On the obsolescence of sampling methodology in healthcare diligence.

A note on the inherited compromises of an analytical regime whose constraints have lifted.

H. Chaudhry · Crescent Intelligence
crescentintel.com/research/obsolescence-of-sampling

14 March 2026
4632 words

abstract

The prevailing methodology for healthcare reimbursement diligence — examining a sample of claims, manually re-adjudicating them under controlling rules, and extrapolating the variance pattern to the full revenue base — has been the industry standard for three decades. We argue that sampling was always a compromise rather than a methodology of first choice, that the compromise was justified by binding technological and informational constraints, and that those constraints have lifted decisively. The methodology persists past its justifying conditions. We develop the structural errors that sampling introduces — inadequate stratification against population heterogeneity, compounding error through the diligence workflow, and information-theoretic waste against a computable underlying function — and we describe the decomposed alternative that current constraints permit. The transition from sampling to population-scale exact computation is not an incremental improvement. It is a methodological transformation that produces categorically different outputs, and the field's continued reliance on sampling is an artifact of inheritance rather than a defensible methodological position.

1 the practice as it stands

A healthcare quality-of-earnings engagement, as practiced today by the major advisory firms and the diligence arms of private equity buyers, follows a recognizable workflow. The diligence team obtains a claim-level extract from the target entity, typically covering twelve to twenty-four months of services rendered. From this extract, a sample is drawn — usually two hundred to one thousand claims, stratified across payer, specialty, and claim type. The sample is then manually re-adjudicated: for each claim, an analyst examines the documentation, applies the controlling fee schedule and contract terms, computes what should have been paid, and compares the result to what was actually paid. The variance pattern observed across the sample is extrapolated to the full revenue base. The extrapolated variance becomes a QofE adjustment, which becomes an input to deal valuation, financing structure, and post-close integration planning.

The methodology has been refined across three decades. The major advisory firms have developed proprietary stratification frameworks, internal coding manuals, and quality-control procedures that govern how samples are drawn, how claims are adjudicated, and how findings are reported. The output is a deliverable that PE buyers, lenders, and acquirers have learned to interpret and to incorporate into their decisions. The methodology has institutional weight. It is not the product of carelessness or lack of rigor; it is the product of decades of careful adaptation by serious practitioners.

This note examines the methodology not as an instance of poor practice but as an instance of inheritance. The methodology was developed under specific constraints. Those constraints have lifted. The methodology persists. We argue that

the persistence is not defensible on current grounds, and that the methodological transformation now available is categorical rather than incremental.

2 sampling was always a compromise

Before developing the historical case, we make a more fundamental observation. Sampling, as a methodology for estimating a population parameter, has always been epistemically inferior to exact computation when exact computation is available. This is not a contentious claim. It is a basic result of statistical theory. A sample produces an estimate with a confidence interval; exact computation produces a value. An estimate is information-poor relative to a value. The confidence interval is not a feature of sampling; it is a measure of the information that sampling fails to recover from the underlying population.

When exact computation is feasible, no statistical practitioner would choose sampling. The choice of sampling over exact computation always reflects a binding constraint — typically, that exact computation is not feasible at the required scale or under the available cost. Sampling is the methodology of necessity, not of preference. Statistics as a discipline developed largely to handle situations in which exact computation is impossible: the population is too large to enumerate, the relevant variable is not observable for every member, or the cost of comprehensive measurement exceeds the value of the resulting precision. Under these conditions, sampling produces the best estimate that is actually achievable. The methodology earns its place.

Healthcare reimbursement was treated as one of these conditions. Claims data, until recently, was difficult to acquire at population scale. Computation was expensive enough that adjudicating millions of claims was infeasible for a diligence engagement bounded by weeks and a fixed budget. Encoding the controlling rule system at the resolution required for exact derivation was a research problem that no commercial entity had solved. Under these conditions, sampling was the right methodology — not because it produced accurate results but because the alternative did not exist.

The methodology's adoption was rational. Its persistence is what we examine.

3 the constraints that justified sampling

To understand why sampling became standard, and to evaluate whether its standing remains defensible, we have to specify the constraints that originally justified it. Three constraints were binding.

