one contrast study showed prompt pouch emptying lost significantly more weight at one year than slow emptiers [111] — the “paradox of the pouch,” paradoxical only under the restrictive model it embarrassed. Dirksen’s association between fast liquid emptying and high hormone secretion supplies the mediating link [10].

The *slow-emptying school* holds nearly the reverse: a longer, narrower pouch empties more gradually, sustains satiety signaling and fullness over the meal cycle, and — by Laplace’s law, wall tension in a narrow tube being lower at equal pressure — resists dilating over the years: long and narrow is better. This is not a straw position; it is the explicit design rationale, invoking Poiseuille and Laplace, of the only extended-pouch randomized trial, whose three-year result favored the *longer* pouch (31% versus 27% TWL, driven by less regain) [13]. At five-to-nine years the difference had attenuated to a non-significant trend (26.1% versus 24.1% TWL; regain 70% versus 47%, $p = 0.07$) in a follow-up sample reduced to 47% of the trial [16] — but the direction never reversed, and the trial remains the literature’s only randomized evidence pointing toward *larger* constructed pouches, in direct opposition to the fast-emptying school and to the small-pouch doctrine alike.

The two schools cannot both be right as stated, and their coexistence is itself informative. What they share is that the operative variable is *geometry and its hydraulic consequence*, not volume: a short wide 20-mL pouch and a long narrow 20-mL pouch are identical in every volume-classified study and may be opposites in emptying behavior. The mediating measurements are also thin — Riccioppo’s volumetry is late (median 47 months), so construction and dilation are

confounded (Section 6.7); the Dutch trial varied length and volume together, so its contrast is not clean in either variable. The emptying-dynamics steelman thus predicts, for Part II: a real but *shape-dependent* effect, inconsistent in sign across studies that classify exposure by volume or length alone — which is, notably, a mechanism that would generate precisely the heterogeneity that defeated pooling in the field’s only systematic review, where nine of fourteen studies found no association, five favored smaller pouches, and none favored larger [15].

5.3.4 The safety gradients: harms in both directions

Safety attaches pouch construction to harm in both directions, and the two directions must be kept separate. The better-documented gradient penalizes *larger* construction: the larger the pouch, the more acid-secreting mucosa remains in the food stream and the longer the staple line, and marginal ulcer — ulceration at the gastrojejunal junction, where jejunal mucosa faces unbuffered acid — is reported to scale accordingly. In the Scandinavian registry, each additional centimeter of stapler used for the pouch raised the relative risk of marginal ulcer by 14% (95% CI 9–20%) [14]; in the large-versus-small cohort, six-year marginal ulcer incidence was 23.7% against 6.4% ($p = 0.023$) [18]. Two qualifications temper the mechanistic neatness: stapler length is a construction proxy, not pouch volume; and marginal ulcer is multifactorial — smoking and diabetes are the best-documented patient-level risk factors, in both an endoscopic case series and a 37,301-patient registry analysis that each also found pouch dimensions predictive [57, 126], while NSAID use, mucosal ischemia, staple-line foreign material, and *Helicobacter* status are classically implicated with less consistent confirmation (the endoscopic series found no NSAID association) [96, 57] — so an unadjusted gradient is plausible mechanism plus plausible confounding, not a guaranteed monotone law. Note what this gradient is *not*: it is not a counterweight to a small-pouch weight-loss benefit. It points the same way — toward smaller construction — and so cannot by itself define a trade-off. The genuine counterweight is the set of harms plausibly attached to very small or tight construction: anastomotic stricture, tension on the gastrojejunostomy with its leak risk, and food intolerance. These are less systematically quantified in the pouch-size literature, which is itself a finding Part II must report rather than paper over; the safety question (Q5) is therefore analyzed per outcome, in both directions, as absolute risks (Section 13.1.4). If Part II finds no weight-loss effect of pouch size, the rational surgical policy does not become indifference — it becomes “small enough for ulcer prevention, and no smaller than the stricture, tension, and intolerance risks allow,” with each side of that sentence resting on its own evidence. Dumping, sometimes proposed as a small-pouch penalty, showed no pouch-size dependence in the one randomized comparison reporting it (Sigstad scores 7.0 versus 7.2 at long-term follow-up) [16].

5.4 From mechanism to hypotheses

The mechanistic account now yields the dissertation’s hypotheses, stated before the data of Part II are seen and scored against them in Chapter 14.

H1 (primary effect). Under the neuroendocrine model that the evidence of this chapter supports, the pooled effect of smaller constructed pouch size on %TWL at one year and beyond (Q1) is expected to be small — clinically negligible against the operation’s ~25–35% total weight loss — and statistically fragile. The legacy model predicts the opposite: a consistent, monotone advantage for small pouches. The primary analysis (Chapter 10) supplies the estimate both predictions are scored against.

H2 (dose–response). If an effect exists, pure neuroendocrine logic predicts no smooth volume gradient (the mechanism offers volume no channel), while emptying-dynamics logic predicts a

gradient organized by geometry rather than volume — and therefore a weak or incoherent gradient when exposure is expressed volumetrically (Q2). A strong, smooth volumetric dose–response would be the single most restrictive-model-friendly result Part II could produce.

H3 (measurement artifact). Under pure neuroendocrine logic the modality analysis (Q3) should show the association, where it appears at all, concentrated in construction proxies and absent under imaged volumetry — the pattern expected if the forty-year dispute is largely a dispute about proxy-correlated confounding rather than about dose. The restrictive model predicts the mirror image: sharpening under better measurement. This is the dissertation’s pivot, and the mechanistic grounds for expecting artifact have now been stated in full — with the caveat, developed in Chapter 6, that the modality comparison is an exploratory effect-modification analysis across studies that also differ in design, era, and size, so a modality-dependent pattern updates confidence in the artifact reading rather than proving it.

H4 (evidence conflict). The mechanistic account predicts that the field’s flagship contradictions (Q4) — registry null at one year against positive cohorts, and the small-pouch doctrine against the extended-pouch trial — are not random noise but structured: early horizons and extreme anatomical contrasts (fundus-retaining large pouches) should generate effects, hormonally similar constructions compared at late horizons should not, and volume-classified exposures should scramble any shape-dependent signal [14, 18, 13].

H5 (safety, both directions). Marginal ulcer (Q5) is expected to increase with larger construction across the evidence base, independent of any weight-loss effect — the acid-load and staple-line mechanism makes the direction plausible, though confounding by smoking, NSAIDs, and technique means the gradient is expected, not guaranteed [14, 18]. The harms plausibly attached to very small or tight construction — stricture, anastomotic tension and leak, food intolerance — are predicted to run the opposite way, and Part II analyzes each safety outcome separately, in both directions, as absolute risks, without collapsing benefit and harm onto a single axis.

H6 (OAGB contrast). One-anastomosis gastric bypass varies pouch length over a far wider range than RYGB while sharing the hindgut re-routing. If the neuroendocrine reading is right, even this widened anatomical variance should couple weakly to weight loss; if emptying dynamics matter, the long narrow OAGB pouch is the field’s inadvertent dose-ranging experiment (Q6). Section 13.2 tests this.

H7 (mechanistic scoring). Finally, the dissertation commits to reading its own results back through this chapter (Q7) — as updates of confidence among the three positions, not as adjudication of mechanism. A study-level observational synthesis cannot identify mechanism: a null under good measurement is consistent with the neuroendocrine model but also with restricted exposure range, limited power, residual measurement error, and shape–outlet variation that volume classification misses; a surviving volumetric effect would favor the emptying-dynamics amendment — and with it a research program on pouch *shape*, compliance, and outlet behavior rather than volume — without uniquely establishing it; an undecidable result specifies the measurement standard the decisive study needs. What Part II can deliver is a shift in the weight of evidence, stated per position. The referee’s scorecard is now written — it records confidence, not verdicts — and Part II plays the match. Before it can be played, though, both of the match’s measuring instruments need calibration: how the exposure is measured (Chapter 6) and how the outcome is reported (Chapter 7) turn out to be problems as deep as the mechanism itself.

Chapter 6

Measuring the pouch

Every chapter of Part II depends on a single variable: the size of the constructed gastric pouch. This chapter examines how that variable is actually produced. The conclusion can be stated at the outset, because the evidence for it is overwhelming and the rest of the chapter simply lays it out: pouch size enters the literature through at least five mutually incompatible measurement channels, none of which has been validated against any other on more than a few dozen patients; the best-validated channels are used almost exclusively in small single-center studies, while the largest cohorts rely on the weakest proxies; and the quantity itself changes by a factor of several over the years in which outcomes are assessed. A field that has argued for forty years about whether pouch size predicts weight loss has, for most of those forty years, not been arguing about the same physical quantity.

This is not a rhetorical flourish but the operational premise of the dissertation’s third research question. Classical measurement error attenuates associations toward null, so noise alone cannot manufacture an effect: if the pouch-size association is concentrated in studies using weak proxies and absent under direct volumetry, the suspect is not the noise itself but whatever travels with the proxy — confounding by surgeon, era, and technique, or selective reporting — while a real effect should be clearest where measurement error is smallest. Examining that pattern requires a defensible, explicitly ordered scale of measurement quality, and this chapter closes by defining one. Two honesty notes attach to it at birth. The ordering was drafted with the preliminary literature scan in view — the same studies quoted throughout this chapter — so the stratified analysis it supports is exploratory and hypothesis-generating, not a confirmatory test; the repository’s version history documents when it was fixed. And because measurement tier travels with study era, design, and size, the between-tier comparison is an exercise in weighing competing explanations (Chapter 11), not a verdict on measurement error. The sections in between treat each modality with the seriousness its users intended: what it measures, what error it introduces, and what its validation record actually contains.

6.1 What quantity is being measured?

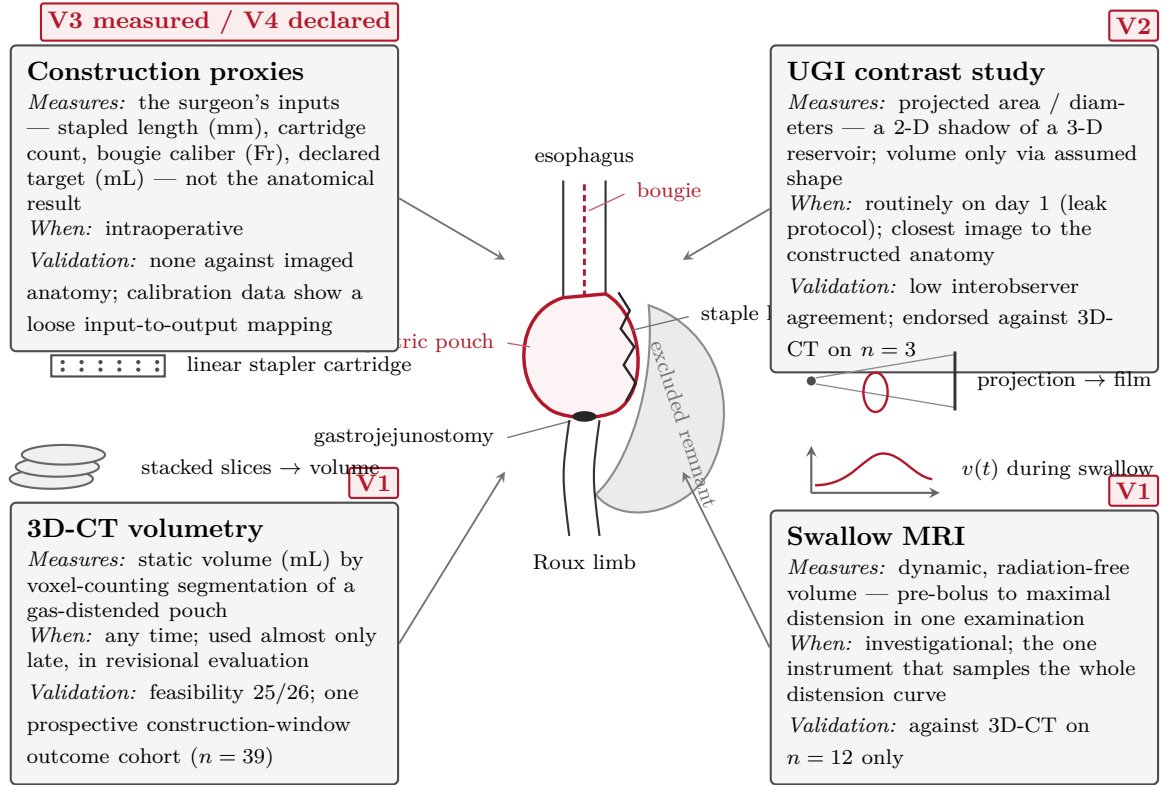
Measurement science begins by defining the measurand, and here the literature fails before any instrument is picked up. “Pouch size” is reported in the primary literature as a volume (milliliters), a length (centimeters of staple line or of pouch axis), an area (square centimeters on a radiograph), a count (stapler cartridges fired), an instrument caliber (the diameter of the calibrating tube the pouch was built around), or a categorical impression (“small,” “dilated”). These are not interchangeable parameterizations of one quantity. A pouch is roughly a tube-like reservoir; for an idealized cylinder the enclosed volume scales linearly with length but with the *square* of the diameter, so two pouches of identical measured length can differ severalfold in volume if one is wide and one is narrow. This is not a hypothetical: when a randomized trial reported that a longer pouch produced *better* maintained weight loss [13], the ensuing

exchange of letters argued explicitly that width, not nominal size, is the operative dimension [107, 144]. Length-based and volume-based studies are therefore measuring partially independent quantities, and any synthesis that pools them is pooling different exposures.

The deeper problem is that the pouch is not a rigid vessel. It is a distensible muscular organ, and the volume of a distensible vessel is not a number but a function — volume at a given distending pressure, the integral of a compliance curve. The single most instructive measurement in the entire validation literature makes this concrete: in the one study that imaged the pouch dynamically, mean pouch volume was 9.5 ± 5.5 mL before a liquid bolus and 40.4 ± 21.0 mL at maximal distension during swallowing — a more than fourfold change within a single examination of the same patients on the same day [19]. Any reported “pouch volume” is therefore meaningful only relative to a distension protocol: what was swallowed, under what instruction, imaged at what moment. Protocols differ across studies and are frequently unreported. In signal-processing terms, the field has been comparing amplitude readings taken at unrecorded points of a periodic waveform.

The error structure matters as much as the error magnitude, because different error structures corrupt an outcome analysis in different ways — and everything in this paragraph is conditional on an assumed error model, which must be stated as an assumption rather than smuggled in. Under *classical* error — additive, nondifferential, independent of the true value and of the outcome — systematic bias shifts group means but leaves within-study contrasts comparatively intact, becoming dangerous only when studies measured under different biases are compared or pooled, which is precisely what a meta-analysis of this literature must do; random per-patient noise leaves means alone but attenuates every correlation and regression slope toward zero, the regression-dilution effect familiar from any errors-in-variables analysis. If pouch size were measured with classical reliability ρ — the fraction of measured variance that is true variance — an observed correlation would understate the true one by roughly a factor $\sqrt{\rho}$, and an observed regression slope by a factor ρ . What this literature supplies, however, is not ρ . The one available between-instrument statistic, an intraclass correlation of 0.594 for gastroscopy against CT volumetry [145], is an agreement coefficient between two imperfect, differently scaled instruments, neither of which is established as error-free; it documents substantial disagreement, but it is not the classical reliability ratio, and no quantitative attenuation factor can be extracted from it. Estimating one would require repeatability data or a validation model against a defensible reference with error terms for both instruments — neither of which exists in this anatomy. The defensible statement is qualitative: instrument disagreement of this order is ample to attenuate a clinically meaningful association substantially, *if* the classical model holds. And the classical assumptions are themselves contestable here: differential filling protocols, segmentation choices, surgeon effects, and outcome-triggered imaging can make the error differential in practice, biasing contrasts in either direction rather than only toward null. Two asymmetric consequences survive with those caveats attached. Null results from noisy modalities are weak evidence of no effect. Positive results from noisy modalities — which classical noise would push toward null — call for explanations beyond the noise: chance, selective reporting, differential error, confounding of the exposure with the surgeon or era that produced it, or a genuinely large effect. Any serious reading of this literature has to carry both halves of that asymmetry, and the hierarchy at the end of this chapter is built to make them explicit.

Two further distinctions organize everything that follows. First, *anatomical* versus *functional* size: the anatomical capacity of the reservoir versus the volume a patient can or does actually use, which also depends on emptying through the gastrojejunostomy (the surgical connection between pouch and jejunum) and on behavior. Second, *construction* size versus *current* size: what the surgeon built versus what exists at the time of measurement, months or years later, after remodeling. Section 6.7 shows that these diverge substantially over time, which is why this dissertation’s exposure is construction size only.



Tier tags V1–V4: the modality-quality hierarchy of Section 6.9 — V1 measured volume, V2 imaged geometry, V3 measured construction, V4 declared technique.

Figure 6.1: Four measurement modalities for gastric pouch size, arranged around a stylized RYGB anatomy (esophagus, calibration bougie, pouch, staple line, excluded remnant, Roux limb). Each panel states what the modality actually measures, when in the clinical course it is applied, and the extent of its validation record as documented in this chapter; tags give the tier each modality occupies in the modality-quality hierarchy of Section 6.9. The fifth channel, endoscopic estimation, measures a collapsed, insufflated cavity from inside and is treated in Section 6.6. Diagram: AI-generated vector illustration (see figure provenance, Appendix E).

6.2 Construction proxies: bougie and stapler

The oldest pouch measurements were made by the surgeon’s eye at the operating table. In the 1982 study that helped fix the small-reservoir doctrine, every patient was described as receiving “a similar 30–45 ml proximal gastric reservoir” — an intraoperative volume estimate, with the study’s inferential weight resting on the assumption that the estimate was accurate and uniform [5]. The 1984 radiologic data reviewed below, showing a fivefold volume spread after a nominally standardized operation [78], suggest how optimistic that assumption was. The modern descendants of the eyeball estimate are more disciplined but share its defining property: the most common “measurement” of pouch size is not a measurement at all but a record of how the pouch was built. Two instruments dominate. The *bougie* (or orogastric calibration tube) is a flexible tube passed through the mouth into the stomach, against which the pouch is cut; its caliber is given in French units (Fr), where 1 Fr equals one-third of a millimeter, so the common 36–40 Fr tubes are 12–13 mm in diameter. The *linear stapler* divides the stomach along a line, placing rows of staples on both sides of the cut; pouch construction takes a sequence of stapler firings, each from a cartridge of fixed length (typically 45 or 60 mm). The number of cartridges fired, or the total stapled length, is then recorded as the size of the pouch.

Both proxies encode real constraints — a pouch cut flush against a 40 Fr tube cannot be arbitrarily wide at the tube, and more stapling generally means a longer pouch — but the chain of inference from proxy to volume compounds error at every link. The bougie fixes, at best, one diameter at one position; how much stomach is left lateral to the tube depends on stapler angulation and traction, which vary between surgeons and between operations. Stapled length is a perimeter measurement, not a volume: it says nothing about width, and by the cylinder geometry above, volume is more sensitive to the unmeasured dimension than to the measured one. Intraoperative conditions add further variance — insufflated abdomen, paralyzed patient, tissue under traction — so even the recorded length is a measurement of tissue in a state that no postoperative image reproduces.

The proxies are also unstandardized in practice. A national survey of 183 bariatric centers in China found that 86% used a specialized bougie, with calibers ranging from 30 to 38 Fr and 36 Fr the most common; 42% of centers reported bougie-related complications [146]. Whether calibration is used at all measurably changes the result: in a case-matched comparison in which the only technical difference was the use of a 40 Fr orogastric calibration tube, the calibrated group achieved 33.3% total weight loss at one year versus 28.6% without calibration ($p = 0.007$) [147]. Read as measurement science rather than as surgical technique, this is a striking result: standardizing the instrument moved the outcome by almost five percentage points of total body weight — larger than most claimed pouch-size effects. One reading is that uncalibrated construction scatters the true exposure widely enough that “pouch” without calibration is barely a defined variable; but the inference is not forced, because calibration is itself a surgical intervention and a case-matched comparison cannot exclude residual confounding by whatever else distinguishes calibrated practice. The result motivates treating construction scatter as a live error source; it does not measure the scatter.

What do stapler-based proxies deliver when used as the exposure in outcome studies? The largest cohort in this literature — and the largest study of pouch size ever conducted — is the Scandinavian registry analysis of 14,168 patients, in which technique was highly standardized and total stapler length used to complete the pouch (mean 145 mm) served as the pouch-size proxy [14]. Stapler length predicted excess-BMI loss at six weeks but not at one year, while each additional centimeter of stapler raised the relative risk of marginal ulcer — an ulcer at the gastrojejunal anastomosis — by 14% (95% CI 9–20%) [14]. A small retrospective cohort comparing pouches built with two versus three 60-mm cartridges found no difference in any six-month weight outcome (total weight loss 28.2 ± 6.0 versus $26.1 \pm 7.5\%$; $p > 0.05$), and its authors explicitly named measurement-method heterogeneity as a driver of the conflicting literature [148]. Note the logical structure of the registry result: the proxy correlates with

outcome early, when anatomy is as constructed, and decorrelates by one year — exactly the signature expected if either (a) pouch size does not govern one-year weight loss, or (b) the proxy has lost its connection to current pouch size through remodeling. Distinguishing (a) from (b) is impossible within a proxy-measured study, which is precisely why the modality question must be examined across studies (Chapter 11).

The most careful proxy work shows what the approach can achieve at its best — and where its ceiling is. In a prospective series of 231 patients, pouch construction was standardized against an orogastric tube, pouch height was recorded as the number of stapler firings, and a per-patient volume of 10–20 cm³ was calculated from the measured dimensions; across that deliberately narrow range, excess weight loss at one year did not differ by pouch size [105]. This is an intraoperative measurement, not a declaration, and its null result is informative — but only within its twofold volume range. A study that restricts its exposure to 10–20 cm³ can say nothing about the 30, 60, or 100 mL pouches that the imaging literature routinely documents, and a narrow exposure range mechanically shrinks any detectable correlation: the signal variance is capped by design while the outcome variance is not. Range restriction and measurement noise thus push in the same direction, and several of the field’s null results operate under both.

Some studies go one step further removed and report only the surgeon’s intended volume — a randomized comparison of “10–20 mL” versus “25–35 mL” pouches, for example, with no per-patient measurement of what was actually created [17]. There the exposure is a declared target, not a measurement: the study design assumes exactly the construction precision that the calibration data above call into question.

6.3 Contrast radiography

The upper gastrointestinal (UGI) contrast study is the traditional postoperative image of bypass anatomy: the patient swallows a radiopaque liquid (barium suspension, or a water-soluble agent when leak is suspected) and the pouch is imaged under fluoroscopy — real-time projection X-ray. It is cheap, ubiquitous, and was for decades performed routinely in the first days after surgery. It is worth remembering why: the routine early UGI study exists to detect anastomotic leak, a life-threatening complication identified in about 5% of patients in a large early series, most within the first postoperative week [149]. Pouch-size measurement on these films is opportunistic — a secondary reading of an image acquired, positioned, and timed for a different clinical question. That opportunism is also the modality’s unique strength for this dissertation: the day-one UGI is the only widely available image of the pouch as constructed, before any remodeling, which makes it the modality closest in time to the exposure of interest. Its physics, however, are unforgiving: a UGI film is a two-dimensional projection of a three-dimensional object, so the observable is a projected area or a pair of diameters, and any volume estimate requires an assumed shape. Filling is incomplete and posture-dependent; the distension state at exposure is uncontrolled.

The UGI record on pouch size versus weight loss is the purest demonstration in this literature that modality alone does not determine the answer — because the same modality, applied at the same postoperative moment, has produced both answers. In 320 patients imaged on postoperative day one, pouch area on the anteroposterior film (mean 63.9 cm²) correlated negatively with excess weight loss at six months ($\beta = -0.241$, $p < 0.01$) and twelve months ($\beta = -0.302$, $p < 0.02$) — larger pouch, less weight loss [103]. In 82 patients imaged on postoperative day one with essentially the same technique, the correlation coefficients between pouch area and excess weight loss at 3, 6, 12, and 24 months were 0.038, 0.110, 0.015, and 0.042 — indistinguishable from zero [118]. A third study had three blinded experts grade pouch and fundus size on ordinal scales from routine UGI films and found no significant weight-loss differences across grades [106], and a barium-swallow assessment at a mean of three years found identical excess weight loss whether the measured pouch was above or below its 50-mL dilation threshold [119].

The reliability record is thinner than the usage record and largely adverse. A retrospective

case-control reassessment of UGI studies in patients with weight-loss failure found pouch dilation in 23% of failures versus 11% of well-losing controls ($p = .024$) — a statistically real but weak discrimination — and the authors concluded that low interobserver agreement “greatly questions the reliability and usefulness” of the modality for pouch assessment [150]. A barium-study series using dimensional criteria (pouch >6 cm long or >5 cm wide) found a “large pouch” in 37% of patients with recurrent weight gain — and in 52% of controls with satisfactory weight maintenance [151]: the feature the examination is ordered to find was more prevalent in the patients doing well. Endoscopic comparison points the same way: in a series of weight-regain evaluations, eleven dilated pouches were diagnosed at endoscopy but only two of them were also detected on the UGI study [152]. And the difficulty is not new: already in 1984, a prospective study that measured pouch volume with a standardized radiologic protocol after a nominally uniform gastroplasty found early volumes ranging from 51 to 244 mL — a nearly fivefold spread in an operation intended to be identical — with no correlation between pouch or stoma size and weight loss [78]. That 1984 finding contains the whole later literature in miniature: either the operation cannot be built to specification, or the radiologic measurement cannot read what was built, and the study cannot tell which.

6.4 Computed tomography volumetry

Three-dimensional computed tomography (3D-CT) volumetry is the closest thing the field has to a reference method. The patient drinks dilute contrast and effervescent granules (which release carbon dioxide, distending the pouch with gas); a multidetector CT acquires an isotropic volume image; and the pouch is segmented in software, its volume computed by voxel counting rather than inferred from shape assumptions. The measurement is genuinely volumetric, observer-independent to the extent segmentation is, and captures anatomy the other modalities miss.

Its validation history is short and its samples are small. The first report imaged three patients on postoperative day one (mean calculated pouch volume 31.6 cm^3) and, in those three patients, found pouch area by 3D-CT statistically indistinguishable from the area on the routine UGI film (15.2 versus 16.9 cm^2 , $p = 0.549$) [153] — a result frequently taken as validating the UGI-area surrogate, on an evidential base of $n = 3$. The first systematic series applied multislice-CT volumetry to 26 post-bariatric patients being evaluated for revision and succeeded in 25 (96%), demonstrating feasibility and clinically useful anatomical detail (pouch volume over 30 mL in ten patients, widened gastrojejunostomy in eight) [154]. Larger series exist only in the revisional-planning context, where 3D-CT is now routine at some centers — 144 patients in one report [155] — and in one-anastomosis gastric bypass (OAGB), where 3D-CT volumetry has been used to characterize pouch volume and intrathoracic pouch migration in series of 50 [156].

Only one study has used CT volumetry prospectively to ask this dissertation’s question, and it deserves emphasis because it is the methodological high-water mark of the exposure literature: 39 RYGB patients underwent 3D gastric CT with gas at 3 and 12 months, with zero loss to follow-up, and the measured pouch volume at 3 months showed no correlation whatsoever with excess-BMI loss at one year ($r = .01$, $p = .94$) [157]. Under the best available measurement, in the only prospective volumetric cohort, the effect is absent — a single small study, but pointing exactly where the measurement-artifact hypothesis predicts.

The modality’s limitations are operational rather than physical: ionizing radiation (a mean dose-length product of $282.6 \pm 149.7 \text{ mGy}\cdot\text{cm}$ in one protocol [19]), dependence on adequate gas distension — which failed in 2 of 30 patients in one comparative series [145] — supine positioning that alters pouch configuration, and a static acquisition that samples one instant of the distension curve. None of these prevents CT volumetry from anchoring a measurement hierarchy; all of them explain why it has never been applied at registry scale.

6.5 Swallow magnetic resonance imaging

Magnetic resonance imaging offers the one thing CT cannot: dynamic, radiation-free acquisition. In swallow MRI the patient drinks water while lying in the scanner and fast sequences image the pouch throughout the swallowing process, so the examiner can read volume at true maximal distension rather than at an arbitrary post-ingestion moment — a genuine solution to the distension-state problem of Section 6.1.

Its entire validation against CT in bypass anatomy rests on twelve patients. In the single comparative study — twelve post-RYGB patients, all female, all previously revised for weight regain, each undergoing both examinations within two weeks — mean pouch volume was 40.4 ± 21.0 mL by swallow MRI versus 43.5 ± 30.2 mL by 3D-CT ($p = 0.83$), anastomotic diameter agreed to within a tenth of a millimeter on average, and both modalities identified intrathoracic pouch migration in the same nine patients [19]. The authors are candid that the sample “is too small to draw any definitive conclusions,” and the practical costs are real: roughly half an hour of scanner time against five minutes for CT, 1.7-fold higher cost in their system, and lower volumetric precision due to the long dynamic acquisition [19]. The study is careful, prospective, and internally consistent — and it is the *only* bridge between the field’s two best instruments. The agreement of the two near-reference methods, on which any claim that “direct volumetry” is a coherent category depends, is a twelve-patient result. This claim was re-verified against the published record during drafting; the study is real, singular, and unreplicated in RYGB as of the corpus snapshot date.

6.6 Endoscopy and functional measurement

Upper endoscopy (gastroscopy: a flexible camera passed into the pouch) is the workhorse of weight-regain evaluation because it sees mucosa, ulcers, and fistulas directly. As a size-measuring instrument, however, it performs poorly, for an obvious physical reason: it estimates the dimensions of a collapsed, insufflated cavity from inside, with a wide-angle lens and no external reference frame. In the one direct comparison against 3D-CT — 30 revised RYGB patients examined with both on the same day — gastroscopic and CT pouch measurements agreed within 40% in only 16 of 28 evaluable patients, the intraclass correlation coefficient was 0.594, and gastroscopy identified *none* of the 20 intrathoracic pouch migrations that CT demonstrated [145]. Endoscopy and UGI radiography detect partially disjoint abnormality sets, which is why the two are described as complementary rather than redundant [152]. An emerging quantitative endoluminal tool, the functional lumen imaging probe (EndoFLIP), measures cross-sectional area and distensibility by impedance planimetry and is beginning to be applied to bariatric anatomy — chiefly the sleeve and the gastrojejunal anastomosis rather than pouch volume — but its role remains investigational [158].

One historical method deserves mention because it measured the variable every other modality approximates: functional capacity. Flanagan’s “cottage cheese test” had patients eat a standard food to satiety under a structured protocol, directly quantifying usable pouch volume [159]. Crude as it is, it operationalizes the functional measurand — and its central finding, discussed next, is one of the field’s most important.

6.7 Dilation over time: the moving target

Everything above concerns how well an instrument reads the pouch at one moment. The second, independent confound is that the pouch does not hold still. Flanagan’s functional measurements showed a statistically significant stepwise increase in usable pouch volume at every interval from three months to two years, with one-year functional volumes ranging from 2.5 to 9.0 ounces (roughly 75–270 mL) — and, within that range, no difference in excess weight loss between

Table 6.1: The complete cross-modality validation record for pouch measurement after gastric bypass, as identified in the assembled corpus. Every between-modality agreement estimate in this literature rests on 30 or fewer patients.

Study	Comparison	n	Key result
Alva 2008 [153]	3D-CT vs. UGI area	3	Pouch area 15.2 vs. 16.9 cm ² ($p = 0.549$); UGI area endorsed as volume surrogate
Karcz 2009 [154]	MSCT volumetry (feasibility)	26	Volumetry evaluable in 25/26 (96%)
Felsenreich 2020 [19]	Swallow MRI vs. 3D-CT	12	Volume 40.4 ± 21.0 vs. 43.5 ± 30.2 mL ($p = 0.83$); identical migration detection
Arnoldner 2020 [145]	Gastroscopy vs. 3D-CT	30	ICC 0.594; 0/20 pouch migrations seen endoscopically
Uittenbogaart 2020 [150]	UGI interobserver	202	Dilation 23% failures vs. 11% controls ($p = .024$); low interobserver agreement
Brethauer 2006 [152]	Endoscopy vs. UGI	30	11 dilated pouches at endoscopy, 2 detected on UGI

smaller and larger pouches [159]. The most dramatic published trajectory comes from sleeve gastrectomy rather than bypass — mean gastric volume in patients with weight regain rising from about 120 mL early after surgery to 524 mL at five years, as summarized in a scoping review of the weight-regain literature [160] — but the direction is general: an operated stomach remodels toward larger capacity, and it does so on the same timescale over which weight-loss outcomes are measured. What the field does not possess is the RYGB analogue of that curve: no serial volumetric trajectory — the same patients’ pouch volumes imaged repeatedly across postoperative years — was identified in the assembled corpus or in a targeted PubMed search (September 2026). The nearest design is the prospective cohort that repeated 3D-CT volumetry at three and twelve months expressly to test whether progressive distension drives weight-loss failure [157]; beyond Flanagan’s functional measurements, the multiyear growth of the bypass pouch rests on inference from cross-sectional revision series, not on repeated measurement of the same anatomy.

Clinical practice treats late enlargement as a disease entity in its own right: “pouch dilatation” is an accepted indication for laparoscopic pouch resizing [161], with published series operating on pouches over 200 mL [162] — volumes an order of magnitude beyond any construction target. The randomized extended-pouch trial adds the sharpest twist: the group constructed with the *longer* (10 cm versus 5 cm) pouch showed no weight regain in the third postoperative year while the standard-pouch group regained, ending with superior total weight loss (31% versus 27%, $p = 0.023$) [13] — construction size and late functional behavior are not even monotonically related.

For measurement, the consequence is decisive. A pouch volume measured at year three is the sum of construction size and years of remodeling, in unknown proportion; correlating it with concurrent weight (as the failure-evaluation literature implicitly does) cannot distinguish “large construction caused poor loss” from “regain and its causes dilated the pouch.” The causal arrow is unidentifiable in cross-section. Only measurements made at or shortly after construction — before remodeling accumulates — expose the surgeon’s actual decision variable. This is the measurement-theoretic justification, argued from the field’s own data, for the construction-only exposure scope fixed in the protocol (Chapter 8).

6.8 What the measurement problem does to the evidence

Before defining the hierarchy, it is worth stating precisely how each of the failures documented above propagates into the outcome literature, because the propagation paths differ and Chapter 11 will need to distinguish them.

Exposure-definition heterogeneity blocks pooling. Two studies that both report “small versus large pouch” but define small by cartridge count in one case and by imaged volume in the other are not estimating the same contrast, and no random-effects model can repair that, because the incompatibility sits in the exposure, not in the effect distribution. The only prior systematic review of this question declined to meta-analyze its fourteen studies on exactly these grounds of quality and heterogeneity [15]. The present dissertation’s wager is that the heterogeneity is not noise but structure: that exposure contrasts become more nearly compatible *within* measurement tiers, and that between-tier differences are themselves informative — as exploratory effect modification to be weighed against competing explanations (era, design, scale), not as a verdict (Chapter 11).

Construction scatter dilutes declared contrasts. Trials that randomize to a target size — 10–20 versus 25–35 mL [17], two versus three cartridges [148] — avoid per-patient measurement error by design, but they inherit a subtler problem: the delivered volumes within each arm are unmeasured and, on the evidence of the calibration and radiologic data above [147, 78], widely scattered. If the scatter is comparable to the separation between targets, the arms overlap in delivered anatomy and the trial estimates a diluted version of its nominal contrast — the same mechanism, at the group level, as regression dilution at the patient level. A positive result surviving this dilution suggests either a real and sizable effect or a confounded one; a null result is close to uninformative about the anatomical question.

Outcome-triggered measurement reverses the inferential arrow. Much of the imaging literature is generated by clinical indication: pouches are imaged *because* the patient regained weight [154, 151, 150]. Datasets assembled this way cannot estimate the effect of size on outcome, because the outcome determined who was measured; they can only compare measured failures with whatever controls were imaged for other reasons. The barium series in which “large pouch” was more prevalent among well-losing controls (52%) than among regainers (37%) [151] is best read not as evidence that large pouches protect weight loss but as a demonstration of how selection plus weak discrimination scrambles a cross-sectional comparison. The construction-window rule in the hierarchy below exists largely to fence this literature off from the causal question.

Instrument disagreement manufactures contradiction. Finally, the field’s flagship discordances shrink when the instruments are kept in view. The day-one UGI studies that disagree [103, 118] share a modality but differ in population size (320 versus 82), in analytic handling of the area variable, and in the outcome horizon at which correlation was sought; with a modality whose interobserver reliability is documented to be low [150], two honest applications of the same protocol can land on opposite sides of significance without either being wrong. The systematic dissection of these conflicts belongs to Chapter 11; the point established here is that measurement alone is sufficient to generate much of the observed disagreement, which is what makes a measurement-stratified reanalysis worth conducting at all.

6.9 The modality-quality hierarchy

The preceding sections support three summary judgments. First, direct volumetric imaging (3D-CT, swallow MRI) measures the quantity of interest, but its cross-validation rests on samples of 3 to 30 patients and it has essentially never been used at scale. Second, planar contrast radiography measures a projection of the quantity, with documented poor interobserver reliability and weak discrimination, but when performed early it at least observes the constructed anatomy. Third, construction proxies — stapler length, cartridge count, bougie caliber, declared target volume

— measure the surgeon’s inputs, not the anatomical output, and the one study that isolated calibration shows the input-to-output mapping is loose. The only prior systematic review of this question found fourteen studies too heterogeneous to pool, with nine null and five positive [15]; the present chapter’s reading is that much of that heterogeneity is measurement heterogeneity, and that it is testable.

The comparison requires the ordering to be fixed before systematic extraction begins, and fixed transparently: the hierarchy was drafted after the preliminary scan that surfaced these studies, so it is a scan-informed stratification — hypothesis-generating, not confirmatory — whose fixing date the repository’s version history records. It is defined here, in Part I, as a pre-commitment in the limited, honest sense available: study-level effect estimates in Chapter 11 will be stratified by these tiers exactly as defined below, and any post-hoc reclassification will be reported as such.

Tier V1 — Measured volume. The exposure is a per-patient pouch volume from cross-sectional volumetric imaging (3D-CT volumetry or MRI volumetry) under a described distension protocol, acquired within the construction window defined below. This is the only tier in which the recorded number is the physical quantity of interest. Exemplar: [157] — the only construction-window volumetric outcome cohort identified. The revisional-planning volumetry literature [154, 155] demonstrates the technique but images late, remodeled anatomy in patients selected for evaluation; it measures current size, not construction, and is excluded from this tier by the timing rule below (Section 6.7).

Tier V2 — Imaged geometry. The exposure is a per-patient quantitative measurement from planar imaging or calibrated intraluminal measurement: UGI pouch area or diameters, radiologic volume estimates from assumed geometry, or instrument-referenced endoscopic measurement. The number measures a projection or section of the pouch, not its volume. Exemplars: [103, 118, 78].

Tier V3 — Measured construction. The exposure is a per-patient intraoperative measurement of the construction itself: total stapler length, measured pouch length in centimeters against a calibrating instrument, or intraoperative volume determination. The number is a real per-patient measurement, but of an input dimension, taken under operative conditions, with volume inferred. Exemplars: [14, 105].

Tier V4 — Declared technique. The exposure is a group-level construction intention with no per-patient measurement: cartridge count as a categorical technique, bougie caliber alone, or a stated target volume or length for a study arm. The number describes the protocol, not any patient. Exemplars: [17, 148].

Four rules govern assignment. (1) A study is classified by the variable its outcome analysis actually used, not by the best measurement it performed somewhere in its methods. (2) *Timing*: a measurement counts toward construction size only if made intraoperatively or within 90 days of surgery, before remodeling dominates (Section 6.7); studies whose only size data are later measurements are analyzed separately as current-size studies, not construction studies, whatever their modality. The 90-day boundary is a pragmatic choice — Flanagan’s data show growth already under way between three months and one year [159] — and its sensitivity will be examined. That examination applies to the tier’s own flagship: the one prospective volumetric cohort measured at three months [157], at the very edge of the window, where remodeling is plausibly already under way — a three-month volume cannot simply be equated with operative anatomy, and the timing sensitivity analysis therefore bears on V1 itself, not only on borderline studies. (3) In comparative studies, both arms must be characterized at the same tier; a contrast between a measured arm and a declared arm takes the lower tier. (4) Within-tier quality —

Table 6.2: The modality-quality hierarchy. Fixed before systematic extraction, after the preliminary scan (see text); Chapter 11 stratifies all effect estimates by these tiers as exploratory effect modification.

Tier	Name	What is recorded	Principal error sources
V1	Measured volume	Volumetric image segmentation (mL)	Distension state; segmentation; timing
V2	Imaged geometry	Projected area / diameters (cm ² , cm)	Projection geometry; filling; observer variation
V3	Measured construction	Intraoperative length (cm, mm of stapler)	Dimension-volume inference; operative tissue state
V4	Declared technique	Protocol target (cartridges, Fr, nominal mL)	No per-patient information; construction scatter

distension protocol reporting, observer blinding, interobserver statistics — is handled by the risk-of-bias assessment (Chapter 8), not by tier promotion or demotion.

