Genesee

A fully labeled synthetic revenue-cycle benchmark

Hammer Labs San Francisco, California hello@hammer.ai

2,400	4,081	16	716
CLAIMS	SERVICE LINES	RECORD TYPES	LABELED FINDINGS

Source: Genesee standard-book manifest and public dataset card, seed 20,260,911.

Abstract

Healthcare revenue-cycle work is the task of following money from a patient's eligibility check, through claim submission, adjudication, settlement and bank deposit, to the moment a payment is posted against the patient's account. Systems that claim to find missing money in this pipeline are hard to evaluate because real hospital records carry no ground truth: nobody knows how much was actually lost. We introduce Genesee, a public synthetic revenue-cycle dataset in which the ground truth is known by construction. Every record is generated, every fault is planted deliberately, and every planted fault is recorded in an answer key. The standard book contains 2,400 professional claims with 4,081 service lines across 16 record types, 200 synthetic patients, 11 payers, and 18 months of activity. The answer key labels 716 findings of 25 kinds, spanning line-, claim-, settlement-, deposit-, and posting-level failures, and includes a negative-control class where the records do not justify a positive finding. The release ships with a schema, declared joins, a generation manifest, an answer format, and a deterministic scorer. Two deliberately weak reference policies establish the floor: flagging nothing scores 0.0% (0/1,391) on dollar-weighted recall, and flagging every denied line scores 36.5% (508/1,391) with 79.8% (796/997) line recall (denominators are in the key's amount units, which this release mislabels; see §4). These are scorer outputs rather than model results. Genesee measures performance on fully observed synthetic episodes; it does not estimate prevalence, recovery, or accuracy on real hospital records.

1. Introduction

Revenue-cycle analysis is a record-linkage and state-reconstruction problem. Eligibility responses precede claims. Acknowledgements and status responses describe transport. Remittances adjudicate individual lines. Settlements group payments, bank deposits establish receipt, and postings establish whether received money reached patient accounting. Each stage is a separate record family with its own identifiers, and the question "where did this dollar go?" can only be answered by joining across them. A claim-only dataset cannot distinguish a payment that was never received from one that was received but never posted; a remittance-only dataset cannot establish whether a settlement reached the bank at all.

Existing public resources do not cover this chain end to end. Synthetic patient generators produce clinical histories, synthetic claims files stop at adjudication, and workflow benchmarks evaluate an agent's actions inside an interface rather than its conclusions about a set of linked records (§2). None of them ship a labeled answer key that says, for each episode, what went wrong, at which stage, and for how much.

Genesee fills that gap with a fully synthetic, fully labeled book. Because every fault is an intervention applied by the generator to an otherwise consistent episode, the local causal truth of each finding is known, and a scorer can credit a system for finding the right money at the right stage rather than for matching a heuristic. The design also treats refusal as part of the task. One planted control class contains services that were billed, adjudicated, paid, deposited, and posted consistently while the clinical indication remains unsettled by the available records; the correct output for those rows is to decline a claim-level accusation. Positive findings and warranted non-findings are therefore evaluated in the same corpus.

The scope claim we make is narrow: to our knowledge, Genesee is the first public synthetic revenue-cycle dataset in which every dollar of leakage, including denials, underpayments, missed charges, timely-filing losses, and contractual adjustments, is labeled in an answer key, and which ships its own scorer. This is a statement about the public resources we surveyed, not a result inferred from the book itself. The current release models professional (837P) claims from one provider; institutional (837I) claims are not included.

Contributions

An end-to-end book. A single package holds source-like records for every stage of the lifecycle, the reference data those records were generated against, the answer key, the scorer, and a manifest. The standard book contains 34 files and occupies 818,613 bytes compressed.

Labels at several grains. The key contains 671 findings attached to claims and 45 findings outside the claim grain (settlements, deposits, and postings). 594 of the 2,400 claims carry at least one planted finding.

Explicit contracts. The schema documents the meaning and source of every field. Join declarations name the keys between record families. The answer schema fixes

the solver's output format. The manifest pins the generator's inputs and the hash of every payload; the public book hash is `de8518bdbdd2fc3bf0d66b2ebc7bb164e000993e5d1d86343c6b556b919bc629`.