The first was *data acquisition*. Claims data in the 1990s lived in proprietary payer systems and provider billing systems that did not communicate. The 837 transaction standard for electronic claim submission and the 835 standard for remittance advice were not formalized until 1996 under HIPAA's Administrative Simplification provisions, and adoption was gradual through the late 1990s and 2000s. Before universal adoption, a provider's full claim history was difficult to extract, often required custom integration work, and was frequently incomplete. Acquiring a population-scale claim extract for a diligence engagement was not impossible but it was sufficiently expen-

sive and time-consuming that sampling — which required only enough claims to characterize the population statistically — was the rational choice.

The second was *computational cost*. Re-adjudicating a single claim requires applying the relevant fee schedule, identifying any applicable NCCI edits, evaluating modifier interactions, and computing the contractual payment under the relevant payer contract. In the 1990s and 2000s, this computation was performed either manually by trained analysts or through proprietary software systems with limited automation. The cost per claim, fully loaded, was high enough that processing the full population for a multi-billion-dollar healthcare entity was infeasible within the time and budget of a diligence engagement. Sampling reduced the computation requirement by two to three orders of magnitude.

The third was *rule-system encoding*. To compute the contractually correct payment for a claim, the diligence team needs an encoded representation of the controlling rules: the Medicare Physician Fee Schedule, the NCCI procedure-to-procedure and medically-unlikely-edit components, the relevant Local Coverage Determinations, the applicable payer medical policies, and the payer-provider contract including its fee schedule attachments. Through the 1990s and 2000s, this encoding existed in partial form across multiple commercial vendors (claim editors, rules engines, payment integrity tools), but no vendor had achieved the resolution required for exact computation at scale. The encoding was partial; the computation it supported was approximate. Under these conditions, even if data acquisition and computational cost were tractable, the underlying function could not be computed exactly, and sampling was the necessary methodology for estimating what the function would have produced.

Each of these constraints, taken alone, would have been sufficient to justify sampling. Together, they made sampling the only feasible methodology. The methodology's three-decade tenure as industry standard is not surprising; it was the rational response to a constraint regime that admitted no alternative.

4 the constraints have lifted

The constraints have not partially lifted. They have lifted decisively, and the lifting can be dated with reasonable precision.

Data acquisition is no longer constraining. The 837/835 standards are now universal. Every contracted provider in the United States generates 837 submissions and receives 835 remittances in standardized form. Claim-level extracts at population scale are available for any contracted entity, with retrieval times measured in hours rather than weeks. The data infrastructure that did not exist in the 1990s is now ambient. A diligence team that wishes to examine a target entity's full claim population can do so without significant marginal cost beyond what sampling would require.

Computational cost is no longer constraining. The full claim population for a multi-billion-dollar healthcare entity, typically several million claims over a relevant diligence window, fits comfortably in working memory on commodity hardware. Per-claim adjudication, when properly automated, runs in milliseconds. Processing the

full population, end to end, can be accomplished in minutes to hours on infrastructure that costs less than a single diligence analyst's salary. The computational constraint that previously made sampling necessary has not eased; it has been eliminated by three orders of magnitude.

Rule-system encoding has become feasible at the required resolution. The work is still hard — capturing the full state space of fee schedules, NCCI edits, payer policies, and contract terms requires sustained engineering investment — but the encoding is no longer rate-limited by the cost of converting regulatory text into structured representation. The proximate enabler is the emergence of large language models capable of extracting structured rules from unstructured regulatory text at scale. The conversion task that previously required hundreds of analyst-years of skilled human labor is now feasible within engineering budgets that any serious organization can sustain. The encoding has shifted from a research problem that no commercial entity had solved to an engineering problem bounded by the resolution of the rule system itself.