Three further consequences of the hierarchy are fixed here. First, the dose-response analysis (research question 2) — exploratory throughout, for the ecological reasons stated in the protocol — is only defensible on the volumetric subset: a gradient analysis requires the exposure to be a continuous physical quantity on a common scale, which only tiers V1 and V2 provide, and V2 only after an explicit geometric conversion whose assumptions must be reported. Cartridge counts and declared targets can support ordered contrasts but not a gradient in milliliters. Second, the one-anastomosis gastric bypass (OAGB) subgroup, in which the pouch is deliberately built much longer than in RYGB, arrives almost entirely through length-declared technique descriptions — its measurement culture is V3–V4, with volumetric characterization only beginning to appear [156] — so the OAGB analysis (Chapter 13) inherits a tier profile that must temper its conclusions. Third, the safety analysis is largely immune to this chapter’s critique, and that asymmetry is itself informative: the marginal-ulcer signal rises with stapler length [14] and with larger declared pouches [18] — proxy-measured in both cases — but ulcer formation plausibly depends on the extent of acid-secreting mucosa retained along the staple line, a quantity for which stapled length is a *good* proxy, arguably better than it is for volume. A variable can be a weak proxy for one causal pathway and a strong proxy for another; the ulcer-gradient side of the safety question (Chapter 13) rests on firmer measurement than the benefit side. The advantage does not extend to the other direction of harm: the complications plausibly attached to very small construction — stricture, anastomotic tension, leak, food intolerance — arrive through their own reporting channels and are assessed per outcome, not on a shared axis with the ulcer gradient.

The hierarchy’s use is specified now so that Chapter 11 inherits an explicit expectation rather than a free parameter. Under a real effect measured with classical error, the association should be strongest where measurement error is smallest — attenuation by classical noise works in only one direction — and thus clearest in V1 and V2. The opposite pattern — an association present, if anywhere, in V3 and V4 and absent in V1 and V2, already suggested by the juxtaposition of the null prospective volumetric cohort [157] with the positive proxy-based and declared- technique studies [17, 18] — would *not* mean that noise manufactured the effect, which classical error cannot do; it would mean that something traveling with the weak proxies did: confounding by surgeon, era, and technique, or selective reporting. The two patterns point at different explanations of the same stratification, which is what makes it worth computing. But the stratification is an exploratory effect-modification analysis, not a hypothesis test with a verdict. Measurement tier is entangled with study design, era, and sample size — V1 is small imaging cohorts, V3 contains the 14,168-patient registry, V4 the randomized technique comparisons — and a between-study contrast across so few studies cannot isolate measurement validity from

those companions. Chapter 11 therefore weighs competing explanations; it does not adjudicate. The dissertation’s reading is deferred to the data; the ruler by which the data will be organized is now fixed.

One caveat closes the chapter honestly. This hierarchy orders measurements of *anatomical* size, because anatomical size is what the surgical decision controls and what the literature reports. If the mechanistic account of Chapter 5 is right that the pouch matters — to the extent it matters at all — as a regulator of nutrient arrival speed rather than as a container, then the causally relevant variable may be functional: emptying dynamics, outlet distensibility, the shape of the volume–pressure curve. No tier of this hierarchy measures that variable; the dynamic phase of swallow MRI [19] and impedance planimetry [158] are the first instruments that could. Should the pouch-size effect vanish under good anatomical measurement, one available reading will be that the field measured the wrong quantity for forty years — a possibility Chapter 14 takes up. And the pouch is only half of the measurement problem: the outcome against which it is correlated, weight loss itself, is reported through its own thicket of competing metrics and definitions — at least eight distinct operationalizations of weight regain alone circulate in this literature [160] — and that half is the subject of the next chapter.

Chapter 7

Measuring the outcome

Chapter 6 established that the exposure side of this dissertation’s question — how large the pouch actually is — reaches the literature through four incompatible modalities with no reference standard. This chapter makes the same argument about the outcome side. Weight loss after bariatric surgery (surgery performed to treat obesity) sounds like the easiest quantity in medicine to measure: a scale, a patient, a date. In practice the field has spent four decades disagreeing about what number to write down, when to write it, and against what reference to normalize it. The disagreement is not cosmetic. The choice of metric changes which patients count as successes, which operations look superior, and — directly relevant here — whether an anatomical variable like pouch size appears to predict outcome at all. A systematic review that pools across studies inherits every one of these choices, usually silently.

The chapter proceeds in five steps. First, the metric family itself: percent excess weight loss (%EWL), percent total weight loss (%TWL), and change in body-mass index, treated as what they are — normalizations of the same raw signal against different references, with derivable and consequential statistical properties. Second, the timepoint problem: the weight trajectory after bypass is nonmonotonic, so an outcome anchored to its minimum (the *nadir*) measures something different from an outcome anchored to a fixed postoperative date. Third, weight regain, an endpoint with no agreed definition and a prevalence that swings by tens of percentage points depending on which definition is applied to the same cohort. Fourth, the field’s attempt to standardize — the 2015 reporting standard of the American Society for Metabolic and Bariatric Surgery (ASMBS) — and the measured extent to which the literature ignores it. Fifth, follow-up attrition: how many patients vanish from bariatric cohorts, the evidence that they do not vanish at random, and the survivorship logic by which non-random attrition specifically distorts correlations between surgical anatomy and late weight outcomes. The chapter closes with the numbered harmonization rules that the meta-analysis of Part II applies, each justified by the material preceding it.

7.1 The metric family

7.1.1 One signal, many normalizations

Every weight-loss metric in this literature is a transformation of the same three measurements: preoperative weight W_0 , follow-up weight W , and height h . Body-mass index (BMI) is weight scaled by height squared, W/h^2 , in kg/m^2 ; it is the field’s standard size coordinate, and a BMI of 25 is the conventional upper bound of the “normal” range. From these, the common metrics are:

- **Absolute loss:** $W_0 - W$ in kilograms, or equivalently ΔBMI in kg/m^2 units.

- **Percent total weight loss:** $\%TWL = 100 \cdot (W_0 - W)/W_0$ — loss normalized to the patient’s own starting weight.
- **Percent excess weight loss:** $\%EWL = 100 \cdot (W_0 - W)/(W_0 - W_{\text{ideal}})$ — loss normalized to the “excess” above an ideal weight, historically read from the actuarial height–weight tables of the Metropolitan Life Insurance Company [163, 164]. The reference is not even unique: with competing ideal-weight conventions and preoperative-weight definitions in circulation, the same weight change can shift the reported $\%EWL$ by up to 17 percentage points [164].
- **Percent excess BMI loss:** $\%EBMIL = 100 \cdot (\text{BMI}_0 - \text{BMI})/(\text{BMI}_0 - 25)$ — the same construction with the reference fixed at BMI 25, which removes the table lookup but keeps the structure.

Van de Laar made the underlying unity explicit by writing the whole family as one parameterized measure, $100 \cdot (\text{BMI}_0 - \text{BMI})/(\text{BMI}_0 - a)$, where the reference point a ranges from 0 (giving $\%TWL$) through the ideal-weight convention (giving $\%EWL$) to 25 (giving $\%EBMIL$) [165]. Seen this way, the field’s metric debate is a dispute about a single scalar: where to put the zero of the scale. That is a familiar problem in measurement generally — the choice of reference does not change the signal, but it changes every ratio computed against it, and a badly chosen reference injects a systematic, covariate-dependent gain into the readout.

7.1.2 The $\%EWL$ problem: a reference correlated with the population

The specific defect of $\%EWL$ is that its denominator, $W_0 - W_{\text{ideal}}$, grows with baseline size while its numerator grows more slowly, so the metric builds a negative dependence on preoperative BMI into the definition itself. Two patients who each shed the same fraction of their body weight receive different scores: the lighter patient scores higher, purely arithmetically.

This is not a hypothetical concern; it has been quantified repeatedly in large cohorts. In 2,420 Roux-en-Y gastric bypass (RYGB) patients at the Cleveland Clinic, preoperative BMI correlated strongly and negatively with 1-year $\%EWL$ ($r = -0.52$, $p < 0.01$) and strongly and positively with BMI units lost ($r = 0.56$), while $\%TWL$ was essentially orthogonal to baseline BMI ($r = 0.04$, $p = 0.3$) [166]. An independent cohort of 846 RYGB patients at Massachusetts General Hospital reproduced the same structure almost coefficient for coefficient: preoperative BMI versus ΔBMI $r = 0.56$ at one year, versus percent excess body weight loss $r = -0.52$, versus percent weight loss $r = 0.04$ ($p = 0.33$) [167]. And in 7,212 women from the Bariatric Outcomes Longitudinal Database (BOLD), splitting the cohort into lighter and heavier halves produced nadir $\%EBMIL$ of 95.3% versus 77.2% and $\%EWL$ of 81.7% versus 70.4% — an apparent 18-percentage-point superiority of surgery in lighter patients — while $\%TWL$ for the same halves was 37.7% versus 40.2%, and reference points near $a = 10$ eliminated the baseline dependence entirely ($p = 0.504$) [165]. The operation removes a nearly constant fraction of body mass across the baseline range; $\%EWL$ manufactures a gradient where the physiology shows none.

For this dissertation the implication is sharper than a general preference for tidy metrics. Any exposure that correlates with baseline BMI — and constructed pouch size plausibly does, because surgeons facing larger stomachs and thicker tissue may staple differently — inherits a spurious association with $\%EWL$ through the metric alone. In statistical terms, $\%EWL$ is a ratio estimator whose denominator is a function of a potential confounder; regressing it on any baseline-correlated exposure produces bias of predictable sign without any physiological effect. A pouch-size literature scored in $\%EWL$ is therefore contaminated at the level of its measuring instrument, in a direction that depends on how pouch size covaries with patient size in each cohort. The per-patient normalization also inflates dispersion: in the Cleveland cohort the 1-year outcome was 30.0 ± 8.5 $\%TWL$ against 68.5 ± 22.9 $\%EWL$ [166] — a coefficient of variation

of 28% versus 33% for the same underlying weights, the extra variance being baseline-BMI heterogeneity re-expressed through the denominator.

7.1.3 %TWL’s rise, and what the older metrics still contaminate

On this evidence the methodological literature has converged: %TWL is the least confounded relative metric and the appropriate default for comparing patients, arms, and studies [166, 167, 165]. The convergence is recent, however, and the stock of evidence this dissertation must synthesize was mostly published before it. The field’s largest pouch-size study — the Scandinavian registry analysis of 14,168 patients — reports percent excess BMI loss [14]; the Dutch extended-pouch randomized trial reports percent total body weight loss [13]; the Chinese small-versus-large-pouch cohort in low-BMI diabetic patients reports BMI change [18]; observational studies report %EWL alone [115], %EWL and %TWL together [147, 148], or bespoke success categories [5]. Table 7.1 assembles the outcome conventions of the landmark studies this dissertation re-analyzes in Part II; no two agree on both metric and timepoint. This is the concrete, local form of the general problem: the only systematic review of pouch size to date found that the “quality and heterogeneity of data precluded a meta-analysis” [15], and outcome-metric heterogeneity is one of the two axes of that failure (the other, exposure measurement, was Chapter 6’s subject).

Table 7.1: Outcome conventions in landmark studies of RYGB pouch size. Metrics and timepoints as reported in each study’s primary weight-loss analysis; follow-up is the completeness at the latest reported weight-loss timepoint, where stated. EP = extended pouch; S = standard pouch.

Study	Year	Design	N	Weight metric(s)	Timepoint(s)
Kirkpatrick [5]	1982	prospective	—	“success” category	—
Heneghan [115]	2012	endoscopy cohort	380	%EWL; success = %EWL > 50 or BMI $\leq$ 30	mean 5.9 $\pm$ 4.0 y
Edholm [14]	2016	registry	14,168	%EBMIL	6 wk; 1 y
Reiber [147]	2018	case-matched	132	%EWL and %TWL	3–24 mo
Boerboom [13]	2020	RCT	132	%TBWL; regain (kg)	1–3 y
Gao [18]	2022	retrosp. cohort	95	Δ BMI	1, 3, 6 y
Haghighat [148]	2022	retrosp. cohort	100	kg; Δ BMI; %EWL; %TWL	6 mo
Harker [16]	2026	RCT follow-up	62	%TWL; recurrent weight gain	5–9 y

7.1.4 Dichotomous “success” and its ceiling

A further layer of the older literature reports outcomes as a binary: success, conventionally at least 50% EWL or a final BMI below 35. The convention circulates under Reinhold’s name: his 1982 critical appraisal of long-term weight loss after gastric bypass [168] is the paper the eponym points to, and the modern literature operationalizes “Reinhold’s criteria” as failure at %EWL of 50 or below and/or a BMI of 35 or more [169]. Whatever its exact original wording, the criterion no longer discriminates: in BOLD gastric-bypass data with two or more years of follow-up, more than 95% of patients surpassed 50% EWL, and the 50th percentile crossed both the 50%-EWL and BMI-35 thresholds 135 days after surgery [170]. A threshold that nearly everyone clears is a saturated detector — it compresses the entire upper range of the outcome distribution into one bin and discards precisely the variance a pouch-size comparison needs. Studies that report only success proportions against such thresholds (for example, defining success as %EWL > 50 or BMI $\leq$ 30 in a cohort assessed at highly variable follow-up times [115]) contribute direction-of-effect information at best, and the meta-analysis treats them accordingly (Rule 8 below).

FIGURE PLACEHOLDER: Metric-comparison plot (data-rendered per spec:011, not TikZ). Panel A: the same postoperative weight trajectory (fixed 30 %TWL at nadir) re-expressed as %TWL, %EWL (ideal weight at BMI 25), %EBMIL, and Δ BMI as a function of baseline BMI from 35 to 60 kg/m², computed directly from the metric definitions — showing %EWL/%EBMIL falling and Δ BMI rising with baseline BMI while %TWL stays flat. Panel B: published cohort correlations of each metric with preoperative BMI ($r = -0.52, +0.56, +0.04$) from Corcelles 2016 (PMID 26803753) and Hatoum 2013 (PMID 23670991). Source data: metric formulas plus the two cited papers’ reported coefficients; rendered from the pipeline’s claims table, AI-generation disclosed in caption per spec:011.

Figure 7.1: The same weight loss under four metrics: baseline-BMI dependence is a property of the metric, not the operation.

7.2 Timepoints: nadir versus fixed

The weight trajectory after RYGB is not monotonic. Weight falls steeply for roughly a year, reaches a minimum — the nadir — and then drifts upward for as long as anyone has measured. In the Longitudinal Assessment of Bariatric Surgery (LABS) cohort, the median maximum weight loss was 37.4% of presurgery weight, reached a median of 2.0 years after surgery (interquartile range 1.0–3.2 years), after which regain proceeded fastest in the first post-nadir year and continued throughout follow-up [171]. BOLD gastric-bypass patients reached nadir BMI at a mean of 603 days [170]. Two consequences follow for outcome definition.

First, *when* an outcome is read off the trajectory changes what is being measured. A 1-year outcome samples the trajectory near its extremum and mostly reflects the operation’s initial potency; a 5-year outcome is the sum of that initial response and several years of regain, processes with plausibly different physiology and different sensitivity to anatomy. A pouch-size effect on early emptying dynamics (Chapter 5) could in principle appear at one horizon and not the other — indeed the Dutch extended-pouch trial found no weight-loss difference for two years and a difference emerging in the third year, driven by differential regain [13]. Comparing a 1-year study against a 6-year study is therefore not a like-for-like comparison even in a common metric, which is why the protocol fixes separate outcome windows (1 year; 3 years; 5 years and beyond) rather than pooling across them — and never uses “longest available follow-up” as a window, since that samples different studies at different phases of the same nonmonotonic curve.

The corpus contains the cleanest possible demonstration that the horizon changes the answer. In the Scandinavian registry analysis — the largest pouch-size study ever conducted — the length of stapler used to construct the pouch predicted percent excess BMI loss at six weeks but not at one year [14]. Had that study reported only an early timepoint, it would stand in the literature as 14,168-patient evidence *for* the pouch-size doctrine; reported at one year, it is the field’s strongest evidence against. Nothing about the patients, the surgery, or the metric differs between those two readings — only the date on which the trajectory was sampled. Any synthesis that mixes follow-up horizons across studies is implicitly averaging over this reversal, and any mechanistic interpretation (Chapter 5) must explain the horizon-dependence itself: an anatomical variable whose predictive power decays within months looks like a transient mechanical constraint, not a durable determinant of the neuroendocrine equilibrium.

Second, the nadir itself is a statistically treacherous anchor. It is an extreme-value statistic of a noisy, sparsely sampled curve: its expected value depends on how often the patient was weighed (denser sampling finds deeper minima), and in retrospective designs it is often reconstructed from the patient’s memory of their lowest weight. The scoping review that cataloged regain definitions identifies “patient recall to estimate nadir weight” alongside loss to follow-up and variable follow-up times as core reasons regain reporting is inconsistent [160].

Nadir-anchored quantities are also, by construction, conditioned on the trajectory’s own optimum — they measure the best state the patient ever reached, not the state the operation durably holds. Cohorts reporting weight loss “at nadir” (e.g., 81.2% excess body weight loss at nadir in the Massachusetts General cohort [167]) therefore report systematically rosier numbers than cohorts reporting fixed-timepoint outcomes, and the two must not be averaged together as if they estimated the same quantity. The harmonization rules below (Rule 5) make this exclusion explicit.

7.3 Weight regain: a definition for every study

Beyond the metric and the timepoint lies a third degree of freedom: studies increasingly report *weight regain* as an endpoint, and the field has never agreed on what the term means. The scoping review of El Ansari and Elhag collects the definitions in use: regain of more than 25% EWL from nadir; regain of $\geq 10\%$ or $> 15\%$ of nadir weight; ≥ 10 kg above nadir; loss of $\geq 10\%$, $\geq 20\%$, or $\geq 25\%$ of maximum weight lost; $\geq 10\%$ of presurgery weight; any regain after diabetes remission; any regain at all; ≥ 5 BMI units above nadir; or a return to BMI over 35 after successful loss [160]. That is ten structurally distinct definitions — differing in unit (kilograms, BMI points, percent), in reference (nadir, maximum loss, presurgery weight), and in form (continuous versus threshold) — before any study-specific variation in the threshold values themselves.

The definitional spread is not an aesthetic problem; it moves the measured prevalence by tens of percentage points within a single cohort. In the LABS gastric-bypass cohort, five years after nadir, 43.6% of patients had regained ≥ 5 BMI points, 50.2% had regained $\geq 15\%$ of nadir weight, and 67.3% had regained $\geq 20\%$ of maximum weight lost — the same 1,406 patients, the same scales, a 24-percentage-point swing from the definition alone [171]. The same study compared the candidate measures head-to-head against clinical endpoints (progression of diabetes, hyperlipidemia, hypertension, quality of life) and found that regain expressed as percent of maximum weight lost had the strongest association and best model fit for nearly all outcomes — a rare piece of empirical guidance in an otherwise conventional choice [171].

The pouch-size literature sits squarely inside this problem. The most influential recent trial in the field concluded that an extended pouch “reduces weight regain” on the basis of a 3-kg regain in the standard arm during the third year [13]; the same surgical group’s 5–9-year follow-up of an extended-pouch randomized trial reported “recurrent weight gain” in 70% versus 47% of patients ($p = 0.07$) under a threshold definition [16]. Whether those findings would survive translation into a different regain definition is exactly the kind of question a reader cannot answer from the published numbers. The meta-analysis therefore does not pool “regain” as an outcome at all: regain-defined results enter the synthesis only where they can be re-expressed as weight at a fixed timepoint, and otherwise contribute to the structured narrative synthesis (Rule 7).

7.4 The ASMBS 2015 standard and its uptake

The field has known all of this for a decade, and in 2015 its main professional society acted: the ASMBS published standardized outcome reporting recommendations for metabolic and bariatric surgery [172], intended — in the words of the study that later audited them — “to achieve consistency in the literature” [173]. The standard’s central weight-metric recommendation — %TWL as the reporting measure — is attested independently of that audit: a 2024 secondary analysis of two randomized trials, re-expressing their outcomes under the standard, names %TWL as the weight-loss measure the ASMBS recommends, and counts only four studies published between 2015 and 2020 that used the complete reporting standard [174].

The audit is sobering. Screening nine PubMed-indexed bariatric journals for weight-loss articles in the first halves of 2015 and 2021, Shahwan and colleagues found overall compliance

with the ASMBS standard rose from 5% to 20% [173]. The composition of reporting shifted in the recommended direction — %TWL reporting rose from 17% to 61% of articles — but %EWL and %EBMIL persisted essentially undiminished (76% to 69%, not significant), the number of *distinct* weight-loss outcome measures in use actually increased from 45 to 54, and 30% of 2021 articles failed to comply even with their own journal’s stated reporting guidance [173]. The audit also flagged a subtler denominator problem: the fraction of studies specifying that “initial weight” meant the immediate preoperative weight *fell* from 15% to 6%, which matters because programs mandating preoperative weight loss can inflate surgical results by anchoring to an earlier referral weight [173].

Two conclusions follow for Part II. First, six years into the standard’s life, four out of five weight-loss papers still did not conform to it; a synthesis cannot simply restrict itself to standard-compliant studies without discarding most of the evidence, including nearly all of it published before 2015. Harmonization — converting what was reported into a common metric — is forced, not optional. Second, the baseline-weight convention is itself a harmonization variable: where a study’s W_0 is a referral weight rather than an operative-day weight, its relative metrics are shifted, and the extraction schema records which convention each study used.

7.5 Attrition and the survivorship problem

7.5.1 Magnitude

Bariatric follow-up decays fast, and the decay is the rule rather than the exception. In one of the earliest laparoscopic RYGB series with decade-long observation, office follow-up was 33% at two years and 7% at ten years, with telephone contact recovering a further 19% at ten years [175]. In BOLD — a national database built expressly for outcome tracking — 49,098 gastric-bypass patients were eligible for two-year follow-up and 8,945 (18%) had a reported weight at two years or beyond [170]. At the level of the literature rather than the cohort, a JAMA systematic review of long-term bariatric outcomes found that of 7,371 clinical studies screened, 29 (0.4%) reported outcomes beyond two years while retaining at least 80% of the initial cohort [176]. The pouch-size literature specifically shows the same spread: the Chinese small-versus-large cohort retained 89.4% at six years [18] — an outlier in the good direction — while the extended-pouch trial’s 5–9-year follow-up reported itself underpowered for its primary comparison because of loss to follow-up [16].

7.5.2 Non-randomness

If patients dropped out at random, attrition would cost precision and nothing else: the retained sample would remain an unbiased, noisier estimate of the cohort. The evidence is that they do not drop out at random. In a series where all patients had reached 3–4 years since surgery, those who had attended every scheduled visit averaged 74% EWL, those lost to follow-up after the first year 61%, and those lost before one year 56% ($p < .05$) — with no difference at the 1-year mark before the groups diverged [177]. In statistical language, bariatric follow-up is informatively censored: the probability of contributing a late data point is correlated with the value that data point would have taken, and the data are missing not at random. Every late-timepoint mean in this literature is therefore an estimate conditioned on continued attendance. What the conditioning selects *for* is less certain than a single cohort can establish. In that series the non-attenders were doing worse, which would make retained means optimistic; but selection can also run the other way — patients with complications, weight regain, or revision in view re-enter follow-up precisely because things went badly — and clinic attendance is not the same thing as successfully ascertained research follow-up, which some studies pursue by telephone or registry linkage. The direction of the bias is therefore cohort-specific and, in general, unknown. What is known is that a reported “57% EWL at ten years” computed on the 7% of a cohort still

attending clinic [175] is not an estimate of the tenth-year outcome of the operated population; it is an estimate of the tenth-year outcome of the operated-and-still-attending population — a different quantity, and on the one direct comparison available, plausibly but not provably a sunnier one.

7.5.3 Why attrition biases anatomical correlations specifically

For this dissertation the damage is more specific than optimistic means, and it is worth stating the survivorship logic exactly, because it applies to every late-timepoint pouch-size correlation in the corpus. A late cross-sectional analysis correlating constructed pouch size with weight loss is computed on survivors of the follow-up process. Attendance is influenced by weight outcome (patients doing poorly disengage, or re-engage selectively when seeking revision) and plausibly by anatomy-linked symptoms: marginal ulcer (ulceration at the surgical junction between pouch and bowel, whose risk rises with pouch stapling length [14]), dumping, reflux, and revisional workup all modulate who is in front of a clinician — and therefore in the dataset — years later. When selection into the analyzed sample depends on both the exposure pathway and the outcome, conditioning on that selection can induce or suppress an exposure–outcome association that does not exist, or does not have that magnitude, in the full cohort. The endoscopy-based pouch-size studies are the extreme case: a cohort assembled from patients referred for endoscopy because of symptoms or weight regain [115] is selected directly on the outcome, and any pouch-size–weight correlation measured within it estimates an association *within the referred stratum*, not in operated patients at large. (Those studies also measure a late, dilated pouch rather than the constructed one — the exposure-side disqualification of Chapter 6 — so they fail this dissertation’s inclusion logic twice over.)

Three practical consequences carry into Part II. First, follow-up completeness is extracted per arm and per analyzed timepoint for every included study and feeds the risk-of-bias assessment (Section 8.8); a between-arm attrition asymmetry is treated as a bias signal, not noise. Second, 1-year outcomes — where completeness is typically still high — are structurally more trustworthy than late outcomes, which is an independent reason (beyond regain physiology) that the primary endpoint sits at one year. Third, no missing-at-random imputation is attempted — under informative censoring it would replace an honest gap with a confident wrong number — but the refusal is not treated as a cure: complete-case estimates remain biased in an unknown direction, so the synthesis reports what the retained samples show, brackets late-timepoint estimates with pre-specified best-case/worst-case bounds (tipping-point-style sensitivity analyses where study-level data permit), and grades the residual uncertainty accordingly (Rule 10).

7.6 Harmonization rules for the meta-analysis

Everything above now compresses into the operating rules the meta-analysis applies. Section 8.7 summarizes them; this section states and justifies them. One preliminary observation bounds what harmonization must accomplish. The primary pooled quantity is a *mean difference in %TWL* between small-pouch and large-pouch arms — an absolute effect in percentage points of body weight, the scale on which clinical negligibility or importance can actually be stated. A standardized mean difference is computed only as a within-metric sensitivity analysis: standardization does not harmonize anything — it rescales each study by its own within-arm dispersion, importing attrition and population heterogeneity into the effect size — and once outcomes genuinely share a scale, the difference of means is the natural measure. Choosing the mean difference raises, rather than lowers, the demands on harmonization, and three limits on what the conversions below can deliver are stated up front. First, they operate on arm-level summary statistics, so they approximate the mean of a per-patient ratio by a ratio of arm means; the two differ in general, and the approximation is a first-order (delta-method) one whose error grows with

the within-arm dispersion of baseline BMI. Second, converting a standard deviation is harder than converting a mean: the SD of %TWL depends on the covariance between baseline BMI and weight lost, which studies do not report, so any converted SD rests on a stated covariance assumption bounded by sensitivity analysis, not on arithmetic alone. Third, randomization balances baseline-BMI distributions in expectation only — in small trials the realized arms can differ, and matched designs only approximate balance — so no within-study conversion is treated as error-free. Converted and native outcomes are always distinguished, and where a defensible conversion is unavailable the metrics are analyzed separately rather than forced onto one axis. The rules:

1. **Target metric.** The common outcome scale is %TWL, because it is the relative metric empirically least confounded by baseline BMI [166, 167, 165]. Outcomes reported natively in %TWL (or %TBWL, an identical quantity under a different name) enter unchanged.
2. **%EWL conversion.** %EWL is converted arm-wise to %TWL by $\%TWL = \%EWL \times (BMI_0 - BMI_{ref})/BMI_0$, using the study’s reported arm-level baseline BMI and its declared ideal-weight convention for BMI_{ref} ; where the convention is undeclared, $BMI_{ref} = 25$ is assumed and the entry flagged. The identity is exact for an individual (Section 7.1) but only approximate at arm level — a first-order delta-method approximation, since the mean of a ratio is not the ratio of means — with error growing in the within-arm spread of baseline BMI. Standard deviations are converted by the same factor, which would be exact only if the factor were common to every patient; because it varies with baseline BMI and covaries with weight lost, the converted SD carries additional error under a stated covariance assumption. Every converted value is flagged in the evidence tables; the pooled analysis is repeated excluding converted outcomes and under sensitivity bounds spanning plausible covariance assumptions; and where an arm’s baseline-BMI dispersion is large or unreported, the study is not converted at all but analyzed in its native metric alongside, not inside, the %TWL pool.
3. **%EBMIL conversion.** Identical to Rule 2 with $BMI_{ref} = 25$ by definition, and subject to the same delta-method caveats; this rule covers, among others, the largest study in the corpus [14] — whose continuous stapler-length exposure is in any case analyzed under the protocol’s separate slope estimand, not forced into the two-arm pool (Chapter 8).
4. **Absolute metrics.** ΔBMI or kilograms lost convert to %TWL only where the arm’s baseline mean BMI or weight is reported ($\%TWL = 100 \cdot \Delta BMI / BMI_0$); the conversion is a group-level (ecological) approximation, its SD conversion inherits the covariance problem of Rule 2, and it is flagged as such. Studies reporting only absolute loss without baseline moments contribute to direction counts, not to pooling.
5. **Fixed timepoints only.** Outcomes are read at fixed windows: the 1-year primary window is 10–14 months from surgery, taking the reported timepoint nearest 12 months; the secondary windows are 3 years (30–42 months, nearest 36) and 5 years or beyond (the reported timepoint at ≥ 54 months nearest 60). “Longest available follow-up” is never used as a window, because it samples different studies at different phases of a nonmonotonic trajectory (Section 7.2). Nadir-anchored outcomes are never pooled with fixed-timepoint outcomes, because the nadir is an extreme-value statistic with sampling-dependent bias; studies reporting only nadir-anchored weight loss enter the narrative synthesis only.
6. **Moment recovery.** Where a study reports medians with interquartile ranges or ranges, means and standard deviations are estimated by the Wan and Luo methods [178, 179], with the Hozo method [180] as fallback where only the range is available; every recovered moment is flagged in the evidence tables.

7. **Regain endpoints.** “Weight regain” and “recurrent weight gain” outcomes are not pooled under any of their competing definitions (Section 7.3). Where the underlying fixed-timepoint weights are recoverable, those are extracted instead; otherwise the regain finding is carried in the structured synthesis with its definition stated verbatim.
8. **Dichotomous outcomes.** Success proportions against thresholds ($\geq 50\%$ EWL, BMI cutoffs) are not converted into continuous effect sizes: the thresholds are near ceiling in bypass populations [170] and the implied distributional assumptions are unverifiable. Such studies contribute direction-of-effect information to the SWiM synthesis.
9. **Baseline-weight convention.** The extraction schema records whether W_0 is the immediate preoperative weight, a referral weight, or undeclared [173]; studies using referral weights are flagged, since mandated preoperative weight loss inflates their relative metrics.
10. **Attrition.** Follow-up completeness is extracted per arm at each analyzed timepoint; it is carried into risk-of-bias assessment (missing-outcome-data domains of RoB 2 and ROBINS-I, Section 8.8) and into GRADE certainty. No missing-at-random imputation is applied — bariatric attrition is demonstrably informative where it has been examined [177] — but refusing imputation does not unbiased complete-case estimates, so late-timepoint pooled estimates are accompanied by pre-specified best-case/worst-case bounds (and tipping-point-style sensitivity analyses where study-level data permit), with the direction of selection treated as unknown rather than assumed optimistic (Section 7.5).

These ten rules are fixed here, in Part I, before the systematic screening and extraction of Part II begin — though, like the modality hierarchy of Chapter 6, they were drafted after, and informed by, the preliminary scan whose studies this chapter quotes; the repository’s version history documents that sequence, and the analyses built on scan-informed rules carry a hypothesis-generating label rather than a confirmatory one. The measurement chapters have now argued that the field’s forty-year disagreement about pouch size rests on an exposure measured four incompatible ways and an outcome reported in dozens of incompatible forms, at inconsistent times, on informatively-thinned samples. Whether any signal survives once both sides of the measurement are disciplined is the empirical question, and Part II exists to answer it.

Part II

The Study

Chapter 8

Methods

Part I closed having fixed the study’s two instruments: the modality-quality hierarchy for the exposure (Section 6.9) and the harmonization rules for the outcome (Section 7.6). This chapter reports the conduct of the systematic review and meta-analysis that consumes them. Conduct follows the methodological guidance of the Cochrane Handbook [181]; reporting follows the PRISMA 2020 statement (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [182], which is a reporting guideline, not a prescription for how a review is conducted — the distinction matters and is kept throughout. The chapter proceeds through protocol, eligibility criteria, information sources and search, study selection, data collection, synthesis methods, and the assessment of bias and certainty. Two features of the conduct are unusual for a systematic review and are therefore described at full depth rather than in passing. First, the search and the evidence corpus are pipeline-driven: every search execution, every retrieved record, and every stored full text lives in a version-controlled catalog, so that the PRISMA flow numbers reported in Chapter 9 are generated by query from that catalog rather than tallied by hand. Second, data extraction is assisted by large language models under a span-verification constraint — no extracted value enters the analysis unless it string-matches a verbatim quote in the stored source text — with human review of every included study and a blinded audit of a pre-specified extraction sample. Section 8.5 discloses this design, its rationale, and its limits precisely, because an extraction method this new must be auditable rather than merely asserted.

8.1 Study design and protocol

The study is a systematic review of the relationship between constructed gastric pouch size and weight loss after Roux-en-Y gastric bypass (RYGB), with a pre-specified meta-analysis attempt. “Attempt” is a deliberate term. The single prior systematic review in this literature found its fourteen included studies too heterogeneous to pool [15], and no later pooled estimate was identified by the searches conducted here — “none identified,” not “none exists,” since the search has stated limits (Section 8.3). The present protocol therefore pre-commits to both branches: a random-effects meta-analysis if pooling meets the defensibility criteria stated in Section 8.9, and a formal Synthesis Without Meta-analysis (SWiM) [183] if it does not — in which case the study-by-study demonstration of *why* pooling fails is itself reported as a primary result, not as an apology.

The protocol of record is a versioned internal protocol maintained under version control in the research repository alongside the corpus and the analysis code. Every deviation arising during screening, extraction, or analysis is recorded in the protocol file as a dated amendment; the protocol history is therefore a complete, timestamped audit trail rather than a snapshot. What that audit trail establishes is stated precisely: it is evidence of *sequence* — which decision was fixed before which analysis — not evidence that any decision was made independently of the preliminary literature scan. The full protocol text and its amendment log are reproduced in

the appendices.

8.1.1 Pre-specification and the preliminary scan

This protocol was written after a preliminary scan of the literature that included reading the results of the field’s landmark studies — the scan that motivated the dissertation and supplied the recall set for search validation. Decisions made after inspecting results cannot be called prospective, whatever their timestamps say, so the protocol distinguishes its decisions honestly rather than claiming a uniform pre-specification it does not have. Fixed on grounds independent of any outcome data, before screening: the PICO criteria, the primary estimand and its fixed timepoint windows, the harmonization and conversion rules, the statistical model, and the pooling-defensibility gate. Informed by the preliminary scan, and therefore labeled hypothesis-generating rather than confirmatory wherever they surface in Part II: the measurement-modality hierarchy and the effect-modification analysis built on it (Section 8.6), and the specific hypotheses of Part I, which were shaped by knowing which landmark studies found what. The dated protocol history documents which decision belongs to which class; no analysis shaped after seeing study results is presented as a confirmatory test.

Registration in PROSPERO, the international prospective register of systematic reviews, was undecided at the time the protocol was finalized. PROSPERO does not accept retrospective registration, so the decision gates the start of formal screening: registration, if elected, precedes any screening beyond search piloting, and the final registration status — registered with an identifier, or not registered with the reason stated — is recorded here and in the limitations of Chapter 14.

8.2 Eligibility criteria

Eligibility was defined in PICO form (population, intervention/exposure, comparator, outcome) before screening began.

8.2.1 Population

Adults (18 years or older) undergoing *primary* bariatric bypass surgery — that is, a first weight-loss operation, not a revision of an earlier one. The core population is RYGB in any technical era: open, laparoscopic, or robotic access, with era retained as a pre-specified subgroup rather than an exclusion. One-anastomosis gastric bypass (OAGB, also called mini gastric bypass) — a simplified variant with a single bowel connection and a deliberately long, narrow pouch — forms a pre-specified subgroup analyzed separately from RYGB throughout, because its literature varies pouch length far more widely than RYGB practice does. The OAGB stratum is a descriptive parallel analysis: it is never pooled with RYGB, and it is not treated as external validation of the RYGB findings, because the operation differs in more than pouch geometry. Revisional operations and adolescent cohorts are excluded: revision confounds the exposure (the pouch being revised is not the pouch that was constructed), and adolescent physiology and outcome trajectories differ enough to break comparability.

8.2.2 Exposure and comparator

The exposure is a smaller *constructed* pouch, as defined by each study: an estimated volume in milliliters, a pouch length in centimeters, a stapler count or cumulative stapler length, or a bougie calibration size (the bougie is the calibration tube passed into the stomach during construction; its diameter constrains, but does not determine, the pouch that is stapled around it). The comparator is a larger constructed pouch under the same study’s definition. The primary

contrast is therefore *study-defined small-versus-large*: each study contributes the contrast it actually made. That honesty has a cost the synthesis must carry rather than hide: contrasts made at different places on different physical scales (20 versus 50 ml is not 5 versus 10 cm) are not automatically estimates of one effect, and no outcome standardization or random-effects machinery makes them so. Pooling is therefore restricted to contrasts that satisfy the compatibility rules of Section 8.9; incompatible contrasts are synthesized without pooling. Where a study reports a continuous pouch-size–outcome association rather than a group contrast, it qualifies if an effect estimate with variance is extractable, and it is analyzed under a separate pre-specified slope estimand (Section 8.9) — never force-dichotomized into the arm-contrast pool.

The restriction to *construction* size is a scope decision carried from the measurement analysis of Chapter 6. Pouches dilate substantially over years of follow-up, so any late-measured size is a mixture of the surgeon’s choice and the tissue’s subsequent behavior; postoperative dilation is excluded as an exposure here and treated only as a discussion topic. Studies whose exposure is measured pouch size at a late timepoint, without anchoring to construction, are excluded from the primary analysis.

8.2.3 Outcomes

The primary outcome is percent total weight loss (%TWL) at one year — weight lost as a percentage of preoperative body weight, the metric whose case is argued in Section 7.1. Secondary outcomes are percent excess weight loss (%EWL — weight lost as a percentage of the amount by which preoperative weight exceeds an ideal-weight reference), retained for comparability with the older literature and converted to %TWL where baseline data allow, though converted values feed only a labeled sensitivity analysis and never the primary pool (Section 8.7); weight loss in fixed windows at three years and at five years — never “longest available follow-up,” which would pool different phases of the weight trajectory into one pseudo-timepoint, and never an unbounded “five years or beyond,” which would do the same at the far end of the trajectory; and the safety outcomes of research question 5, which runs in both directions along the exposure. On one side, marginal ulcer — ulceration at or near the gastrojejunal anastomosis, the surgical junction between pouch and small bowel: the largest registry analysis in this literature reported the risk rising by 14% (95% confidence interval 9–20%) per additional centimeter of stapler used in constructing the pouch [14], a gradient that penalizes *larger* construction and therefore points the same way as any weight-loss benefit of smaller construction — it is not a counterweight to that benefit. On the other side, the harms plausibly attached to very small construction: anastomotic stricture, tension or leak at the gastrojejunostomy, and food intolerance. Each safety outcome is extracted and analyzed separately, on an absolute-risk scale, and the two directions are never combined into a single benefit–harm axis. Baseline body-mass index (BMI) is always extracted, whatever the reported outcome metric, because it is the conversion factor between %EWL and %TWL.

8.2.4 Study designs

Randomized controlled trials and comparative observational studies are eligible. A study qualifies if it compares at least two pouch-size groups, or reports a pouch-size–outcome association with extractable effect data. Case series without any internal comparison are excluded: a single-arm series can describe a pouch size but cannot estimate the effect of choosing it. No restriction is placed on publication date or language; non-English records are flagged for translation at screening rather than dropped by filter.

8.3 Information sources and search strategy

8.3.1 Electronic searches

The primary source is PubMed/MEDLINE, searched programmatically through the National Library of Medicine’s E-utilities interface by the repository’s corpus pipeline. Executing searches through the pipeline rather than the web interface has a methodological consequence worth stating: every search execution is stored — query string, execution timestamp, and the full PMID hit list — in a relational catalog, so identification counts are reproducible by query and the provenance of every candidate record is explicit. Records are deduplicated by identifier (PMID, with DOI as secondary key), not by title similarity: the same paper reached by two searches is one catalog row with two search attributions.

Four search strings were executed on 31 August 2026 (Table 8.1; full strings in Appendix A). S1 is the core query, intersecting four concept blocks: the operation (gastric bypass and its synonyms), the structure (pouch, reservoir), a size vocabulary (size, volume, length, dimension, small, large, extended, capacity), and a weight-loss outcome vocabulary. S2 targets the OAGB literature, whose terminology (one-anastomosis, mini gastric bypass) the core string does not reach. S3 targets construction-technique reports that never use a size word — studies describing stapler sequence, staple-line length, or bougie calibration — because construction proxies are exactly the exposure the protocol admits. S4 is an early-literature supplement added after piloting: records from the founding era often lack indexed abstracts, so title-and-abstract field tags see titles only, and the 1982 reservoir-versus-stoma literature [5] was reachable only through a stoma-vocabulary string. No date filter was applied (the 1982 literature must be reachable); no language filter was applied; PubMed’s humans filter was deferred to screening because it silently loses recent records not yet indexed.

Table 8.1: Electronic search executions, 31 August 2026 (PubMed via E-utilities). Full query strings in Appendix A. “New” counts records not returned by any earlier search; deduplication is by PMID.