An executable evaluation. The bundled scorer reports dollar-weighted recall, false dollars on clean episodes, explanation correctness, line precision, line recall, evidence completeness, latency, and the number of broken cases. Every metric is reported with its numerator and denominator.

2. Related resources

Workflow-agent evaluation

HealthAdminBench [1] evaluates computer-use agents in four deterministic administrative GUI environments, with 135 expert-defined tasks and 1,698 evaluation points across prior authorization, appeals and denials management, and durable medical equipment ordering. Its unit of evaluation is an agent's action in an interface; Genesee's unit is a linked revenue-cycle record and a finding in the answer key. The two are complementary: a record-level solver evaluated on Genesee could supply the evidence an agent acts on in a workflow environment.

Synthetic patient and claims data

Synthea [2] simulates synthetic patient histories from probabilistic disease modules and exports them in standard health-record formats. The CMS Synthetic Medicare Claims Public Use Files [3] provide enrollment and claims for 8,671 synthetic beneficiaries in the Research Identifiable File format; CMS states that the files are intended to build familiarity with the data and to support software development, not inference about Medicare beneficiaries. Genesee differs from both in purpose and in unit: it models the payment lifecycle after the claim leaves the provider, plants faults as named interventions, attaches recoverable amounts to them, and defines a scoring contract.

Dataset metadata

Croissant [4] defines a shared metadata vocabulary for a dataset's attributes, resources, structure, and semantics. Genesee's manifest and schema pursue the same machine-readable goal within a single domain package. The current release does not claim Croissant conformance.

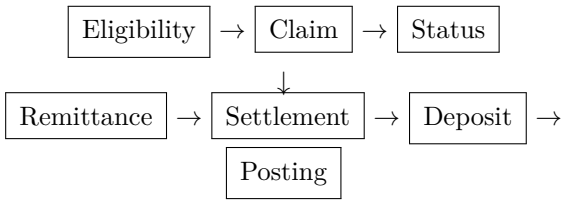
3. Dataset construction

Synthetic, calibrated, and declared

Every patient, provider, claim, remittance, settlement, deposit, posting, and clinical note in Genesee is fabricated. The provider's identifier is deliberately check-digit invalid so that it cannot be mistaken for a real NPI. The shape of the population is drawn from named public aggregates: the code mix, age bands, payer mix, fee-schedule values, published rates, and edit rules are each recorded inside the book together with the source they were taken from and an evidence grade. Fault rates are declared generator parameters, not measured prevalence.

The standard configuration uses seed 20,260,911 and produces 2,400 claims over 18 months for 200 patients, 11 payers, and 17 procedure codes. The manifest is intended to regenerate the same bytes from the same inputs; the book hash and per-file hashes make any drift visible.

Lifecycle and intervention point



The generator first creates an internally consistent episode and then applies a named fault mutation to it. A mutation may alter the claim, the remittance, the settlement, the deposit, the posting, the coverage policy version, the authorization, or the clinical note. Each answer-key row records the affected identifier, the kind of fault, the interpretation a solver is expected to reach, and, where the fault has a price, the amount. Because the intervention is known, the local causal truth of each finding is available without any inference from the records themselves.

Finding distribution

The answer key contains 716 rows across 25 kinds (Appendix A). The largest classes are not interchangeable. `improper_payment_findable` is a signal visible in the claims data alone. `improper_payment_needs_chart` is the negative control: the records are consistent and the correct output is restraint. `precedent_break` is a longitudinal deviation from the provider's own history. The remaining classes cover coding edits, authorization, documentation, timely filing, remittance gaps, deposits, takebacks, and postings.

Record inventory

Appendix B lists every file in the book with its row count. The corpus families follow the X12 transactions they stand in for: 837P claims, 835 remittances, 271 eligibility responses, 277CA acknowledgements, and 276/277 status responses, together with settlements, deposits, postings, prior authorizations, coverage policies with versions, a contract with its terms, a charge master, a roster, and clinical notes. Reference data (expected-amount bases, payer published medians, and NCCI edits) ships alongside so that pricing is reproducible from the package alone.

4. Task and scoring

Input and output

The episode builder presents one case per claim, holding every record in the book that links to it, with each line priced against a named basis. The solver returns an episode-level reason, line-level findings keyed by line control number, the outstanding amount and the stage at which it went missing, the first date at which the fault was detectable, and the identifiers of the records cited as evidence. The solver may refuse. The answer schema distinguishes a missing or unknown value from zero.