The reader who notes the role of language models in rule extraction may reasonably ask about hallucination — the well-documented tendency of language models to generate plausible but incorrect output. The concern is appropriate but it applies to a different application of language models than the one relevant here. There are two ways to use a language model. In the first, the model is the production system: queries are submitted, the model generates answers, and the answers depend on the model's reliability at the moment of generation. In this mode, hallucination is a serious and persistent concern. In the second, the model is an extraction tool whose outputs are subject to deterministic verification before being committed to a downstream system. Proposed rule representations are verified against the source regulatory text, against test cases derived from known-correct adjudications, and against cross-references between rules in the encoded system. Only verified rules enter the production rule base. The production system, once compiled, is deterministic — it does not call out to a language model at runtime, and its rule applications do not depend on language model reliability at the moment of any particular adjudication.[1]

The hallucination concern is real for the first mode and not present in the second. The encoding work described here uses language models exclusively in the second mode, with verification as the structural constraint that prevents extraction errors from reaching the production system. The methodological discipline of extraction-with-verification distinguishes responsible use of language models from the more common production-system use that motivates the hallucination concern. The encoding constraint has lifted not because language models are reliable enough to be trusted as production systems but because they are reliable enough to be useful as extraction tools whose outputs are subject to verification at a cost dramatically lower than the cost of producing the same extractions through human labor.

All three constraints have lifted. The methodology that they justified is now a methodology in search of a justification.

[1] marginalia

We date the lifting of the data-acquisition constraint to roughly 2010, the computational constraint to roughly 2015, and the encoding constraint to the present. Sampling was the right methodology through approximately 2010 and an increasingly indefensible methodology since.

5 how methodologies survive past their justifying conditions

The persistence of sampling methodology in healthcare diligence is not a unique pathology. It is an instance of a more general pattern: professional methodologies tend to outlive the constraints that originally justified them.

The mechanism is straightforward. A methodology is developed in response to specific technological and informational conditions. The methodology becomes embedded in training programs, professional norms, software tools, regulatory frameworks, and client expectations. By the time the underlying conditions change, the methodology has accumulated infrastructure that is independent of the conditions. Updating the methodology requires re-training, re-tooling, re-norming, and re-educating clients — all of which is costly. The methodology persists because the cost of inertia is borne diffusely by everyone (in the form of slightly worse analytical work) while the cost of updating is borne acutely by whoever attempts to update first.

The pattern is general across professional fields. Clinical trial methodology developed under constraints (limited patient populations, slow data collection, expensive statistical computation) that produced specific conventions about sample sizes, statistical tests, and trial duration. The constraints have lifted with the advent of electronic health records, biobanks, and computational genomics, but the conventions have updated only partially. Financial accounting developed under constraints (paper-based recordkeeping, annual audit cycles, limited inter-company data exchange) that produced conventions about reporting periods, accrual treatments, and consolidation rules. The constraints have lifted with continuous data infrastructure and real-time financial systems, but the conventions persist. Actuarial science developed under constraints (limited longitudinal data, computationally expensive life tables, infrequent rate revision) that produced conventions about credibility weighting, experience rating, and pricing cycles. The constraints have lifted with modern data infrastructure, but the conventions update slowly.

In each case, the methodology survives not because it remains optimal but because the infrastructure surrounding it has become independent of the conditions that originally justified it. The methodology is no longer a response to current reality; it is a cultural artifact of its historical origins.

Healthcare diligence methodology is one such artifact. The QofE workflow, the stratification frameworks, the sample-size conventions, the proprietary coding manuals, the manual re-adjudication protocols, the report templates — all of this infrastructure was built under the three original constraints. The infrastructure has accumulated weight. The constraints have lifted. The infrastructure persists.

The recognition that one is operating within an inherited methodology is itself uncomfortable. It implies that decades of accumulated expertise are organized around constraints that no longer bind. It implies that the methodological refinements developed over the past two decades — the increasingly sophisticated stratification frameworks, the increasingly granular sampling protocols — have been improvements within a regime that should have been replaced rather than refined. The discomfort is real, and the field's incentive to avoid it is correspondingly strong. The methodology persists in part because confronting its obsolescence requires the practitioners to confront the obsolescence of their own accumulated expertise.