ID	Concept target	Matches	New	Cumulative unique
S1	Core RYGB pouch–size–outcome	421	421	421
S2	OAGB / mini gastric bypass	59	14	435
S3	Construction technique (stapler, bougie)	131	64	499
S4	Early literature (reservoir/stoma vocabulary)	85	27	526
Total				526

The four searches identified 526 unique records, which constitute the identification stage of the PRISMA flow (Figure 8.1).

8.3.2 Search validation

The strings were piloted against a recall criterion before being frozen: every in-scope landmark study known from the preliminary literature scan had to appear in the union of the hit lists. The recall set spans the eras and designs the review must reach — the 1982 reservoir-versus-stoma study [5] (captured by S4), the 14,168-patient Scandinavian registry analysis [14] (S1 and S3), the prior systematic review [15] (S1 and S4), the extended-pouch randomized trial [13] (S1), and the six-year small-versus-large pouch comparison in low-BMI diabetic patients [18] (S1). All in-scope landmarks were captured; one deliberately out-of-scope seed (a systematic review of gastrojejunostomy stoma diameter [184], whose exposure is the outlet rather than the pouch) was correctly absent from the hit lists and is retained solely as a citation-chasing seed. Precision

was visibly imperfect at piloting — the early-literature string admits 1970s case reports — and is deliberately left to screening, which is the stage built to discard noise.

8.3.3 Other sources

The Cochrane Central Register of Controlled Trials (CENTRAL) was checked manually with the same concept blocks, since its trial records do not fully overlap PubMed’s. Citation chasing is conducted backward (reference lists) and forward (citing articles) from two seed reviews — the pouch-size review of Mahawar and colleagues [15] and the stoma-diameter review of Al-Fagih and colleagues [184] — and from every study ultimately included. Records reached by citation chasing are accounted on a separate PRISMA identification line from database records.

Embase was not searched: no institutional access was available. Nor were Web of Science or Scopus, trial registries, dissertation indexes, or conference proceedings searched systematically. These are real limitations, stated here and weighed in Chapter 14: Embase indexes European surgical journals beyond MEDLINE’s coverage and its absence plausibly costs observational studies more than trials; conference-only and registry-only evidence is invisible to this search; and citation chasing, however diligently done, is not a reproducible substitute for a database search. The review therefore claims a reproducible, validated PubMed-centered search — not comprehensive coverage of the field — and treats publication and database bias as a standing threat to any null finding or small-study-effect conclusion drawn from the included set.

8.4 Study selection

Screening proceeds in the standard two stages — title/abstract, then full text — against the eligibility criteria of Section 8.2. Every full-text exclusion is recorded with a single primary reason (wrong exposure, no comparison, revisional population, wrong outcome, not primary research), and the excluded-study log with reasons is reported in Appendix D.

Screening is performed by a single reviewer. Formal systematic-review methodology expects dual independent screening, and a single-author dissertation cannot honestly claim it. The mitigation is structural rather than cosmetic: the screening decisions are recorded per record in the catalog with the criterion that decided them, so the entire screening pass is auditable and re-runnable by a second person after the fact, and the extraction stage that follows applies machine verification to every value a human might have miskeyed. The residual risk — a relevant study wrongly discarded at title/abstract stage by one tired reader — is not eliminated by any of this, and is carried as a stated limitation.

FIGURE PLACEHOLDER (F4): PRISMA 2020 flow diagram. Boxes: records identified by database searching (n=526, per Table 8.1) and by citation chasing / CENTRAL (separate line, n pending); records after deduplication; records screened at title/abstract; full texts assessed; studies included in qualitative synthesis; studies included in meta-analysis (or SWiM). Source data: generated by query from the corpus catalog (`data/catalog.db`) by the pipeline — counts are never hand-tallied. Rendering: data-rendered per spec:011 (PRISMA flow is a data plot, not an illustration); not TikZ-by-hand.

Figure 8.1: PRISMA 2020 flow diagram of study identification, screening, and inclusion. The diagram is generated programmatically from the corpus catalog; identification counts correspond to Table 8.1.

8.5 Data collection: LLM-assisted, span-verified extraction

Data extraction in this review is assisted by large language models (LLMs). That sentence, undisclosed or vaguely disclosed, would poison a systematic review, because language models fabricate plausible numbers under exactly the conditions extraction creates: a long source document, a structured target schema, and pressure to fill every field. The extraction pipeline is therefore designed around a single inviolable rule, and this section describes the machinery that enforces it.

8.5.1 The span-verification rule

No span, no citation. Every extracted value — an effect estimate, a group size, a follow-up duration, a pouch volume — must be accompanied by a verbatim quote from the source document, and that quote must string-match the stored source text exactly. The match is mechanical: an exact substring test against the immutable snapshot of the source (Section 8.10), not a semantic-similarity judgment by another model. A value whose quote fails the match is blocked from the claims table regardless of how confident the model was; it can only enter after a human locates the true supporting span or discards the claim. In signal-processing terms, span verification is a matched filter against the source text: it verifies *presence*, not semantic correctness. It cannot certify that the source’s number is true, nor that a faithfully quoted number has been attached to the correct arm, denominator, analysis population, timepoint, or adjusted-versus-crude status — a model can quote a real value and file it under the wrong comparison, and the match will pass. What the rule removes is one specific failure mode, the one LLM assistance introduces and hand extraction does not: a number synthesized in transit between the paper and the analysis. The semantic failure modes are the business of the human review and the blinded audit described below, and of nothing else in the pipeline.

8.5.2 Well-formedness gating

Extracted claims pass a second, independent gate before entering analysis: structural well-formedness. A claim missing its unit, its denominator or cohort size, its population, or its time window is blocked — flagged with a machine-readable list of the missing fields — whatever the model’s confidence score. Confidence is recorded but never gates; completeness gates. The distinction matters because model confidence measures fluency, not evidential adequacy: a fluent claim that “weight loss was greater in the small-pouch group” with no denominator and no time window is not evidence, and a gate keyed on confidence would pass it. Claims blocked for incompleteness are kept with a corroboration key rather than discarded, so that a second source reporting the missing context can complete them.

8.5.3 Model selection and adversarial checking

The model tier used for extraction is selected by a gold-case evaluation, not by assumption: extractions are scored against hand-curated gold cases drawn from real corpus papers, jointly on value *and* unit *and* scope (a correct number with the wrong denominator scores zero), and the cheapest tier meeting the accuracy bar runs the bulk extraction, with escalation to a stronger tier where the evaluation shows it is needed. The gold cases and per-tier scores are reported in Appendix B. Synthesized text generated over the corpus (as opposed to extracted values) additionally receives a fabrication check by a model from a different family than the one that wrote it; the check is advisory and logged, never silently applied.

8.5.4 Human review

Every included study’s extraction is reviewed in full by the author against the source document, with a recorded sign-off per study. The division of labor is deliberate: the machine reads everything and is never trusted; the human trusts nothing untraced and cannot read everything at uniform attention. Span verification makes the human review tractable — the reviewer checks highlighted quotes in context rather than re-deriving every number from a blank page — and the review in turn catches what span verification cannot: a quote that is faithful to the source but attached to the wrong comparison arm, or a true value extracted into the wrong schema field. Because the reviewing author is also the pipeline’s operator, per-study review alone is not an independent check of the machinery; in addition, a blinded audit is performed on a pre-specified random sample of pipeline extractions — critical quantitative fields re-extracted from the source without sight of the pipeline’s output, discrepancies adjudicated against the source, and the resulting error rates reported in Appendix B alongside the gold-case scores. Extraction from abstract-only sources (where no full text could be legitimately obtained) is permitted but flagged, and the evidence tables of Chapter 9 mark such studies explicitly; abstract-only extraction is the most vulnerable case, because the methods, attrition, and adjusted-versus-crude status that give a number its meaning live in the full text.

What this design does not do is also stated plainly. Span verification establishes provenance, not validity: a faithfully extracted number from a biased study is still biased, which is what the risk-of-bias assessment (Section 8.8) is for. And single-author review plus a blinded audit sample is not dual independent extraction of every study; the protocol records this as a limitation, weighed in Chapter 14.

8.5.5 Data items

From every included study the following are extracted: study design and country; enrollment period and surgical era (open, laparoscopic, robotic); sample size per arm; population characteristics (age, sex, baseline BMI, diabetes prevalence); the pouch-size definition, its numerical value per arm, and the measurement modality by which it was ascertained (Section 8.6); comparator logic (randomized assignment, surgeon preference, era comparison); the outcome metric as reported, with mean and dispersion per arm at each timepoint; follow-up duration and attrition per arm; safety-event counts per arm for each outcome separately — marginal ulcer, anastomotic stricture, leak, and food intolerance or reoperation for intolerance; funding source and declared conflicts. Baseline BMI is mandatory wherever any weight outcome is extracted, to enable metric conversion.

8.6 Exposure classification: the measurement-modality hierarchy

Each included study is classified, at extraction time and before any outcome data are analyzed, into the measurement-modality hierarchy defined in Section 6.9, using its tier names verbatim: V1, measured volume (per-patient volumetric imaging within the construction window); V2, imaged geometry (planar contrast measurements or instrument-referenced endoscopic measurement); V3, measured construction (per-patient intraoperative measurement — stapler length, measured pouch length, intraoperative volume determination); V4, declared technique (group-level construction targets with no per-patient measurement). The classification is the instrument of research question 3, and what it can deliver is stated with care. It drives an *exploratory effect-modification analysis*: a single pre-stated meta-regression of effect size on modality tier and study design — one model, fixed in advance, with no covariates added or dropped after the data are seen, whose deliverable is the interaction estimate with its confidence interval rather

than any significance verdict. It cannot deliver a causal verdict of measurement artifact, because modality is confounded at the study level with era, design, sample size, exposure range, and patient selection — the direct-imaging tier is populated by small cohorts, the proxy tiers by the largest registry and the randomized technique comparisons, so a between-tier difference could reflect power, confounding, or intervention extremity rather than measurement validity. The direction of inference also runs against the naive reading: classical nondifferential measurement error attenuates associations, so an effect that appears *only* under weak proxies is not explained by noise — it suggests proxy-correlated confounding or selective reporting, which is a different and testable worry. The analysis is therefore framed throughout Part II as competing-explanations evidence: design enters the pre-stated model as a covariate, while era and study size — which the expected study count cannot support as additional model terms — are argued explicitly in the interpretation. And because the hierarchy, though built on measurement-science grounds, was defined after the preliminary literature scan (Section 8.1.1), the analysis is labeled hypothesis-generating, not confirmatory.

8.7 Outcome harmonization

The literature reports weight loss in incompatible metrics (%TWL, %EWL, BMI change, excess-BMI loss), at heterogeneous timepoints, with attrition handled inconsistently. The harmonization rules applied here are derived and justified in Section 7.6 and only summarized: %TWL is the target metric, and the primary pool admits only values that reach it without leaning on a distributional assumption — native %TWL, or a direct arm-level conversion from reported BMI change and baseline BMI, which is algebra rather than approximation. %EWL and excess-BMI loss are still converted to %TWL where baseline weight or BMI and the ideal-weight convention are recoverable, with the conversion flagged in the evidence tables, but the converted values never enter the primary pool; they feed a separately labeled sensitivity analysis that repeats the pooling with the converted studies added. The demotion is deliberate, because the conversion is not exact and cannot be made so from aggregate data. Arm-level conversion through mean baseline BMI substitutes a ratio of means for the mean of a ratio and ignores the covariance between baseline BMI and weight loss — it biases both the converted mean and, more severely, the converted standard deviation. Within the sensitivity analysis, conversions are therefore performed by the delta method under explicitly stated covariance assumptions, with sensitivity bounds spanning the plausible assumption range carried through the converted-inclusive pooling; where no defensible assumption set exists, the study is not converted and its metric is analyzed separately. Converted values are never presented as equivalent to natively reported %TWL, and no primary result rests on them. Outcomes are anchored to fixed timepoint windows: one year is 10–14 months (nearest 12), three years is 30–42 months (nearest 36), and five years is 54–66 months (nearest 60). Observations beyond 66 months are reported study by study and never pooled, and where a study reports more than one eligible observation inside a window, the observation nearest the window center is taken. Nadir-anchored results are not mixed with fixed-timepoint results in any pooled estimate. Where a study reports medians and ranges, standard approximations for mean and standard deviation are applied and flagged. Attrition at each analyzed timepoint is extracted per arm and carried into the risk-of-bias assessment rather than corrected by imputation.

8.8 Risk of bias and certainty of evidence

Risk of bias is assessed per study, per outcome: randomized trials with the revised Cochrane tool RoB 2 [185], and comparative observational studies with ROBINS-I [186], which frames each observational comparison against the target trial it imperfectly emulates — the correct frame here, where the usual assignment mechanism is surgeon habit and the central confounding worry

is that surgeons who build smaller pouches differ systematically in ways that also affect weight loss. Assessment is by the single author, with the same auditability mitigation as screening: every domain judgment is recorded with its supporting quote. Certainty of evidence is rated per outcome with the GRADE framework [187], with observational evidence entering at low certainty and rated up or down by its standard criteria; the GRADE evidence profiles close Part II in Chapter 13.

8.9 Statistical synthesis

8.9.1 Primary analysis

The primary estimand is the *mean difference in %TWL at one year* between smaller- and larger-construction arms, pooled under a random-effects model across the studies whose contrasts pass the compatibility rules of the defensibility gate below. The choice of effect measure follows from the harmonization work of Section 7.6: once outcomes are on a single clinically meaningful scale, the natural measure is the difference in that scale’s own units — percentage points of total body weight — which is what any claim of clinical negligibility or importance must be stated in. A standardized mean difference is deliberately *not* the primary measure: standardization divides by each study’s within-arm standard deviation, making the result depend on population heterogeneity, attrition, and measurement practice rather than on the effect alone, and — the decisive point — it standardizes the outcome, not the exposure. It does nothing to make a 20-versus-50 ml contrast commensurable with a 5-versus-10 cm contrast, and treating it as if it did would launder exposure incommensurability through an outcome transform. Hedges’ g with the small-sample correction is computed only as a within-metric sensitivity analysis. The pool itself admits only native %TWL and direct arm-level conversions from BMI change and baseline BMI (Section 8.7); %EWL-converted values appear only in the labeled converted-inclusive sensitivity analysis, and studies whose outcomes cannot be defensibly brought to %TWL at all are analyzed separately on their native metric and reported alongside the pool, never forced in.

The random-effects model treats the true effect as a distribution with between-study variance τ^2 as an explicit parameter — appropriate here because even compatible contrasts differ in where they cut the exposure scale. What the model cannot do is also stated: a random-effects distribution over genuinely different estimands has no interpretable mean, which is why the compatibility gate precedes pooling rather than being delegated to τ^2 . Because the number of poolable studies is expected to be small, the estimation machinery is chosen for small- k behavior: τ^2 is estimated by restricted maximum likelihood, with the Paule–Mandel estimator as a pre-specified sensitivity analysis [188] — not by the classical DerSimonian–Laird moment estimator [189], which systematically underestimates heterogeneity when studies are few. Confidence intervals use the Knapp–Hartung adjustment in its *modified* form [190, 191]: the unmodified adjustment can, with few and similarly sized studies, produce intervals narrower than the conventional Wald interval, and the modification guards against exactly that. Results are reported across estimators, so that any conclusion resting on the choice of machinery is visible as such. The sign convention is fixed in advance: a positive difference favors the smaller pouch. Trials framed in the opposite direction — the extended-pouch trial randomizes to *longer* construction [13] — are entered with the contrast reversed, so that every study answers the same question with the same sign.

Studies reporting a continuous pouch-size–outcome association rather than a two-arm contrast — the registry’s stapler-length gradient [14] is the paradigm — are analyzed under a separate pre-specified *slope estimand*: effect per unit of the study’s own exposure scale, synthesized across studies only where the scales agree. They are never dichotomized into the arm-contrast pool; a forced median split would manufacture a contrast the study did not make. Effects in the fixed windows at three years and at five years are pooled separately under the same rules as the one-year primary; observations beyond the five-year window are reported study by study, never pooled. Each safety outcome — marginal ulcer, anastomotic stricture, leak, intolerance

— is analyzed separately, as absolute risks by pouch-size group. Safety pooling is gated exactly as efficacy pooling is, and more: at least three studies must report the outcome’s events, within a comparable follow-up window, and within the same ascertainment stratum — protocol endoscopy and symptomatic capture are different detectors with order-of-magnitude different efficiencies, and they are never mixed in one pooled estimate. Zero-event studies are reported raw, as observed proportions; no continuity-corrected pooled risk difference is computed from them. Where the gate fails, the outcome is synthesized narratively against the registry gradient [14]; the two directions of harm are reported per outcome and never collapsed onto one axis or mixed across effect scales.

8.9.2 Heterogeneity, subgroups, and dose-response

Heterogeneity is quantified primarily by τ^2 , the between-study variance in the effect’s own units. I^2 is reported with its correct interpretation [192]: the estimated proportion of observed variance attributable to between-study differences rather than sampling error — a relative quantity that rises with study precision and is imprecise at small k , which is why no analysis decision in this protocol keys on an I^2 threshold. A prediction interval for the effect in a new study is computed and reported only when at least five studies contribute; below that, τ^2 is too weakly identified for the interval to be more than decoration, so none is computed — and at any k , no prediction interval serves as a decision rule. Four subgroup analyses are pre-specified: measurement modality (the exploratory effect-modification analysis of Section 8.6, with design, era, and study size examined as competing explanations), study design (randomized versus observational), surgical era (open versus laparoscopic), and procedure (RYGB versus OAGB, the OAGB stratum a descriptive parallel never pooled into the primary estimate). Subgroup claims rest on the estimated between-subgroup difference with its confidence interval, never on within-subgroup significance patterns and never on a significance trigger that promotes one presentation over another; stratified estimates are reported only where a stratum contains at least three studies, and descriptively otherwise; and all subgroup comparisons are study-level and observational — labeled as such wherever Part II leans on them.

The dose-response question is addressed by a purely descriptive cross-study display of the volumetric subset: study-level outcomes set against study-level pouch volumes, with no pooled slope and no inferential confidence interval attached to any trend. An earlier draft of this protocol promised two-stage generalized least-squares trend estimation [193]; that promise is withdrawn, because the method requires within-study dose contrasts with their covariance structure, which this literature is not expected to supply at the anticipated study count — and a slope fitted across study-level means instead would be an ecological comparison, vulnerable to aggregation bias, dressed in inferential clothing it has not earned. The gradient-versus-threshold question of research question 2 is therefore addressed descriptively, by whether the extreme-contrast and narrow-range studies are mutually consistent, not by trend or spline machinery the data cannot carry. Small-study effects are assessed by funnel plot and Egger’s regression test [194] only if the primary meta-analysis contains at least ten studies, below which the test is uninformative — and where run, reported as suggestive of small-study effects, never as an adjudication of their cause.

8.9.3 Defensibility criteria and the SWiM fallback

Whether to pool is decided *before* any pooled estimate is inspected, on compatibility grounds — clinical and physical compatibility of the contrasts, not statistical-heterogeneity cutoffs. Pooling proceeds for a given outcome and window only if all of the following pre-stated criteria, fixed verbatim in the protocol of record, hold. (1) At least three studies contribute genuine two-arm construction contrasts — three being the floor for *any* pooled estimate in this dissertation, efficacy or safety; below it the presentation is stratified or narrative, never pooled. (2) All

pooled contrasts sit within one exposure family, volume with volume or length with length — never across families, because no statistical model makes 20-versus-50 ml and 5-versus-10 cm estimates of one quantity. (3) Within each study, the arms are orderable and their nominal construction targets do not overlap on the study’s own scale — non-overlap of the *stated targets*, the only quantity available at study level, not of confidence intervals or of unobserved per-patient anatomy. (4) The outcome is native %TWL, or reaches it by direct arm-level conversion from BMI change and baseline BMI (Section 8.7), within a common fixed timepoint window — %EWL-converted values never enter the primary pool. (5) A variance is extractable per arm from the study’s own report; no variance is imputed from external studies. These criteria are checkable from the evidence tables before any effect is computed, which is what makes them a gate rather than a post hoc verdict; no “uninterpretability” judgment applied after seeing the pooled result substitutes for them. When pooling proceeds, the pooled estimate is interpreted in percentage points of %TWL, with its τ^2 and, at five or more studies, its prediction interval alongside — heterogeneity statistics describe the pooled set; they do not retroactively decide whether pooling was defensible. One reporting rule attaches to the volume family: because even same-family contrasts sit at different distances on the exposure scale, every volume-family pooled mean difference is reported alongside the per-study exposure contrast distances, and where those distances vary by more than a factor of two within the pool, the pooled mean is presented as a range-averaged summary with the per-study forest as the primary display, not as a single clinical number.

If the criteria fail — globally or within a planned subgroup — synthesis follows the SWiM reporting guideline [183]: per-study effect sizes with their uncertainty, reported on each study’s defensible metric and grouped by exposure family and modality tier, with a transparent statement of which synthesis question each grouping can and cannot answer. Direction-of-effect or p-value vote counting is explicitly excluded — counting signs discards magnitude and precision, which are the only things the studies actually estimate. Either way, the heterogeneity itself is dissected and reported in Chapter 10; the fallback is a different reporting discipline, not a lesser one.

8.10 Software, provenance, and reproducibility

The evidence infrastructure is a purpose-built Python pipeline over a single SQLite catalog, developed in the same repository as this text. Sources are stored as immutable snapshots — raw bytes and extracted text, addressed by SHA-256 content hash — so the text a claim was verified against is fixed at fetch time and cannot drift under the citation. The catalog records, per source: identifiers (PMID, PMCID, DOI), citation metadata, retraction status as exposed by PubMed (claims from retracted sources carry the flag into every table that cites them), language, and text depth (metadata only, abstract, or full text) — so that a claim extracted from an abstract can never silently masquerade as full-text evidence. At the time of search execution the corpus held 526 cataloged records, 107 of them with full text from the PubMed Central open-access subset. Access is legitimate-only by design: the pipeline uses the E-utilities interface and plain HTTP against open-access sources, and no paywall circumvention is performed; where full text was inaccessible, the study is analyzed from its abstract and flagged as such. All analysis figures in Part II (forest, funnel, dose-response, PRISMA flow) are rendered from pipeline data by code in the repository, never drawn or generated by a model; anatomical and schematic figures elsewhere in this dissertation are LLM-authored vector graphics and are disclosed as such in their captions, with full provenance in Appendix E. The pipeline itself is documented in Appendix F.

8.11 Ethics

This study is secondary research on published aggregate data; no human participants were recruited and no individual-level data were accessed. No ethics approval was required.

8.12 Protocol deviations

Deviations from the protocol are recorded as dated amendments in the protocol file and summarized here in the final text. At the time of this draft, one amendment predates screening: the addition of the fourth search string (S4) after piloting revealed that early-era records without indexed abstracts were unreachable by the original three strings (Section 8.3). The amendment was made before any screening decisions and affects identification only.

Chapter 9

The evidence base: corpus and study characteristics

This chapter maps the evidence on which Part II’s analyses stand. It has two layers. The first is available now and is reported here in full: the candidate corpus — the 526 unique records that four systematic searches retrieved from PubMed, cataloged with their metadata, abstracts, and, where openly available, full texts. A candidate corpus is itself an object worth studying, provided its nature is kept in view: every count in this layer is a property of *what the search retrieved* — four deliberately over-inclusive queries, noise included by design — not of the eligible literature, which does not exist as a defined set until screening completes. Read that way, the corpus’s shape in time tracks the history told in Chapter 3; its concentration in a handful of journals maps where the retrieved records are published; and its census of record types describes the raw material the screen must sort. Claims about the eligible evidence base itself are deferred to the screened set. The second layer — the included set that survives screening, with per-study structured evidence profiles — does not exist yet. Where it belongs, this chapter carries explicit placeholders rather than provisional numbers: per the protocol (Chapter 8), no screening or inclusion count is reported until screening has actually been performed and logged.

9.1 From search to catalog

The search strategy is specified in Chapter 8 and reproduced verbatim, with dates and hit counts, in Appendix A. Its structure matters for reading the corpus, so it is summarized here. Four PubMed queries were executed through the E-utilities pipeline on August 31, 2026. S1 is the core query: gastric bypass terms crossed with pouch/reservoir terms, size terms, and weight-outcome terms. S2 targets one-anastomosis gastric bypass (OAGB), the single-connection variant of the operation in which the pouch is deliberately long and pouch dimensions vary far more widely — the pre-specified contrast subgroup of research question Q6. S3 drops size words entirely and instead crosses pouch terms with construction terms (stapler, staple line, bougie, calibration), because a study that fixed its pouch by technique may never use the word “size” in its title or abstract. S4 is the early-literature supplement: records from the 1970s and 1980s frequently lack indexed abstracts, so title-and-abstract queries see only their titles, and the founding papers phrase the question as “reservoir versus stoma” rather than “pouch size.” S4 therefore pairs reservoir/pouch terms with stoma terms and no others.

Table 9.1 gives the yield. The four searches returned 696 hits resolving to 526 unique records. The overlap structure is worth a moment: 373 records (70.9%) were found by exactly one search, 136 by two, 17 by three, and none by all four. Each query therefore contributed a genuinely distinct slice — S3’s 64 unique construction-technique records and S4’s 27 early records are papers the core query cannot see — which is the intended behavior of a multi-pronged strategy and a warning against single-string searches in a literature whose vocabulary shifted over forty

Table 9.1: Search yield, PubMed via E-utilities, executed 2026-08-31. “New” counts records not returned by any earlier search in the sequence. Full query strings in Appendix A.

Search	Focus	Matches	New	Cumulative unique
S1	Core RYGB pouch–size–outcome	421	421	421
S2	OAGB subgroup	59	14	435
S3	Construction technique	131	64	499
S4	Early literature (reservoir/stoma)	85	27	526

years. The recall check built into the pilot confirmed that every in-scope landmark study known from the preliminary scan was captured, including the 1982 reservoir-versus-stoma paper[5] that only S4 reaches.

The catalog records what the pipeline could legitimately obtain for each hit. Of the 526 records, 521 carry an indexed abstract; five are metadata-only (title, authors, journal, year), all from the early literature. For 107 records (20.3%), the full text was retrieved from PubMed Central under open-access terms and stored as an immutable, hash-verified snapshot; these full texts are the substrate for the span-verified extraction described in Chapter 8. Full texts of included studies not available through PMC will be obtained through legitimate access channels during extraction, and the final full-text coverage of the included set will be reported here. Language: 511 records (97.1%) are in English; the remaining 15 comprise six German, three Spanish, three Chinese, and one each in French, Polish, and Hungarian — flagged, per protocol, rather than excluded at the search stage. No record in the corpus is flagged as retracted in its PubMed metadata.

9.2 Screening and the PRISMA flow

Screening had not begun when this chapter was drafted; the PRISMA flow diagram (Figure F4, Chapter 8) and every count downstream of the 526 cataloged records will be inserted from the screening log, not estimated. The reporting contract is fixed now: the flow will state records identified per source (the four PubMed searches, Cochrane CENTRAL, and backward/forward citation chasing, each on its own accounting line), records screened on title and abstract, records excluded with tallied reasons, full texts assessed, full texts excluded with per-study reasons (Appendix D), and studies included — separately for the RYGB core and the OAGB subgroup. Noise is expected and visible already at the catalog level: the yield includes 1970s pediatric case reports, trials of jejunal pouch reconstruction after cancer gastrectomy, and revision-surgery series whose exposure is a dilated pouch rather than a constructed one. That is what screening is for; none of it was excluded by search design, because precision bought at the search stage is recall silently lost.

PLACEHOLDER — PENDING SCREENING: PRISMA counts from 526 cataloged records to the included set (title/abstract exclusions with reasons; full-text assessments; inclusions, RYGB core vs. OAGB subgroup). To be generated from the screening log by the spec:002 pipeline; no number here may be written by hand.

9.3 The corpus in time

The corpus spans 1975 to 2026 with a median publication year of 2014; 85.6% of the records postdate 2000 (Table 9.2). The distribution describes what the search retrieved — relevant and irrelevant alike — and is read here as a retrieval profile, not as a census of the eligible evidence; even so, its two ends repay attention.

Table 9.2: Candidate corpus by decade of publication (n = 526; catalog query, 2026-09-01). The 2020s row spans 2020–2026.

Decade	Records	Share
1975–1979	7	1.3%
1980–1989	27	5.1%
1990–1999	42	8.0%
2000–2009	120	22.8%
2010–2019	164	31.2%
2020–2026	166	31.6%

The oldest record in the corpus is Mason’s 1975 report in *Annals of Surgery*, and it is the doctrine at its source. Reviewing 442 patients over nine years, Mason’s group distilled the operation into three components: bypass of stomach and duodenum, “a small fundic segment,” and a small gastroenterostomy stoma — and predicted that their observed rate of stomal ulceration “may be lowered with the present emphasis on a smaller fundic pouch”[4]. Both halves of this dissertation’s question — does small-pouch construction drive weight loss, and what does it cost in ulcers — are present in the corpus’s very first document, stated as convictions rather than findings. Seven years later Kirkpatrick and Siegel ran the first deliberate test of the doctrine’s internal ranking: with a consistently tiny reservoir (under 45 ml), stoma size had little effect on success, elevating the reservoir to the “critical determinant” of the operation[5]. The pouch-size question is thus older than almost everything else in bariatric surgery, and it entered the literature already answered — by assertion.

The middle of the distribution follows the history of Chapter 3: a trickle through the open-surgery decades (76 records to 1999), then a step change as the operation went laparoscopic — by 2008, 98% of 215 surveyed American bariatric surgeons reported performing the bypass laparoscopically[6] — with the 2000s nearly tripling the 1990s. The most striking feature, however, is at the recent end: the retrieval is still accelerating. The 2020s already match the entire 2010s with more than three years of the decade unpublished, and the two largest single years in the whole corpus are the two most recent — 26 records in 2025 and 27 in 2026. A fifty-year-old question about one of medicine’s most standardized operations is generating more candidate records per year now than at any prior point. How much of that growth belongs to the eligible literature rather than to the search’s deliberate over-capture — the OAGB wave re-opening pouch geometry, revision surgery forcing retrospective interest in construction, case reports of complications, or simple publication inflation — is a question only the screened set can answer; but at minimum, the retrieval profile is not that of a question the field has stopped writing about.

9.4 The corpus in print and in place

The retrieved records are concentrated in few venues. A single journal, *Obesity Surgery*, carries 177 of the 526 records (33.7%); with *Surgical Endoscopy* (42) and *Surgery for Obesity and Related Diseases* (39), the top three account for 49.0% of the corpus (Table 9.3). The remainder scatters across more than a hundred titles, with general surgery journals (*Annals of Surgery*, *Archives of Surgery*, *The British Journal of Surgery*) contributing mostly the older records and gastroenterology/endoscopy journals appearing with the revision-era literature. The concentration is a property of the retrieval, and on its own it carries no inference: a specialty question concentrating in specialty journals may reflect nothing more than appropriate indexing. Whether publication bias or small-study effects operate in this literature is not answerable from journal counts; Chapter 13 weighs those threats with the instruments built for them, on the included

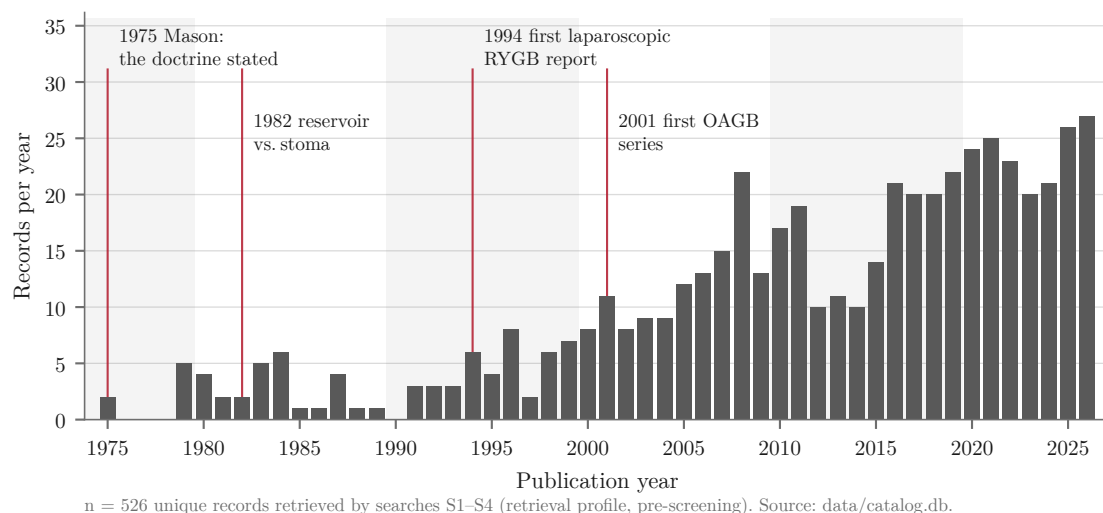


Figure 9.1: The candidate corpus in time: annual counts of the 526 unique records retrieved by searches S1–S4, 1975–2026. Every count is a property of the search retrieval — four deliberately over-inclusive queries, pre-screening — not of the eligible literature. Vertical marks: the doctrine’s origin (Mason 1975; the 1982 reservoir-versus-stoma study), the first laparoscopic RYGB report (1994), and the first OAGB series (2001). Rendered from `data/catalog.db` by `analysis/corpus_figs/make_ch09_figs.py` (spec:011 data-figure policy; provenance in Appendix E); underlying decade totals in Table 9.2. A companion panel restricted to the included set will be added once screening completes.

set.

Geography is reported here with a caveat the catalog imposes. PubMed record metadata as stored does not include a normalized country of origin, so no country distribution can be honestly queried from the catalog today; author affiliations will be extracted from the stored full texts and abstracts during the extraction phase, and the geographic map will be rendered from that table. Two weak signals are available now. Language: 97.1% of records are in English, with the non-English tail (German, Spanish, Chinese, French, Polish, Hungarian) concentrated in the pre-2000 literature. And the anchor studies of Section 9.6 already sketch the poles of the coming map: the largest registry evidence is Scandinavian, the randomized evidence is Chinese and Dutch, and the survey evidence spans 47 countries — a geographic split that will matter in Chapter 11 because it is confounded with population, baseline body-mass index, and surgical convention.

FIGURE PLACEHOLDER — PENDING AFFILIATION EXTRACTION: geographic distribution of the corpus (records per country of first-author affiliation; included set highlighted). Source data: affiliation table to be extracted by the spec:002 pipeline from stored snapshots; not derivable from `catalog.db` as it stands. Data-rendered per spec:011.

Figure 9.2: The corpus in space (pending affiliation extraction; see text for why this cannot yet be queried).

Table 9.3: Journals contributing five or more records to the candidate corpus (catalog query, 2026-09-01).

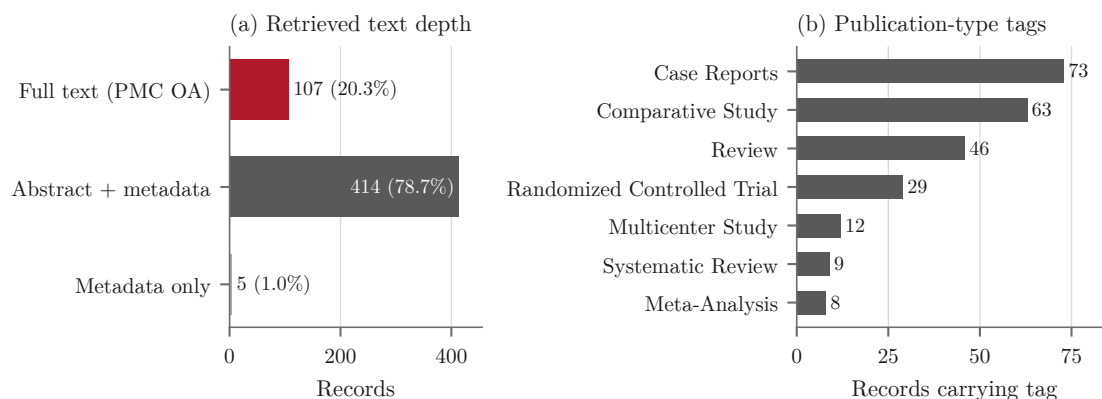
Journal	Records
Obesity Surgery	177
Surgical Endoscopy	42
Surgery for Obesity and Related Diseases	39
Cureus	14
Gastrointestinal Endoscopy	10
Annals of Surgery	10
International Journal of Surgery Case Reports	9
JSLs	8
British Journal of Surgery	6
Journal of Gastrointestinal Surgery	6
Archives of Surgery	6
American Journal of Surgery	6
Surgical Clinics of North America	5
The American Surgeon	5
J. Laparoendoscopic & Adv. Surgical Techniques A	5
BMC Surgery	5

9.5 A census of record types

PubMed publication-type tags are crude — they are assigned by indexers, not by methodological review — and this census, like every count in this chapter, describes the retrieval, not the eligible evidence base: the searches deliberately over-capture, so the tallies include records that screening exists to remove. With that caveat fixed, the composition of the candidate set is worth recording. Among the 526 records: 73 case reports (13.9%), 63 comparative studies, 46 reviews, 29 records tagged as randomized controlled trials, 12 multicenter studies, nine systematic reviews, and eight meta-analyses (Figure 9.3b; panel a shows the text-depth breakdown of Section 9.1). Two features of this census matter for what follows.

First, the top of the pyramid is thinner than it looks, in both directions. Of the 29 RCT-tagged records, a large fraction are not about the question at all: the tag sweeps in randomized trials of jejunal pouch reconstruction after total gastrectomy for cancer — a different pouch in a different disease — along with trials of gastric banding, staple height at the gastrojejunostomy, and revision procedures. The tags also miss: the six-year Chinese cohort comparing large and small pouches, described in secondary accounts as randomized, states plainly in its methods that it is a retrospective comparative cohort[18], and it carries no RCT tag. Neither the tag nor the abstract impression substitutes for full-text screening, which is why the protocol requires design classification from methods sections, with every included study human-reviewed.

Second, no pooled estimate of the pouch-size effect was identified within the candidate set. Of the nine systematic-review-tagged records, exactly one addresses pouch or stoma size against weight loss: Mahawar and colleagues’ 2020 review, which found 14 studies, judged that “quality and heterogeneity of data precluded a meta-analysis,” and proceeded without pooling[15]. Of the eight meta-analysis-tagged records, none pools pouch size against weight loss; they concern ghrelin dynamics, revision procedures, hiatus hernia repair, gastrectomy reconstruction, and pediatric surgery. What this establishes is stated with its limits: within a 526-record candidate set built by four overlapping searches designed to over-capture, no pooled estimate exists — a queryable, reproducible property of this PubMed-derived retrieval. It cannot establish that the worldwide synthesis layer is empty: Embase, Scopus, and non-indexed sources were not



n = 526 cataloged records (retrieval profile, pre-screening); (b) tags are indexer-assigned, not mutually exclusive. Source: data/catalog.db

Figure 9.3: A census of the candidate corpus (n = 526; retrieval profile, pre-screening). (a) Retrieved text depth: full text stored from PubMed Central open access for 107 records (20.3%, accented — the substrate for span-verified extraction), indexed abstract plus metadata for 414 (78.7%), metadata only for five (1.0%). (b) PubMed publication-type tags among the 526 records; indexer-assigned, not mutually exclusive, and describing the retrieval rather than the eligible evidence base. Rendered from data/catalog.db by analysis/corpus_figs/make_ch09_figs.py (spec:011 data-figure policy; provenance in Appendix E).

searched (Chapter 8), and the field-wide claim remains “none identified,” to be finalized only after CENTRAL checking and citation chasing complete.

The base of the pyramid, meanwhile, is growing fastest in the retrieval: case reports rise from 12 in the 2000s to 24 in the 2010s to 33 already in the 2020s, largely pouch complications and revision anatomy — much of it destined for exclusion at screening. The recent acceleration of Section 9.3 is thus disproportionately anecdote at the retrieval level; whether the eligible literature shares that shape is deferred to the screened set.

9.6 Anchor studies: the fixed points visible before screening

Screening will define the included set, but the field’s load-bearing studies are already known — they motivated the dissertation (Chapter 1) and served as the recall check for the search. Table 9.4 profiles them from their retrieved texts, in the format the full per-study evidence tables (Appendix C) will use. They are presented here not as results — Chapters 10 and 11 do that work — but as a demonstration, on the best-known corner of the corpus, of what the heterogeneity that defeated prior synthesis actually consists of.

Read down the exposure column first. The largest study measures the pouch in millimeters of stapler cartridge consumed — a construction proxy, since stapler length tracks staple-line length, not enclosed volume[14]. The Chinese studies define it in milliliters, estimated intraoperatively from stapler-scale width and height through a geometric formula[18, 17]. The Dutch trial defines it in centimeters of length, and its “extended” pouch is long and narrow — a shape contrast at least as much as a size contrast[13]. The imaging studies measure actual volume, but after surgery, in revision-era patients, when dilation has already decoupled measured volume from constructed volume[154, 150]. These are not four precisions of one measurement; they are four different physical quantities sharing a word. Chapter 6 built the quality hierarchy for exactly this reason, and the anchor table is its first application.

Now read the direction column. Within these ten studies alone, smaller is better at one year in a 76-patient trial[17], irrelevant at one year in a 14,168-patient registry[14], and *worse* than

Table 9.4: Anchor studies of the pouch-size literature, profiled from retrieved abstracts and full texts (extraction pending pipeline verification). EBMIL = excess body-mass-index loss; TWL = total weight loss; EWL = excess weight loss; MU = marginal ulcer (an ulcer at the gastrojejunal anastomosis).