Metrics

Dollar-weighted recall credits missing money found at the correct stage. *False dollars on clean episodes* sums priced false findings on episodes the key calls clean, and becomes unknown, rather than understated, when any positive finding cannot be priced. *Explanation correctness* compares the episode reason code to the key. *Line precision* and *line recall* operate at the control-number grain. *Evidence completeness* checks that the cited record identifiers exist and are relevant. Broken payloads are counted and excluded from every other denominator.

Known defect in this release

Version 2026.09.12 carries a unit defect in the money metrics, which we report rather than relabel. The episode builder truncates the key's dollar fields with `int()` into a field named `missing_minor_units`, so dollar-weighted recall in this release runs over truncated dollars labeled as cents. In addition, the key's amount fields mix units by record family: claim-level rows are in dollars while settlement-level rows are in integer minor units, as their source records are. The ratios reported below are scorer outputs and are unaffected in kind; the unit attached to their numerators and denominators is not reliable and should not be read as cents or as dollars. Both defects are fixed in the next generator release, after which the reference policies are re-scored.

5. Reference policies

Policy	DWR	Line recall
Flag nothing	0.0% (0/1,391)	0.0% (0/997)
Every denied line	36.5% (508/1,391)	79.8% (796/997)

Table 1: Reference policy outputs from the bundled scorer, $n = 2,400$ cases. DWR is dollar-weighted recall; its numerator and denominator are in the key's amount units, which this release mislabels (§4). Line recall is over 997 lines the key marks as unpaid.

We report two policies that no one would deploy, to show what the scorer rewards. The *flag-nothing* policy answers every case as fully reconciled. It finds no missing money and no unpaid lines, and its 84.7% (2,033/2,400) explanation correctness is the share of cases that are in fact clean. The *every-denied-line* policy flags each line whose remittance carries a denial. It reaches 99.7% (796/798) line precision and 79.5% (1,907/2,400) explanation correctness, but emits 715 findings it cannot price, so its false-dollar total is reported as unknown rather than as zero. Neither policy reads clinical notes, longitudinal precedent, policy versions, deposits, or postings.

Three properties of the evaluation follow. Denial detection is not revenue-cycle reconstruction: the denied-line policy recovers most unpaid lines and little more than a third of the missing money. High precision can coexist with incomplete recall. And an accounting quantity that cannot be computed is reported as unknown, not as its lower bound.

6. Limitations

The generator's tests compare the same seeded book under distinct hash seeds and pin fixture hashes; the dataset page checks its counts against the embedded card; and a bundled self-check verifies manifest hashes, schema expectations, and the presence of the scorer. These establish package integrity and determinism within the tested implementation.

They do not establish clinical realism, prevalence calibration, transfer to other markets, or accuracy on real records. The public book contains one synthetic provider shaped after one market. Claims are professional (837P) analogues; institutional (837I) claims, which carry most of a hospital's revenue, are not modeled in this release. Payer identities were chosen for the coverage of their published rates, not for any inferred contract relationship. The amounts in the key are exposures within a synthetic episode, not recovered cash.

Regeneration on an independent machine has not been performed, so we do not claim cross-machine byte reproducibility. A held-out evaluation requires an unpublished seed and a separately retained answer key; the public standard book alone cannot prevent a system from being tuned to its key.

7. Responsible use

Genesee is suitable for scorer development, record-linkage testing, retrieval evaluation, abstention analysis, regression testing, and agent research. Results should be reported with the dataset version, seed, book hash, generator commit, task configuration, numerator, denominator, and broken-case count. No figure derived from Genesee describes a real hospital, payer, or patient, and none should be presented as if it did.

Comparisons should hold the scorer version fixed. A new generator build that changes the package but not the

modeled outside world is a build, not a new vintage; a change to the external calibration is a new vintage.

8. Availability and citation

The dataset page at https://benchmark.hammer.ai/genesee generates the standard book on request and publishes the schema, sample records, task contract, reference-policy table, version notes, and citation text. The package includes the scorer and the answer schema. This paper and its LaTeX source are built from the same public JSON card as the page; the generated macros record the card's SHA-256 as `e5b60d5e9d78b2d52ae999937fe15bbb61ea42d1edb918c3aa6d387c422ef762` and the generator commit as `v0.3.0-295-g0805190`.