6 the structural errors of sampling

We now turn to the mathematical case. Sampling, as the diligence industry practices it, introduces three structural errors that the prevailing methodology does not adequately address. Each error compounds the others. Together they explain why the methodology produces systematically misleading results even when applied with full competence.

First, structural bias from inadequate stratification. Sampling produces unbiased estimates of population parameters only when the sample is representative of the population. Standard practice uses stratified random sampling, drawing claims across strata defined by payer, specialty, and claim type to ensure representativeness across these dimensions. The variance of the resulting estimator decomposes into two components:

$$\mathrm{SE}\big(\hat{\theta}\big) = \sqrt{\mathrm{Var}_{\mathrm{sampling}}\big(\hat{\theta}\big) + \mathrm{Bias}^2_{\mathrm{structural}}}$$

The first term shrinks with sample size. The second term does not. Structural bias arises when the population contains heterogeneity along dimensions that the stratification does not capture. For healthcare claims, the relevant strata are not just payer, specialty, and claim type. They include contract version, policy effective date, NCCI edit applicability, modifier interaction, place of service, provider taxonomy, patient demographic, prior authorization status, and dozens of other dimensions that affect the contractual computation. The number of meaningful strata, expressed as the Cartesian product of these dimensions, exceeds practical sample sizes by several orders of magnitude. Standard sampling protocols, working with samples in the hundreds or low thousands, cannot achieve representativeness across this heterogeneity. The samples are unbiased with respect to the strata they explicitly capture and biased with respect to everything else.

The bias does not appear in the reported confidence interval. The confidence interval is computed from the sampling variance — the first term — and the second term is assumed to be zero. This assumption is rarely tested and is almost always wrong. The reported standard error understates the true error, and the practitioner who acts on the reported figure is acting on a confidence interval that does not reflect the actual uncertainty of the estimate.

This is the first error. Sampling-based diligence findings are wrong, not just imprecise, and the magnitude of the wrongness is not captured by the methodology's own self-assessment.

Second, compounding error through the diligence workflow. A QofE finding is not a standalone fact. It is an input to a deal valuation, which is an input to a financing decision, which is an input to a forward-looking forecast of the combined entity's revenue cycle performance. Each downstream use of the sampling-based estimate inherits the estimate's error. The error compounds.

The compounding can be expressed compactly. If the QofE finding has standard error σ_0, and the deal valuation model has its own modeling uncertainty σ_v that is independent of the QofE error, and the financing decision has its own further uncertainty σ_f, the total uncertainty of the final decision is bounded by:

$$\sigma_{\text{total}} \geq \sqrt{\sigma_0^2 + \sigma_v^2 + \sigma_f^2 + \ldots}$$

This lower bound assumes the errors are independent. In practice, they are correlated — the deal valuation model uses the QofE finding as a baseline, so its modeling assumptions are not independent of the QofE error — and the actual compounded error is larger than the independent-error bound suggests. By the time a sampling-based finding has propagated through the deal model into the financing decision and the forward forecast, the compounded uncertainty is substantial. The decision that the buyer ultimately makes — to acquire at price *P*, to finance at structure *F*, to forecast cash flow *C* — rests on a chain of estimates each of which inherited and amplified the original sampling error.[2]

[2] marginalia

The compounding error argument has been noted occasionally in the diligence literature but rarely incorporated into deliverables. The reported confidence interval almost always reflects only the first-order sampling variance.

This is the second error. Sampling-based findings do not produce localized uncertainty that can be quarantined within the diligence report. They produce uncertainty that propagates through every subsequent step of the deal, and the propagation is not visible in the diligence deliverable.

Third, information-theoretic waste. When a population parameter is the output of a deterministic function over inputs that are themselves observable, sampling to estimate the parameter is throwing away information that is already available. This is the deepest of the three errors, and it is the error that § M.01 articulated for the convolutional approach generally. We restate it here in the diligence-specific case.