Study	Design	N	Pouch exposure	Outcome, horizon	Direction
Edholm 2016[14]	Registry cohort (SOREg)	14,168	Total stapler length (mean 145 mm)	%EBMIL, 6 wk / 1 y; MU	Predicts 6 wk, not 1 y; MU risk +14%/cm
Mahawar 2020[15]	Systematic review	14 studies	Mixed (as reported)	Weight loss	9 null, 5 favor smaller, 0 favor larger
Ren 2015[17]	RCT	76	10–20 ml vs. 25–35 ml	%EWL, 1 y	Favors smaller
Yi 2015[195]	RCT	60	Undivided-stomach gastrojejunostomy vs. small pouch	BMI, glycemia, 1 y	Favors small pouch
Gao 2022[18]	Retrospective cohort	95	>50 ml subtotal vs. 10–20 ml	BMI change, 6 y; MU	Favors smaller; MU 23.7% vs 6.4% in <i>large</i>
Boerboom 2020[13]	RCT	132	Length 10 cm vs. 5 cm	%TWL, 3 y	Favors <i>longer</i> (31% vs 27%)
Harker 2026[16]	RCT follow-up	62	Extended vs. standard	%TWL, 5–9 y	26.1% vs 24.1%, n.s.; attrition-limited
Hany 2026[7]	Survey, 47 countries	245 surgeons	Practice patterns	—	53.5% measure pouch; median 6 cm
Karcz 2009[154]	Imaging series	26	MSCT volumetry	Feasibility	Volumetry evaluable in 96%
Uittenbogaart 2020[150]	Case-control	202	UGI contrast (dilation)	Weight-loss failure	23% vs 11% dilated; low interobserver agreement

a longer pouch at three years in a 132-patient trial[13] — whose own extended follow-up then decays to a nonsignificant trend under 5–9 years of attrition[16]. The contradiction is not a subtle statistical tension; it is a sign disagreement among the field’s best studies. Meanwhile the ulcer signal is the most consistent finding on the table, and it runs one way along the exposure: in the registry, each additional centimeter of staple line raises marginal-ulcer risk by 14% (95% CI 9–20%)[14], and in the low-BMI diabetic cohort the *large* pouch carried 23.7% ulcer incidence against 6.4% for the small[18] — both penalizing larger construction, though “larger” means more staple line in one study and more retained acid-secreting mucosa in the other. The harms attached to the small end of the scale — stricture, anastomotic tension, intolerance — are not visible in this table at all, an absence the bidirectional safety analysis of research question 5 must carry rather than ignore. And the survey column quantifies the doctrine’s present condition: barely half of 245 surgeons across 47 countries routinely measure the pouch they construct at all, and most describe it with unstandardized words rather than numbers[7]. The exposure variable of this entire literature is, in the majority of current practice, unrecorded.

9.7 The anatomy of heterogeneity, pre-stated

Mahawar’s review declined to pool on the grounds of quality and heterogeneity[15] but did not anatomize that heterogeneity quantitatively. This section fixes the axes along which the included set will be profiled, so that the profiling cannot be accused of being fitted to the data it later describes. Six axes, each already visible in the anchor table: (1) *exposure metric* — volume, length, stapler consumption, bougie caliber, or subjective descriptor, mapped to the Section 6.9 hierarchy; (2) *contrast definition* — what “small” and “large” mean in each study, and the distance between them in each study’s own units; (3) *outcome metric and horizon* — %TWL, %EWL, %EBMIL, or BMI change, at timepoints from six weeks to nine years, harmonized by the Section 7.6 rules; (4) *population* — baseline BMI and diabetes status, which separate the East Asian metabolic-surgery literature from the Euro-American weight-loss literature; (5) *design and era*; and (6) *attrition* — magnitude and handling, given that follow-up loss is the mechanism by which a nine-year randomized comparison can arrive underpowered[16]. For each axis the evidence profile will record the study-level value verbatim (span-verified) and a harmonized value with its conversion rule.

PLACEHOLDER — PENDING SCREENING AND EXTRACTION: (i) full per-study structured evidence profiles for the included set (design, N, pouch definition and modality tier, comparator logic and contrast distance, outcome metric, follow-up, attrition, funding), summarized here with complete versions in Appendix C; (ii) the quantitative heterogeneity table over the six pre-stated axes, which Chapter 10 consumes as the basis for pooling and subgroup decisions. All values pipeline-extracted with verbatim span verification and human review; nothing in these tables may be drafted ahead of the data.

What can be said ahead of the data is the shape of the problem the data must answer. The candidate corpus assembled here is large, old, accelerating, concentrated in three journals, without an identified pooled synthesis, and anchored by well-known studies that measure different quantities, on different populations, against different outcomes, and disagree in sign. That configuration is consistent with a proxy-contaminated literature in which associations survive only under weak, confounding-prone measures — and equally consistent with a real but shape-dependent effect diluted across incommensurable exposures. Chapters 10 and 11 weigh those competing readings against each other; neither can be adjudicated outright by a study-level synthesis, and the analysis plan of Chapter 8 is built to say which evidence bears on which reading rather than to declare a verdict the design cannot deliver.

Chapter 10

Primary analysis: does pouch size predict weight loss?

This chapter answers the dissertation’s first two questions, restated from Section 1.6. Q1: in adults undergoing primary Roux-en-Y gastric bypass (RYGB), does a smaller constructed pouch produce greater weight loss (percent total weight loss at 1 year and at ≥ 3 years)? Q2: is there a gradient, or is any effect threshold-like? No pooled answer has been identified in this literature. The single systematic review of the question, by Mahawar and colleagues in 2020, located fourteen studies, judged their quality and heterogeneity incompatible with meta-analysis, and stopped at a narrative tally: nine studies found no significant association between pouch size and weight loss, five found larger pouches associated with poorer weight loss, and none found larger pouches associated with better weight loss [15]. Everything in this chapter is therefore either the first pooled estimate identified in this literature or a rigorous demonstration of why no defensible pooled estimate exists. Both analyses were specified before any effect data were extracted; the specification is reproduced here in full so that the results, when they arrive in the slots marked below, can be read against a commitment that predates them. One honesty note governs that claim: the specification predates *extraction*, but it was written after a preliminary scan of the landmark studies’ abstracts, and the protocol (spec:010) records, with dated version history, which design choices preceded the scan and which followed it. Analyses whose form was informed by the scan are labeled hypothesis-generating wherever they appear; nothing shaped after seeing study results is called confirmatory.

10.1 The estimand

The quantity being estimated must be stated with care, because the included studies do not share one. Each study defines its own contrast: a “small” pouch versus a “large” one, where the labels mean whatever the study’s construction protocol or measurement made them mean. In the randomized trial of Ren and colleagues, small meant an estimated 10–20 mL pouch and large meant 25–35 mL [17]. In the Scandinavian registry cohort, pouch size was a continuous proxy — the total length of surgical stapler fired to construct the pouch, mean 145 mm [14]. In the Dutch extended-pouch trial, the contrast was a 10 cm pouch versus a 5 cm pouch [13]. These are not points on one scale; they are different physical quantities (volume, staple-line length, pouch length) measured in different ways at different times.

The primary analysis therefore pools *study-defined small-versus-large contrasts*: within each study, the group constructed with the smaller pouch by that study’s own definition is compared with the group constructed with the larger one, on percent total weight loss (%TWL) at one year. The effect measure is the *mean difference* in %TWL, in percentage points — an absolute, clinically readable quantity on the scale the field’s targets are stated in. A study enters the primary pool only if it reports %TWL natively or reaches it by direct arm-level conversion from

BMI change and baseline BMI (Section 7.6), which is algebra on reported quantities rather than an approximation. %EWL-derived conversions never enter the primary pool, because an arm-level %EWL→%TWL conversion uses the group’s mean baseline BMI where the definition uses each patient’s own, so it ignores the within-arm covariance between baseline weight and follow-up weight — an error the aggregate data cannot bound away. Such conversions are still performed, flagged, and propagated with delta-method sensitivity bounds under stated covariance assumptions, but they feed only the labeled sensitivity analysis of Section 10.4, which repeats the pooling with the converted studies added and shows whether the native-%TWL conclusion survives their inclusion. Where a study’s metric cannot be defensibly brought to %TWL at all — absolute loss without baseline moments, nadir-anchored outcomes — it is analyzed within its own metric and reported separately, not forced onto the common scale. The standardized mean difference (SMD) — the between-group difference divided by the pooled standard deviation, a signal in units of noise σ — is computed only as a within-metric sensitivity analysis; it is dimensionless by construction, so no result in this dissertation is labeled an “SMD in %TWL,” and it is not the primary effect measure, because SMDs absorb each study’s outcome reliability and population spread into the denominator and can standardize materially different constructs into false agreement. Sign convention throughout: positive differences favor the smaller pouch.

One class of study cannot enter this estimand at all, and pretending otherwise would be the analysis’s first error. Studies whose exposure is *continuous* — the registry’s stapler length [14], Nishie’s continuous pouch area [118] — report no small-versus-large contrast, and manufacturing one by dichotomizing at a median or an arbitrary cut point would create an effect estimate the study never made, at a comparator distance chosen by this analysis rather than by any design. The protocol (spec:010) therefore specifies a *separate slope estimand* for continuous-exposure studies: the change in outcome per unit of the study’s native exposure scale, extracted as reported, synthesized across such studies only where their exposure scales genuinely align, and otherwise presented study by study alongside the arm-contrast pool. The two estimands are reported side by side and never merged.

Two features of the estimand deserve emphasis. First, it is an estimate of the effect of *constructed* size — the surgeon’s decision at the operating table — not of pouch size at any later time. Postoperative dilation is a different variable with a different causal structure and is excluded by protocol (spec:010); Section 6.7 explains why conflating the two has corrupted much of the field’s reasoning. Second, the estimand inherits each study’s *comparator distance*: a trial comparing 15 mL with 30 mL pouches and a cohort comparing arms a few milliliters apart are testing doses of different magnitude, and a pooled mean difference averages over that distance. The protocol therefore attaches an honesty rule to the volume family: every pooled mean difference is reported alongside the per-study exposure contrast distances, and if those distances vary by more than a factor of two within the pool, the pooled mean is presented as a range-averaged summary with the per-study forest as the primary display — not as a single clinical number. This is a real limitation of the small-versus-large design, and it is what the exploratory dose-response analysis of Section 10.6 exists to probe — though, as that section states, not to repair.

One coding problem must be settled before any study enters the pool, because it determines the sign of two of the field’s most prominent effect estimates. The Dutch trial randomized patients to an *extended* pouch — deliberately longer (10 cm versus 5 cm) but narrow, constructed tight to a calibration bougie — on the hypothesis that a longer, narrower pouch dilates less and empties more slowly [13, 16]. Along the length axis the extended pouch is the *larger* exposure; along the axis most volume-based studies intend, it is not clearly larger at all, and the trial’s own physiological rationale (Laplace’s law: wall tension, and hence dilation propensity, grows with radius) concerns geometry, not volume. Coding this trial as “large pouch” alongside Ren’s 25–35 mL arm would subtract quantities measured on different axes. The pre-stated rule:

pooling happens within one exposure family only — volume-defined contrasts with volume-defined, length-defined with length-defined, never across — and every pooled contrast must be orderable within its family, with non-overlapping nominal construction targets. The extended-pouch design is a length-family contrast with volume held ambiguous; it is analyzed within the length family and is never merged into the volume pool, under any sign coding, because no coding choice can turn a length difference into an estimate of a volume difference.

10.2 The model, and how to read the forest plot

Pooling uses a random-effects model [189]. The model treats the true effect in study i as a draw $\theta_i \sim \mathcal{N}(\mu, \tau^2)$: the studies are not assumed to measure one identical constant with different noise, but to measure genuinely different true effects — different populations, comparator distances, techniques — distributed around a mean μ with between-study standard deviation τ . Given the exposure disorder just described, a fixed-effect model (all studies share one θ) would be a physically false description of this literature; it is run only as a sensitivity check. Each study is weighted by the inverse of its total variance (within-study sampling variance plus τ^2), and τ^2 itself is estimated from the data by restricted maximum likelihood, with the Paule–Mandel estimator as a sensitivity comparison — the estimator choice matters at small k and the classical DerSimonian–Laird estimator is known to misbehave there [188]. The confidence interval for μ uses the modified Knapp–Hartung (Hartung–Knapp–Sidik–Jonkman) adjustment rather than a normal Wald interval, because a Wald interval after any τ^2 estimate ignores the uncertainty in the heterogeneity estimate itself and is anti-conservative at the study counts this literature will supply [190]; the unadjusted interval is reported alongside as a sensitivity check, and disagreement between the two is itself reported as a small-sample warning.

Three statistics describe the spread. τ^2 is the variance of the true effects — the heterogeneity itself, in squared percentage-point units of %TWL. I^2 re-expresses it relative to noise: the estimated proportion of the variability in observed effects attributable to between-study heterogeneity rather than within-study sampling error [192]. It must be read for what it is — a relative, not an absolute, quantity: because its denominator contains the within-study variances, I^2 rises as the included studies become more precise even when τ^2 is unchanged, so it measures heterogeneity relative to the precision of *this* set of studies, not a stable property of the literature; τ^2 and the prediction interval, which are on the outcome scale, carry the absolute information. The 95% prediction interval is the most clinically honest of the three: the range in which the true effect of a new study, drawn from the same distribution of populations and techniques the model assumes, is expected to fall [196]. A pooled mean excluding zero while the prediction interval straddles it means the *average* true effect favors smaller pouches while individual study settings — particular populations, techniques, comparator distances — may not; it is a statement about study-level variation under the model, not a direct forecast of what any one surgeon or patient will experience. That distinction, and its limit, need stating in every summary of a literature whose single studies have contradicted one another for decades. Per protocol, the prediction interval is computed and reported only at five or more studies; below that, τ^2 — and hence the interval — is too weakly identified to bear weight, so no interval is computed, and at no k does a prediction interval serve as a decision input (spec:010).

The forest plot (Figure 10.1) is read as follows. Each row is one study’s contrast: the square marks the study’s mean difference in %TWL, its area proportional to the study’s weight in the pool; the horizontal whiskers span the 95% confidence interval. The vertical line at zero is the null — no difference in %TWL between small and large construction; effects to the right favor the smaller pouch. The diamond at the bottom is the pooled estimate, its width the 95% confidence interval of μ ; the extended horizontal bar beneath it is the 95% prediction interval. Heterogeneity statistics print below the diamond. Where rows cluster tightly around the diamond, the studies agree; where confidence intervals barely overlap the diamond, the pooled mean is an average of

disagreement, and the subgroup structure of Chapter 11 becomes the more informative display.

FIGURE PLACEHOLDER (F5): Forest plot of study-defined small-versus-large pouch contrasts, mean difference in %TWL (percentage points) at 1 year. One row per included study (author, year, N, exposure definition, modality tag), squares sized by inverse-variance weight, 95% CIs, pooled random-effects diamond, 95% prediction interval bar (drawn only if $k \geq 5$, per spec:010), τ^2/I^2 annotation. Data-rendered from the pipeline extraction table per spec:011 (never hand-drawn, never LLM-generated); source data file and rendering script logged in Appendix E.

Figure 10.1: Primary meta-analysis: constructed pouch size and percent total weight loss at one year. [RESULT PENDING]

10.3 Defensibility criteria for pooling, stated in advance

Mahawar’s review declined to pool on the grounds of quality and heterogeneity [15] but did not state criteria under which pooling would have been acceptable. This analysis does, so that “we pooled” and “we could not pool” are both verifiable outcomes rather than judgment calls made after seeing the data. The primary pooled estimate is reported as a defensible synthesis only if all of the following hold:

The criteria are of two kinds, and the distinction is part of the pre-commitment: criteria 1–3 are *veto*es — failure of any one means no pooled estimate is claimed and the analysis routes to the Synthesis Without Meta-analysis framework of Section 10.8 — while criteria 4–6 are *reporting rules* that change how a computed pooled estimate is presented without suppressing it.

1. **Minimum evidence (veto).** At least three studies contribute a small-versus-large contrast on %TWL (native, or by direct arm-level ΔBMI /baseline-BMI conversion) within the one-year window of 10–14 months, taking the observation nearest 12 months where a study reports more than one inside the window. Three is the floor for *any* pooled estimate in this dissertation, efficacy or safety; below it the presentation is stratified or narrative. At $k = 3$ the pooled mean and its modified Knapp–Hartung interval are reported, but τ^2 is acknowledged as weakly identified and, per protocol, no prediction interval is computed below $k = 5$ (spec:010).
2. **One exposure family, orderable arms (veto).** Every pooled contrast sits within a single exposure family — volume-defined with volume-defined, length-defined with length-defined, never across — and is orderable on that family’s physical quantity with non-overlapping nominal construction targets, per the coding rule of Section 10.1; length-defined contrasts are never merged into the volume pool, and continuous-exposure studies remain in the slope estimand.
3. **Extractable and verified inputs (veto).** Group means and standard deviations (or statistics from which they are recoverable by standard identities) are extracted through the pipeline and span-verified against the stored source text — no imputation of unreported variances from external studies enters the primary analysis.
4. **Dominance (reporting rule).** If a single study carries more than 60% of the pooled weight, the pooled estimate is reported alongside, not instead of, that study’s own result, and the leave-one-out analysis is promoted to the main text. Dominance does not veto pooling; it changes the presentation, and the 60% threshold is computed from the model weights, not judged by eye.

5. **Interpretable center (reporting rule).** The heterogeneity statistics are reported unconditionally, and there is no I^2 veto — a heterogeneous pool is reportable. Whether the study-level effects split by an identifiable structural feature is carried by the pre-specified subgroup machinery of Chapter 11, whose deliverable is the interaction estimate with its confidence interval — there is no significance trigger, and no nominal p -value promotes stratified estimates over the overall presentation. Stratified estimates are themselves reported only where a stratum holds at least three studies, and descriptively otherwise.
6. **Comparator-distance honesty (reporting rule).** Within the volume family, the pooled mean difference is reported alongside each study’s exposure contrast distance; if those distances vary by more than a factor of two within the pool, the pooled mean is presented as a range-averaged summary with the per-study forest as the primary display, not as a single clinical number (Section 10.1).

Either way, every computed statistic is reported; the criteria govern what the dissertation *claims*, not what it shows.

10.4 The pooled estimate at one year

The contrasts the analysis must reconcile are best introduced by the studies that created the controversy. Table 10.1 profiles the landmark candidates known before screening — the studies around which the forty-year disagreement has organized. It is illustrative, not the included set; the included set, with full evidence profiles, is established by the PRISMA screening of Chapter 9.

Table 10.1: Landmark small-versus-large pouch contrasts (pre-screening illustration; directions as reported by each study’s abstract or full text, before any pipeline extraction). RCT = randomized controlled trial; UGI = upper-gastrointestinal contrast study; %EWL = percent excess weight loss; %TWL = percent total weight loss; EBMI = excess-BMI loss.

Study	Design, N	Exposure contrast (modality)	Outcome, follow-up	Reported
Roberts 2007[103]	Cohort, 320	pouch area, mean 63.9 cm ² (UGI day 1)	%EWL, 12 mo	smaller b
Nishie 2007[118]	Cohort, 82	pouch area, continuous (UGI day 1)	%EWL, to 24 mo	null
O’Connor 2008[105]	Cohort, 231	10–20 cm ³ range (staple fires)	%EWL, 12 mo	null
Topart 2011[119]	Cohort, 107	≤50 vs >50 mL (barium swallow)	%EWL, ~36 mo	null
Heneghan 2012[115]	Case-control, 380	pouch length/volume (endoscopy)	%EWL, ~6 y	mixed; st
Ren 2015[17]	RCT, 76	10–20 vs 25–35 mL (construction est.)	%EWL, 12 mo	smaller b
Edholm 2016[14]	Registry, 14,168	stapler length, continuous (proxy)	%EBMI, 12 mo	null at 1
Riccioppo 2018[48]	Cohort, 67	≤40 vs >40 mL (3D volumetry)	%EWL, ~47 mo	smaller/f
Boerboom 2020[13]	RCT, 132	10 vs 5 cm length (construction)	%TWL, 36 mo	<i>longer</i> be
Gao 2022[18]	Retrospective cohort, 95	small vs large (construction est.)	BMI change, 6 y	smaller b
Haghighat 2022[148]	Retrospective cohort, 100	2 vs 3 stapler loads	%TWL, 6 mo	null
Harker 2026[16]	RCT follow-up, 62	10 vs 6 cm length (construction)	%TWL, 5–9 y	null (und

The table already displays the anatomy of the conflict the pooled analysis must resolve or explain. The null results concentrate in the studies with the largest samples and the most objective exposures — the 14,168-patient registry, whose stapler-length proxy predicted excess-BMI loss at six weeks but not at one year [14], and the imaging cohorts of Nishie, O’Connor, and Topart [118, 105, 119]. The positive results come from a small randomized trial in a low-BMI diabetic population [17], a UGI cohort measuring day-one pouch area [103], a retrospective diabetic cohort [18], and a volumetric study whose effect ran through emptying rate rather than volume alone [48] — while the only other randomized evidence found the *longer* pouch superior at three years [13]. Whether these strands average to a real effect, to zero, or to an

uninterpretable mixture is precisely what the pooled estimate, the separately reported slope estimands (the registry and the continuous imaging cohorts never enter the arm-contrast pool), and the heterogeneity decomposition will state.

[RESULT PENDING: primary pool.] k studies met the inclusion and pooling criteria, contributing n patients. Pooled mean difference in %TWL at one year: [MD in percentage points, 95% CI, modified Knapp–Hartung], $\tau^2 = [\text{value}]$, $I^2 = [\text{value}]$, 95% prediction interval [range, computed only if $k \geq 5$]. Forest plot: Figure 10.1.

[RESULT PENDING: slope estimand.] Continuous-exposure studies, per study: exposure scale, reported slope with 95% CI, adjustment set; synthesized only if exposure scales align (Section 10.1).

[RESULT PENDING: sensitivity analyses.] (a) RCT-only pool; (b) leave-one-out influence analysis, reported in full if any study exceeds the 60% weight threshold (criterion 4); (c) fixed-effect comparison; (d) the labeled converted-inclusive pool — delta-method %EWL→%TWL conversions added to the native pool — and within-metric SMDs as a scale-free cross-check; (e) Paule–Mandel versus REML τ^2 , and unadjusted Wald versus modified Knapp–Hartung intervals.

10.5 Weight loss at three years and beyond

The secondary timepoints ask whether any one-year effect survives, attenuates, or reverses — and they are *fixed windows*, not “longest available follow-up.” Two windows are pre-specified: three years (reported timepoint nearest 36 months, within 30–42 months) and five years (nearest 60 months, within 54–66 months), pooled separately; observations beyond 66 months are reported study by study and never pooled, and where a study reports more than one eligible observation inside a window, the one nearest the window center is taken (spec:010). The distinction is not pedantry: a “longest follow-up ≥ 3 years” composite would pool outcomes measured at three, six, and nine years — materially different phases of the postoperative course — and make temporal change indistinguishable from between-study heterogeneity, which is exactly the confusion this chapter’s own mechanistic framing predicts matters. The Dutch program’s median 109-month follow-up is the working example: it falls beyond the five-year window and is reported study by study below, never pooled. The mechanistic stakes are set in Part I: if constructed size acts at all, the restrictive model expects its effect early, while the pouch is the binding constraint, with attenuation as intake adapts and anatomy dilates; the emptying-dynamics model permits a late-emerging effect on weight *regain* rather than on nadir loss. The extended-pouch trial is the canonical example of the latter pattern: no weight-loss difference for two years, then divergence in the third driven by regain in the standard-pouch arm, 31% versus 27% total weight loss at 36 months [13].

The long-term evidence is thin and attrition-damaged, and this section will say so with numbers rather than adjectives. The same Dutch trial illustrates the decay: of 134 randomized patients, 62 were assessed at median 109 months, and the 5–9-year difference (26.1% versus 24.1% TWL, favoring the extended pouch) was no longer statistically significant — by the authors’ own account an underpowered comparison from which no definitive long-term advantage can be concluded [16]. Section 7.5 establishes that bariatric follow-up losses of this magnitude are the rule and are non-random; every late-timepoint estimate below therefore carries its per-study follow-up fraction alongside the effect.

[RESULT PENDING: late-window pools.] Three-year window: k studies, per-study follow-up fractions [table], pooled mean difference in %TWL [value, 95% CI, modified Knapp–Hartung], heterogeneity [values]. Five-year window (54–66 months): same statistics, pooled separately. Observations beyond 66 months: per-study listing, never pooled. Comparison against the one-year estimate and against the regain-specific findings, if extractable.

10.6 Dose–response: gradient or threshold? (descriptive)

A pooled two-group contrast cannot distinguish “smaller is better” from “below some size, all pouches are equivalent.” The distinction matters clinically — it is the difference between a construction target and a construction ceiling — and mechanistically. If the pouch works as a container, effect should scale with capacity across its range; if it works as a signaling antechamber whose influence runs through emptying dynamics and hormone timing (the Part I steelman), volume per se may matter only above a threshold where geometry changes the emptying regime, and the volumetric dose–response should be flat where the two-group contrasts, which juxtapose extremes, still show a difference.

What the available data permit, however, is stated up front, because it is less than the question deserves. This analysis is *purely descriptive by protocol* (spec:010), for three reasons that no amount of care removes. First, it is ecological: study-level mean outcomes are set against study-level volume midpoints, so dose is entangled with everything else that varies between studies — population, baseline BMI, era, technique bundle, outcome metric — and a between-study pattern is not a within-study dose effect. A valid dose-response meta-analysis would need within-study dose contrasts with their covariance structure (including correlated comparisons from any multi-arm study); this literature does not supply them, which is why the protocol withdraws the two-stage generalized least-squares trend estimation an earlier draft promised. Second, the expected study count ($k \approx 5$) cannot carry inference at all: no pooled slope is fitted, no confidence interval is attached to any trend, and no spline or threshold model is entertained, because a curve estimated from five ecological points would be shape without information. Threshold behavior is therefore addressed only qualitatively, by whether the extreme-contrast studies and the narrow-range studies are mutually consistent. Third, the dose axis itself is noisy: a day-one contrast-study area and a CT volumetric measurement are not interchangeable estimates of the same milliliter (Section 6.9).

Within those limits, the display is restricted to the volumetric subset — studies reporting an estimated or imaged pouch volume with extractable outcomes — for however many such studies qualify, since a descriptive plot, unlike a fitted slope, needs no minimum k to avoid false precision. Volume, not length or stapler count, is the dose axis, because it is the only quantity on which the studies’ exposures can be placed on one scale. The response axis is %TWL at one year, harmonized per Section 7.6; studies whose outcomes cannot be defensibly harmonized are listed but not plotted, because raw %TWL and standardized differences cannot share an axis. The candidate volumetric anchors span roughly 10 to 70 mL across the studies profiled above [17, 48, 119, 103] — about a factor of seven in dose, which is more leverage than the field’s rhetoric of a uniform “small pouch” suggests. Whatever pattern emerges is reported as hypothesis-generating description, not as a test of gradient or threshold.

FIGURE PLACEHOLDER: Descriptive dose–response display, volumetric subset. X-axis: study-level pouch volume (mL, midpoint of assigned range or reported mean, modality-tagged); y-axis: harmonized %TWL at 1 year only (non-harmonizable studies listed, not plotted); point size by study precision; no fitted slope, no confidence band, no spline or threshold fit — labeled descriptive/ecological per spec:010. Data-rendered from the pipeline extraction table per spec:011; provenance in Appendix E.

Figure 10.2: Descriptive (ecological) dose–response display of constructed pouch volume against weight loss, volumetric subset. [RESULT PENDING]

[**RESULT PENDING: dose–response, descriptive.**] Volumetric subset size (k); the cross-study display of Figure 10.2, with no pooled slope and no inferential interval; qualitative consistency of extreme-contrast versus narrow-range studies, reported as ecological, hypothesis-generating description; studies excluded from the display listed with the reasons each failed

extraction or harmonization.

10.7 Heterogeneity, quantified and dissected

Whatever the pooled means show, the heterogeneity statistics are a result in their own right, because heterogeneity is the field’s stated reason for never having pooled [15]. Reporting τ^2 , I^2 , and the prediction interval turns that assertion into a measurement: how much of the forty-year disagreement is sampling noise around a common effect, and how much is real structural variation between studies? The decomposition is deferred to its own chapters where it carries the dissertation’s argument: the modality stratification — does the effect live only in weak proxies? — is Chapter 11’s question, and design, era, and RYGB/OAGB subgroups follow there and in Section 13.2. Small-study effects are assessed by funnel plot and Egger’s regression test [194] only if at least ten studies pool, and — per protocol — reported as *suggestive* evidence: near $k = 10$, funnel asymmetry cannot distinguish selective publication from genuine small-study heterogeneity, and Egger’s test is never treated as an adjudicator of cause (spec:010). The funnel is displayed alongside the risk-of-bias material in Chapter 13. One asymmetry deserves flagging in advance — with an equally advance statement of what it cannot show. In Mahawar’s tally, five studies favored smaller pouches, nine found nothing, and none favored larger [15]. But that is a tally of *significance categories*, not of signs: the nine “null” studies have directions that the tally discards, so the sign distribution of the literature is currently unknown, and a symmetric-sign expectation is not established under heterogeneous effects, unequal standard errors, and doctrine-driven hypotheses. The 5–9–0 split is consistent with a modest true effect diluted by underpowered nulls, with directional publication of results that agree with surgical doctrine, or with neither. This analysis therefore extracts and displays the *signed* effect estimate of every included study — converting the tally into a distribution — and lets the signed distribution, not the significance categories, carry whatever inference is available; the funnel geometry is read alongside it as suggestive context only.

[RESULT PENDING: heterogeneity.] τ^2 , I^2 with 95% CI, prediction interval (computed only if $k \geq 5$), Egger intercept if $k \geq 10$ (suggestive only); signed-estimate distribution across all included studies; proportion of τ^2 absorbed by the pre-specified subgroup structure (reported here, dissected in Chapter 11).

10.8 If pooling fails: synthesis without meta-analysis

If any veto criterion of Section 10.3 fails, the chapter does not retreat to an impressionistic narrative. The fallback is a formal Synthesis Without Meta-analysis conducted and reported to the SWiM guideline [183]: the synthesis question and grouping structure stated up front (the same modality and design strata as the meta-analysis), and a defined synthesis display that preserves what vote counting destroys — each study’s *effect estimate with its confidence interval*, on its native metric, grouped by strata, so that magnitude and uncertainty remain visible even when no pooled number is defensible. Direction is summarized across studies, but no conclusion rests on counting significant results: a tally of P values recreates the significance-based vote counting that synthesis guidance exists to prevent, and formatting it as SWiM would not redeem it. The display replaces the forest plot, and the report adds an explicit statement of which studies drive each conclusion and why pooling was indefensible — criterion by criterion, study by study. The distinction between this and what the field has had is the difference between a measurement that fails with a documented error budget and one that was never attempted: Mahawar’s review established *that* the studies resist combination [15]; the SWiM fallback would establish *which* incompatibilities are responsible, and therefore what a study would have to report for pooling to become possible — a specification that feeds directly into the future-directions program of Chapter 15.

10.9 What carries forward

This chapter produces the pooled small-versus-large estimates at one year and in the fixed late windows, the slope estimands of the continuous-exposure studies, the exploratory dose–response description of the volumetric subset, and the heterogeneity budget — or the documented demonstration that these quantities cannot yet be defensibly computed. It deliberately stops short of interpretation. The estimate’s meaning depends on whether it survives stratification by measurement quality, which is Chapter 11’s question; its clinical import depends on the bidirectional safety analysis of Section 13.1; and its verdict on mechanism is delivered against the predictions derived in Section 5.3 only in the discussion. What Part II’s remaining chapters inherit from this one is a number with an honest uncertainty — or an honest account of why the field’s oldest technical question still lacks one.

Chapter 11

The measurement-artifact test and the evidence conflict

This chapter carries the dissertation’s pivot: research questions Q3 — does the answer depend on how pouch size was measured (imaged volume vs. length vs. stapler/bougie proxies)? — and Q4 — why do the randomized trials disagree with the largest registry, and with each other? It is written in two registers, and the distinction matters. The *designs* — the modality-stratified estimator and the design-level dissection protocol — are fixed here, before any pooled number exists, so that the reading rules cannot be adjusted to fit the results. The *facts* about the conflicting studies, by contrast, can already be written in full, because they depend only on the published characteristics of those studies, which have been extracted from the corpus; where a pooled or stratified estimate belongs, the text says [RESULT PENDING] and nothing else. The chapter therefore reads as a results chapter with its quantitative core deliberately left open — which is the honest shape of a pre-specified analysis.

11.1 The modality-stratification design

Section 6.9 established the modality-quality hierarchy this chapter consumes, and its tier names are used here verbatim: tier V1, *measured volume* (per-patient volumetric imaging — 3D-CT or MRI — within the construction window); tier V2, *imaged geometry* (planar contrast radiography, radiologic volume estimates from assumed geometry, or instrument-referenced endoscopic measurement); tier V3, *measured construction* (per-patient intraoperative measurement of the construction itself — total stapler length, measured pouch length, intraoperative volume determination); and tier V4, *declared technique* (a group-level construction intention with no per-patient measurement — cartridge count as a categorical technique, bougie caliber alone, or a stated target volume or length for a study arm). The hierarchy was defined on measurement-science grounds — validation evidence, operator dependence, and the distance between what is recorded and the physical quantity of interest — before any outcome data were examined, and it is not revisited here.

The analysis it feeds is a subgroup decomposition of the primary random-effects model of Chapter 10. Every included contrast is assigned one tier, by the modality used to *define the exposure* (not the modality used anywhere in the paper); the pooled small-versus-large effect is then re-estimated within each tier that holds at least three studies — below that floor a tier’s studies are presented descriptively, with no stratified estimate claimed — and a single pre-stated meta-regression of effect size on tier and design, with no model selection and no covariates added or dropped after the data are seen, asks whether measurement quality predicts the answer. Its deliverable is the between-tier interaction estimate with its confidence interval; there is no significance trigger, and no nominal *p*-value promotes the stratified estimates over the overall presentation. Assignment happens at extraction time, blind to each study’s direction

of effect, and is audited by the span-verification pipeline: the tier call must be supported by a verbatim quote from the study’s methods section.

What this design can deliver must be stated as plainly as what it tests. The analysis is an *exploratory effect-modification* analysis, per protocol (spec:010), because tier is not randomly assigned to studies: it is confounded with study era (imaging studies cluster early, registry proxies late), with design (the trials sit in V4, the largest cohort in V3), with sample size, and with time of outcome assessment. A between-tier difference is therefore evidence that *something* co-varying with measurement modality predicts the answer — it cannot, by itself, adjudicate that measurement quality is that something. Design is the one competing covariate the pre-stated model carries; era and study size, which the expected study count cannot support as additional model terms, are argued explicitly in the interpretation rather than silently absorbed into a verdict — and no covariate is added or removed once the data are seen. One further caveat bounds even the best tier. The imaging tiers measure *achieved* early anatomy, not the surgeon’s construction act: by the time any postoperative image is acquired — even within the 90-day construction window of Section 6.9 — remodeling, emptying adaptation, and early weight loss have begun to operate, so an imaged volume is a post-treatment descendant of the construction, and the causal exposure throughout this dissertation remains the constructed size, with early imaging read as its best available measurement rather than its replacement. The protocol therefore splits the window itself: a V1 or V2 measurement acquired within 30 days of surgery is classified *construction-proximal*, while one acquired at 31–90 days is carried under a separate *achieved-anatomy* flag — the two are never silently merged, and every analysis touching the later flag discusses it as what it is, a post-treatment descendant of the exposure rather than a measurement of the construction (spec:010). A day-one contrast study and a ten-week volumetric study are both “within the window,” but they are not measurements of the same thing, and the flag keeps that distinction from disappearing into a tier label.

11.1.1 What measurement error can and cannot manufacture

The artifact hypothesis is often stated loosely — “the effect is only there because the measurements are bad” — and the loose form is statistically incoherent. Classical, nondifferential measurement error in an exposure does not create associations; it destroys them. If a stapler-length proxy is simply a noisy channel carrying the true pouch volume, the noise attenuates any true regression slope toward zero — regression dilution — exactly as an impedance-mismatched transducer loses signal power without inventing signal. Under pure noise, the *weakest* tiers should show the *weakest* association, and volumetry the strongest. So the naive artifact story has the physics backwards, and the stratified analysis must be read against explicitly stated signatures rather than a vague intuition. Four patterns are pre-committed here, each with its leading reading *and* the rival explanations that the design cannot exclude — because tier is confounded with era, design, sample size, and assessment timing (Section 11.1), no pattern licenses a verdict by itself.

Attenuation pattern. Effect size increases monotonically from tier V4 to tier V1. Leading reading: a real effect exists and proxies dilute it; the field’s null results were produced by measuring the exposure badly. Rival readings: the higher tiers may also differ in population, exposure range, or outcome timing in ways that independently enlarge the effect.

Inversion pattern. An association is present in the construction tiers V3–V4 and absent in the imaging tiers V1–V2. Classical nondifferential noise alone cannot produce this, so it demands a non-noise explanation — but several compete. The measurement-artifact reading is that the proxy is correlated with something other than pouch volume that itself affects outcome — surgeon, era, institutional technique bundle, patient selection — so the proxy association is carried by a proxy-correlated confounder rather than by the anatomy. (Two distinct mechanisms are deliberately kept apart here: *differential* measurement error — error whose magnitude or direction depends on the outcome or its causes — and *confounding of the proxy* are different

FIGURE PLACEHOLDER (F7): Modality-subgroup forest plot. One panel per measurement tier (V1 measured volume, V2 imaged geometry, V3 measured construction, V4 declared technique), each showing the per-study effects (mean difference in %TWL where harmonizable), a within-tier random-effects pooled estimate only for tiers holding at least three studies (descriptive rows otherwise), V1/V2 studies flagged construction-proximal (≤ 30 days) versus achieved-anatomy (31–90 days), and the between-tier interaction estimate with its confidence interval annotated and labeled as an exploratory effect-modification estimate. Data-rendered from the pipeline extraction table per spec:011 — never hand-drawn. [RESULT PENDING]

Figure 11.1: Effect of constructed pouch size on weight loss, stratified by the measurement-modality hierarchy of Section 6.9.

pathways, though either can leave this footprint.) The rival readings are mundane: the V1–V2 studies are fewer, smaller, older, and differently timed than the V3–V4 studies, so an inversion could reflect differences in power, population, exposure range, or design rather than measurement per se. The pattern is the one the measurement-artifact hypothesis predicts; observing it would make that hypothesis a live leading explanation of the forty-year disagreement, to be weighed against the rivals — not a demonstrated artifact.

Uniform null and uniform effect. No tier shows an association, or all do with comparable magnitude. Either reading is also informative: the first weighs against the doctrine across every measurement quality the literature has produced; the second says the proxies, crude as they are, track enough of the true exposure to preserve the signal.

Two complications are handled in the design rather than discovered later. First, the tiers are confounded with time of assessment: imaged volumetry is usually performed months to years after surgery, when dilation has already decoupled measured volume from constructed volume (Section 6.7). The protocol therefore restricts the exposure to construction size, and any study whose “pouch size” is a late measurement is either re-classified (if a construction-time record also exists) or excluded from the exposure set, per the criteria of Section 8.2. An instructive boundary case from the corpus: Topart and colleagues measured pouch volume by barium swallow a mean of three years after surgery in 107 patients and found no correlation with percent excess weight loss — a finding about the dilated pouch, not the constructed one, and classified accordingly [119]. Second, the tiers are unevenly populated: the literature is dominated by tiers V3 and V4, and a stratified estimate from a two-study tier settles nothing — which is why the three-study floor of Section 11.1 demotes any thinner tier to descriptive presentation. The defensibility criteria of Section 8.9 apply within tiers exactly as they do to the overall pool.

The corpus already illustrates how the tiers will populate. At tier V2, Roberts and colleagues measured pouch area on day-one contrast radiographs in 320 patients and found a significant negative correlation with excess weight loss at 6 and 12 months [103]. At tier V3, O’Connor and Carlin recorded staple fires against a calibration tube in 231 patients, all with pouches of 10–20 mL, and found no association with one-year excess weight loss — a result as consistent with a true null as with restriction of range in a deliberately uniform exposure [105]. Heneghan and colleagues graded pouch and stoma dimensions endoscopically in 380 symptomatic or weight-regaining patients — a V2-class instrument applied years after surgery, outside the construction window — and found pouch length, pouch volume, and stoma diameter all inversely correlated with excess weight loss, with stoma diameter the only independent predictor in multivariable analysis [115]. And Riccioppo and colleagues combined three-dimensional volumetric reconstruction — V1-class imaging, though performed at a median of 47 months — with emptying scintigraphy in 67 patients (median volume 28 mL) and associated small, *fast-emptying* pouches with better late weight loss [48]. These four studies already span the full range of reported directions — which is precisely why the stratified estimate, not the study list, is the deliverable.

The stratified estimates themselves: [RESULT PENDING].