Suggested citation: Hammer Labs. *Genesee: A Fully Labeled Synthetic Revenue-Cycle Benchmark.* Dataset version 2026.09.12, seed 20,260,911, generator `v0.3.0-295-g0805190`. https://benchmark.hammer.ai/genesee.

References

[1] S. Bedi et al. *HealthAdminBench: Evaluating Computer Use Agents on Healthcare Administration Tasks.* arXiv:2604.09937, 2026. https://arxiv.org/abs/2604.09937

[2] J. Walonoski et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record." *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018. https://doi.org/10.1093/jamia/ocx079

[3] Centers for Medicare & Medicaid Services. *Synthetic Medicare Enrollment, Fee-for-Service Claims, and Prescription Drug Event Data: User Guide.* Version 1.0. https://www.cms.gov/data-research/statistics-trends-and-reports/medicare-claims-synthetic-public-use-files

[4] M. Akhtar et al. "Croissant: A Metadata Format for ML-Ready Datasets." *NeurIPS Datasets and Benchmarks Track*, 2024. NeurIPS proceedings

Appendix A. Answer-key distribution

Finding kind	Rows	Share	Relative count
improper payment findable	162	22.6%	
improper payment needs chart	150	20.9%	
precedent break	106	14.8%	
ncci units cut	58	8.1%	
ncci pair denied	35	4.9%	
silent zero	24	3.4%	
deposit missing	18	2.5%	
provider level adjustment	15	2.1%	
member id stale	14	2.0%	
chart units unsupported	13	1.8%	
trace mismatch	12	1.7%	
auth missing	11	1.5%	
chart dx missing	10	1.4%	
modifier underpay	10	1.4%	
duplicate denial	9	1.3%	
policy wrong version	9	1.3%	
chart signed after billing	8	1.1%	
posting missing	8	1.1%	
sequencing	8	1.1%	
auth units short	7	1.0%	
takeback	7	1.0%	
filed after timely filing	6	0.8%	
posting misapplied	6	0.8%	
auth expired	5	0.7%	
claim replaced	5	0.7%	

Source: Public answer-key aggregation, $n = 716$ rows. Bar length is relative to the largest class; the share denominator is 716.

Appendix B. Record inventory

Record family	File	Rows
Claims (837P)	corpus/claims_837.json	2,400
Remittance advice (835)	corpus/remits_835.json	2,407
Settlements (BPR/TRN)	corpus/settlements.json	1,629
Bank deposits (CCD+)	corpus/deposits.json	1,611
Cash postings	corpus/postings.json	2,392
Answer key	corpus/planted.json	716
Eligibility responses (271)	corpus/eligibility_271.json	2,400
Acknowledgements (277CA)	corpus/acknowledgements_277ca.json	2,400
Claim statuses (276/277)	corpus/statuses_277.json	2,400
Prior authorisations	corpus/authorizations.json	1,185
Authorisation requirements	corpus/authorization_requirements.json	256
Clinical notes	corpus/charts.json	2,400
Coverage policies, with versions	corpus/coverage_policies.json	1,265
Enforced policies	corpus/enforced_policies.json	695
Contracted rates	corpus/contract.json	1
Contract terms	corpus/contract_terms.json	11
Charge master	corpus/charge_master.json	690
Rendering providers	corpus/roster.json	14
Provider	corpus/provider.json	1
Patients	corpus/patients.json	200
Population summary	corpus/population.json	1
Expected-amount bases	expected_bases.json	155
Payer published medians	calibration_pairs.json	510
NCCI edits	ncci/ptp-edits.json	115 codes; 108 MUE; 787 pairs
Manifest	manifest.json	1

Source: Public schema in the dataset card, version 2026.09.12. Counts are rows or objects as defined by each record family.

Appendix C. Episode reason distribution

Reason	Cases
claim replaced	5
fully reconciled	2,033
line adjudicated zero	290
line not adjudicated	24
payment not received	34
posting misapplied	6
posting missing	8

Source: Truth reason codes from the public scorer, $n = 2,400$ cases.