The contractual payment C for a claim is the output of a function $f : X \to C$, where X is the claim's input state (codes, modifiers, place of service, contract terms, applicable policies) and f is the controlling rule system. The function is deterministic; given X, the value of C is uniquely determined. The conditional entropy of C given X is zero:

$$H(C \mid X) = 0$$

When the diligence team samples claims to estimate the population's contractual variance, they are estimating a quantity whose value is computable exactly from the claim-level inputs. The estimation introduces error where no error need exist. The sampling produces a confidence interval around C that should be a point. The methodology is information-theoretically wasteful — it discards the information that the rule system encodes and reconstructs an approximation of that information from a sample, with attendant error.

This is the third error, and it is the one that no refinement of sampling methodology can address. The problem is not that sampling is being practiced poorly. The problem is that sampling is the wrong methodology for any quantity whose value is exactly computable from observable inputs. The contractual payment C is such a quantity. The diligence industry's continued use of sampling for C reflects a failure to recognize that the underlying function is computable, which is the same failure that § M.01 identified in the convolutional approach.

The three errors compound. Structural bias means the sampling estimate is wrong; workflow propagation means the wrongness amplifies through the deal; information-theoretic waste means the error was avoidable in principle. Each error alone would be sufficient cause to question the methodology. Together they constitute a structural case for replacement.

7 the methodological transformation

When the constraints lift and the alternative becomes feasible, the diligence methodology that emerges is categorically different from sampling. It is not a more sophisticated sample. It is not a larger sample. It is a different kind of analytical object entirely.

The transformation operates as follows. The rule system is encoded at the resolution required for exact computation. The full claim population is examined rather than sampled. For each claim, the controlling rules are applied to derive the contractual payment C exactly. The behavioral residual B is observed in full across the population rather than estimated from a sample. The QofE finding is no longer an extrapolation from a sample; it is a derivation from the full population.

The output is a different kind of object than what sampling produces. The contractual component is a proof: dollar-exact, verifiable by re-derivation, with zero residual uncertainty on the rule-governed portion. The behavioral component is a forecast: calibrated against the full population, with uncertainty bounds that reflect actual signal-to-noise rather than sampling artifacts. The two components are reported separately and admit different downstream uses. The deal model that takes a decomposed diligence finding as input is working with materially different information than the deal model that takes a sampling-based finding as input.

The transformation has specific consequences that deserve naming.

The confidence interval on contractual truth collapses to zero. Under the prevailing methodology, the QofE finding has a confidence interval that reflects the sampling variance of the estimator. Under the decomposed methodology, the contractual finding has no confidence interval, because it is a derivation rather than an estimate. The downstream uses that depend on this finding — deal valuation, contract negotiation, working capital sizing — operate with point values rather than ranges. The propagation of uncertainty through the workflow shrinks accordingly.

The behavioral forecast is calibrated against the full population. Under the prevailing methodology, the behavioral component is implicit in the variance pattern of the sample and is extrapolated to the full revenue base. Under the decomposed method-

ology, the behavioral component is observed across every claim in the population. The forecast's calibration is bounded only by the underlying signal-to-noise ratio of payer behavior, not by sampling artifacts. The forecast can be decomposed by payer, by specialty, by region, by claim type, by any feature of interest, and the decomposition is exact rather than extrapolated.

The diligence deliverable becomes substantively different in content. Under sampling, the deliverable reports a QofE adjustment with a confidence interval. Under decomposition, the deliverable reports a contractual variance (exact) and a behavioral variance (calibrated forecast). The two are not commensurable with the single-number output of sampling. The buyer who has learned to interpret QofE deliverables under the sampling paradigm must learn to interpret deliverables that report two different epistemic objects, each with its own appropriate downstream use.

The methodological transformation is not an incremental refinement. It produces categorically different outputs from categorically different inputs through a categorically different process. The diligence industry's continued reliance on sampling is not a marginal choice between competing methodologies of comparable adequacy. It is a choice to operate under a methodological regime whose constraints have lifted.

8 implications across the diligence workflow

The methodological transformation has consequences that extend well beyond the QofE engagement itself. We name them briefly. Each deserves its own development in future notes.