11.2 A framework for dissecting evidence conflicts

Two studies “conflict” when they are treated as estimates of the same quantity and the *difference* between their effects is larger than sampling error plausibly allows — a judgment made, wherever the data permit, by estimating that difference directly with its own confidence interval, or by a subgroup interaction or meta-regression test, never by eyeballing whether two separate 95% intervals overlap: interval overlap is not a valid test of between-study difference (two intervals can overlap while the difference is significant, and vice versa), and a framework built on it would misclassify conflicts in both directions. The dissection framework also refuses the shared-quantity premise until it has been checked: before declaring a contradiction, establish that the studies asked the same question of the same population about the same exposure measured comparably, on the same outcome scale, over the same horizon, in samples that survived follow-up the same way. Each axis on which they differ is a candidate explanation, and the verdict for a given conflict is one of three: *explained* (a named axis, by itself, plausibly reproduces the disagreement), *reduced* (axes account for part of it), or *irreducible* (the studies are commensurable on every axis and still disagree — a genuine inconsistency that heterogeneity statistics must carry). The axes, fixed in advance:

Population — baseline BMI, diabetes status, region; a trial optimizing glycemic control in a BMI < 35 cohort is not estimating the same estimand as a registry of standard bariatric candidates. *Exposure definition and contrast distance* — what was actually varied (volume, length, caliber, fundus inclusion) and how far apart the arms sit on that dimension; a continuous proxy observed within one standardized technique has whatever contrast the technique’s residual variance happens to supply, which may be a fraction of an engineered two-arm contrast. *Measurement modality* — the tier of Section 11.1. *Co-intervention bundle* — everything else that differed with the exposure: limb lengths, anastomotic technique, era, surgeon experience. *Outcome metric* — Chapter 7 showed that %EWL, %EBMIL (excess-BMI loss referenced to BMI 25), and %TWL rank the same patients differently; a conflict stated across metrics is not yet a conflict. *Time horizon* — an exposure can have a real, transient effect; a six-week association and a one-year null are jointly consistent with a restriction effect that decays as neuroendocrine adaptation dominates (Section 5.3), though attenuation is only *established* by directly comparing the timepoint-specific effects (an effect-by-time interaction), never by significance at one timepoint and not the other. *Attrition* — differential or heavy loss to follow-up converts an estimate into a statement about the retained subsample. *Design and analysis* — randomized allocation versus adjustment, and whether the exposure groups are separated in calendar time.

The remainder of the chapter applies this instrument to the field’s two flagship contradictions.

11.3 The registry versus the trials

The largest body of evidence on either side of the question comes from a single study. Edholm, Ottosson, and Sundbom analyzed 14,168 primary laparoscopic RYGB patients in the Scandinavian Obesity Surgery Registry (SOReg), all operated with linear-stapled gastrojejunostomy under Sweden’s highly standardized national technique, using the total length of stapler fired to construct the pouch as the exposure — tier V3, a per-patient measurement of the surgeon’s input rather than the anatomical result, with a mean of 145 mm [14]. Stapler length predicted percent excess-BMI loss at six weeks but not at one year, at which point BMI had fallen from a mean of 42.4 to 28.9 kg/m² and the significant predictors were female sex, low preoperative BMI, young age, and absence of diabetes. The same exposure predicted harm monotonically:

Table 11.1: The conflicting studies: extracted design characteristics. Tier refers to the measurement-modality hierarchy of Section 6.9. MU = marginal ulcer; f/u = follow-up.

	Edholm 2016 [14]	Ren 2015 [17]	Gao 2022 [18]	Boerboom 2020 / Harker 2026 [13, 16]
Design	registry cohort (SOREg)	RCT	retrospective cohort	RCT (single-blind) + long-term f/u
N	14,168	76	95 (53 small / 42 large)	132 randomized; 62 at long-term f/u
Population	Swedish primary LRYGB, mean BMI 42.4	Chinese T2DM, BMI 30–35	Chinese T2DM, BMI < 35	Dutch bariatric, mean BMI 42.9 (f/u subset)
Exposure contrast	continuous stapler length (mean 145 mm)	pouch 10–20 vs. 25–35 mL	10–20 mL lesser-curve pouch vs. >50 mL fundus-containing subtotal pouch	pouch length 10 vs. 5 cm at 40 Fr caliber
Modality tier	V3 (measured stapler length)	V4 (declared target volume)	V4 (declared technique, two eras)	V4 (declared length target)
Primary metric	%EBMIL	%EWL, weight	Δ BMI, %EWL, T2D remission	%TWL, regain
Horizon	1 year	12 months	6 years	3 years; 5–9 years
F/u completeness	complete-data inclusion criterion	91% at 12 mo	89.5% at 6 y	47% at median 109 mo
Direction	significant at 6 wk; NS at 1 y; MU risk +14%/cm	smaller better ($p = 0.033$)	smaller better; MU 23.7% vs. 6.4% in large	<i>longer</i> better at 3 y ($p = 0.023$); NS trend at 9 y

each additional centimeter of stapler raised the relative risk of marginal ulcer — an ulcer at the gastrojejunal anastomosis — by 14% (95% CI 9–20%) [14].

Set against this registry null stand the randomized programs summarized in Table 11.1. Mahawar and colleagues’ systematic review — the only one in this literature — found fourteen studies of pouch size, of which two were randomized, judged the body of evidence too heterogeneous to pool, and could conclude only that a larger pouch “may be associated with adverse weight loss outcomes,” with no study finding a larger pouch better [15]. Since that review closed, both randomized programs have extended their follow-up, and one of them now reports in the direction Mahawar’s review said no study had found.

Run the axes across the registry–trial comparison and the conflict thins considerably before any pooling is attempted. *Contrast distance*: the registry’s exposure is the residual variation left over after national standardization — variation the Swedish technique exists to suppress. The abstract reports the mean stapler length but not its dispersion; extracting that distribution from the full text is a pipeline task, because without it the registry’s effective contrast cannot be compared to the trials’ engineered ones. If the standardization worked, the registry estimated the slope of outcome on pouch size over a narrow exposure range — and restriction of range *reduces* the information each patient contributes to a slope. Whether the registry’s 14,168 patients compensate cannot be asserted from the mean stapler length alone; the detectable slope depends on the exposure variance, which is exactly the dispersion not yet extracted, so no claim about the registry’s power to detect or exclude an effect is made until it is. What the narrow range does establish is the design’s reach: whatever slope the registry can or cannot see, it says almost nothing about what happens two tiers of pouch size away. The Chinese contrasts (10–20 mL versus a fundus-containing reservoir several times larger) and the Dutch contrast (a doubling of pouch length) live outside that range. *Population*: both Chinese studies enrolled diabetic patients at BMI below 35 — a cohort the Swedish registry mean of 42.4 barely

overlaps — with glycemic remission, not weight, as the clinical driver. *Metric*: %EBMIL, %EWL against a Chinese ideal-BMI reference of 24, Δ BMI, and %TWL appear across the four programs; the harmonization rules of Section 7.6 govern the conversion, and no two of these studies published on the same scale. *Time horizon*: the registry reports a statistically significant stapler-length association at six weeks and a non-significant one at one year. Read with the framework’s own discipline, that is weaker than it looks: the two coefficients were not directly compared, so attenuation is not established by the significance flip alone; and a six-week association in an observational registry could also reflect residual confounding or the multiplicity of predictors examined, or be too small to matter clinically even if real. The pattern is *consistent with* — not proof of — the signature Section 5.3 assigns to a restrictive mechanism giving way to a neuroendocrine one, and extracting the timepoint-specific effect estimates for a direct comparison is a stated pipeline task. The registry and the trials are therefore not estimating one quantity with discordant answers; they are sampling different segments of a dose-response curve, in different populations, on different scales, at different times. Whether the segments are even mutually consistent is what the exploratory dose-response analysis of Chapter 10 will describe — within its stated ecological limits, and without claiming to test the joint curve: [RESULT PENDING].

One axis, meanwhile, shows no conflict at all. The registry’s marginal-ulcer gradient (relative risk +14% per centimeter of stapler) and the Chinese cohort’s ulcer excess in the large-pouch group (23.7% versus 6.4% at six years, $p = 0.023$) point the same way: more stapled pouch, more ulcers [14, 18]. The safety evidence agrees while the efficacy evidence disputes — a pattern Section 13.1.4 takes up in the benefit–harm analysis, alongside the harms plausibly attached to the other end of the construction range.

11.4 Trial versus trial

The two randomized programs are routinely cited as contradicting each other: the Chinese randomized trial of Ren and colleagues found the *smaller* pouch superior, the Dutch extended-pouch trial found the *longer* pouch superior. Stated at that altitude the contradiction is real. Dissected, it largely dissolves, because the two programs did not vary the same physical quantity.

The Chinese program’s contrast is volumetric, and anatomically radical. In the randomized study of Ren and colleagues, 76 diabetic patients with BMI 30–35 received pouches of 10–20 mL versus 25–35 mL; at twelve months the smaller-pouch group had lost more weight (15.8 versus 13.7 kg) with higher %EWL (68.9% versus 59.1%, $p = 0.033$) and better glycemic control [17]. In the same clinical territory, the Central South University cohort reported by Gao and colleagues compared a 10–20 mL lesser-curve pouch against its center’s earlier technique: a subtotal gastric reservoir extending from the greater curvature to the angular incisure, over 50 mL, anastomosed end-to-side to the posterior gastric wall — a construct that retains the fundus, the stomach’s ghrelin-secreting territory, which the standard small pouch excludes by design (Section 4.2). At six years, with 89.5% follow-up, the small-pouch group showed greater BMI reduction (−4.25 versus −2.06 kg/m²), higher diabetes remission, and one quarter the marginal-ulcer incidence [18]. Two design facts about this study matter more than its direction. It is a retrospective cohort, not a randomized trial — the study itself says so plainly, though secondary citations (including this project’s own preliminary scan notes) have mislabeled it. And its exposure groups are separated in calendar time: the large pouch was the center’s early technique from 2008 and the small pouch its later one, so the pouch contrast is inseparable from a decade of learning curve. On the dissection axes, Gao versus anything is *explained* before it is contradictory: era-confounded design, and an exposure contrast (fundus-containing reservoir versus fundus-excluding pouch) that no other study in the field shares.

The Dutch program’s contrast is geometric, not volumetric. Boerboom and colleagues randomized 132 patients to a standard pouch — first stapler fired 5 cm below the angle of His,

completed with two further 60 mm loads against a 40 French calibration tube — versus an extended pouch begun 10 cm below the angle of His and completed with three loads against the same tube [13, 16]. Both arms are narrow, lesser-curve-based, fundus-excluding pouches; what was doubled is length at fixed caliber. For the first two years the arms were indistinguishable; in year three the standard-pouch group regained roughly 3 kg while the extended group held, yielding three-year total weight loss of 31% versus 27% ($p = 0.023$) [13]. At five to nine years (median 109 months), with 47% of the cohort retrievable, the difference was no longer significant — 26.1% versus 24.1% total weight loss — with a trend toward less recurrent weight gain after the extended pouch (47% versus 70%, $p = 0.07$); the authors state the follow-up is underpowered [16].

Now put the two programs on the same axes. *Exposure*: the Chinese studies varied volume, with the large arm crossing an anatomical boundary (fundus inclusion); the Dutch trial varied length at constant caliber, both arms staying inside the standard anatomical envelope. These contrasts are nearly orthogonal, and both directions of result are jointly satisfiable by a single sentence: narrow, fundus-excluding pouches of modest volume outperform large fundus-containing reservoirs, and added *length* of a narrow pouch is at worst neutral and possibly protective against late regain. On this reading there is no RCT-versus-RCT contradiction to explain — there is a two-dimensional exposure (volume, geometry) that the literature keeps projecting onto a single axis called “size.” The pooled analysis operationalizes this by coding volume-contrasts and length-contrasts as distinct exposure families before any combination is attempted (Section 10.1); whether the reconciling sentence survives quantification: [RESULT PENDING].

Mechanism, though, does not reconcile so easily. The Dutch program’s stated rationale is that a longer, narrower pouch empties *more slowly* (Poiseuille), dilates less (Laplace), and sustains satiety [16]. Riccioppo’s volumetric data associate small pouches with *faster* emptying and faster emptying with better late weight loss [48] — the emptying-dynamics steelman of Section 5.3 run in the opposite direction. Both camps invoke emptying speed as the mediator and disagree about its sign. This is a genuinely open mechanistic question, flagged here and adjudicated against the pooled results in Chapter 14.

Attrition, finally, is not a footnote to the Dutch program; it is demonstrably capable of moving the answer within a single trial. At 36 months the full randomized cohort showed a significant 4-point total-weight-loss advantage for the extended pouch [13]; the long-term report, restricted to the 62 patients (47%) reachable at nine years, recomputes the same 36-month timepoint in the retained subset as 30.6% versus 27.3% — nonsignificant [16]. Same trial, same timepoint, different answer, and the only thing that changed is who remained. This is the survivorship logic of Section 7.5 caught in the act, and it calibrates how much weight any long-horizon pouch-size claim in this literature can bear. The program’s reporting also illustrates the exposure-imprecision problem at its most basic: across the primary and follow-up reports the standard pouch is described as 5 cm and as 6 cm, and the randomized denominator as 132 and as 134 [13, 16]. If a flagship randomized trial cannot hold its own exposure definition fixed to the centimeter across two of its own publications, the field’s forty years of construction proxies should be read with corresponding humility.

11.5 Adjudication, pre-committed

The dissection above leaves the field’s conflicts in three different states before pooling. The registry-versus-trial conflict is provisionally *reduced*: contrast distance, population, metric, and time horizon each carry part of the disagreement, and the exploratory dose-response analysis will describe, within its ecological limits, whether the segments are mutually consistent. The Gao-versus-anything conflict is *explained* at the design level: era confounding and an incommensurable exposure. The Dutch-versus-Chinese RCT conflict is provisionally *explained* by

exposure orthogonality — subject to the quantitative check — while its mechanistic sub-conflict over emptying speed remains *irreducible* on current evidence. Q3’s own answer waits on the interaction estimate and stratified displays of Section 11.1: whether the pouch-size effect survives measurement quality, attenuates with it, or inverts against it — read, per Section 11.1.1, as competing explanations weighed against the tier’s confounders, not as a predetermined artifact ruling. [RESULT PENDING]. What is already fixed is how each pattern will be read — and that no result can now change the rules by which it is judged. The oldest technical question in bariatric surgery has been argued for forty years almost entirely in the construction tiers; every conflict dissected in this chapter lives within tiers V3 and V4, and not one of the flagship studies measured the pouch it actually built. Whatever the stratified estimates return, two boundaries still have to be drawn around them — the safety trade-off that runs in both directions (the ulcer gradient taxing larger construction, and the harms plausibly attached to very small construction), and the OAGB anatomy that stretches the exposure range — and drawing them is the work of Chapter 13.

Chapter 12

The detectability floor: a measurement-fidelity simulation

Chapter 11 asked whether the pouch-size answer depends on how the pouch was measured, and had to leave its quantitative core open until the synthesis runs. This chapter asks a question that does not have to wait, because it does not use the outcome data at all: given the documented error of the field’s measurement tiers and the sizes of the studies actually conducted, what true effect could this literature ever have detected? The answer is a number — the detectability floor — and computing it is the second clause of the dissertation’s aim: not only whether constructed pouch size predicts weight loss, but what minimum measurement fidelity the field would need for that question to be answerable at all.

12.1 Motivation: a question the existing data can already answer

The simulation’s inputs are two things the preceding chapters have already assembled: error models for the measurement tiers of Section 6.9, calibrated to the validation literature laid out in Chapter 6, and the realized designs — per-tier study counts, sample sizes, arm contrasts, and outcome standard deviations — of the screened corpus. It uses no observed effect estimate from any included study. That independence has two consequences. First, the analysis is immune to the pre-specification concern that Chapter 6 attached to the modality hierarchy: nothing in a power computation can be tuned toward a preferred result, because the results it is computed from are never consulted. Second, it makes the synthesis outcome symmetric. If the pooled analysis of Chapter 10 finds an effect, the floor says whether the finding is surprising; if pooling fails or returns null, the floor says whether forty years of null and conflicting results are evidence about physiology or merely evidence about instruments. A literature whose detectability floor sits above every physiologically plausible effect size was never going to resolve its question — the stalemate becomes a mathematical inevitability, not a physiological mystery — and demonstrating that is a conclusion, not a failure of synthesis.

The positioning is deliberate: the simulation answers “why did forty years fail,” and its final deliverable — the fidelity requirement of Section 12.3 — answers “what would it take.” Those are the two halves of the extended aim, and Chapter 15 builds its registry recommendation on the second.

12.2 Error models

Each tier of the Chapter 6 hierarchy receives an explicit error model with three separately stated components — classical (random, nondifferential, per-patient), systematic (between-instrument

or between-protocol shift), and differential (error correlated with the outcome or with what predicts it) — so that no generic appeal to regression dilution is allowed to stand in for a mechanism. The calibration sources are the validation record of Table 6.1, and their thinness is part of the model, not a footnote to it.

Tier V1 — measured volume. The best-calibrated tier and still thinly calibrated: the agreement of the field’s two near-reference instruments rests on twelve patients [19], so the V1 error variance enters as a wide band, not a point value. Distension state contributes a protocol-dependent component — the same pouches read fourfold differently within one examination [19] — and the timing rule of Section 6.9 contributes a remodeling term even inside the construction window [157].

Tier V2 — imaged geometry. Calibrated from the planar record: the projection-to-volume inference, the documented low interobserver agreement of UGI pouch assessment [150], and the modality’s weak discrimination in case-control use — dilation in 23% of failures versus 11% of controls [150], “large pouch” more prevalent in well-losing controls than in regainers [151]. The 1984 gastropasty series, with its fivefold spread of measured volumes after a nominally uniform operation [78], anchors the joint scatter of construction plus planar reading.

Tier V3 — measured construction. A real per-patient measurement of an input dimension [14, 105]: its classical component is small, and its dominant error is structural — the loose length-to-volume mapping, entering as a proxy-validity term rather than instrument noise.

Tier V4 — declared technique. No per-patient information at all [17, 148]: the model is construction scatter around a declared target, calibrated from the calibration-tube comparison [147] and the radiologic spread above [78], so that arm overlap in delivered anatomy dilutes the nominal contrast.

Timing error is modeled for all tiers from the dilation trajectories of Section 6.7: Flanagan’s stepwise functional growth [159] and the published remodeling course rising from roughly 120 mL early after surgery toward 524 mL at five years [160]. Every parameter carries its provenance — source, value, uncertainty band — in a versioned parameter file, and every thinly calibrated input (the $n = 12$ agreement study above all) is swept over its band in sensitivity analyses rather than fixed at a best guess. Where a source supplies only an agreement statistic — the intraclass correlation of 0.594 between gastroscopy and CT [145] — the model does not launder it into a reliability ratio; it brackets the classical component consistent with it and sweeps the bracket.

12.3 Design

The data-generating model is patient-level. A true constructed pouch volume is drawn from a distribution matched to the realized designs; the outcome, %TWL at the study’s horizon, is generated under a hypothesized true effect drawn from a grid running from zero to a generous upper bound — bounded by the largest effects claimed anywhere in this literature, so that no plausible effect lies outside the grid. The observed exposure is the true value degraded by the tier’s error model, classical and systematic components applied separately; differential-error scenarios run as sensitivity analyses, not as the base case. Each simulated study then analyzes its observed data exactly as the real studies did — arm contrasts at the realized k and N , with the corpus-extracted outcome standard deviations — and the fraction of simulated literatures that detect the effect is that tier’s power.

Three deliverables are fixed. First, the *detectability surfaces*: per tier, power as a function of true effect size at the realized designs, with the minimum true effect detectable at 80% power read off each surface — the detectability floor itself, rendered as the dissertation’s flagship

Table 12.1: Error-model parameters by tier: classical, systematic, and differential components with calibration source (PMID) and uncertainty band per parameter, copied from the versioned parameter file (`analysis/simulation/parameters.yaml`), which carries the full provenance note for every row. ^a explicit assumption — no source quantifies the value; ^t thin calibration, swept over its band in the sensitivity analyses. Volumes are modeled on the log scale, so CV-type and log-SD values are approximately error SD as a fraction of true volume.

Tier	Component	Value	Band	Source (PMID)
Population	volume median (mL)	30.0	20.0–45.0	7091134
Population	volume log-SD ^t	0.41	0.25–0.60	6474086
Outcome	mean %TWL	28.0	25.0–33.0	36536734
Outcome	SD of %TWL ^t	7.0	5.5–8.5	36536734
V1	classical CV ^t	0.25	0.10–0.50	32564309
V1	systematic shift (log-SD) ^{a,t}	0.15	0.05–0.40	32564309
V1	differential correlation ^a	0.0	0.0–0.3	—
V2	classical CV ^t	0.40	0.25–0.60	32917519
V2	systematic shift (log-SD) ^{a,t}	0.20	0.05–0.40	31203729
V2	differential correlation ^a	0.0	0.0–0.3	—
V3	proxy correlation $r^{a,t}$	0.60	0.40–0.80	26194258
V3	classical CV ^a	0.05	0.02–0.12	18294920
V3	systematic shift (log-SD) ^a	0.05	0.0–0.15	26194258
V3	differential correlation ^a	0.0	0.0–0.3	—
V4	target attainment ^{a,t}	0.70	0.40–1.00	—
V4	construction scatter (log-SD) ^t	0.40	0.25–0.60	6474086
V4	systematic shift (log-SD) ^a	0.10	0.0–0.30	40488443
V4	differential correlation ^a	0.0	0.0–0.3	—

figure set. Second, the *fidelity requirement*, the design question inverted: for a given true effect and sample size, the measurement precision — error standard deviation as a fraction of true volume — a future study or registry would need, which is the concrete number the registry recommendation of Chapter 15 cites. Third, the *sensitivity sweeps* over every thin calibration input, reported alongside the primary surfaces rather than in an appendix, because the width of the floor’s uncertainty is part of the finding. The simulation is seeded and configuration-driven in `analysis/simulation/`; every figure in this chapter regenerates from one command.

12.4 Results

Every number in this section is **provisional**: the simulation ran on the pre-extraction design grid (`analysis/simulation/designs.yaml`, provisional flag), which encodes the realized designs from the known literature rather than from the screened corpus. When the spec:010 extraction lands, the grid is replaced wholesale and this section regenerates; the error models and the shape of the result do not depend on the swap, but the digits below do.

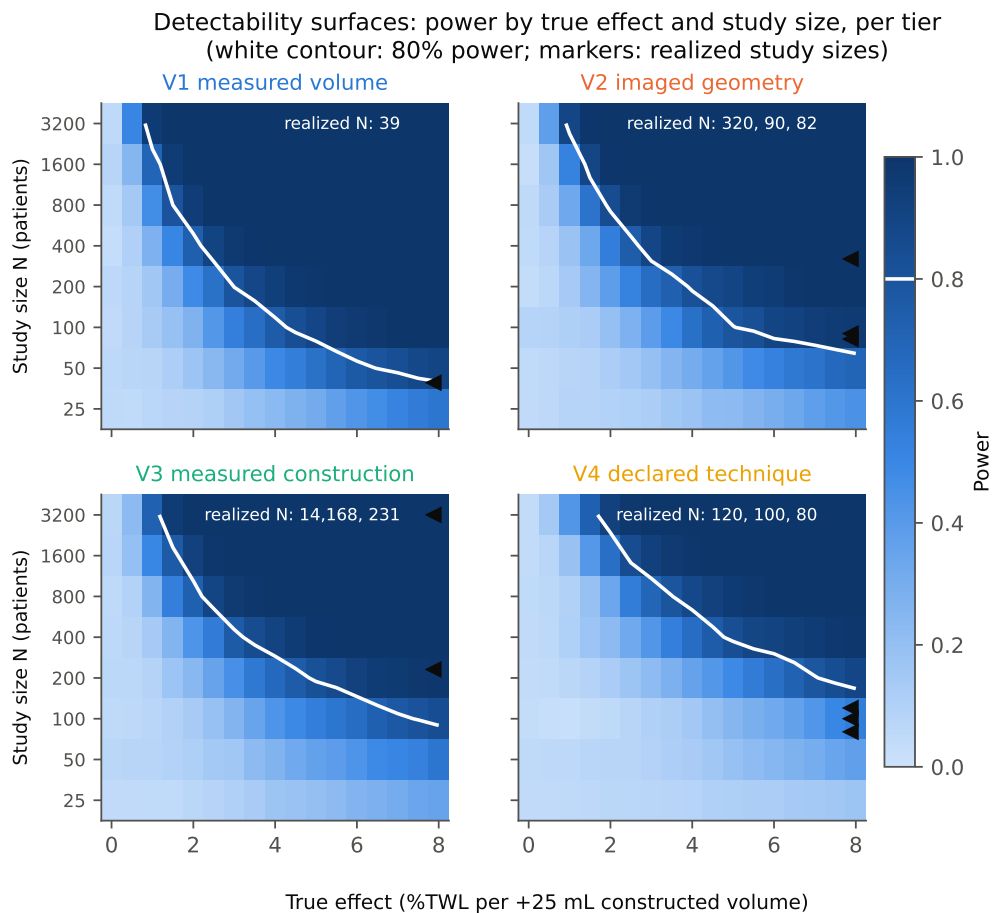
Detectability surfaces (provisional). Per-tier minimum detectable effects at 80% power, in %TWL per 25 mL of true constructed volume, at the pre-extraction grid’s realized designs (Figure 12.1): Tier V1 — one 39-patient volumetric cohort — has a floor of 7.69; Tiers V2 ($k = 3$, $N = 492$) and V4 ($k = 3$, $N = 300$) never reach 80% power anywhere on the effect grid, whose 8-unit upper bound was set beyond the largest effects claimed in this literature; Tier V3 has a floor of 0.56, carried almost entirely by the Scandinavian registry cohort (SOReg) with its 14,168 patients. The V3 floor is the one number that moves materially under the calibration sweeps: because no study measures the length-to-volume proxy correlation, sweeping the assumed r over its 0.40–0.80 band moves the floor between 0.86 and 0.41.

The floor against plausibility (provisional). Placed against physiologically plausible effects, the provisional floors already carry the stalemate-inevitability result in outline. The declared-technique and imaged-geometry tiers — V4 and V2, the tiers that dominate this literature by study count — have floors above the entire simulated effect grid: at their realized designs and documented error, no effect this literature has ever claimed would have been detectable at 80% power. The one V1 cohort’s floor of 7.69 sits at the top of the plausible range; only V3, with the registry’s five-figure sample, reaches floors below 1 %TWL per 25 mL — and V3 is precisely the tier whose exposure is a proxy, so its floor is bought at the price of the proxy-validity assumption the sweep above quantifies. The sensitivity sweeps over the remaining thin inputs (the $n = 12$ V1 agreement study above all) widen the bands without changing this ordering.

The fidelity requirement (provisional). Inverting the design question (Figure 12.2): to detect a true effect of 2 %TWL per 25 mL at $N = 500$, a study needs measurement error no larger than 28.5% of true volume — demanding, but within reach of protocolized volumetric imaging. A true effect of 1 %TWL per 25 mL is unattainable at $N = 100$ or $N = 500$ at any simulated error level; it first becomes reachable around $N \approx 2000$, where the requirement is an error SD of at most 30.5% of true volume. This precision-versus-sample frontier is the concrete requirement the registry recommendation of Chapter 15 cites.

12.5 What a simulation can and cannot show

A simulation demonstrates the consequences of stated assumptions; it cannot prove the assumptions. The error models are calibrated to a validation literature that Chapter 6 showed to rest on samples of 3 to 30 patients, and the honest response — wide bands, swept — widens the floor’s uncertainty rather than removing it. The classical-error base case is contestable for the reasons already given in Section 6.1, which is why the differential scenarios exist; and the outcome model inherits the metric problems of Chapter 7. What survives these caveats is the question’s shape:



PROVISIONAL design grid (pre-extraction; spec:010 replaces designs.yaml)

Figure 12.1: Detectability surfaces by measurement tier V1–V4: power to detect each true pouch-size effect on %TWL (per 25 mL of constructed volume) at the realized study designs, with the 80%-power floor marked. Rendered from simulation output (spec:011); provenance in Appendix E. Provisional: computed on the pre-extraction design grid (designs.yaml, provisional flag); final figures regenerate from the screened corpus.

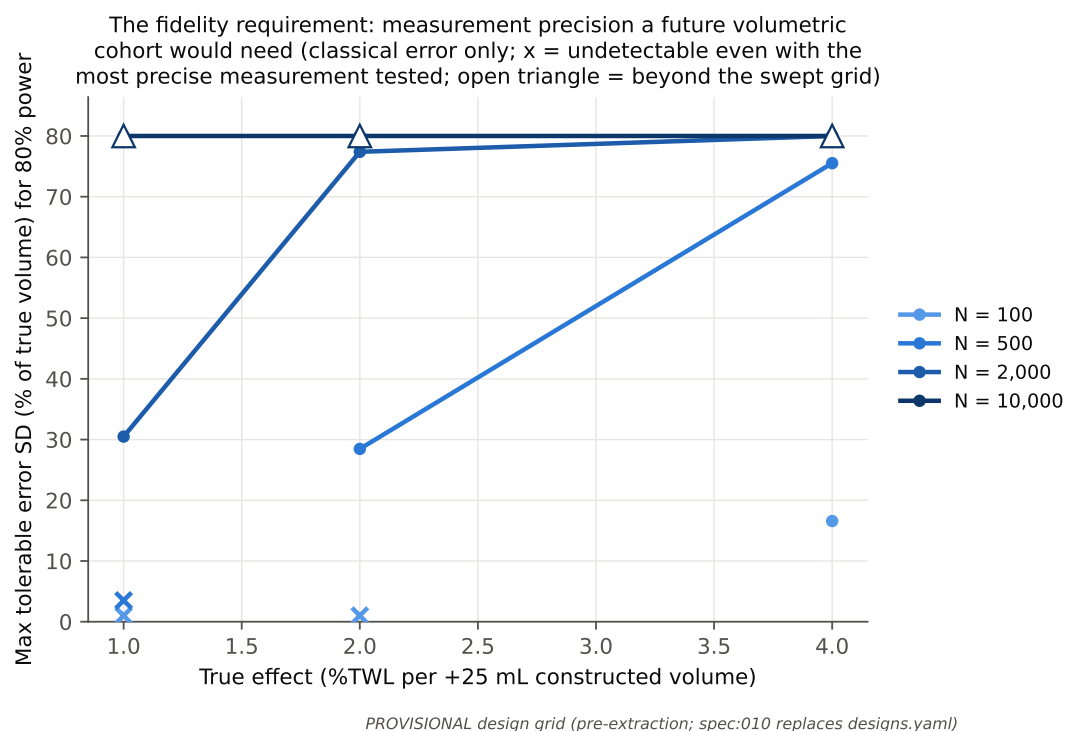


Figure 12.2: The fidelity requirement: the measurement precision (error SD as % of true volume) a future study or registry would need to detect a given true effect at a given sample size at 80% power. Rendered from simulation output (spec:011); provenance in Appendix E. Provisional: computed on the pre-extraction design grid (designs.yaml, provisional flag); final figures regenerate from the screened corpus.

the floor is a property of instruments and sample sizes, both documented, and no reading of the outcome literature can move it.

Chapter 13

Boundaries: safety trade-off and the OAGB contrast

Chapters 10 and 11 ask whether pouch size predicts weight loss and whether the answer is an artifact of measurement. This chapter draws the boundaries around whatever those chapters find. Two boundaries matter. The first is safety: an effect estimate for weight loss, even a positive one, is not a construction recommendation until it is set against the harms attached to the same anatomical choice at *both* ends of its range — the best-measured harm in this literature, marginal ulceration, runs with larger construction, a direction that complicates the field’s intuitions rather than confirming them, while the harms plausibly attached to very small construction (stricture, anastomotic tension, leak, intolerance) are sparsely measured and must be sought separately (research question 5). The second is anatomical: the one-anastomosis gastric bypass (OAGB), in which the pouch is deliberately built three times longer than the Roux-en-Y standard, supplies the variance in pouch geometry that the RYGB literature lacks, and is therefore analyzed as a descriptive parallel literature alongside whatever the RYGB core analysis concludes (research question 6). The chapter closes Part II with the risk-of-bias and certainty assessments for the whole evidence base, and ends with a fenced section whose ground rules are stated now, before any data have been mined, precisely so that nothing found later can masquerade as having been pre-specified.

13.1 The safety counterweight: marginal ulceration, and the harms at the other end

13.1.1 The harm, anatomically

A marginal ulcer is a peptic ulcer at or just beyond the gastrojejunostomy — the surgically created junction between the gastric pouch and the small bowel — usually on the jejunal side, where intestinal mucosa that never evolved to face gastric acid suddenly does. Its pathogenesis is the same acid-versus-defense imbalance familiar from ordinary peptic ulcer disease, plus two surgical contributions: local ischemia along the staple line, and whatever acid-secreting stomach the surgeon left inside the pouch. This last point is the mechanistic hinge of the whole safety question — and it must be stated as the hypothesis it is, not as an established chain. The acid-producing parietal-cell mass of the stomach lies predominantly in the fundus and body, below the small cardia segment from which a standard pouch is cut (the histological geography laid out in Section 4.1); a short pouch should therefore contain little acid-secreting mucosa, and a longer construction plausibly retains more — though stapler length is not the same variable as pouch length or volume, additional stapling along the lesser curvature need not annex fundic mucosa monotonically, and marginal-ulcer causation is multifactorial in any case: smoking, NSAID use, diabetes, and mucosal ischemia are all implicated alongside acid exposure [125, 57]. The

hypothesis nonetheless yields a directional prediction worth testing: if retained acid-secreting mucosa matters, ulcer risk should *rise*, not fall, with construction size — with the caveat that an observed association between a construction proxy and ulceration is evidence of association, not a demonstration of acid-mass mediation. Clinically, marginal ulcers present with pain, bleeding, or — rarely — perforation, and in a systematic review of 41 studies covering 16,987 RYGB patients they occurred in 4.6% overall, with per-study incidence spanning 0.6% to 25%, and with the position and size of the pouch, smoking, and NSAID use identified as the associated factors [125]. That fortyfold spread across studies is not biological; it is ascertainment. Registries that count only symptomatic, clinically surfaced ulcers sit at the bottom of the range; cohorts that put every patient through protocol endoscopy (direct visual inspection of the anastomosis with a flexible camera) sit at the top. The measurement problem that Chapter 6 diagnosed on the exposure side of this literature has an exact counterpart on the safety-outcome side, and every comparison in this section must be read with the detector efficiency in mind.

13.1.2 The registry signal, verified

The anchor evidence for the safety trade-off comes from the same Scandinavian registry cohort whose null weight-loss result anchors Chapter 10: 14,168 laparoscopic RYGB patients in the Scandinavian Obesity Surgery Registry (SOReg), with the total length of linear stapler used to construct the pouch — mean 145 mm — serving as the construction proxy for pouch size [14]. Symptomatic marginal ulcers were recorded in 0.9% of patients at one year. The relative risk of marginal ulcer rose by 14% (95% CI 9–20%) for each additional centimeter of stapler used for the pouch. Meanwhile, stapler length predicted percent excess-BMI loss (%EBMIL, the outcome metric dissected in Section 7.1) at six weeks but not at one year. The authors’ conclusion is the entire safety trade-off in one sentence: a smaller pouch does not predict better weight loss at one year, but it does reduce marginal ulcers, so additional stapling should be avoided [14].

The structure of this result deserves emphasis, because it is unusual and analytically convenient: the benefit estimate and the harm estimate come from the *same* cohort, the same exposure proxy, and the same follow-up window — which rules out the between-study explanations (era, population, design) that dog every other comparison in this literature. It does not rule out everything. Weight loss and marginal ulceration have different causal structures and different ascertainment processes, so confounding and measurement error need not act on the two outcomes identically; and stapler length could predict ulceration through routes that have nothing to do with its validity as a volume proxy — ischemia along a longer staple line, more foreign material, a larger anastomotic surface — as readily as through retained acid-secreting mucosa. A dose–response gradient on the harm side alongside a vanishing gradient on the benefit side is *consistent with* pouch construction genuinely governing acid load while only transiently governing intake, but it is not unique to that reading. The safety signal is therefore, in signal-processing terms, at most a suggestive positive control: it shows the stapler-length variable carries enough information to detect *some* real anatomical effect at this sample size, which weakens — without defeating — the dismissal of the weight-loss null as pure proxy noise. Chapter 11 develops this point, with its limits, in the artifact analysis.

13.1.3 Corroboration, and the limits of it

The direction of the SOReg signal — larger construction, more ulcers — recurs wherever it has been looked for with adequate ascertainment. Table 13.1 assembles the verified evidence. In a retrospective two-technique comparison from a single Chinese center, 95 type-2-diabetic patients with BMI below 35 kg/m² received RYGB with either a large pouch (subtotal, over 50 mL) or a small lesser-curvature pouch (approximately 10–20 mL); at six years, with 89.4% follow-up, marginal ulcers had occurred in 23.7% of the large-pouch group against 6.4% of the small-pouch group ($p = 0.023$) [18]. The same paper reports an odds ratio of 1.13 (95% CI

1.08–1.34, $p = 0.047$) for this comparison — a figure quoted here verbatim from the source but flagged as internally inconsistent with its own crude rates, which imply a crude odds ratio near 4.5; the published OR most plausibly belongs to an adjusted or differently scaled model, and this review treats the crude event rates, not the OR, as the extractable effect data from this study. The pouch contrast there is extreme and the design retrospective, but the systematic follow-up makes the incidence figures unusually trustworthy. In an endoscopic series of 103 marginal-ulcer patients, pouch length was associated with ulcer formation on univariate analysis (OR 1.2 per centimeter, $p = 0.02$; mean pouch length 5.6 cm), alongside smoking and diabetes, with diabetes the strongest multivariate predictor (OR 5.6) [57]. The authors’ reading — increased acid exposure and mucosal ischemia jointly at work — matches the anatomical prediction above.

The corroboration has limits, and they are the ascertainment limits already flagged. A two-center Iranian comparison of long (approximately 80 mL, narrow, bougie-calibrated) versus short (approximately 30 mL) pouches in 219 patients found *zero* marginal ulcers in either group on protocol endoscopy at twelve months, with no significant difference in percent total weight loss [108] — a reminder that with routine proton-pump-inhibitor prophylaxis (standard acid-suppression therapy after RYGB) and a narrow calibrated construction, absolute ulcer rates can be driven low enough that no gradient is observable at $n \approx 200$ and one year. And the only randomized evidence bearing on the question runs through the Dutch extended-pouch trial, which reported no differences in complications between a 10 cm extended pouch and a standard short pouch through three years [13] — but that trial was powered for weight loss, not for an outcome with single-digit percent incidence. An absence of ulcer signal at $n = 132$ is compatible with relative risks that a 14,000-patient registry resolves easily. The safety analysis of this review therefore weights ascertainment method explicitly: symptomatic capture, protocol endoscopy, and mixed designs are extracted as separate strata that are never combined in any pooled estimate, mirroring the exposure-side modality hierarchy of Section 6.9.

Table 13.1: Verified evidence on pouch construction and marginal ulceration. Values extracted from the cited sources; span verification per spec:002 pending pipeline pass. PPI = proton pump inhibitor; EGD = esophagogastroduodenoscopy (protocol endoscopy).

Source	Design	N	Exposure contrast	Ulcer ascertainment	Finding
Edholm 2016 [14]	Registry cohort (SOREg)	14,168	stapler length, mean 145 mm	symptomatic, 1 y	0.9% ulcers; RR +14% per cm (95% CI 9–20%)
Gao 2022 [18]	Retrospective, two techniques	95	pouch > 50 mL vs. 10–20 mL	clinical, 6 y (89.4% follow-up)	23.7% vs. 6.4% ($p = 0.023$)
Azagury 2011 [57]	Endoscopic case series	103	pouch length (mean 5.6 cm)	diagnostic EGD	OR 1.2 per cm, univariate ($p = 0.02$)
Coblijn 2014 [125]	Systematic review, 41 studies	16,987	pouch position/size among factors	mixed	4.6% pooled crude incidence; range 0.6–25%
Safari 2024 [108]	Retrospective, two centers	219	~80 mL narrow vs. ~30 mL	protocol EGD, 12 mo	0 ulcers in both groups (routine PPI)
Boerboom 2020 [13]	RCT	132	10 cm vs. 5 cm pouch	clinical, 3 y	no complication differences (not powered for ulcer)

13.1.4 The safety trade-off, in both directions

Research question 5, as reframed in the protocol, is bidirectional: construction size must be weighed against harm at *both* ends of its range. On the large end, the harm is measured:

marginal-ulcer risk rises with larger, longer construction, taxed at roughly 14% relative risk per centimeter of stapler in the registry [14] — a direction that, it must be said plainly, runs against the field’s habitual framing, in which harm was assumed to attach to aggressive smallness. On the small end, the candidate harms — named here on anatomical reasoning, not on measured comparative evidence — are anastomotic tension (a gastrojejunostomy fashioned on a very short pouch may have to reach farther), stomal stenosis and stricture, anastomotic leak, and food intolerance or dysphagia. These are mechanistically plausible but sparsely measured: the comparative pouch-size literature rarely reports them per arm. Where per-arm data exist, they are so far unremarkable — the two-center Iranian comparison recorded zero leaks, as well as zero ulcers, in both arms at twelve months [108], and the Dutch trial reported no differences in complications through three years while being powered for weight loss, not for any safety outcome [13]. Absence of per-arm reporting is not evidence of absence, and this review extracts every one of these outcomes wherever a study reports it. If the primary analysis of Chapter 10 confirms that smaller construction buys no durable weight-loss benefit, the trade-off question does not become moot: it becomes whether *larger* construction, which the ulcer data tax, purchases anything at all — and, symmetrically, whether very small construction incurs harms that its advocates have not measured.

One randomized result claims that it does. The Dutch extended-pouch trial randomized 132 patients to a 10 cm versus a standard short pouch and found, at three years, total weight loss of 31% versus 27% ($p = 0.023$), driven by less weight regain in the extended-pouch arm [13]. Its five-to-nine-year follow-up, however, retained only 47% of the cohort, and at a median of 109 months the difference had shrunk to 26.1% versus 24.1% total weight loss — non-significant — with a non-significant trend toward less recurrent weight gain in the extended-pouch group (47% vs. 70%, $p = 0.07$); the authors state plainly that the study is underpowered and that no definitive long-term advantage can be concluded [16]. The result sits oddly against the rest of the field: of the fourteen studies in the only prior systematic review, five associated *larger* pouches with poorer weight loss and none with better [15] — the extended-pouch trial is the lone comparative evidence pointing the other way, which is precisely the kind of design-level conflict Chapter 11 dissects. Notably, the extended-pouch hypothesis is not restrictive but hydrodynamic — a longer, narrower conduit should empty more slowly and dilate less, by appeal to Poiseuille’s and Laplace’s laws [16] — which places this trial’s logic squarely inside the emptying-dynamics steelman that Chapter 5 identified as the one mechanistic route by which pouch geometry could matter in a neuroendocrine model.

The trade-off, then, has candidate harms at both ends — ulceration rising with length on one side; tension, stricture, leak, and intolerance plausibly attached to shortness on the other — and one disputed benefit. The analysis contract is as follows, and it deliberately refuses a single “frontier” metric: different safety outcomes have different denominators, time horizons, and ascertainment processes, and risk differences and risk ratios cannot share an axis. Each safety outcome — marginal ulcer; stomal stenosis or stricture; anastomotic leak; food intolerance or dysphagia; tension-attributed reoperation — is analyzed *separately*, as absolute risks per study arm. Pooling per outcome, as risk differences, requires the same three-study floor as every pooled estimate in this dissertation, and two comparability conditions besides: the contributing studies must report the outcome in both arms within a comparable follow-up window, and within the same ascertainment stratum — protocol endoscopy and symptomatic capture are different detectors, and they are never mixed in one pooled estimate. Where fewer than three comparable studies exist, the outcome is presented stratified or narratively, with absolute risks per arm and no pooled number. Zero-event studies — the two-center comparison with no ulcers in either arm [108] is already in hand — are reported raw, as observed proportions; no continuity correction manufactures a pooled risk difference from them. The time horizon and ascertainment stratum are stated for every estimate, and no composite is constructed across outcomes. The per-outcome panels (Figure 13.1) are then set beside the pooled weight-loss

estimates of Chapter 10 for a benefit–harm reading that is a juxtaposition, not a synthesis: a benefit indistinguishable from zero next to a harm gradient monotonic in size is the quantitative form of the SOReg conclusion, while benefit persisting under good measurement would justify the field’s precision effort. Either pattern is a finding.

FIGURE PLACEHOLDER: Per-outcome safety panels. One panel per safety outcome (marginal ulcer; stomal stenosis/stricture; anastomotic leak; intolerance/dysphagia; tension-attributed reoperation): absolute risks per study arm, with per-outcome pooled risk differences and CIs only where ≥ 3 studies share a comparable follow-up window and ascertainment stratum; zero-event studies shown as raw proportions, never entered into a continuity-corrected pooled RD; annotation gives time horizon and ulcer-ascertainment stratum (symptomatic / protocol endoscopy / mixed, never pooled across strata); point size encodes study N . No composite safety metric, no shared effect axis across outcomes, per spec:010. Data-rendered from the pipeline claims table per spec:011 (never LLM-generated); source data: extraction rows for the safety outcomes defined in spec:010.

Figure 13.1: Per-outcome absolute-risk safety analyses for pouch construction size (data-rendered; placeholder pending pipeline extraction).

13.2 The OAGB contrast

13.2.1 A longer lever arm

The one-anastomosis gastric bypass (also mini-gastric bypass, OAGB/MGB) reroutes the intestine with a single gastrojejunal anastomosis rather than the two of a Roux-en-Y: a jejunal loop is brought up and joined directly to the pouch, with no separate biliary diversion. Its pouch is the anatomical point of interest here. Where the RYGB pouch is a short chamber cut from the cardia, the OAGB pouch is a long, narrow, bougie-calibrated tube running most of the lesser curvature — in the original Rutledge description, a conduit from below the crow’s foot (the terminal fan of the anterior vagal nerve on the lesser curve, a landmark introduced in Section 4.2) up to the angle of His [91]; in Carbajo’s variant, an estimated 25 mL divided pouch built parallel to the lesser curvature from the esophagogastric junction to the crow’s-foot level [92]; in the current mainstream technique, an explicitly “long and narrow” pouch [197]. The first international consensus exercise on the procedure reached 94% agreement that pouch construction should start in the horizontal portion of the lesser curvature [93].

The quantitative contrast with RYGB practice is stark, and it is the design rationale for this subgroup. In the 2026 IFSO global survey of RYGB technique — 245 surgeons across 47 countries — only 53.5% routinely measured pouch length at all, and those who did reported a median of 6 cm (IQR 5–7 cm), with 67.3% describing their pouch as “short or small” [7]. In the corresponding Israeli national survey of OAGB technique — the one country where OAGB is the most prevalent bariatric operation — 57.4% of surgeons fashion a pouch of 10–15 cm [198]. In statistical terms, the RYGB literature estimates a slope over a lever arm of about two centimeters of between-surgeon variation, while OAGB practice spans double to triple the RYGB length as its *routine* range. Restriction of range attenuates any estimable association, so the wider OAGB range could in principle support a better-conditioned dose–response analysis on pouch length — with the caveat that these surveyed lengths are between-surgeon and between-study variation, not within-study exposure contrasts, and only the latter carry inferential weight. That is why the protocol (spec:010) pre-specifies OAGB as a separately analyzed subgroup rather than excluding it as a different operation.

13.2.2 What the contrast can and cannot show

Separately analyzed, because the contrast is confounded by construction: OAGB differs from RYGB simultaneously in loop reconstruction, in a typically longer biliopancreatic limb, in the pouch's exposure to bile, in anastomotic configuration and emptying, and in patient selection — so OAGB effect estimates can never be pooled into the RYGB core without dissolving the exposure definition, and for the same reason the OAGB literature cannot *validate* an RYGB result, in either direction. The subgroup is a descriptive parallel, not an external-validation arm. The mechanistic intuition remains suggestive: if pouch geometry drives weight loss through emptying dynamics, the effect should be easier to detect across the OAGB length range than within the compressed RYGB range. But the inferences available from the parallel are weaker than that intuition implies. A null length–outcome pattern across the OAGB literature would be *consistent with* an RYGB null while remaining equally compatible with uncontrolled technique variation across OAGB studies washing out a real gradient; it cannot strengthen the external validity of the RYGB conclusion. A clean gradient in OAGB beside a null in RYGB would be *consistent with* the range restriction and proxy noise diagnosed in Chapters 6 and 11, but also with bile exposure, limb length, or selection doing the work. The analysis contract: random-effects pooling of study-defined pouch-length contrasts within OAGB; direction and magnitude set beside the RYGB core estimate as a descriptive juxtaposition, with the competing explanations named; and a purely descriptive cross-study dose–response display on length — no pooled slope, no inferential interval — per the same rule as the volumetric RYGB subset (Section 8.9).

13.2.3 The safety gradient, reversed

The OAGB contrast also sharpens the safety analysis, because in OAGB the safety pressure on pouch length runs in the opposite direction from RYGB. The loop reconstruction leaves the pouch exposed to bile: in a prospective scintigraphic series (a nuclear-medicine tracer study of where bile actually goes), 31.6% of OAGB patients had bile reflux into the gastric pouch at six months, and 2.6% into the esophagus [70]. In a multicenter retrospective series of 2,678 OAGB patients, postoperative duodeno-gastro-esophageal reflux was associated with pre-existing reflux disease and with a gastric pouch *shorter* than 9 cm ($p = 0.001$) [199]: the long pouch is not an incidental preference but the operation's built-in spacer, holding the biliary stream away from the esophagus. Long-term endoscopic surveillance is reassuring on the ulcer side — in a four-center cohort of 1,172 patients at five years, anastomotic ulceration occurred in 2.6%, predominantly in heavy smokers, with erosive esophagitis in 7.7% and no Barrett's esophagus [200]; a five-year single-center MGB series reported marginal ulcers in 4% and incapacitating biliary reflux requiring conversion in 1.6% [201]. And the two design pressures can be married deliberately: the “long pouch gastric bypass” grafts an OAGB-style long pouch onto a Roux-en-Y diversion precisely to keep the reflux protection while eliminating bile exposure, reporting clinically relevant entero-biliary reflux in 1% at four years [109].

The synthesis point for this dissertation is symmetry: in RYGB, the best-measured harm taxes every added centimeter of pouch [14]; in OAGB, the best-understood harm taxes every centimeter removed below about nine [199]. Two operations, one anatomical variable, opposite safety gradients — and in both cases the convention (short pouches in RYGB, long pouches in OAGB) crystallized before comparative evidence existed, the pattern Chapter 3 traced across the field's history. If the weight-loss analysis returns null in both operations, the picture that emerges is of a variable whose rational optimization is driven by safety alone, with different optima in different operations — a reading the descriptive parallel can motivate but, for the reasons above, not prove; it would be no less clinically consequential than a weight-loss effect, and it would be stated as exactly that: a convergent descriptive pattern.

13.3 Risk of bias and certainty of evidence

Part II closes with the formal quality assessment that every synthesis in Chapters 10–13 inherits. Randomized trials are assessed with RoB 2, the revised Cochrane tool that grades each result (not each study) across five bias domains from randomization through selective reporting [185]. Observational comparisons are assessed with ROBINS-I, which evaluates a non-randomized study against the hypothetical target trial it emulates, across seven domains beginning with confounding [186]. Certainty per outcome — the pooled %TWL estimates at one year and in the fixed three-year and five-year windows, each safety outcome separately (marginal ulcer foremost), and the OAGB subgroup — is rated with GRADE, which starts randomized evidence high and observational evidence low, then downgrades for risk of bias, inconsistency, indirectness, imprecision, and publication bias, and upgrades for large effects and dose–response gradients [187]. Should the primary pooling fail its pre-stated defensibility criteria, the synthesis-without-meta-analysis fallback [183] changes the presentation of results, not the obligation to rate their certainty.

Three structural features of this literature are already visible from the evidence assembled above and will dominate the assessment regardless of how individual judgments fall. First, the exposure is proxy-measured in most studies (Section 6.9), which in ROBINS-I terms is misclassification of intervention and in GRADE terms indirectness. Second, follow-up attrition is severe even in randomized evidence — the only long-term RCT follow-up in the field retained 47% of its cohort [16] — which bears on missing-data domains and on imprecision. Third, the ulcer outcome’s ascertainment heterogeneity (Section 13.1.3) is a detection-bias problem that no pooling can average away and must instead stratify. The dose–response gradient of the SOReg safety signal [14] is, conversely, the one feature in this literature that GRADE rewards with an upgrade. Small-study effects are assessed by funnel-plot asymmetry with Egger’s test where at least ten studies contribute (Figure 13.2), and the result is reported as suggestive of small-study effects only — asymmetry has several causes, and the test is never treated as an adjudicator of publication bias.

Table 13.2: Risk of bias in randomized trials (RoB 2), per outcome.

TABLE PLACEHOLDER: RoB 2 domain judgments for each randomized result contributing to the pooled analyses (randomization process; deviations from intended interventions; missing outcome data; outcome measurement; selective reporting; overall). Rendered as a traffic-light figure/table from pipeline-recorded judgments via robvis or equivalent [202]; data-rendered per spec:011.
--

Table 13.3: Risk of bias in non-randomized studies (ROBINS-I), per outcome.

TABLE PLACEHOLDER: ROBINS-I domain judgments for each comparative observational study (confounding; selection; intervention classification — the modality hierarchy of Chapter 6 enters here; deviations; missing data; outcome measurement — ulcer ascertainment stratum enters here; selective reporting; overall). Data-rendered from pipeline-recorded judgments.

13.4 Exploratory corpus-mined patterns

This section is fenced. Everything above this point was pre-specified in the protocol (spec:010) before screening; everything reported below this point, in the final version, will be the product of unplanned exploration of the assembled claims table and is hypothesis-generating only.

The ground rules, stated before any mining has occurred:

Table 13.4: GRADE summary of findings, all outcomes.

TABLE PLACEHOLDER: GRADE summary-of-findings table: %TWL at 1 year (primary), %TWL at 3 years and at 5 years (fixed windows; observations beyond 66 months per-study only), each safety outcome separately (marginal ulcer; stenosis/stricture; leak; intolerance), OAGB subgroup. Columns: studies (n), design mix, pooled estimate (or SWiM summary), downgrades/upgrades with reasons, final certainty. Generated from the claims table; every cell traceable to span-verified extraction rows.

FIGURE PLACEHOLDER (F6): Funnel plot — study-level effect estimates (small-vs-large %TWL contrast) against their standard errors, with pseudo-95%-confidence funnel and Egger regression line; drawn only if ≥ 10 studies contribute, per spec:010. Data-rendered from pipeline output per spec:011; never LLM-generated.

Figure 13.2: Funnel plot for small-study effects in the primary contrast (data-rendered; placeholder pending pooled analysis).

1. Exploratory analyses run only after the pre-specified analyses of Chapters 10–13 are complete and locked; no exploratory result feeds back into any pre-specified analysis or its interpretation.
2. Candidate directions are named now, in advance: effect-by-era, effect-by-baseline-BMI, and the joint pattern of benefit and per-outcome safety estimates beyond the pre-specified analyses. Anything else that surfaces is reported as fully unplanned.
3. Every reported pattern carries an explicit multiplicity caveat: the claims table supports a large number of possible cross-study contrasts, and patterns found by searching it inherit the selection effect of that search. No p -value from this section is interpretable at face value.
4. No GRADE certainty rating is assigned to exploratory findings, and none is carried into the conclusions of Chapter 15 except explicitly labeled as hypothesis-generating, per the original contribution requirement of spec:000.
5. Each exploratory claim obeys the same span-verification discipline as every other quantitative claim in this dissertation: no span, no citation.

[This section is intentionally empty in this draft. It fills, or is reported as empty, only after the pipeline’s cross-study mining pass.]

Part III

Synthesis

Chapter 14

Discussion

14.1 Interpretation committed before data

This chapter was written before the analyses of Chapters 10–13 returned their numbers, and that is its design, not its defect. A discussion section drafted after the results are known is an exercise in accommodation: the author knows the answer and constructs the interpretation that makes the answer look inevitable. The epidemiological literature has a name for the analytic version of this failure — outcome-dependent analysis — and a remedy for it: pre-registration, the commitment to hypotheses and analysis plans before the data can influence them. There is no established convention for pre-registering an *interpretation*, but nothing prevents one. The protocol (spec:010, Chapter 8) fixed the analyses; this chapter fixes, in advance, what each possible outcome would mean, so that when the numbers arrive they select among interpretations already written rather than summoning new ones.

The commitment takes the form of three lenses along the question’s principal axis — does the effect survive measurement quality? Either the pouch-size effect on weight loss vanishes when measurement is sound (Lens I); or it survives measurement quality (Lens II); or the evidence base cannot distinguish the two (Lens III). The lenses are conditional in form — “if the results show X , then the field should conclude Y ” — and each is argued in full below, at the same depth it would receive if it were the one selected. When the results chapters are final, the selected lens will be marked and the others retained: the unselected branches are part of the scientific record of what this dissertation was prepared to conclude, and they document that the selected interpretation was not built to fit the data.

Three lenses do not exhaust the outcome space, and pretending they do would invite exactly the accommodation this chapter exists to prevent — an ambiguous result forced into the nearest pre-written narrative. Three subsidiary branches are therefore committed alongside them. *Direction reversal*: if the pooled estimate favors larger or longer construction — the direction of the Dutch extended-pouch trial [13] — the result is handled within Lens II as a surviving effect whose sign contradicts the doctrine; the doctrine is not rescued by relabeling, and the geometric dissection of Section 14.3 carries the interpretive weight. *Clinically important heterogeneity*: if effects are real but differ materially by population, follow-up phase, or geometry, no single-effect lens applies; the honest summary is then the effect-modification analysis itself, reported without a pooled-headline shorthand. *Efficacy–safety discordance*: the efficacy and safety answers may disagree in either arrangement — null efficacy alongside real per-outcome harms, or real efficacy priced by them — and the per-outcome safety analyses of Section 13.1.4 are kept separate precisely so that no lens below is permitted to trade them against each other in a single currency.

Two fixed points constrain all three lenses and should be stated first, because they hold regardless of which branch the data select. First, the question is real and the field has not answered it: the only prior systematic review found fourteen studies, of which two were randomized, and declared that the quality and heterogeneity of the data precluded a meta-analysis

[15]. Second, the choice at issue is made without instrumentation: in the 2026 IFSO worldwide survey of Roux-en-Y gastric bypass technique, only 53.5% of 245 responding surgeons across 47 countries routinely measured the length of the pouch they had just constructed, while 67.3% described their pouches qualitatively as “short” or “small,” against variable anatomical endpoints, with calibration bougies ranging from 28 to 46 French [7]. A field that does not measure the parameter cannot learn from it; that fact bounds every conclusion below.

14.2 Lens I: no effect under sound measurement

If the pooled estimate of Chapter 10 is null — or if it is nominally positive but the modality-stratified analysis of Chapter 11 shows the association confined to the construction tiers (V3–V4) and absent under imaged volumetry (V1–V2), with the competing explanations that chapter puts on the table — era, design, cohort size, timing of assessment — examined and found wanting — then the pouch-size doctrine is a legacy of the founding model, and the field should say so plainly.

The doctrine’s pedigree is not evidentiary. It was born with the operation, in the era that understood gastric bypass mechanically: Kirkpatrick and Siegel’s 1982 prospective series concluded that a tiny reservoir — under 45 mL — was the critical determinant of success, with stoma size contributing little once the reservoir was small [5]. That conclusion was an inference from the restrictive model, not a demonstration of dose-response; the study varied the stoma, not the reservoir. Within two years, prospective radiological measurement in gastroplasty patients was already failing to find any correlation between pouch volume and weight loss across a 51–244 mL range [78], and the null kept returning wherever the pouch was actually measured: day-one contrast studies [118], barium volumetry at three years [119]. Meanwhile the mechanistic literature relocated the operation’s effect from the staple line to the gut–brain axis — to exaggerated postprandial GLP-1 and PYY release, altered bile-acid signaling, and central appetite rewiring rather than mechanical restriction of intake [9] — and in the rodent model, where pouch size can be varied far beyond the surgical range, bypass improved weight and glucose tolerance equally with a large pouch, a small pouch, or no pouch at all [124].

Under this lens, the largest dataset in the literature is not an anomaly to be explained but the expected result finally observed at adequate power. In 14,168 Scandinavian registry patients, the length of stapler used to construct the pouch predicted excess-BMI loss at six weeks and not at one year [14]. A six-week effect that decays by one year is precisely the signature the neuroendocrine model predicts: a transient mechanical phase, swamped once hormonal adaptation dominates. The positive studies, on this reading, are what a proxy-measured, confounded observational literature produces around a true null. And here the direction of the measurement argument matters and deserves precision, because it is the opposite of the intuitive one. Classical random error in an exposure attenuates association — regression dilution biases toward the null — so if construction proxies were merely *noisy* versions of true volume, the proxy studies should show *weaker* effects than the volumetric ones, not stronger. An association that appears only under weak proxies therefore cannot be rescued as an attenuated glimpse of a real effect; it implies the proxy carries information other than volume. Stapler count and staple-line length travel with surgeon, institution, era, and technique; a “small pouch” defined by construction habit is a bundle of correlated surgical choices, any of which can generate the association. Chapter 6 built the modality hierarchy on exactly this logic, and Lens I is the outcome in which the hierarchy does its work. Even then the claim must be sized honestly: tier travels with era, design, and cohort size, so the stratified contrast is exploratory effect modification, not an adjudication. Under Lens I the measurement account is the most economical explanation of forty years of disagreement — not a proven one.

The practical conclusion under Lens I is not that pouch construction is a free parameter. The safety gradient survives the efficacy null, and it is the same registry that supplies it: each

additional centimeter of stapler used for the pouch raised the relative risk of marginal ulcer — an ulcer at or near the gastrojejunal anastomosis, the operation’s signature late complication — by 14% (95% CI 9–20%) [14], and the retrospective large-versus-small-pouch cohort observed marginal ulcers in 23.7% of large-pouch patients against 6.4% with small pouches at six years [18], with pouch length also associated with ulceration in endoscopic series [57]. But that gradient is observational registry evidence about stapler length, not randomized evidence about construction targets, and it does not license a directional construction policy: shorter staple line is not the same variable as smaller volume, the per-centimeter model has not been tested beyond its observed range, and the harms plausibly attached to very small construction — stricture, anastomotic tension, leak, food intolerance — are not priced by an ulcer gradient that happens to run the other way (Section 13.1.4). What Lens I does establish is a misallocation of precision effort: calibrating, measuring, and standardizing pouch *volume* in pursuit of weight loss spends surgical attention on a variable that does not pay in the currency it is spent for. The defensible reformulation of the doctrine is accordingly more modest than a target: *avoid stapling that serves no anatomical purpose, expect nothing from pouch volume for weight loss, and treat the comparative safety of construction sizes as an open question rather than a settled gradient.*

14.3 Lens II: the effect survives measurement quality

If instead the pooled estimate is positive and the Chapter 11 stratification shows it persisting — or strengthening — in the volumetric and imaging strata, then the registry null was wrong for identifiable reasons (a proxy too coarse, an outcome window too early, a comparator spread too narrow), surgical habit was accidentally right, and two consequences follow: the mechanistic account requires amendment, and technique standardization becomes an evidence question rather than a preference.

The amendment is the emptying-dynamics bridge built in Section 5.3: even in a fully hormonal model of the operation, the pouch is the antechamber that sets the rate of nutrient arrival at the jejunum, so geometry can modulate the hormonal signal without any appeal to restriction. The empirical support is real: scintigraphic studies show the post-bypass pouch empties far faster than the intact stomach, and that fast emptying of liquid nutrient is associated with high GLP-1 and PYY_{3–36} secretion [10]; smaller measured pouches empty faster [48]; and prompt emptying on early contrast study predicted greater weight loss at one year in 304 patients [111], with the small-volume, fast-emptying pouch associated with better maintained weight loss at four years [48].

But a surviving effect would force the field to confront a problem this literature has so far been able to ignore: the emptying amendment is internally contradictory about sign. The fast-emptying literature just cited argues that small pouches work because they empty quickly and amplify the hindgut hormone response. The only randomized trial to show a between-arm difference at three years argued the reverse: its 10-cm extended pouch was designed, with explicit appeal to Poiseuille’s law (flow through a tube falls as the tube lengthens) and Laplace’s law (wall tension, and hence dilation propensity, rises with radius), to empty *more slowly* and dilate less — and the extended pouch won, 31% versus 27% total weight loss at three years ($p = 0.023$), driven by less weight regain [13]. At five to nine years the difference was no longer significant (26.1% versus 24.1% total weight loss) in a follow-up that retained only 47% of the cohort and was, by its own account, underpowered, but the direction held, and recurrent weight gain remained more prevalent in the standard-pouch arm (70% versus 47%, $p = 0.07$) [16]. A two-hospital cohort comparing a long narrow 80-mL pouch against a short 30-mL pouch found no significant difference at one year [108]. If Lens II is selected, these results cannot all be read as one effect: “pouch size” is being used as a scalar for what is actually a geometry — volume, length, width, and outlet behavior are separate variables with separate physics. The reconciliation the mechanism chapters anticipate is that the effect, if real, is *shape-dependent*

rather than volume-dependent: fast early emptying (small volume, favorable outlet) may drive the hormonal amplitude that sets nadir weight loss, while resistance to dilation (narrow, long geometry; low wall tension by Laplace) governs the late regain phase. The two randomized signals — the Chinese volume-contrast trial of Ren and colleagues, in which the 10–20-mL pouch beat the 25–35-mL pouch on excess weight loss at twelve months [17], and the Dutch length-contrast trial, in which the extended pouch showed its advantage in the regain phase at three years [13] — would then not be in conflict at all: they varied different physical quantities, volume versus length, and they measured different phases of a two-phase process with different geometric determinants. The six-year small-pouch cohort of Gao and colleagues [18] corroborates the volume direction, but as a retrospective comparison of sequential technique eras against a substantially different large-pouch anatomy it is cohort corroboration, not randomized evidence; Chapter 11 dissects it.

Under Lens II, standardization stops being a matter of taste. If a measured construction parameter moves a primary outcome, then the finding that barely half of surgeons measure it [7] becomes a quality-of-care problem, and the dose-response analysis of Chapter 10 supplies the target: either a gradient, implying volume (or length) should be titrated and recorded like a drug dose, or a threshold, implying a floor below which further reduction buys nothing. The safety counterweight then bites in the opposite direction from Lens I: the extended-pouch design that showed the regain benefit consumes *more* staple line — three 60-mm loads against two [16] — and the registry prices every additional centimeter at 14% relative marginal-ulcer risk [14]. Section 13.1.4 prices the two directions separately, outcome by outcome, and under Lens II the trade-off becomes a genuine optimization with a nonzero interior solution — though one that must be argued per outcome and per time horizon, not read off a single frontier curve.

14.4 Lens III: undecidable — and exactly what would decide it

If pooling fails its pre-stated defensibility criteria, or if the pooled interval is too wide to distinguish a clinically meaningful effect from zero, or if the modality strata are too thin to run the artifact test with power, the dissertation’s obligation is not to shrug but to specify. Undecidability, demonstrated rigorously, is a finding with an actionable complement: the exact description of the study that would decide the question. The ingredients are all in evidence.

Exposure. The decisive study measures constructed pouch volume directly, at or shortly after surgery, with an imaging method whose error is characterized. The candidates exist — three-dimensional CT volumetry has been feasible since 2009 [154], and swallow MRI agreed with 3D-CT to within measurement noise (mean pouch volume 40.4 ± 21.0 versus 43.5 ± 30.2 mL, $p = 0.83$) — but that agreement study enrolled twelve patients [19], which is a calibration anecdote, not a validation. Step one is therefore a proper method-comparison study (Bland–Altman design, prespecified limits of agreement) of sufficient size; step two is deploying the validated method at construction, not years later, because late imaging measures dilation — a different variable, on a different causal pathway, systematically enriched in weight-regain populations [150].

Contrast and size. The trial randomizes to volumetrically verified small versus large construction, wide enough apart to matter. The existing trial data put the standard deviation of percent total weight loss near 9–11 points [13, 16]; taking $\sigma = 9.3$ (the value the extended-pouch trial itself used for its power calculation [16]), a two-arm trial detecting a 3-point difference in %TWL at 80% power and two-sided $\alpha = 0.05$ requires roughly 150 patients per arm before attrition, and about 55 per arm for a 5-point difference. These are unremarkable numbers — single-network scale, an order of magnitude smaller than the registry cohort that currently anchors the null [14]. The decisive study has been affordable for years; it has not been done because the exposure was never instrumented, not because the sample was out of reach.

Outcome and duration. Percent total weight loss at one year, with prespecified follow-up through at least five, and regain defined from nadir in advance — because the one consistent

lesson of the long-term data is that the interesting divergence is late: nadir weight loss was indistinguishable between extended and standard pouches (35.5% versus 34.6%) while the regain trajectories separated [16]. A one-year primary endpoint alone would have called that trial null at every timepoint it reported early.

Retention. The same trial’s arc — significant at three years, inconclusive at nine with 47% of the cohort retained [13, 16] — is the attrition lesson in miniature. A decisive study budgets for retention as a design parameter, not an aspiration.

Under Lens III, the forward statement for registries (developed in Chapter 15) follows directly: a single measured-volume field, captured at construction, would convert every future registry cohort into the observational arm of this study design. The field’s oldest technical question is unanswerable from construction proxies; it is cheaply answerable from one number nobody records.

14.5 This review against the existing synthesis

The only prior systematic review of this question is Mahawar, Sharples, and Graham (2020) [15], and the comparison scaffolding in Table 14.1 is the frame within which this dissertation’s results should be read. Three of its findings deserve emphasis before the contrasts. First, its vote count was asymmetric: of fourteen pouch studies, nine found no association and five found larger pouches worse — and *none* found larger pouches better [15]. That asymmetry is a lead, not a demonstration: the categories are significance-based, so the point-estimate directions of the nine “null” studies are unreported — for all the tally shows, several may lean larger-better without reaching significance — and a symmetric sign expectation under the null is itself an assumption in a heterogeneous, doctrine-driven literature. This review therefore extracts and displays the signed estimates the tally discards; the small-study-effects analysis of Chapter 13, at a study count near the conventional minimum, is read as suggestive context for that display, never as an adjudicator of cause. Notably, the first study to report even a trend in the opposite direction — the extended-pouch follow-up, which describes itself as exactly that first [16] — appeared after Mahawar’s search window closed. Second, the review pooled two exposures this dissertation deliberately separates: pouch size and stoma size were reviewed together, and in its stoma analysis the same asymmetric pattern recurred (six null, four larger-worse, none larger-better) [15]. The protocol here treats stoma diameter as a confounder to be extracted and adjusted for, not an exposure — a scoping decision defended in Chapter 8, and material because the best-measured endoscopic evidence finds stoma diameter, not pouch dimensions, to be the independent correlate of weight regain [115]. Third, and centrally: Mahawar and colleagues did not decline to meta-analyze out of methodological conservatism; they judged the data unpoolable [15]. This dissertation’s primary analysis is, in effect, a formal test of that judgment. If pooling succeeds under the harmonization rules of Section 7.6, the disagreement was tractable and the field waited five years longer than it needed to; if pooling fails against pre-stated criteria, this work converts their qualitative verdict into a quantitative anatomy of *why* — which studies, which axes of heterogeneity, which measurement strata.

14.6 Scoring the mechanistic predictions

Chapter 5 closed by deriving testable predictions from each mechanistic model, and this section is where they are scored — conditionally now, finally when the results chapters are final — in the sense fixed by H7: as updates of confidence among competing positions, not as adjudication of mechanism. A study-level observational synthesis cannot identify mechanism, and the subgroup patterns that bear on these models are confounded with era, design, and study size, so no pattern below selects a model outright; each shifts weight among them. Table 14.2 restates the predictions next to the observables that bear on them.

Table 14.1: Comparison scaffolding: the prior systematic review versus the present work. Mahawar column extracted from the published review [15]; present-work column per protocol (spec:010, Chapter 8).

	Mahawar et al. 2020	Present work
Exposure	Pouch and/or stoma size at surgery, reviewed jointly	Constructed pouch size only; stoma diameter and limb lengths treated as confounders
Search	PubMed online search	PubMed/PMC E-utilities pipeline (526 records cataloged), Cochrane CENTRAL, backward/forward citation chasing; Embase absence declared
Included studies	14 pouch studies (2 randomized); 10 stoma studies	Section 9.2
Synthesis	Narrative; “quality and heterogeneity of data precluded a meta-analysis”	Random-effects pooling attempted; SWiM fallback with pre-stated failure criteria
Direction of findings	9 null / 5 larger-worse / 0 larger-better (pouch)	Pooled estimate with confidence interval, Chapter 10
Measurement modality	Not stratified	Pre-committed modality hierarchy (Section 6.9); artifact test in Chapter 11
Dose-response	None	Volumetric-subset gradient analysis
OAGB	Not addressed	Pre-specified, separately analyzed subgroup
Safety counterweight	Not assessed	Marginal ulcer as co-primary safety outcome

The legacy restrictive model — the pouch as a small container whose capacity limits intake — predicts a monotone volume dose-response, strongest under accurate volume measurement, stable across follow-up. Its problems predate this dissertation: measured pouch volumes showed no correlation with weight loss as early as 1984 [78], the same null recurs under direct measurement in the modern literature [118, 119], and the mechanism review literature has formally moved on [9]. If Chapter 10 nonetheless returns a clean volumetric gradient, the restrictive model earns partial rehabilitation — as an early-phase mechanism at minimum — and the field’s forty-year instinct was better calibrated than its evidence.

The pure neuroendocrine model predicts no volume dependence within the surgical range, permits a transient early mechanical effect, and — this is its sharpest, most falsifiable signature — predicts that any persistent association should concentrate in weak-proxy strata, since proxies carry technique and era rather than physiology. The registry’s six-week-yes, one-year-no pattern [14] and the pouch-size-independence of the rodent bypass [124] are its existing support. It is the only model for which the Chapter 11 modality analysis — exploratory effect modification, with its confounders on the table — counts as direct support rather than a nuisance adjustment: Lens I is this model’s predicted outcome.

The emptying-dynamics amendment predicts an effect that is real but geometric — residing in emptying rate and dilation propensity, hence in shape and outlet, not in volume as such. Its distinctive signatures: volume dose-response flat while length/width contrasts are not; effects that differ by follow-up phase (hormone-amplitude effects at nadir, dilation-resistance effects in the regain years); and an association between measured emptying and outcome wherever emptying was measured [10, 48, 111]. Its embarrassment, argued in Section 14.3, is the unresolved sign of the emptying–outcome relationship — fast emptying as hormonal amplifier [10] versus slow emptying as satiety extender [13, 16] — and any reading that favors this model must also commit to one of those signs, or partition them by phase, explicitly.

Table 14.2: Mechanistic predictions (committed in Section 5.3) and the analyses that weigh them. The confidence-shift column records how each model’s standing moves — not a verdict; per H7 the synthesis updates confidence among competing positions — and is completed from Chapters 10 and 11 when final.

Model	Distinctive prediction	Weighed by	Confidence shift
Restrictive (legacy)	Monotone volume dose-response; strongest under imaged volumetry; stable over follow-up	Dose-response analysis; modality strata	—
Neuroendocrine (pure)	Null within surgical range under sound measurement; early transient effect only; residual association confined to proxy strata	Modality artifact test; 6-week vs. 1-year contrast	—
Emptying-dynamics amendment	Shape- not volume-dependent effect; divergence between nadir and regain-phase outcomes; geometry contrasts positive while volume gradient flat	Dose-response on volumetric subset vs. length-based contrasts; ≥ 3 -year outcomes	—

14.7 The incretin-era stakes, revisited

Section 2.4 argued that pharmacology raises rather than retires this dissertation’s question; the argument can now be closed with the results framework in place. Once-weekly semaglutide produced a mean 14.9% body-weight reduction at 68 weeks against 2.4% with placebo [11];

tirzepatide at its highest dose produced 20.9% at 72 weeks, with 57% of participants losing at least a fifth of their body weight [12]. Contemporary RYGB cohorts in this literature report one-year total weight loss around 30% [108] and nadir losses near 35% in trial populations [16]. The pharmacological numbers now overlap the lower range of the surgical ones, and the drugs work by the same signaling channels — GLP-1 receptor agonism, with GIP co-agonism in tirzepatide — that the neuroendocrine model credits for the operation’s effect [9].

This reframes what surgery is *for*, and the reframing runs through this dissertation’s question in both directions. If Lens I holds, the convergence is complete on the hormonal account: the operation is an anatomical delivery system for a neuroendocrine effect that drugs now approximate, its constructed geometry matters for safety and durability but not efficacy, and the surgical value proposition rests on what pharmacology cannot yet offer — a one-time intervention with multi-decade durability, established remission and mortality evidence, and freedom from indefinite adherence. If Lens II holds, the opposite and more interesting conclusion follows: there exists a mechanical parameter with an efficacy effect size that no drug can replicate, because no drug can change the rate at which nutrients arrive at the jejunum. In a treatment landscape where drugs replicate the hormones, the effect sizes of the anatomy are precisely the quantities the field needs on the table — they are surgery’s non-replicable margin. Either way, the era in which a surgeon could shrug at the pouch-size question because the whole operation was self-evidently superior is over; the comparison with pharmacotherapy prices every component of the operation, and components without measurable effect must justify themselves on other grounds.

14.8 Clinical implications, drawn conservatively

Pending the final numbers, the implications that survive all three lenses are few and should be stated at exactly their earned strength. First, nothing in this literature supports *enlarging* the pouch beyond convention in pursuit of better weight loss; the single trial trend in that direction is underpowered, self-labeled exploratory, and concerns a long *narrow* pouch, not a capacious one [16]. Second, the marginal-ulcer gradient with staple-line length is the best-powered construction–outcome association in the field [14]; it remains an observational association that cannot found a directional construction policy, but it supports one conservative behavior today, independent of the efficacy question: avoiding stapling that serves no anatomical purpose, where technique allows. Third, no result of this dissertation will license abandoning calibration discipline — under every lens, the case for *knowing* what was built is strengthened, whether the knowledge serves efficacy (Lens II), safety accounting (Lens I), or the decisive future study (Lens III). What no reader should take from this work, under any lens, is a specific target volume in milliliters; the current guidance apparatus of the field — whose 2022 joint ASMBS/IFSO statement on surgical indications prescribes when to operate, not how to build the pouch [8] — is silent on that number, and this dissertation’s contribution is to determine whether the silence reflects missing evidence or a genuinely inert parameter.

14.9 Limitations

These are stated without cushioning; each names a way this dissertation’s answer could be wrong.

Embase was not searched — and neither were Web of Science, Scopus, trial registries, dissertation indexes, or conference proceedings. The search relied on PubMed/PMC via the E-utilities pipeline, Cochrane CENTRAL, and citation chasing. Embase indexes European and conference literature underrepresented in MEDLINE; systematic reviews that omit it miss eligible studies, and a bariatric-surgical literature with strong European registries is a plausible worst case for that omission; conference-only and registry-only evidence is invisible to this search entirely. The

direction of the resulting bias is unknown and unknowable from within this review, and the consequence is stated at full strength: missing studies could change the pooled estimate, the heterogeneity picture, the population of the modality strata, and the novelty claims of Chapter 15.2. This work is accordingly presented as a first pooled *attempt*, not a definitive synthesis.

Screening was performed by a single human reviewer with pipeline verification, not dual independent human review. Cochrane Handbook conduct guidance expects two independent screeners because single-screener recall is imperfect and idiosyncratic (PRISMA, which governs reporting rather than conduct, requires only that the process be disclosed — and it is). The pipeline’s second pass is a mitigation with a different error profile, not an equivalent; its recall against a human second reviewer has not been benchmarked beyond the gold cases reported in Appendix B.

The exposure data are dominated by construction proxies. The modality hierarchy of Section 6.9 was built precisely because stapler loads and bougie sizes are weak surrogates for volume, but a hierarchy does not conjure volumetric data into existence: the strata this dissertation most needs — imaged volume at construction — are the thinnest, and the flagship validation behind “near-reference” volumetry rests on twelve patients [19]. The artifact test is therefore asymmetric: it can weigh heavily against the proxies, but its power to vindicate a small effect in the volumetric stratum is limited by that stratum’s size.

The source cohorts lose patients non-randomly, and this review inherits every loss. The central long-term randomized comparison retained 47% of its cohort at nine years [16]. Attrition in bariatric follow-up is not missing-at-random — patients with poor outcomes disengage differently from patients with good ones — and any anatomical correlation computed on completers is exposed to survivorship distortion. Meta-analysis pools the published estimates; it cannot repair what the source cohorts never observed.

Extraction was LLM-assisted. Every extracted value is required to string-match a verbatim span in the stored source text, and every included study receives human review; this design catches fabrication and transcription error but not shared blind spots — a value that is correctly quoted but wrongly contextualized (a per-protocol number read as intention-to-treat, a nadir read as a fixed timepoint) can pass span verification, because span matching verifies presence, not semantic correctness. The audit that addresses this — a blinded human re-extraction of a pre-specified sample of pipeline extractions, with error rates reported alongside the gold-case evaluation in Appendix B — bounds, but does not eliminate, this risk.

The outcome metrics required harmonization, and harmonization is modeling. The conversion rules of Section 7.6 — between %EWL and %TWL, across regain definitions, across timepoints — inject assumptions into every pooled number. They are stated and justified there, but a reader who rejects them rejects the pooling.

Interpretation was pre-committed, but by the same hands that ran the analyses. The three lenses constrain hindsight; they do not constitute independence. A pre-registered interpretation written by the analyst is weaker than one adjudicated by parties blind to the data, and this dissertation had one analyst.

What remains after these subtractions is deliberately modest: a transparent, reproducible accounting of what the pouch-size literature can and cannot support, with its interpretation on record before its results. Chapter 15 states the answers to the seven questions at the confidence that accounting earns.

Chapter 15

Conclusions and future directions

This chapter closes the dissertation in three movements. First, it returns to the seven research questions posed in Section 1.6 and answers each one at the confidence the evidence has earned — no more, no less; every answer slot below that depends on the analyses of Chapters 10–13 carries an explicit [PENDING] marker until the corresponding result is final and GRADE-rated. Second, it names the dissertation’s original contributions as such, and separates them cleanly from confirmation of what was already known. Third, it states the forward program: a registry recommendation and the specification of the definitive trial. These last two are written in full here, deliberately — they follow from the structure of the evidence, not from its direction, and they hold under every result scenario pre-committed in Chapter 14.

15.1 The seven questions, answered

Table 15.1 summarizes the answers; the sections that follow give each question one page: what was asked, what the evidence showed, and the confidence rating with its reasons.

Table 15.1: The seven research questions and their answers at earned confidence. GRADE ratings per the methods of Chapter 8. All entries pending completion of the Part II analyses.