For PE diligence: every QofE finding produced under sampling methodology should be reproducible under decomposed methodology, and the gap between the two will be material in most cases. PE buyers who have built portfolio-level analytical capabilities on top of sampling-based diligence inputs are operating on a foundation that contains the three structural errors described above. The portfolio-level analytics inherit and compound the errors. The transition to decomposed diligence inputs is not optional in the long run; it is the only path to portfolio-level analysis that is not structurally compromised.

For contract negotiation: payer-provider contracts negotiated against sampling-based historical baselines have been settled on incomplete information. The provider that knows its exact contractual entitlement under the existing contract has a different negotiating position than the provider that knows its sampling-based estimate of the entitlement. The renegotiation opportunity, aggregated across the provider population, is large.

For receivables finance: the underwriting of healthcare-receivable-backed instruments — factoring lines, asset-backed loans, securitizations — has been priced against sampling-based estimates of receivable performance. The mispricing is structural and bidirectional. Some receivables have been overpriced relative to their actual contractual entitlement; some have been underpriced. The market clears, but it clears on incomplete information, and the clearing prices reflect methodological noise as much as underlying credit quality.

For regulatory enforcement: audit programs that rely on sampling-based extrapolation should transition to population-scale verification, which the technology now supports. The False Claims Act enforcement framework, the Medicare administrative contractor audit programs, and the commercial payer recovery audit contractor programs all operate under sampling-based extrapolation regimes that produce findings with the same structural errors as PE diligence. The regulatory framework will eventually update; the providers and payers who anticipate the update will adapt earlier than those who wait.

For internal provider operations: the revenue cycle management function within healthcare provider organizations has been built around sampling-based monitoring — periodic audits of denial patterns, sampling-based review of underpayment recovery, exception reporting against statistical thresholds. The decomposed methodology enables continuous, population-scale monitoring that identifies underpayments and contractual violations in near-real-time rather than retrospectively. The operational implications for the provider's working capital and cash conversion cycle are substantial.

Each of these implications is a domain of work in itself. The diligence industry's transition to decomposed methodology is the leading edge, because the diligence context has the highest information density per analytical hour. But the transition will propagate through every adjacent function that has historically relied on sampling-based estimates of healthcare receivables.

9 closing

Sampling was the rational methodology for healthcare diligence under three binding constraints: difficult data acquisition, expensive computation, and infeasible rule-system encoding. Those constraints justified the compromise. The compromise was always epistemically inferior to exact computation; it was adopted because exact computation was not available, not because sampling was preferred on its merits.

The constraints have lifted. Data is available at population scale. Computation is cheap. Rule-system encoding is feasible at the required resolution, enabled by extraction tools that did not exist a decade ago and that operate under verification discipline appropriate to their place in the methodology. The conditions that justified the compromise no longer obtain. The methodology persists not because it remains the right response to current conditions but because the infrastructure built around it — training programs, professional norms, software tools, client expectations, accumulated expertise — has become independent of the conditions that originally justified it. The methodology is now an inherited artifact rather than a defensible methodological position.

The structural errors of sampling — inadequate stratification against population heterogeneity, compounding error through the diligence workflow, information-theoretic waste against a computable underlying function — are present in every sampling-based diligence finding produced under current practice. The errors are not failures of competence. They are inherent to the methodology under conditions

where exact computation is available and is not being performed. The field's most careful sampling work is still subject to all three errors.

The decomposed methodology — encoded rule system, full population computation, exact derivation of the contractual component, statistical modeling of the behavioral residual — is the diligence methodology that current constraints permit and that current standards of analytical rigor require. The transition is not a question of whether but of when. The firms that recognize the methodological obsolescence earliest will produce work of categorically different quality than the firms that continue to refine the inherited methodology. The PE buyers, lenders, and acquirers who receive the work will eventually update their expectations to match what current methodology can produce. The lag between the methodological transformation and the market's adjustment to it is the window in which the recognition is most valuable.

This note has named the obsolescence. The methodology is not wrong because it is poorly practiced. It is wrong because the conditions that justified it have lifted, and the practitioners who continue to refine it are refining the wrong thing.

---

see also

§ M.01 On the deterministic decomposition of healthcare payment.
§ M.03 On the architecture of decomposable targets.