Question	Answer	GRADE
Q1 Smaller pouch → greater weight loss?	[PENDING]	[PENDING]
Q2 Dose-response gradient?	[PENDING]	[PENDING]
Q3 Measurement-modality dependence?	[PENDING]	[PENDING]
Q4 Why do the study designs conflict?	[PENDING]	[PENDING]
Q5 Safety trade-off in both directions?	[PENDING]	[PENDING]
Q6 Different relationship in OAGB?	[PENDING]	[PENDING]
Q7 Which model’s predictions match?	[PENDING]	[PENDING]

15.1.1 Q1 — Does a smaller constructed pouch produce greater weight loss?

This is the question the field has answered by habit since 1982, when a small reservoir was declared the critical determinant of a successful gastric bypass on a single-center series [5]. Forty years later the direct evidence still points in both directions: the largest cohort ever assembled — 14,168 registry patients with stapler length as the pouch-size proxy — found that pouch size predicted weight loss at six weeks but not at one year [14]; the Chinese randomized trial of Ren and colleagues — a volume contrast — found the *smaller* pouch ahead at twelve months [17]; the randomized length contrast found the *longer* pouch ahead at three years (31% versus 27% total weight loss, $p = 0.023$) [13]; and a six-year retrospective cohort in low-BMI diabetic patients

corroborated the small-pouch direction on both weight and glycemic control [18]. The two trials varied different physical quantities — volume versus length (Chapter 11) — so the conflict is partly definitional; the pooled estimate of Chapter 10 is the first quantitative adjudication of what remains.

Answer: [PENDING — pooled mean difference in %TWL, or SWiM synthesis if pooling failed its pre-stated criteria, from Chapter 10.] **Confidence:** [PENDING]

15.1.2 Q2 — Is there a gradient, or is any effect threshold-like?

A real causal effect of pouch size should show a gradient — or a defensible threshold — across the volumetric subset, the studies in which pouch size was actually measured rather than inferred from instruments. The prior literature offers a caution here: within a deliberately uniform small-pouch series — 231 patients whose pouches were construction-estimated at 10–20 mL against a calibration tube, not imaged — variation in size made no difference to one-year weight loss [105], which is what a threshold below the operative range, restriction of range in a compressed exposure, or no effect at all would look like.

Answer: [PENDING — dose-response analysis, Chapter 10.] **Confidence:** [PENDING]

15.1.3 Q3 — Does the answer depend on how pouch size was measured?

The measurement-artifact hypothesis is this dissertation’s signature test. Pouch size reaches the literature through four incompatible modalities with no reference standard, and the validation chain is thin: the agreement study between swallow MRI and 3D-CT — the two best candidates for a reference method — rests on twelve patients [19]. If the pouch-size effect lives only in the weak construction proxies and vanishes under imaged measurement — as the prospective volumetric cohort and the day-one imaging series already hint [157, 118] — then the most economical reading of forty years of disagreement is a measurement problem. The modality contrast that tests this is exploratory effect modification, not an adjudication: tier travels with era, design, and cohort size, and Chapter 11 weighs those competing explanations alongside the artifact reading.

Answer: [PENDING — modality-stratified estimates, Chapter 11.] **Confidence:** [PENDING]

15.1.4 Q4 — Why do the randomized trials disagree with the largest registry, and with each other?

The registry, the randomized trials, and the cohorts do not merely scatter around a small effect; they disagree in sign [14, 13, 18] — though the two randomized programs varied different physical quantities, volume versus length, so part of the apparent contradiction is definitional rather than empirical. The feature-by-feature dissection of Chapter 11 tested whether population, pouch definition, comparator distance, follow-up length, or attrition explains the disagreement — or whether it is irreducible on current data.

Answer: [PENDING — Chapter 11.] **Confidence:** [PENDING]

15.1.5 Q5 — How does construction size trade off against harm in both directions?

The best-measured safety result in this literature has a tight interval: in the Scandinavian registry, each additional centimeter of stapler used for the pouch raised the relative risk of marginal

ulcer — an ulcer at the gastrojejunal connection — by 14% (95% CI 9–20%) [14], and the small-pouch cohort likewise saw more ulcers with larger pouches [18]. Note the direction: on this evidence the ulcer risk scales with staple-line *length*, so it penalizes larger construction, not smaller — pointing the same way as any small-pouch benefit, which means it is no counterweight by itself. The harms plausibly attached to very small construction run through different outcomes entirely — anastomotic stricture, tension, leak, and food intolerance — and Section 13.1.4 analyzes each safety outcome separately, as absolute risks per outcome and time horizon, rather than folding both directions into a single frontier metric.

Answer: [PENDING — Section 13.1.4.] **Confidence:** [PENDING]

15.1.6 Q6 — Does the relationship differ in one-anastomosis gastric bypass?

OAGB varies pouch length far more widely than RYGB ever has, which makes it a natural stress test: an effect too small to see across RYGB’s compressed size range should widen with the anatomical variance.

Answer: [PENDING — Section 13.2.] **Confidence:** [PENDING]

15.1.7 Q7 — Which model’s predictions do the findings match?

Section 5.3 derived opposing predictions before the data were seen: under a purely hormonal model of the operation, constructed volume should not matter and any residual association should concentrate in the weak-proxy strata; under the emptying-dynamics amendment, a shape- or emptying-dependent effect survives even when volume per se does not. The empirical literature contains a direct observation on this hinge: in 67 patients with both 3D volumetry and scintigraphic emptying studies, it was the pouch’s *emptying rate*, not its volume as such, that tracked long-term weight loss [48].

Answer: [PENDING — Section 5.3 scored against Part II results in Chapter 14.]

Confidence: [PENDING]

15.2 Original contributions

Three contributions of this dissertation are intended to be new to the literature. Each is stated here as an intended contribution pending completed searching and synthesis — not as a finding already in hand — and each novelty claim rests on “none identified” by the searches of Chapter 8, within their stated limits (Embase was not searched), not on a guarantee:

1. **The first identified pooled attempt.** No meta-analysis of pouch size and weight loss after RYGB was identified; the single prior systematic review found its 14 studies too heterogeneous to pool [15]. Chapter 10 either supplies the missing estimate or converts that qualitative verdict into a quantitative, criterion-by-criterion anatomy of why pooling fails — new knowledge only to the extent that the anatomy is genuinely more specific than the verdict it extends.
2. **The measurement-modality analysis.** No prior analysis was identified that stratifies the pouch-size effect by measurement quality using a hierarchy fixed before the data were seen (Section 6.9).
3. **The dose-response analysis on the volumetric subset.** [PENDING — confirm no prior dose-response synthesis surfaced during screening.]

Any pattern found by mining the cross-study claims table beyond these pre-specified analyses is reported in Section 13.4 as exploratory and hypothesis-generating, and none of it is restated here as a conclusion.

15.3 The registry recommendation

Bariatric registries should record a measured pouch volume. This recommendation does not depend on which way the results fell, because it follows from the structure of the evidence rather than its direction. The field’s oldest technical question is currently unanswerable from what registries collect: the largest cohort analysis ever performed had to use total stapler length as its exposure — an instrument count standing in for an anatomical volume [14] — and the current state of practice makes even length unreliable. In the 2026 IFSO survey of 245 surgeons across 47 countries, only 53.5% routinely measured pouch length at all; the rest described their pouches subjectively, with terms like “short” and “small” anchored to no consistent anatomical landmark, and the survey’s own authors called for registries to capture pouch length and calibration details [7]. We endorse that call and sharpen it: length is still a proxy. Direct volumetry is feasible at clinical scale — 3D-CT pouch volumetry has been in clinical use since 2009 [154], has been applied prospectively with complete follow-up [157], and agrees with swallow MRI [19] — but that agreement study enrolled twelve patients, which is the entire foundation of the modality cross-validation. A registry field for measured volume at a fixed early timepoint (before dilation dominates the signal), plus a properly powered validation study of the imaging methods, would let the next cohort analysis ask the real question directly. How precise that field must be is a number, not an aspiration: the fidelity requirement — the measurement error a study or registry can tolerate and still detect a given true effect at a given scale — is derived by simulation in Chapter 12, from the documented tier errors and the realized designs of this literature. Its figures are provisional until the extraction-final design grid replaces the pre-extraction one, and they are therefore cited here by reference, not restated. Until then, every registry-scale answer inherits the proxy’s error.

15.4 The definitive trial

If the question merits a final answer — and the annual volume says it does: the IFSO worldwide survey counted 159,543 RYGB procedures in 2021, reported by only 57 of its 74 national societies and therefore an undercount [2] — the trial that settles it needs the following properties. Each requirement is forced by an identified failure of the existing evidence.

Population and design. Adults undergoing primary RYGB; parallel-group randomization to two construction targets. The prior review’s central conclusion was precisely that non-randomized data cannot resolve this [15].

Widely separated arms, volumetrically verified. The one existing pouch-length RCT randomized 5 cm versus 10 cm and showed the contrast is safe and surgically achievable [13]; the definitive trial should keep at least that separation. The causal exposure remains the *randomized construction target*, analyzed by intention to treat: a measured volume months after surgery is already downstream of dilation, emptying, adherence, and early weight loss — a post-randomization mediator that cannot replace assignment. Volumetry enters in two supporting roles: a standardized measurement as early as technically feasible, to verify that the randomized arms achieved separated volumes — with a measurement protocol whose error meets the fidelity requirement derived in Chapter 12 — and imaging at approximately three months — the timepoint at which prospective volumetry has been shown practical [157] — as a measured mediator for the mechanistic analysis. Without that verification the trial reproduces the proxy problem it exists to escape; with the exposure redefined as achieved volume it would acquire a worse one.

Outcomes. Percent total weight loss at one year, with a pre-specified co-primary analysis at three years or beyond: the existing RCT found no difference during the first two years and a significant divergence only in year three [13], so a one-year trial can be null and still miss the clinically operative effect. Reporting should follow the standardized bariatric outcome definitions

of the 2015 ASMBS reporting standard [172].

Safety. Marginal ulcer as a powered secondary endpoint, not an afterthought: the per-centimeter ulcer gradient is the best-measured effect in this literature [14], and it is what turns a null weight-loss result into an actionable conclusion — if size does not buy weight loss, shorter staple lines are pure gain.

Mechanistic substudy. An embedded emptying-scintigraphy or dynamic-MRI arm, because the strongest current signal is that emptying dynamics, not static volume, may carry the effect [48]; a trial that measures only volume could be null for the wrong reason.

Sample size. [PENDING — powered on the variance components and plausible effect size from Chapter 10; the pooled analysis exists in part to make this number computable for the first time.]

15.5 Closing

The gastric pouch is the most deliberately constructed artifact in the most studied operation in metabolic medicine, and its size has been chosen by inheritance since a 1982 declaration that was never tested at scale [5]. This dissertation set out to replace that inheritance with an estimate — pooled, stratified by measurement quality, and bounded by safety. [PENDING — one paragraph: the answer in three sentences, written when Part II is final.] What can be said already, and independently of the result, is this: the question is answerable, the field has simply never instrumented itself to answer it (a failure Chapter 12 quantifies as a detectability floor, and inverts into the fidelity requirement the instruments must meet), and the two instruments it needs — a registry that records what was actually built, and one trial that verifies what it randomized — are both specified above and both within reach.

References

1. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. *Lancet* 2024; 403:1027–50. DOI: 10.1016/S0140-6736(23)02750-2
2. Angrisani L, Santonicola A, Iovino P, Palma R, Kow L, Prager G, Ramos A, and Shikora S. IFSO Worldwide Survey 2020–2021: Current Trends for Bariatric and Metabolic Procedures. *Obesity Surgery* 2024; 34:1075–85. DOI: 10.1007/s11695-024-07118-3
3. Mason EE and Ito C. Gastric bypass in obesity. *Surgical Clinics of North America* 1967; 47:1345–51. DOI: 10.1016/s0039-6109(16)38384-0
4. Mason EE, Printen KJ, Hartford CE, and Boyd WC. Optimizing results of gastric bypass. *Annals of Surgery* 1975; 182:405–14. DOI: 10.1097/00000658-197510000-00006
5. Kirkpatrick JR and Siegel T. Critical determinants of a successful gastric bypass: reservoir versus stoma. *American Journal of Gastroenterology* 1982; 77:464–6
6. Madan AK, Harper JL, and Tichansky DS. Techniques of laparoscopic gastric bypass: on-line survey of American Society for Bariatric Surgery practicing surgeons. *Surgery for Obesity and Related Diseases* 2008; 4:166–72. DOI: 10.1016/j.soard.2007.08.006
7. Hany M, Zidan MH, Altabbaa H, Amgad A, Al-Momani H, Mosaad A, Bessa A, Hanna F, Zayed N, Hassan AY, Badawy O, Heih O, El-Masry H, Mourad S, El-Shamarka A, Abdou ME, Abouelnasr AA, Elkeleny MR, Torensma B, Alhashash M, Jaheen HMZ, Abokhozima A, and Oviedo RJ. Global Technical Variations in Roux-en-Y Gastric Bypass: A Worldwide Survey Among IFSO Members. *Obesity Surgery* 2026; 36:3535–57. DOI: 10.1007/s11695-026-08724-z
8. Eisenberg D, Shikora SA, Aarts E, Aminian A, Angrisani L, Cohen RV, De Luca M, Faria SL, Goodpaster KPS, Haddad A, Himpens JM, Kow L, Kurian M, Loi K, Mahawar K, Nimeri A, O’Kane M, Papasavas PK, Ponce J, Pratt JSA, Rogers AM, Steele KE, Suter M, and Kothari SN. 2022 American Society for Metabolic and Bariatric Surgery (ASMBS) and International Federation for the Surgery of Obesity and Metabolic Disorders (IFSO): Indications for Metabolic and Bariatric Surgery. *Surgery for Obesity and Related Diseases* 2022; 18. Co-published in *Obesity Surgery* 2023;33(1):3–14 (PMID 36336720):1345–56. DOI: 10.1016/j.soard.2022.08.013
9. Pucci A and Batterham RL. Mechanisms underlying the weight loss effects of RYGB and SG: similar, yet different. *Journal of Endocrinological Investigation* 2019; 42:117–28. DOI: 10.1007/s40618-018-0892-2
10. Dirksen C, Damgaard M, Bojsen-Møller KN, Jørgensen NB, Kielgast U, Jacobsen SH, Naver LS, Worm D, Holst JJ, Madsbad S, Hansen DL, and Madsen JL. Fast pouch emptying, delayed small intestinal transit, and exaggerated gut hormone responses after Roux-en-Y gastric bypass. *Neurogastroenterology and Motility* 2013; 25:346–e255. DOI: 10.1111/nmo.12087

11. Wilding JPH, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, McGowan BM, Rosenstock J, Tran MTD, Wadden TA, Wharton S, Yokote K, Zeuthen N, and Kushner RF. Once-Weekly Semaglutide in Adults with Overweight or Obesity. *New England Journal of Medicine* 2021; 384:989–1002. DOI: 10.1056/NEJMoa2032183
12. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, Kiyosue A, Zhang S, Liu B, Bunck MC, and Stefanski A. Tirzepatide Once Weekly for the Treatment of Obesity. *New England Journal of Medicine* 2022; 387:205–16. DOI: 10.1056/NEJMoa2206038
13. Boerboom A, Cooiman M, Aarts E, Aufenacker T, Hazebroek E, and Berends F. An Extended Pouch in a Roux-En-Y Gastric Bypass Reduces Weight Regain: 3-Year Results of a Randomized Controlled Trial. *Obesity Surgery* 2020; 30:3–10. DOI: 10.1007/s11695-019-04156-0
14. Edholm D, Ottosson J, and Sundbom M. Importance of pouch size in laparoscopic Roux-en-Y gastric bypass: a cohort study of 14,168 patients. *Surgical Endoscopy* 2016; 30:2011–5. DOI: 10.1007/s00464-015-4432-2
15. Mahawar K, Sharples AJ, and Graham Y. A systematic review of the effect of gastric pouch and/or gastrojejunostomy (stoma) size on weight loss outcomes with Roux-en-Y gastric bypass. *Surgical Endoscopy* 2020; 34:1048–60. DOI: 10.1007/s00464-019-07277-w
16. Harker MJR, Okkema S, Schuurman M, Heusschen L, Vugts G, and Hazebroek EJ. Extended Versus Standard Pouch in Roux-en-Y Gastric Bypass: Five To Nine Year Follow-Up Results of a Randomized Controlled Trial. *Obesity Surgery* 2026; 36:1271–80. DOI: 10.1007/s11695-026-08528-1
17. Ren Y, Yang W, Yang J, and Wang C. Effect of Roux-en-Y gastric bypass with different pouch size in Chinese T2DM patients with BMI 30–35 kg/m². *Obesity Surgery* 2015. DOI: 10.1007/s11695-014-1411-y
18. Gao X, Dai S, Wang G, Li W, Song Z, Su Z, Zhu S, Zhu L, and Li P. Large Versus Small Gastric Pouch for Roux-en-Y Gastric Bypass in Individuals With Type 2 Diabetes and a Body Mass Index < 35 kg/m²: Six-Year Outcomes. *Frontiers in Endocrinology* 2022; 13:913062. DOI: 10.3389/fendo.2022.913062
19. Felsenreich DM, Arnoldner MA, Langer FB, Bichler C, Vock N, Steinlechner K, Gachabayov M, Rojas A, Beitzke D, Mang T, Prager G, and Kulinna-Cosentini C. Swallow Magnetic Resonance Imaging Compared to 3D-Computed Tomography for Pouch Assessment and Hiatal Hernias After Roux-en-Y Gastric Bypass. *Obesity Surgery* 2020; 30:4192–7. DOI: 10.1007/s11695-020-04758-z
20. GBD 2015 Obesity Collaborators, Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine* 2017; 377:13–27. DOI: 10.1056/NEJMoa1614362
21. Ward ZJ, Bleich SN, Cradock AL, Barrett JL, Giles CM, Flax C, et al. Projected U.S. State-Level Prevalence of Adult Obesity and Severe Obesity. *New England Journal of Medicine* 2019; 381:2440–50. DOI: 10.1056/NEJMsa1909301
22. Rubino F, Cummings DE, Eckel RH, Cohen RV, Wilding JPH, Brown WA, et al. Definition and diagnostic criteria of clinical obesity. *Lancet Diabetes & Endocrinology* 2025; 13:221–62. DOI: 10.1016/S2213-8587(24)00316-4
23. Clapp B, Ponce J, Corbett J, Ghanem OM, Kurian M, Rogers AM, et al. American Society for Metabolic and Bariatric Surgery 2022 estimate of metabolic and bariatric procedures performed in the United States. *Surgery for Obesity and Related Diseases* 2024; 20:425–31. DOI: 10.1016/j.soard.2024.01.012

24. Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, and Friedman JM. Positional cloning of the mouse obese gene and its human homologue. *Nature* 1994; 372:425–32. DOI: 10.1038/372425a0
25. Schwartz MW, Seeley RJ, Zeltser LM, Drewnowski A, Ravussin E, Redman LM, et al. Obesity Pathogenesis: An Endocrine Society Scientific Statement. *Endocrine Reviews* 2017; 38:267–96. DOI: 10.1210/er.2017-00111
26. Leibel RL, Rosenbaum M, and Hirsch J. Changes in energy expenditure resulting from altered body weight. *New England Journal of Medicine* 1995; 332:621–8. DOI: 10.1056/NEJM199503093321001
27. Fothergill E, Guo J, Howard L, Kerns JC, Knuth ND, Brychta R, et al. Persistent metabolic adaptation 6 years after “The Biggest Loser” competition. *Obesity* 2016; 24:1612–9. DOI: 10.1002/oby.21538
28. Sumithran P, Prendergast LA, Delbridge E, Purcell K, Shulkes A, Kriketos A, and Proietto J. Long-term persistence of hormonal adaptations to weight loss. *New England Journal of Medicine* 2011; 365:1597–604. DOI: 10.1056/NEJMoa1105816
29. Look AHEAD Research Group, Wing RR, Bolin P, Brancati FL, Bray GA, Clark JM, et al. Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. *New England Journal of Medicine* 2013; 369:145–54. DOI: 10.1056/NEJMoa1212914
30. Yanovski SZ and Yanovski JA. Long-term drug treatment for obesity: a systematic and clinical review. *JAMA* 2014; 311:74–86. DOI: 10.1001/jama.2013.281361
31. Gasoyan H, Pfoh ER, Schulte R, Le P, and Rothberg MB. Early- and later-stage persistence with antiobesity medications: A retrospective cohort study. *Obesity* 2024; 32:486–93. DOI: 10.1002/oby.23952
32. Abu Dayyeh BK, Bazerbachi F, Vargas EJ, Sharaiha RZ, Thompson CC, Thaemert BC, et al. Endoscopic sleeve gastrectomy for treatment of class 1 and 2 obesity (MERIT): a prospective, multicentre, randomised trial. *Lancet* 2022; 400:441–51. DOI: 10.1016/S0140-6736(22)01280-6
33. Salminen P, Helmiö M, Ovaska J, Juuti A, Leivonen M, Peromaa-Haavisto P, et al. Effect of Laparoscopic Sleeve Gastrectomy vs Laparoscopic Roux-en-Y Gastric Bypass on Weight Loss at 5 Years Among Patients With Morbid Obesity: The SLEEVEPASS Randomized Clinical Trial. *JAMA* 2018; 319:241–54. DOI: 10.1001/jama.2017.20313
34. Peterli R, Wölnerhanssen BK, Peters T, Vetter D, Kröll D, Borbély Y, Schultes B, Beglinger C, Drewe J, Schiesser M, Nett P, and Bueter M. Effect of laparoscopic sleeve gastrectomy vs laparoscopic Roux-en-Y gastric bypass on weight loss in patients with morbid obesity: the SM-BOSS randomized clinical trial. *JAMA* 2018; 319:255–65. DOI: 10.1001/jama.2017.20897
35. Jastreboff AM, Kaplan LM, Frías JP, Wu Q, Du Y, Gurbuz S, et al. Triple-Hormone-Receptor Agonist Retatrutide for Obesity — A Phase 2 Trial. *New England Journal of Medicine* 2023; 389:514–26. DOI: 10.1056/NEJMoa2301972
36. Sjöström L, Narbro K, Sjöström CD, Karason K, Larsson B, Wedel H, et al. Effects of bariatric surgery on mortality in Swedish obese subjects. *New England Journal of Medicine* 2007; 357:741–52. DOI: 10.1056/NEJMoa066254
37. Sjöström L, Lindroos AK, Peltonen M, Torgerson J, Bouchard C, Carlsson B, et al. Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. *New England Journal of Medicine* 2004; 351:2683–93. DOI: 10.1056/NEJMoa035622
38. Adams TD, Davidson LE, Litwin SE, Kim J, Kolotkin RL, Nanjee MN, et al. Weight and Metabolic Outcomes 12 Years after Gastric Bypass. *New England Journal of Medicine* 2017; 377:1143–55. DOI: 10.1056/NEJMoa1700459

39. Lincoff AM, Brown-Frandsen K, Colhoun HM, Deanfield J, Emerson SS, Esbjerg S, et al. Semaglutide and Cardiovascular Outcomes in Obesity without Diabetes. *New England Journal of Medicine* 2023; 389:2221–32. DOI: 10.1056/NEJMoa2307563
40. Longitudinal Assessment of Bariatric Surgery (LABS) Consortium, Flum DR, Belle SH, King WC, Wahed AS, Berk P, et al. Perioperative safety in the longitudinal assessment of bariatric surgery. *New England Journal of Medicine* 2009; 361:445–54. DOI: 10.1056/NEJMoa0901836
41. Wilding JPH, Batterham RL, Davies M, Van Gaal LF, Kandler K, Konakli K, et al. Weight regain and cardiometabolic effects after withdrawal of semaglutide: The STEP 1 trial extension. *Diabetes, Obesity and Metabolism* 2022; 24:1553–64. DOI: 10.1111/dom.14725
42. Aronne LJ, Sattar N, Horn DB, Bays HE, Wharton S, Lin WY, et al. Continued Treatment With Tirzepatide for Maintenance of Weight Reduction in Adults With Obesity: The SURMOUNT-4 Randomized Clinical Trial. *JAMA* 2024; 331:38–48. DOI: 10.1001/jama.2023.24945
43. Agboola F, Fahim SM, Lin GA, Lee W, Richardson M, and Rind D. Semaglutide and tirzepatide for obesity: a summary from the Institute for Clinical and Economic Review’s New England Comparative Effectiveness Public Advisory Council. *Journal of Managed Care & Specialty Pharmacy* 2026; 32:631–6. DOI: 10.18553/jmcp.2026.32.5.631
44. Dhippayom T, Meraz M, Lee H, Hur C, Inadomi JM, Veettil SK, Dunn JD, and Chaiyakunapruk N. GLP-1 receptor agonists for treating obesity without diabetes: a systematic review and meta-analysis of economic evaluations. *Diabetes, Obesity and Metabolism* 2026; 28:1339–49. DOI: 10.1111/dom.70322
45. Cummings DE, Weigle DS, Frayo RS, Breen PA, Ma MK, Dellinger EP, and Purnell JQ. Plasma ghrelin levels after diet-induced weight loss or gastric bypass surgery. *New England Journal of Medicine* 2002; 346:1623–30. DOI: 10.1056/NEJMoa012908
46. Roux CW le, Aylwin SJ, Batterham RL, Borg CM, Coyle F, Prasad V, et al. Gut hormone profiles following bariatric surgery favor an anorectic state, facilitate weight loss, and improve metabolic parameters. *Annals of Surgery* 2006; 243:108–14. DOI: 10.1097/01.sla.0000183349.16877.84
47. Salehi M and D’Alessio DA. Mechanisms of surgical control of type 2 diabetes: GLP-1 is the key factor—Maybe. *Surgery for Obesity and Related Diseases* 2016; 12:1230–5
48. Riccioppo D, Santo MA, Rocha M, Buchpiguel CA, Diniz MA, Pajecki D, Cleva R de, and Kawamoto F. Small-Volume, Fast-Emptying Gastric Pouch Leads to Better Long-Term Weight Loss and Food Tolerance After Roux-en-Y Gastric Bypass. *Obesity Surgery* 2018; 28:693–701
49. Schauer PR, Bhatt DL, Kirwan JP, Wolski K, Aminian A, Brethauer SA, et al. Bariatric Surgery versus Intensive Medical Therapy for Diabetes — 5-Year Outcomes. *New England Journal of Medicine* 2017; 376:641–51. DOI: 10.1056/NEJMoa1600869
50. Mingrone G, Panunzi S, De Gaetano A, Guidone C, Iaconelli A, Capristo E, et al. Metabolic surgery versus conventional medical therapy in patients with type 2 diabetes: 10-year follow-up of an open-label, single-centre, randomised controlled trial. *Lancet* 2021; 397:293–304. DOI: 10.1016/S0140-6736(20)32649-0
51. Courcoulas AP, Patti ME, Hu B, Arterburn DE, Simonson DC, Gourash WF, et al. Long-Term Outcomes of Medical Management vs Bariatric Surgery in Type 2 Diabetes. *JAMA* 2024; 331:654–64. DOI: 10.1001/jama.2024.0318
52. Carlsson LMS, Sjöholm K, Jacobson P, Andersson-Assarsson JC, Svensson PA, Taube M, et al. Life Expectancy after Bariatric Surgery in the Swedish Obese Subjects Study. *New England Journal of Medicine* 2020; 383:1535–43. DOI: 10.1056/NEJMoa2002449

53. Adams TD, Gress RE, Smith SC, Halverson RC, Simper SC, Rosamond WD, et al. Long-term mortality after gastric bypass surgery. *New England Journal of Medicine* 2007; 357:753–61. DOI: 10.1056/NEJMoa066603
54. Syn NL, Cummings DE, Wang LZ, Lin DJ, Zhao JJ, Loh M, et al. Association of metabolic-bariatric surgery with long-term survival in adults with and without diabetes: a one-stage meta-analysis of matched cohort and prospective controlled studies with 174 772 participants. *Lancet* 2021; 397:1830–41. DOI: 10.1016/S0140-6736(21)00591-2
55. Parrott J, Frank L, Rabena R, Craggs-Dino L, Isom KA, and Greiman L. American Society for Metabolic and Bariatric Surgery Integrated Health Nutritional Guidelines for the Surgical Weight Loss Patient 2016 Update: Micronutrients. *Surgery for Obesity and Related Diseases* 2017; 13:727–41. DOI: 10.1016/j.soard.2016.12.018
56. Bjerkan KK, Sandvik J, Nymo S, Græslie H, Johnsen G, Mårvik R, Hyldmo ÅA, Kulseng BE, Sommersteth S, Høydal KL, and Hoff DAL. Vitamin and Mineral Deficiency 12 Years After Roux-en-Y Gastric Bypass: a Cross-Sectional Multicenter Study. *Obesity Surgery* 2023; 33:3178–85. DOI: 10.1007/s11695-023-06787-w
57. Azagury DE, Abu Dayyeh BK, Greenwalt IT, and Thompson CC. Marginal ulceration after Roux-en-Y gastric bypass surgery: characteristics, risk factors, treatment, and outcomes. *Endoscopy* 2011; 43:950–4. DOI: 10.1055/s-0030-1256951
58. Schäfer A, Gehwolf P, Kienzl-Wagner K, Cakar-Beck F, and Wykypiel H. Linear or circular: anastomotic ulcer after gastric bypass surgery. *Surgical Endoscopy* 2022; 36:3011–8. DOI: 10.1007/s00464-021-08597-6
59. Kremen AJ, Linner JH, and Nelson CH. An experimental evaluation of the nutritional importance of proximal and distal small intestine. *Annals of Surgery* 1954; 140:439–48. DOI: 10.1097/00000658-195409000-00018
60. Payne JH and DeWind LT. Surgical treatment of obesity. *American Journal of Surgery* 1969; 118:141–7. DOI: 10.1016/0002-9610(69)90113-5
61. Alden JF. Gastric and jejunoileal bypass. A comparison in the treatment of morbid obesity. *Archives of Surgery* 1977; 112:799–806. DOI: 10.1001/archsurg.1977.01370070011001
62. Griffen WO, Young VL, and Stevenson CC. A prospective comparison of gastric and jejunoileal bypass procedures for morbid obesity. *Annals of Surgery* 1977; 186:500–9. DOI: 10.1097/00000658-197710000-00012
63. Buckwalter JA. Clinical trial of jejunoileal and gastric bypass for the treatment of morbid obesity: four-year progress report. *The American Surgeon* 1980; 46:377–81
64. Dobbins JW and Binder HJ. Importance of the colon in enteric hyperoxaluria. *New England Journal of Medicine* 1977; 296:298–301. DOI: 10.1056/NEJM197702102960602
65. Kollmann L, Lauerer J, Vladimirov M, Germer CT, and Seyfried F. Postoperative long-term complications after intestinal bypass surgery: internal hernia, anastomotic ulcer, chronic malabsorption. *Chirurgie* 2025; 96. German:989–95. DOI: 10.1007/s00104-025-02389-7
66. Scopinaro N, Gianetta E, Civalleri D, Bonalumi U, and Bachi V. Bilio-pancreatic bypass for obesity: II. Initial experience in man. *British Journal of Surgery* 1979; 66:618–20. DOI: 10.1002/bjs.1800660906
67. Hess DS and Hess DW. Biliopancreatic diversion with a duodenal switch. *Obesity Surgery* 1998; 8:267–82. DOI: 10.1381/096089298765554476
68. Halverson JD and Koehler RE. Gastric bypass: analysis of weight loss and factors determining success. *Surgery* 1981; 90:446–55

69. Smith CD, Herkes SB, Behrns KE, Fairbanks VF, Kelly KA, and Sarr MG. Gastric acid secretion and vitamin B12 absorption after vertical Roux-en-Y gastric bypass for morbid obesity. *Annals of Surgery* 1993; 218:91–6. DOI: 10.1097/00000658-199307000-00014
70. Saarinen T, Pietiläinen KH, Loimaala A, Ihalainen T, Sammalkorpi H, Penttilä A, and Juuti A. Bile reflux is a common finding in the gastric pouch after one anastomosis gastric bypass. *Obesity Surgery* 2020; 30:875–81. DOI: 10.1007/s11695-019-04353-x
71. Ben Amor I, Petrucciani N, Kassir R, Al Munifi A, Piche T, Debs T, and Gugenheim J. Laparoscopic conversion of one anastomosis gastric bypass to a standard Roux-en-Y gastric bypass. *Obesity Surgery* 2017; 27:1398. DOI: 10.1007/s11695-017-2646-1
72. Halverson JD, Zuckerman GR, Koehler RE, Gentry K, Michael HE, and DeSchryver-Kecskemeti K. Gastric bypass for morbid obesity: a medical–surgical assessment. *Annals of Surgery* 1981; 194:152–60. DOI: 10.1097/00000658-198108000-00007
73. Knecht BH. Mason gastric bypass. Long-term follow-up and comparison with other gastric procedures. *American Journal of Surgery* 1983; 145:604–8. DOI: 10.1016/0002-9610(83)90102-2
74. Torres JC, Oca CF, and Garrison RN. Gastric bypass: Roux-en-Y gastrojejunostomy from the lesser curvature. *Southern Medical Journal* 1983; 76:1217–21
75. Printen KJ and Mason EE. Gastric surgery for relief of morbid obesity. *Archives of Surgery* 1973; 106:428–31. DOI: 10.1001/archsurg.1973.01350160046007
76. Mason EE. Morbid obesity: use of vertical banded gastroplasty. *Surgical Clinics of North America* 1987; 67:521–37. DOI: 10.1016/s0039-6109(16)44230-1
77. Pace WG, Martin EW, Tetirick T, Fabri PJ, and Carey LC. Gastric partitioning for morbid obesity. *Annals of Surgery* 1979; 190:392–400. DOI: 10.1097/00000658-197909000-00015
78. Andersen T and Pedersen BH. Pouch volume, stoma diameter, and clinical outcome after gastroplasty for morbid obesity. A prospective study. *Scandinavian Journal of Gastroenterology* 1984; 19:643–9
79. Mason EE. Vertical banded gastroplasty for obesity. *Archives of Surgery* 1982; 117:701–6. DOI: 10.1001/archsurg.1982.01380290147026
80. Näslund I, Wickbom G, Christoffersson E, and Agren G. A prospective randomized comparison of gastric bypass and gastroplasty. Complications and early results. *Acta Chirurgica Scandinavica* 1986; 152:681–9
81. Wezenbeek MR van, Smulders JF, Zoete JPJGM de, Luyer MD, Montfort G van, and Nienhuijs SW. Long-term results of primary vertical banded gastroplasty. *Obesity Surgery* 2015; 25:1425–30. DOI: 10.1007/s11695-014-1543-0
82. NIH Consensus Development Conference Panel. NIH conference. Gastrointestinal surgery for severe obesity. *Annals of Internal Medicine* 1991; 115:956–61
83. Wittgrove AC, Clark GW, and Tremblay LJ. Laparoscopic gastric bypass, Roux-en-Y: preliminary report of five cases. *Obesity Surgery* 1994; 4:353–7. DOI: 10.1381/096089294765558331
84. Ali MR, Sugerman HJ, and DeMaria EJ. Techniques of laparoscopic Roux-en-Y gastric bypass. *Seminars in Laparoscopic Surgery* 2002; 9:94–104
85. Sundbom M and Gustavsson S. Hand-assisted laparoscopic Roux-en-Y gastric bypass: aspects of surgical technique and early results. *Obesity Surgery* 2000; 10:420–7. DOI: 10.1381/096089200321594282

86. Regan JP, Inabnet WB, Gagner M, and Pomp A. Early experience with two-stage laparoscopic Roux-en-Y gastric bypass as an alternative in the super-super obese patient. *Obesity Surgery* 2003; 13:861–4. DOI: 10.1381/096089203322618669
87. Alexandrou A, Felekouras E, Giannopoulos A, Tsigris C, and Diamantis T. What is the actual fate of super-morbid-obese patients who undergo laparoscopic sleeve gastrectomy as the first step of a two-stage weight-reduction operative strategy? *Obesity Surgery* 2012; 22:1623–8. DOI: 10.1007/s11695-012-0718-9
88. Brown WA, Liem R, Al-Sabah S, Anvari M, Boza C, Cohen RV, Ghaferi A, Våge V, Himpens J, Kow L, Morton J, Musella M, Pattou F, Sakran N, Clapp B, Prager G, Shikora S, and IFSO Global Registry Collaboration. Metabolic bariatric surgery across the IFSO chapters: key insights on the baseline patient demographics, procedure types, and mortality from the Eighth IFSO Global Registry Report. *Obesity Surgery* 2024; 34:1764–77. DOI: 10.1007/s11695-024-07196-3
89. Jeon D, Yoo M, Kang SH, Kim TH, Suh YS, and Park YS. Beyond the procedure: revisiting the technical determinants of sleeve gastrectomy and Roux-en-Y gastric bypass. *Journal of Metabolic and Bariatric Surgery* 2026; 15:20–7. DOI: 10.17476/jmbs.2026.15.1.20
90. Rutledge R. The mini-gastric bypass: experience with the first 1,274 cases. *Obesity Surgery* 2001; 11:276–80. DOI: 10.1381/096089201321336584
91. Rutledge R, Kular K, and Manchanda N. The Mini-Gastric Bypass original technique. *International Journal of Surgery* 2019; 61:38–41. DOI: 10.1016/j.ijssu.2018.10.042
92. Carbajo M, García-Caballero M, Toledano M, Osorio D, García-Lanza C, and Carmona JA. One-anastomosis gastric bypass by laparoscopy: results of the first 209 patients. *Obesity Surgery* 2005; 15:398–404. DOI: 10.1381/0960892053576677
93. Mahawar KK, Himpens J, Shikora SA, Chevallier JM, Lakdawala M, De Luca M, Weiner R, Khammas A, Kular KS, Musella M, Prager G, Mirza MK, Carbajo M, Kow L, Lee WJ, and Small PK. The first consensus statement on one anastomosis/mini gastric bypass (OAGB/MGB) using a modified Delphi approach. *Obesity Surgery* 2018; 28:303–12. DOI: 10.1007/s11695-017-3070-2
94. McCulloch P, Altman DG, Campbell WB, Flum DR, Glasziou P, Marshall JC, Nicholl J, and Balliol Collaboration. No surgical innovation without evaluation: the IDEAL recommendations. *The Lancet* 2009; 374:1105–12. DOI: 10.1016/S0140-6736(09)61116-8
95. Kassir R, Blanc P, Lointier P, Tiffet O, Breton C, Ben Amor I, Iannelli A, and Gugenheim J. An analysis of surgical anatomy of the gastric fundus in bariatric surgery: why the gastric pouch expands? A point of technique. *International Journal of Surgery* 2014; 12:1151–6. DOI: 10.1016/j.ijssu.2014.08.409
96. Sapala JA, Wood MH, Sapala MA, and Flake Jr. TM. Marginal ulcer after gastric bypass: a prospective 3-year study of 173 patients. *Obesity Surgery* 1998; 8:505–16. DOI: 10.1381/096089298765554061
97. Siilin H, Wanders A, Gustavsson S, and Sundbom M. The proximal gastric pouch invariably contains acid-producing parietal cells in Roux-en-Y gastric bypass. *Obesity Surgery* 2005; 15:771–7. DOI: 10.1381/0960892054222849
98. Hedberg J, Hedenstrom H, Nilsson S, Sundbom M, and Gustavsson S. Role of gastric acid in stomal ulcer after gastric bypass. *Obesity Surgery* 2005; 15:1375–8. DOI: 10.1381/096089205774859380

99. Fakhry J, Stebbing MJ, Hunne B, Bayguinov Y, Ward SM, Sasse KC, Callaghan B, McQuade RM, and Furness JB. Relationships of endocrine cells to each other and to other cell types in the human gastric fundus and corpus. *Cell and Tissue Research* 2019; 376:37–49. DOI: 10.1007/s00441-018-2957-0
100. Kim HH, Jeon TY, Park DY, Kim YJ, Lee SY, Lee JY, Lee JG, Jeong DW, Yi YH, Cho YH, Im SJ, Bae MJ, and Choi EJ. Differential expression of ghrelin mRNA according to anatomical portions of human stomach. *Hepato-Gastroenterology* 2012; 59:2217–21. DOI: 10.5754/hge11865
101. Capella JF and Capella RF. Gastro-gastric fistulas and marginal ulcers in gastric bypass procedures for weight reduction. *Obesity Surgery* 1999; 9:22–7, discussion 28. DOI: 10.1381/096089299765553674
102. Sapala JA, Wood MH, Sapala MA, Schuhknecht MP, and Flake Jr. TM. The micropouch gastric bypass: technical considerations in primary and revisionary operations. *Obesity Surgery* 2001; 11:3–17. DOI: 10.1381/096089201321454042
103. Roberts K, Duffy A, Kaufman J, Burrell M, Dziura J, and Bell R. Size matters: gastric pouch size correlates with weight loss after laparoscopic Roux-en-Y gastric bypass. *Surgical Endoscopy* 2007; 21:1397–402. DOI: 10.1007/s00464-007-9232-x
104. Campos GM, Rabl C, Mulligan K, Posselt A, Rogers SJ, Westphalen AC, Lin F, and Vittinghoff E. Factors associated with weight loss after gastric bypass. *Archives of Surgery* 2008; 143:877–83, discussion 884. DOI: 10.1001/archsurg.143.9.877
105. O'Connor EA and Carlin AM. Lack of correlation between variation in small-volume gastric pouch size and weight loss after laparoscopic Roux-en-Y gastric bypass. *Surgery for Obesity and Related Diseases* 2008; 4:399–403. DOI: 10.1016/j.soard.2007.12.007
106. Madan AK, Tichansky DS, and Phillips JC. Does pouch size matter? *Obesity Surgery* 2007; 17:317–20. DOI: 10.1007/s11695-007-9058-6
107. Mahawar KK. It's the Width Not the Size of the Pouch That Matters. *Obesity Surgery* 2020; 30:1132–3. DOI: 10.1007/s11695-019-04315-3
108. Safari S, Ekramnia I, Chehresonbolli Y, Ahmadi V, and Kermansaravi M. Outcomes and complications after long versus short gastric pouch Roux-en-Y gastric bypass in patients with severe obesity. *Scientific Reports* 2024; 14:31012. DOI: 10.1038/s41598-024-82200-5
109. Ribeiro R, Pouwels S, Parmar C, Pereira J, Manacas L, Guerra A, Borges N, Ribeiro J, and Viveiros O. Outcomes of Long Pouch Gastric Bypass (LPGB): 4-Year Experience in Primary and Revision Cases. *Obesity Surgery* 2019; 29:3665–71. DOI: 10.1007/s11695-019-04051-8
110. Stano S, Alam F, Wu L, Dutia R, Ng SN, Sala M, McGinty J, and LaFerrere B. Effect of meal size and texture on gastric pouch emptying and glucagon-like peptide 1 after gastric bypass surgery. *Surgery for Obesity and Related Diseases* 2017; 13:1975–83. DOI: 10.1016/j.soard.2017.09.004
111. Akkary E, Sidani S, Boonsiri J, Yu S, Dziura J, Duffy AJ, and Bell RL. The paradox of the pouch: prompt emptying predicts improved weight loss after laparoscopic Roux-Y gastric bypass. *Surgical Endoscopy* 2009; 23:790–4. DOI: 10.1007/s00464-008-0069-8
112. Deden LN, Cooiman MI, Aarts EO, Janssen IMC, Gotthardt M, Hendrickx BW, and Berends FJ. Gastric pouch emptying of solid food in patients with successful and unsuccessful weight loss after Roux-en-Y gastric bypass surgery. *Surgery for Obesity and Related Diseases* 2017; 13:1840–6. DOI: 10.1016/j.soard.2017.07.031

113. Horowitz M, Cook DJ, Collins PJ, Harding PE, Hooper MJ, Walsh JF, and Shearman DJ. Measurement of gastric emptying after gastric bypass surgery using radionuclides. *British Journal of Surgery* 1982; 69:655–7. DOI: 10.1002/bjs.1800691108
114. Naslund I and Beckman KW. Gastric emptying rate after gastric bypass and gastroplasty. *Scandinavian Journal of Gastroenterology* 1987; 22:193–201. DOI: 10.3109/00365528708991879
115. Heneghan HM, Yimcharoen P, Brethauer SA, Kroh M, and Chand B. Influence of pouch and stoma size on weight loss after gastric bypass. *Surgery for Obesity and Related Diseases* 2012; 8:408–15. DOI: 10.1016/j.soard.2011.09.010
116. Yimcharoen P, Heneghan HM, Singh M, Brethauer S, Schauer P, Rogula T, Kroh M, and Chand B. Endoscopic findings and outcomes of revisional procedures for patients with weight recidivism after gastric bypass. *Surgical Endoscopy* 2011; 25:3345–52. DOI: 10.1007/s00464-011-1723-0
117. Abu Dayyeh BK, Jirapinyo P, and Thompson CC. Plasma Ghrelin Levels and Weight Regain After Roux-en-Y Gastric Bypass Surgery. *Obesity Surgery* 2017; 27:1031–6. DOI: 10.1007/s11695-016-2418-3
118. Nishie A, Brown B, Barloon T, Kuehn D, and Samuel I. Comparison of size of proximal gastric pouch and short-term weight loss following routine upper gastrointestinal contrast study after laparoscopic Roux-en-Y gastric bypass. *Obesity Surgery* 2007; 17:1183–8. DOI: 10.1007/s11695-007-9200-5
119. Topart P, Becouarn G, and Ritz P. Pouch size after gastric bypass does not correlate with weight loss outcome. *Obesity Surgery* 2011; 21:1350–4. DOI: 10.1007/s11695-011-0460-8
120. Browning KN. Stress-induced modulation of vagal afferents. *Neurogastroenterology and Motility* 2019; 31:e13758. DOI: 10.1111/nmo.13758
121. Lin E, Gletsu N, Fugate K, McClusky D, Gu LH, Zhu JL, Ramshaw BJ, Papanicolaou DA, Ziegler TR, and Smith CD. The effects of gastric surgery on systemic ghrelin levels in the morbidly obese. *Archives of Surgery* 2004; 139:780–4. DOI: 10.1001/archsurg.139.7.780
122. Carrasco F, Rojas P, Csendes A, Codoceo J, Inostroza J, Basfi-fer K, Papapietro K, Watkins G, Rojas J, and Ruz M. Changes in ghrelin concentrations one year after resective and non-resective gastric bypass: associations with weight loss and energy and macronutrient intakes. *Nutrition* 2012; 28:757–61. DOI: 10.1016/j.nut.2011.11.004
123. Xu HC, Pang YC, Chen JW, Cao JY, Sheng Z, Yuan JH, Wang R, Zhang CS, Wang LX, and Dong J. Systematic Review and Meta-analysis of the Change in Ghrelin Levels After Roux-en-Y Gastric Bypass. *Obesity Surgery* 2019; 29:1343–51. DOI: 10.1007/s11695-018-03686-3
124. Xu R, Zhu C, Pierre JF, and Yin DP. Gastric Bypass Improves Obesity and Glucose Tolerance Independent of Gastric Pouch Size. *Obesity Surgery* 2020; 30:1635–41. DOI: 10.1007/s11695-020-04403-9
125. Coblijn UK, Goucham AB, Lagarde SM, Kuiken SD, and Wagensveld BA van. Development of ulcer disease after Roux-en-Y gastric bypass, incidence, risk factors, and patient presentation: a systematic review. *Obesity Surgery* 2014; 24:299–309. DOI: 10.1007/s11695-013-1118-5
126. Wennerlund J, Gunnarsson U, Strigard K, and Sundbom M. Acid-related complications after laparoscopic Roux-en-Y gastric bypass: risk factors and impact of proton pump inhibitors. *Surgery for Obesity and Related Diseases* 2020; 16:620–5. DOI: 10.1016/j.soard.2020.01.005

127. Finks JF, Carlin A, Share D, O'Reilly A, Fan Z, Birkmeyer J, and Birkmeyer N. Effect of surgical techniques on clinical outcomes after laparoscopic gastric bypass—results from the Michigan Bariatric Surgery Collaborative. *Surgery for Obesity and Related Diseases* 2011; 7:284–9. DOI: 10.1016/j.soard.2010.10.004
128. Ahmad J, Martin J, Ikramuddin S, Schauer P, and Slivka A. Endoscopic balloon dilation of gastroenteric anastomotic stricture after laparoscopic gastric bypass. *Endoscopy* 2003; 35:725–8. DOI: 10.1055/s-2003-41579
129. Khoraki J, Funk LM, Greenberg JA, Levenson G, and Campos GM. The Effect of Route of Anvil Insertion on Stricture Rates with Circular Stapled Gastrojejunostomy During Laparoscopic Gastric Bypass. *Obesity Surgery* 2016; 26:517–24. DOI: 10.1007/s11695-015-1782-8
130. Scarpellini E, Arts J, Karamanolis G, Laurenus A, Siquini W, Suzuki H, Ukleja A, Van Beek A, Vanuytsel T, Bor S, Ceppa E, Di Lorenzo C, Emous M, Hammer H, Hellstrom P, Laville M, Lundell L, Masclee A, Ritz P, and Tack J. International consensus on the diagnosis and management of dumping syndrome. *Nature Reviews Endocrinology* 2020; 16:448–66. DOI: 10.1038/s41574-020-0357-5
131. Frank AL, Mohli I, Kaur N, and Arnold W. Severe Dumping and Persistent Obesity in a Post-Roux-en-Y Gastric Bypass Patient With a Non-discernible Gastric Pouch and Symptomatic Improvement on Tirzepatide: A Case Report. *Cureus* 2026; 18:e113316. DOI: 10.7759/cureus.113316
132. Tase A, Aly M, Adil MT, Munasinghe A, Rashid F, Jambulingam P, Whitelaw D, Jain V, Al-Ta'an O, and Askari A. Use of transoral outlet reduction endoscopy (TORE) in the management of resistant dumping syndrome. *Surgical Endoscopy* 2026; 40:3304–8. DOI: 10.1007/s00464-026-12620-z
133. Odstreil EA, Martinez JG, Santa Ana CA, Xue B, Schneider RE, Steffer KJ, Porter JL, Asplin J, Kuhn JA, and Fordtran JS. The contribution of malabsorption to the reduction in net energy absorption after long-limb Roux-en-Y gastric bypass. *The American Journal of Clinical Nutrition* 2010; 92:704–13. DOI: 10.3945/ajcn.2010.29870
134. Stenström B, Furnes MW, Tømmerås K, Syversen U, Zhao CM, and Chen D. Mechanism of gastric bypass-induced body weight loss: one-year follow-up after micro-gastric bypass in rats. *Journal of Gastrointestinal Surgery* 2006; 10:1384–91. DOI: 10.1016/j.gassur.2006.07.005
135. Hao Z, Mumphy MB, Townsend RL, Morrison CD, Münzberg H, Ye J, and Berthoud HR. Body composition, food intake, and energy expenditure in a murine model of Roux-en-Y gastric bypass surgery. *Obesity Surgery* 2016; 26:2173–82. DOI: 10.1007/s11695-016-2062-y
136. Ryan KK, Tremaroli V, Clemmensen C, Kovatcheva-Datchary P, Myronovych A, Karns R, Wilson-Pérez HE, Sandoval DA, Kohli R, Bäckhed F, and Seeley RJ. FXR is a molecular target for the effects of vertical sleeve gastrectomy. *Nature* 2014; 509:183–8. DOI: 10.1038/nature13135
137. Dutia R, Embrey M, O'Brien CS, Haeusler RA, Agénor KK, Homel P, McGinty J, Vincent RP, Alaghband-Zadeh J, Staels B, Roux CW le, Yu J, and Laferrère B. Temporal changes in bile acid levels and 12 α -hydroxylation after Roux-en-Y gastric bypass surgery in type 2 diabetes. *International Journal of Obesity* 2015; 39:806–13. DOI: 10.1038/ijo.2015.1
138. Liou AP, Paziuk M, Luevano Jr. JM, Machineni S, Turnbaugh PJ, and Kaplan LM. Conserved shifts in the gut microbiota due to gastric bypass reduce host weight and adiposity. *Science Translational Medicine* 2013; 5:178ra41. DOI: 10.1126/scitranslmed.3005687

139. Scholtz S, Miras AD, Chhina N, Prectl CG, Sleeth ML, Daud NM, Ismail NA, Durighel G, Ahmed AR, Olbers T, Vincent RP, Alagband-Zadeh J, Ghatei MA, Waldman AD, Frost GS, Bell JD, Roux CW le, and Goldstone AP. Obese patients after gastric bypass surgery have lower brain-hedonic responses to food than after gastric banding. *Gut* 2014; 63:891–902. DOI: 10.1136/gutjnl-2013-305008
140. Olbers T, Björkman S, Lindroos A, Maleckas A, Lönn L, Sjöström L, and Lönroth H. Body composition, dietary intake, and energy expenditure after laparoscopic Roux-en-Y gastric bypass and laparoscopic vertical banded gastroplasty: a randomized clinical trial. *Annals of Surgery* 2006; 244:715–22. DOI: 10.1097/01.sla.0000218085.25902.f8
141. Seyfried F, Miras AD, Bueter M, Prectl CG, Spector AC, and Roux CW le. Effects of preoperative exposure to a high-fat versus a low-fat diet on ingestive behavior after gastric bypass surgery in rats. *Surgical Endoscopy* 2013; 27:4192–201. DOI: 10.1007/s00464-013-3020-6
142. Roux CW le, Bueter M, Theis N, Werling M, Ashrafi H, Löwenstein C, Athanasiou T, Bloom SR, Spector AC, Olbers T, and Lutz TA. Gastric bypass reduces fat intake and preference. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 2011; 301:R1057–R1066. DOI: 10.1152/ajpregu.00139.2011
143. Wilson-Pérez HE, Chambers AP, Sandoval DA, Stefater MA, Woods SC, Benoit SC, and Seeley RJ. The effect of vertical sleeve gastrectomy on food choice in rats. *International Journal of Obesity* 2013; 37:288–95. DOI: 10.1038/ijo.2012.18
144. Boerboom AB, Berends FJ, and Aarts EO. It is the width not the size of the pouch that matters. *Obesity Surgery* 2020; 30. Letter (reply):1134–5. DOI: 10.1007/s11695-019-04316-2
145. Arnoldner MA, Felsenreich DM, Langer FB, Weber M, Mang T, Kulinna-Cosentini C, and Prager G. Pouch volume and pouch migration after Roux-en-Y gastric bypass: a comparison of gastroscopy and 3 D-CT volumetry: is there a "migration crisis"? *Surgery for Obesity and Related Diseases* 2020; 16:1902–8. DOI: 10.1016/j.soard.2020.07.024
146. Zhang J, Lin Y, Dong Z, Wang C, Huang W, Cheng Z, Xia Z, Li Y, Wei Z, and Yang W. A national, multicenter, survey-based analysis of clinical practice and utilization of bougie in metabolic and bariatric surgery. *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques* 2025; 35:e1385. DOI: 10.1097/SLE.0000000000001385
147. Reiber BMM, Tenhagen M, Hunfeld MAJM, Cense HA, and Demirkiran A. Calibration of the gastric pouch in laparoscopic Roux-en-Y gastric bypass: does it matter? The influence on weight loss. *Obesity Surgery* 2018; 28:3400–4. DOI: 10.1007/s11695-018-3352-3
148. Haghighat N, Kamran H, Moaddeli MN, Hosseini B, Karimi A, Hesameddini I, Amini M, Hosseini SV, Vahidi A, and Moeinvaziri N. The impact of gastric pouch size, based on the number of staplers, on the short-term weight outcomes of Roux-en-Y gastric bypass. *Annals of Medicine and Surgery* 2022; 84:104914. DOI: 10.1016/j.amsu.2022.104914
149. Carucci LR, Turner MA, Conklin RC, DeMaria EJ, Kellum JM, and Sugerman HJ. Roux-en-Y gastric bypass surgery for morbid obesity: evaluation of postoperative extraluminal leaks with upper gastrointestinal series. *Radiology* 2006; 238:119–27. DOI: 10.1148/radiol.2381041557
150. Uittenbogaart M, Leclercq WKG, Smeele P, Linden AN van der, Luijten AAPM, and Dielen FMH van. Reliability and usefulness of upper gastro intestinal contrast studies to assess pouch size in patients with weight loss failure after Roux-en-Y gastric bypass. *Acta Chirurgica Belgica* 2020; 120:329–33. DOI: 10.1080/00015458.2019.1631625

151. Wang B, Levine MS, Rubesin SE, Williams NN, Dumon K, and Raper S. Utility of barium studies for patients with recurrent weight gain after Roux-en-Y gastric bypass. *Clinical Radiology* 2015; 70:67–73. DOI: 10.1016/j.crad.2014.09.018
152. Brethauer SA, Nfonso V, Sherman V, Udomsawaengsup S, Schauer PR, and Chand B. Endoscopy and upper gastrointestinal contrast studies are complementary in evaluation of weight regain after bariatric surgery. *Surgery for Obesity and Related Diseases* 2006; 2:643–8. DOI: 10.1016/j.soard.2006.07.012
153. Alva S, Eisenberg D, Duffy A, Roberts K, Israel G, and Bell R. Virtual three-dimensional computed tomography assessment of the gastric pouch following laparoscopic Roux-Y gastric bypass. *Obesity Surgery* 2008; 18:364–6. DOI: 10.1007/s11695-008-9438-6
154. Karcz WK, Kuesters S, Marjanovic G, Suesslin D, Kotter E, Thomusch O, Hopt UT, Felmerer G, Langer M, and Baumann T. 3D-MSCT gastric pouch volumetry in bariatric surgery-preliminary clinical results. *Obesity Surgery* 2009; 19:508–16. DOI: 10.1007/s11695-008-9776-4
155. Hanssen A, Sabbag DJ, Hanssen RA, and Hanssen DA. 3D CAT scan and gastric volumetry in the planning of revisional bariatric surgery. *Journal of Laparoendoscopic & Advanced Surgical Techniques. Part A* 2022; 32:49–53. DOI: 10.1089/lap.2021.0547
156. Felsenreich DM, Arnoldner MA, Wintersteller L, Mrekva A, Jedamzik J, Eichelster J, Langer FB, and Prager G. Intrathoracic pouch migration in one-anastomosis gastric bypass with and without hiato-plasty: a 3-dimensional-computed tomography volumetry study. *Surgery for Obesity and Related Diseases* 2023; 19:492–9. DOI: 10.1016/j.soard.2022.11.009
157. Robert M, Pechoux A, Marion D, Laville M, Gouillat C, and Disse E. Relevance of Roux-en-Y gastric bypass volumetry using 3-dimensional gastric computed tomography with gas to predict weight loss at 1 year. *Surgery for Obesity and Related Diseases* 2015; 11:26–31. DOI: 10.1016/j.soard.2014.05.014
158. Ezeani C, Malik A, Yousaf MN, Awad MM, and Bazarbashi AN. Role of endoluminal functional lumen imaging probe in bariatric surgery and bariatric endoscopy. *Obesity Surgery* 2025; 35:614–23. DOI: 10.1007/s11695-024-07616-4
159. Flanagan L. Measurement of functional pouch volume following the gastric bypass procedure. *Obesity Surgery* 1996; 6:38–43. DOI: 10.1381/096089296765557240
160. El Ansari W and Elhag W. Weight regain and insufficient weight loss after bariatric surgery: definitions, prevalence, mechanisms, predictors, prevention and management strategies, and knowledge gaps—a scoping review. *Obesity Surgery* 2021; 31:1755–66. DOI: 10.1007/s11695-020-05160-5
161. Müller MK, Wildi S, Scholz T, Clavien PA, and Weber M. Laparoscopic pouch resizing and redo of gastro-jejunal anastomosis for pouch dilatation following gastric bypass. *Obesity Surgery* 2005; 15:1089–95. DOI: 10.1381/0960892055002257
162. Amor IB, Petrucciani N, Kassir R, Malyshev E, Mazoyer C, Korkmaz C, Debs T, and Gugenheim J. Midterm outcomes of gastric pouch resizing for weight regain after Roux-en-Y gastric bypass. *Obesity Surgery* 2020; 30:2723–8. DOI: 10.1007/s11695-020-04560-x
163. Metropolitan Life Foundation. 1983 metropolitan height and weight tables. *Statistical Bulletin (Metropolitan Life Foundation)* 1983; 64. No indexed abstract; cited as the actuarial ideal-weight reference itself, identity verified from the PubMed record. The link to %EWL practice is carried by pmid21159563:3–9

164. Montero PN, Stefanidis D, Norton HJ, Gersin K, and Kuwada T. Reported excess weight loss after bariatric surgery could vary significantly depending on calculation method: a plea for standardization. *Surgery for Obesity and Related Diseases* 2011; 7:531–4. DOI: 10.1016/j.soard.2010.09.025
165. Laar A van de. Bariatric Outcomes Longitudinal Database (BOLD) suggests excess weight loss and excess BMI loss to be inappropriate outcome measures, demonstrating better alternatives. *Obesity Surgery* 2012; 22:1843–7. DOI: 10.1007/s11695-012-0736-7
166. Corcelles R, Boules M, Froylich D, Hag A, Daigle CR, Aminian A, Brethauer SA, Burguera B, and Schauer PR. Total Weight Loss as the Outcome Measure of Choice After Roux-en-Y Gastric Bypass. *Obesity Surgery* 2016; 26:1794–8. DOI: 10.1007/s11695-015-2022-y
167. Hatoum IJ and Kaplan LM. Advantages of percent weight loss as a method of reporting weight loss after Roux-en-Y gastric bypass. *Obesity* 2013; 21:1519–25. DOI: 10.1002/oby.20186
168. Reinhold RB. Critical analysis of long term weight loss following gastric bypass. *Surgery, Gynecology & Obstetrics* 1982; 155. No indexed abstract; full text not retrievable 2026-09-01. Cited for identity, year, and subject of the eponymous criteria; the %EWL-50/BMI-35 content is attributed via pmid33097448:385–94
169. Lunel T, Iceta S, Pasquer A, Pelascini E, Perinel J, Poncet G, Disse E, and Robert M. Third bariatric procedure for insufficient weight loss or weight regain: how far should we go? *Surgery for Obesity and Related Diseases* 2021; 17. Abstract states Reinhold’s criteria verbatim: failure = $\%EWL \leq 50\%$ and/or $BMI \geq 35 \text{ kg/m}^2$:96–103. DOI: 10.1016/j.soard.2020.08.032
170. Laar AW van de and Acherman YI. Weight loss percentile charts of large representative series: a benchmark defining sufficient weight loss challenging current criteria for success of bariatric surgery. *Obesity Surgery* 2014; 24:727–34. DOI: 10.1007/s11695-013-1130-9
171. King WC, Hinerman AS, Belle SH, Wahed AS, and Courcoulas AP. Comparison of the Performance of Common Measures of Weight Regain After Bariatric Surgery for Association With Clinical Outcomes. *JAMA* 2018; 320:1560–9. DOI: 10.1001/jama.2018.14433
172. Brethauer SA, Kim J, Chaar M el, Papasavas P, Eisenberg D, Rogers A, Ballem N, Kligman M, Kothari S, and ASMBS Clinical Issues Committee. Standardized outcomes reporting in metabolic and bariatric surgery. *Surgery for Obesity and Related Diseases* 2015; 11:489–506. DOI: 10.1016/j.soard.2015.02.003
173. Shahwan S, Oochit K, Campbell E, and Kourounis G. Reporting of weight loss outcomes in bariatric surgery following introduction of 2015 ASMBS guidelines. *Surgery for Obesity and Related Diseases* 2022; 18:1195–8. DOI: 10.1016/j.soard.2022.06.296
174. Peterli R, Hurme S, Bueter M, Grönroos S, Helmiö M, and Salminen P. Standardized assessment of metabolic bariatric surgery outcomes: secondary analysis of 2 randomized clinical trials. *JAMA Surgery* 2024; 159:306–14. DOI: 10.1001/jamasurg.2023.6254
175. Higa K, Ho T, Tercero F, Yunus T, and Boone KB. Laparoscopic Roux-en-Y gastric bypass: 10-year follow-up. *Surgery for Obesity and Related Diseases* 2011; 7:516–25. DOI: 10.1016/j.soard.2010.10.019
176. Puzziferri N, Roshek TB, Mayo HG, Gallagher R, Belle SH, and Livingston EH. Long-term follow-up after bariatric surgery: a systematic review. *JAMA* 2014; 312:934–42. DOI: 10.1001/jama.2014.10706
177. Gould JC, Beverstein G, Reinhardt S, and Garren MJ. Impact of routine and long-term follow-up on weight loss after laparoscopic gastric bypass. *Surgery for Obesity and Related Diseases* 2007; 3:627–30. DOI: 10.1016/j.soard.2007.07.005

178. Wan X, Wang W, Liu J, and Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology* 2014; 14:135. DOI: 10.1186/1471-2288-14-135
179. Luo D, Wan X, Liu J, and Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Statistical Methods in Medical Research* 2018; 27:1785–805. DOI: 10.1177/0962280216669183
180. Hozo SP, Djulbegovic B, and Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Medical Research Methodology* 2005; 5:13. DOI: 10.1186/1471-2288-5-13
181. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, and Welch VA, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. version 6.5. Cochrane, 2024. Available from: <https://training.cochrane.org/handbook>
182. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; 372:n71. DOI: 10.1136/bmj.n71
183. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* 2020; 368:l6890. DOI: 10.1136/bmj.l6890
184. Al-Fagih OS, Zuberi S, Niaz O, et al. Impact of gastrojejunostomy anastomosis diameter on weight loss following laparoscopic gastric bypass: a systematic review and meta-analysis. *Obesity Surgery* 2024; 34:2227–36. DOI: 10.1007/s11695-024-07237-x
185. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; 366:l4898. DOI: 10.1136/bmj.l4898
186. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016; 355:i4919. DOI: 10.1136/bmj.i4919
187. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, and Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336:924–6. DOI: 10.1136/bmj.39489.470347.AD
188. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JPT, Langan D, and Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2016; 7:55–79. DOI: 10.1002/jrsm.1164
189. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7:177–88. DOI: 10.1016/0197-2456(86)90046-2
190. IntHout J, Ioannidis JPA, and Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology* 2014; 14:25. DOI: 10.1186/1471-2288-14-25
191. Röver C, Knapp G, and Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology* 2015; 15:99. DOI: 10.1186/s12874-015-0091-1
192. Higgins JPT, Thompson SG, Deeks JJ, and Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327:557–60. DOI: 10.1136/bmj.327.7414.557
193. Greenland S and Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; 135:1301–9. DOI: 10.1093/oxfordjournals.aje.a116237

194. Egger M, Davey Smith G, Schneider M, and Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315:629–34. DOI: 10.1136/bmj.315.7109.629
195. Yi B, Jiang J, Zhu L, Li P, Im I, and Zhu S. Comparison of the effects of Roux-en-Y gastrojejunostomy and LRYGB with small stomach pouch on type 2 diabetes mellitus in patients with BMI < 35 kg/m². *Surgery for Obesity and Related Diseases* 2015; 11:1061–8. DOI: 10.1016/j.soard.2014.12.029
196. IntHout J, Ioannidis JPA, Rovers MM, and Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* 2016; 6:e010247. DOI: 10.1136/bmjopen-2015-010247
197. Felsenreich DM, Bichler C, Langer FB, Gachabayov M, Eichelster J, Gensthaler L, Vock N, Artemiou E, and Prager G. Surgical Technique for One-Anastomosis Gastric Bypass. *Surgical Technology International* 2020; 37:57–61
198. Abu-Abeid A, Yuval JB, Keidar A, Nizri E, Lahat G, Eldar SM, and ISMBS collaborative group. Technical Considerations in One Anastomosis Gastric Bypass — the Israeli Society of Metabolic and Bariatric Surgery Experience. *Obesity Surgery* 2024; 34:2356–62. DOI: 10.1007/s11695-024-07223-3
199. Musella M, Susa A, Manno E, De Luca M, Greco F, Raffaelli M, Cristiano S, Milone M, Bianco P, Vilardi A, Damiano I, Segato G, Pedretti L, Giustacchini P, Fico D, Veroux G, and Piazza L. Complications Following the Mini/One Anastomosis Gastric Bypass (MGB/OAGB): a Multi-institutional Survey on 2678 Patients with a Mid-term (5 Years) Follow-up. *Obesity Surgery* 2017; 27:2956–67. DOI: 10.1007/s11695-017-2726-2
200. Pizza F, Braun A, Lucido FS, Cristiano S, D’Antonio D, Giambavicchio LL, Precisano S, Bruscianno L, Tolone S, Massa C, Santi HT, Paolicelli M, Dell’isola C, Docimo L, and Gambardella C. Five-Year Endoscopic Outcomes After One-Anastomosis Gastric Bypass: A Multicenter Cohort with Bile Reflux, Marginal Ulcers, and Symptom Correlation. *Obesity Surgery* 2026. DOI: 10.1007/s11695-026-08784-1
201. Bruzzi M, Rau C, Voron T, Guenzi M, Berger A, and Chevallier JM. Single anastomosis or mini-gastric bypass: long-term results and quality of life after a 5-year follow-up. *Surgery for Obesity and Related Diseases* 2015; 11:321–6. DOI: 10.1016/j.soard.2014.09.004
202. McGuinness LA and Higgins JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Research Synthesis Methods* 2021; 12:55–61. DOI: 10.1002/jrsm.1411

Appendix A

Search strings and hit counts

Four PubMed searches were executed programmatically through the NCBI E-utilities interface by the repository's corpus pipeline (spec:001) on 31 August 2026, between 20:13 and 20:14 UTC. Each execution is stored in the catalog — query string, execution timestamp, total match count, and the full PMID hit list — so every count below is reproducible by query. Deduplication is by PMID: a record returned by two searches is one catalog row with two search attributions. Retrieval caps (`retmax` 600 for S1–S3, 200 for S4) exceeded every total match count, so no hit list was truncated. The strings are reproduced verbatim from the frozen search-strategy note (`notes/search-strategy-draft.md`).

S1 — core RYGB pouch–size–outcome query

Intersects four concept blocks: the operation, the structure (pouch/reservoir), a size vocabulary, and a weight-outcome vocabulary. Executed 2026-08-31T20:13:17Z; 421 matches.

```
("gastric bypass"[MeSH Terms] OR "gastric bypass"[tiab] OR "Roux-en-Y"[tiab]
OR RYGB[tiab] OR RYGBP[tiab])
AND (pouch[tiab] OR reservoir[tiab])
AND (size[tiab] OR volume[tiab] OR length[tiab] OR dimension*[tiab]
OR small[tiab] OR large[tiab] OR extended[tiab] OR capacity[tiab])
AND ("weight loss"[MeSH Terms] OR "weight loss"[tiab] OR weight[tiab]
OR outcome*[tiab] OR EWL[tiab] OR TWL[tiab] OR BMI[tiab])
```

S2 — OAGB/mini-bypass subgroup

Targets the one-anastomosis literature, whose terminology the core string does not reach (pre-specified subgroup, research question Q6). Executed 2026-08-31T20:13:32Z; 59 matches, 14 new.

```
("one anastomosis gastric bypass"[tiab] OR "one-anastomosis"[tiab]
OR "mini gastric bypass"[tiab] OR "mini-gastric"[tiab]
OR OAGB[tiab] OR MGB[tiab])
AND (pouch[tiab])
AND (size[tiab] OR volume[tiab] OR length[tiab] OR long[tiab]
OR short[tiab])
AND ("weight loss"[tiab] OR weight[tiab] OR outcome*[tiab]
OR EWL[tiab] OR TWL[tiab] OR BMI[tiab])
```

S3 — construction-technique variants

Catches studies that fix the pouch by technique and never use a size word. Executed 2026-08-31T20:13:34Z; 131 matches, 64 new.

```
("gastric bypass"[tiab] OR RYGB[tiab])
AND (pouch[tiab])
AND (stapler*[tiab] OR "staple line"[tiab] OR bougie[tiab]
OR calibrat*[tiab] OR construct*[tiab] OR "surgical technique"[MeSH Terms])
AND (weight[tiab] OR outcome*[tiab])
```

S4 — early-literature supplement

Added after piloting: founding-era records often lack indexed abstracts, so [tiab] tags see titles only, and the 1982 reservoir-versus-stoma literature phrases the question without size words. Executed 2026-08-31T20:14:01Z; 85 matches, 27 new.

```
"gastric bypass"[tiab] AND (reservoir[tiab] OR pouch[tiab])
AND (stoma[tiab] OR stomal[tiab])
```

Filters

No date filter (the 1975–1982 literature must be reachable); no language filter (non-English records flagged at cataloging, not dropped); PubMed’s humans filter deferred to screening, because it silently loses recent records not yet indexed.

Yield

Table A.1: Search yield, PubMed via E-utilities, 2026-08-31. “New” counts records not returned by any earlier search in the sequence; deduplication by PMID.

Search	Focus	Matches	New	Cumulative unique
S1	Core RYGB pouch–size–outcome	421	421	421
S2	OAGB subgroup	59	14	435
S3	Construction technique	131	64	499
S4	Early literature (reservoir/stoma)	85	27	526

The four searches returned 696 hits resolving to 526 unique records: 373 records (70.9%) found by exactly one search, 136 by two, 17 by three, none by all four.

Recall check

The strings were frozen only after a recall criterion was met: every in-scope landmark study named in the preliminary literature scan had to appear in the union of the hit lists. Table A.2 lists the landmark records with the searches that captured each, verified directly against the stored hit lists; all were captured. One deliberately out-of-scope seed — the 2024 systematic review of gastrojejunostomy *stoma* diameter, PMID 38652437, whose exposure is the outlet rather than the pouch — was correctly absent from all four hit lists and is retained solely as a citation-chasing seed.

Table A.2: Recall-check set: landmark records of the pouch-size literature and the searches that captured each (verified against the catalog hit lists, 2026-09-01).

Study	PMID	Role in the literature	Captured by
Kirkpatrick & Siegel 1982	7091134	Reservoir-vs.-stoma founding test	S4
Edholm 2016	26194258	SOReg registry cohort (n = 14,168)	S1, S3
Ren 2015	25236397	Randomized volume contrast	S1
Mahawar 2020	31745633	The only prior systematic review	S1, S4
Boerboom 2020	31444776	Dutch extended-pouch trial (length)	S1
Gao 2022	36120423	Six-year retrospective cohort	S1
Haghighat 2022	36536734	Stapler-count exposure study	S1, S3, S4
Al-Fagih 2024	38652437	Out-of-scope stoma review (seed only)	none (correct)

Pending sources

Two identification lines of the protocol (spec:010) had not been executed when this appendix was compiled and will be recorded here, each with dates and per-source counts, before screening is finalized: (i) the manual Cochrane CENTRAL check using the same concept blocks; (ii) backward and forward citation chasing from the two seed reviews (PMIDs 31745633 and 38652437) and from every included study. Each enters the PRISMA flow on its own accounting line, separate from the database searches above. Embase, Web of Science, Scopus, trial registries, and conference proceedings were not searched; that limitation is stated in Chapter 8 and weighed in Chapter 14.

Appendix B

Extraction schema and gold cases

[Pending.]

Appendix C

Per-study evidence tables

[Pending.]

Appendix D

Excluded studies

[Pending.]

Appendix E

Figure provenance

Per the figure policy (spec:011), every figure in this document carries a provenance note in the repository at `dissertation/figures/provenance/`, one Markdown file per figure, recording what the figure depicts, the source of every labeled fact (chapter section or PMID), the rendering or authoring date, and the generator. Figures fall in two classes. *Data figures* are rendered from pipeline data by code in the repository — never hand-drawn, never generated by an image model — and are regenerable by one command. *Schematic figures* (anatomy, mechanisms) are AI-authored vector illustrations (TikZ), reviewable line by line, with an AI-generation disclosure in the caption. This appendix is the collation of the provenance directory; notes for figures added after this compilation date are appended by the same rule.

Provenance notes on file (collated 2026-09-01)

Eleven provenance notes, one per rendered figure, all dated 2026-09-01. *AI-authored schematic* below means an LLM-authored TikZ vector illustration (Claude, model `claude-fable-5`), reviewed line by line, compiled standalone and visually inspected before inclusion, with the AI-generation disclosure in the figure caption. *Data figure* means rendered from pipeline data by repository code (matplotlib or the simulation engine); no image model and no hand-drawing anywhere in either class.

Figure 3.1 (`ch03-timeline`, Chapter 3) AI-authored schematic. Horizontal timeline 1954–2026 with eleven milestones of bariatric history, each with a one-line fate. All facts restate Chapter 3 sections §jib, §mason, §gastroplasty, §lap, §oagb, §sleeve, §consolidation; per-node citations include PMIDs 13198079, 6073761, 5798387, 907395, 7073493, 1952493, 10742801, 11433900, 14738671, 38592648, 42176157.

Figure 3.2 (`ch03-operations`, Chapter 3) AI-authored schematic. Six panels, one per principal bariatric operation (JIB, Mason loop, VBG, RYGB, sleeve, OAGB), in one visual grammar (reservoir shaded, staple lines, excluded tissue hatched, food/biliopancreatic streams). Anatomy restates Chapter 3 and STYLE.md’s locked pouch-composition framing; PMIDs 5798387, 880023, 907395, 6073761, 1180579, 7268623, 31853864, 7073493, 3296251, 6623129, 11054246, 9678194, 14738671, 42124786, 11433900, 30476553, 15826476.

Figure 4.1 (`ch04-rygb-anatomy`, Chapter 4) AI-authored schematic. Post-RYGB anatomy in stylized anterior view — pouch, bougie, staple lines, excluded remnant, limbs — with callouts marking the dissertation’s exposure set and confounders. Labeled values from Chapter 4 (§anatomy, §construction, §variation, §emptying) and PMIDs 42176157, 31444776, 10065576.

Figure 4.2 (`ch04-pouch-construction`, Chapter 4) AI-authored schematic. The laparoscopic pouch-construction sequence in four panels: lesser sac entry at the crow’s foot, first trans-

verse fire, vertical fires along the bougie, completed divided pouch. Facts from Chapter 4 §construction and PMIDs 42176157, 29243145, 31444776, 25224832, 10065576.

Figure 5.1 (`ch05-two-models`, Chapter 5) AI-authored schematic. Two-panel contrast of the mechanistic debate’s poles: the legacy restrictive–malabsorptive model versus the multicomponent neuroendocrine model, each with its prediction for pouch volume. Sources: Chapter 5 throughout; PMIDs 6073761, 7091134, 23360316, 29055668, 16371744, 29730732, 12023994, 30684171, 24670636, 23536013.

Figure 5.2 (`ch05-hormone-response`, Chapter 5) AI-authored schematic. Causal-chain schematic of the RYGB hormonal response (nutrient pulse → GLP-1/PYY surge; excluded fundus → flattened ghrelin; convergence on satiety). Time-course sketches are explicitly labeled illustrative, not plotted data, per spec:011. Sources: Chapter 5; PMIDs 23360316, 29055668, 16371744, 27568473, 12023994, 29730732.

Figure 6.1 (`ch06-modalities`, Chapter 6) AI-authored schematic. Central stylized RYGB anatomy surrounded by four measurement-modality panels (construction proxies, UGI contrast, 3D-CT, swallow MRI), each with quantity measured, timing, validation record, and V1–V4 tier tag. Sources: Chapter 6 (§proxies, §ugi, §ct, §mri, §hierarchy); PMIDs 29915973, 6474086, 16373763, 31203729, 18274830, 19104904, 34935476, 25500226, 32564309.

Figure 9.1 (`ch09-corpus-timeline`, Chapter 9) Data figure. Annual publication counts, 1975–2026, of the 526-record candidate corpus, queried live from `data/catalog.db` at render time by `analysis/corpus_figs/make_ch09_figs.py` (script AI-authored; data values queried, never model-generated); era annotations sourced to PMIDs 1180579, 7091134, 10742801, 11433900.

Figure 9.3 (`ch09-corpus-census`, Chapter 9) Data figure. Text-depth and publication-type breakdown of the 526 cataloged records, queried live from `data/catalog.db` by the same script; tag caveat (indexer-assigned, non-exclusive) per Chapter 9 §types.

Figure 12.1 (`detectability_surfaces`, Chapter 12) Data figure. Per-tier (V1–V4) power to detect each true pouch-size effect at the realized designs, 80%-power floor marked; rendered by the spec:012 simulation engine (`analysis/simulation/`, seeded, configuration-driven) from `config.yaml`, `parameters.yaml` (per-parameter provenance), and the provisional `designs.yaml`; regenerated when the screened-corpus extraction lands. Displayed numbers verified against `results/mde.csv`.

Figure 12.2 (`fidelity_requirement`, Chapter 12) Data figure. Maximum measurement error permitting 80% power for a given true effect and sample size, from the same simulation and configuration; displayed numbers verified against `results/fidelity.csv`.

One declared placeholder remains in the document (Figure 9.2, the corpus in space): it renders no data and carries no provenance note until the spec:002 affiliation extraction supplies its source table. Notes for figures added after this compilation date are appended by the same rule.

Appendix F

Pipeline documentation

The dissertation is backed by a research workbench in the same repository, built spec-first (every behavior change starts in `specs/`). Two Python tools, both uv-managed and run as `uv run ...`, generate the evidence artifacts; nothing quantitative enters the text except from their outputs.

Corpus tool (spec:001)

`src/corpus_tool` acquires and catalogs the literature. `corpus search` executes a PubMed query through NCBI E-utilities and, with `-save`, stores the execution — query string, timestamp, total count, and full PMID hit list — in a single-file SQLite catalog (`data/catalog.db`); records deduplicate by identifier (PMID, DOI secondary), never by title similarity. `corpus fetch` retrieves full texts for the PubMed Central open-access subset and writes immutable snapshots under `data/raw/` — raw bytes, SHA-256 hash, and extracted text — with the catalog row recording fetch method and achieved text depth (`full_text_pmc`, `abstract`, or `metadata_only`), so no claim can silently pretend abstract-derived text is full text. Locally supplied open-access PDFs enter through `corpus intake-pdf` by the same snapshot rules. `corpus screen` runs the staged screening workflow of the protocol (spec:010): per-record decisions at title/abstract and full-text stages (a pilot stage included), each logged in the catalog with its reason, with batch import, queue, statistics, and export subcommands; the PRISMA flow is generated from this log, never written by hand. Access is legitimate-channels only: an effort ladder of API first, then plain HTTP — no paywall circumvention — with rate limiting and circuit-breaker backoff on every source. The catalog carries retraction status and language flags per record. Raw corpus data are never committed; the catalog and snapshots stay outside version control.

Simulation tool (spec:012)

`src/sim_tool` implements the measurement-fidelity simulation of Chapter 12: a seeded, configuration-driven engine (`analysis/simulation/`) whose every input lives in versioned configuration — error-model parameters with per-parameter provenance and uncertainty in `parameters.yaml`, the study-design grid in `designs.yaml` (provisional until the screened-corpus extraction replaces it), and run settings in `config.yaml`. `sim run` produces the detectability results, `sim figures` renders the figure set from stored results, and `sim sensitivity` sweeps the thin calibration inputs. Every figure is regenerable by one command from the committed configuration.

Downstream and document build

The extraction stage (spec:002, in draft at compilation) turns snapshots into span-verified clinical claims — verbatim source quote, four-way adjudication status, gold-case evaluation to select model tiers — and is the gate under which every citation in this document stands: chapter bibliography fragments carry `keywords={pending-verification}` until the span-verification pass clears them. The document itself builds with `latexmk` (biblatex/biber, Vancouver numeric); data figures are rendered by repo code from the catalog and simulation outputs, per the figure policy of spec:011 and the provenance rules of Appendix E.