309

A Red-Teaming Framework Using Synthetic Healthcare Claims to Evaluate Data Product Accuracy at Scale

Kian Jalaleddini, Chris Ware, Shae Schenk, Xiaoyan Wang, Gabriel Centeno, and Paul Gurney
Komodo Health, New York, NY, and San Francisco, CA

Abstract

Evaluating the accuracy of large-scale healthcare data products is challenging without ground truth. We present a red-teaming methodology for assessing the Komodo Patient Insurance (KPI) data product, which estimates insurance coverage using de-identified open claims data. This approach generates high-fidelity synthetic patient claims to simulate real-world complexity with known ground truth for benchmarking.

Clinical Relevance

This method could be of interest to those involved in evaluating the accuracy and reliability of healthcare data used in clinical and research settings.

Introduction

- Evaluating large-scale healthcare data product accuracy is difficult without ground truth
- Red-teaming involves forming an independent "red team" to systematically challenge an organization's plans, assumptions, and vulnerabilities from an adversarial perspective
- Red-teaming is being increasingly adopted across various domains, including healthcare, to rigorously test systems and identify weaknesses before they can be exploited[1]
- This poster presents a red-teaming methodology to assess the KPI data product, which estimates insurance coverage using de-identified open claims data
- Our approach leverages high-fidelity synthetic patient claims, which are artificially generated datasets designed to mimic the complexity of real-world patient data
- The use of synthetic data in healthcare is gaining significant traction due to its ability to overcome the challenges associated with real data (including privacy) and has been demonstrated as viable through several initiatives[2]

***KPI** is a comprehensive data asset that provides a longitudinal view of the primary and secondary insurance benefits for over 200 million de-identified patients. It was built to solve a core challenge in healthcare analytics: Insurance details extracted from individual medical or pharmacy claims often reflect temporary billing processes rather than a patient's true, continuous coverage. To create a more accurate picture, KPI integrates two key data sources: payer-derived enrollment files, which serve as the ground truth for when a patient is covered; and open-source events data, which enriches this information and fills in gaps. By combining and validating these sources, KPI produces a detailed, patient-centric timeline of insurance status, enabling a deeper and more accurate understanding of the healthcare journey.*

Methods

- **Cohort generation:** We simulated 1,000 synthetic patients, each with a unique RNG seed and distinct demographics (sex, year of birth/death, race, ethnicity), plus geographic variation and utilization patterns calibrated to Komodo's Healthcare Map® distributions
- **Insurance enrollment:** For each patient, the number of insurance plans followed a Poisson distribution. Plan spans do not overlap and have interval start/end dates that are sampled uniformly from January 1, 2016 to January 1, 2024
- **Claims and sources:** We generated medical and pharmacy claims where the number of encounters per patient is drawn from a Gamma–Poisson distribution (i.e., negative binomial) to induce realistic overdispersion. Each claim was then observed by a random subset of data sources; sources are heterogeneous and partially overlapping and can be error-prone, producing imperfect, discordant views of the same underlying claim
- **Temporality:** Service/fill dates were sampled conditional on the Gamma–Poisson intensity, yielding temporally plausible sequences that naturally produce rare edge cases (e.g., dense bursts, sparse tails)
- **Ground truth:** For every patient, we retain the canonical claim stream and enrollment spans (payer entity IDs and dates) as ground truth; source-level claims carry mapped payer identifiers (e.g., BIN/PCN or source-payer IDs) and injected errors
- **KPI evaluation:** We assessed KPI plan inference against ground truth on a patient-month basis (patient–time unit where enrollment is well-defined). For each patient-month:
  - **True positive (TP):** Patient is enrolled, and KPI correctly identifies the plan
  - **True negative (TN):** Patient is not enrolled, and KPI correctly identifies no plan
  - **False positive (FP):** Patient is enrolled, but KPI identifies a wrong plan; OR patient is not enrolled, and KPI wrongly identifies any plan
  - **False negative (FN):** Patient is enrolled, but KPI fails to identify any plan

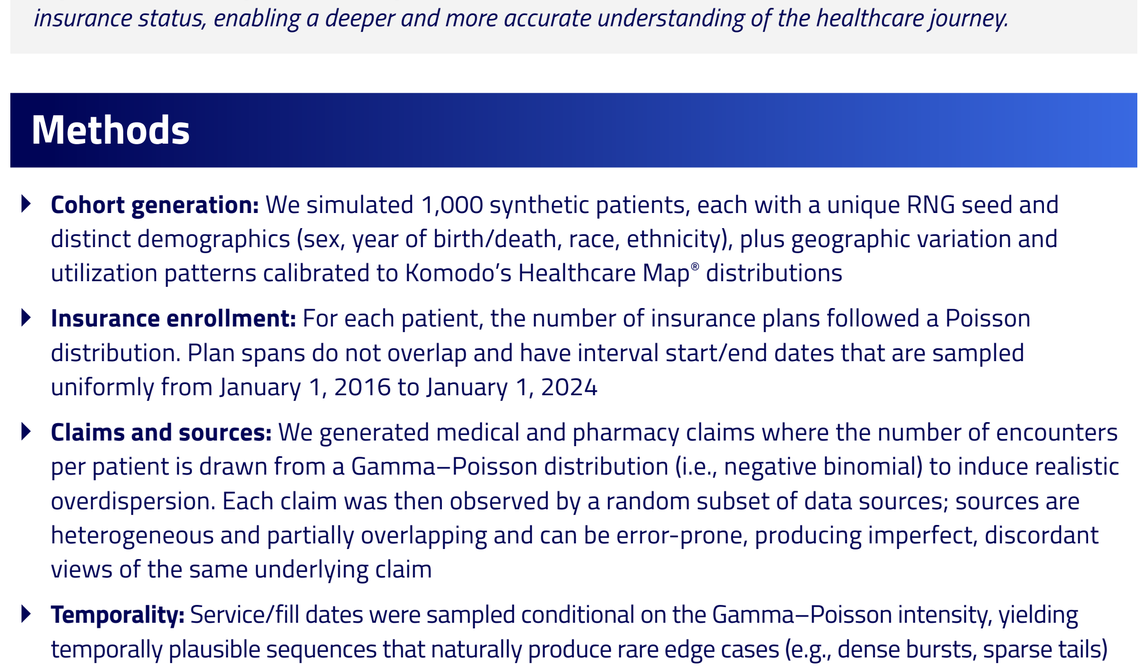

Figure 1. Gamma–Poisson model for patient utilization. A Gamma (α=0.16, β=0.02) distribution was used to model heterogeneity in patients' claim frequency rates (λ). Each patient's λ was sampled from this Gamma distribution, and their total number of claims was drawn from a corresponding Poisson (λ). The resulting Poisson distributions represent realistic variations in healthcare service use across patients.

Results

- Accuracy of inferred insurance plans was quantified using classification metrics (utilization-weighted accuracy and F1-score)
- Early findings show a strong correlation between patient utilization rates and prediction accuracy
- The red-teaming framework identified product limitations affecting specific subpopulations that are slated for future enhancements

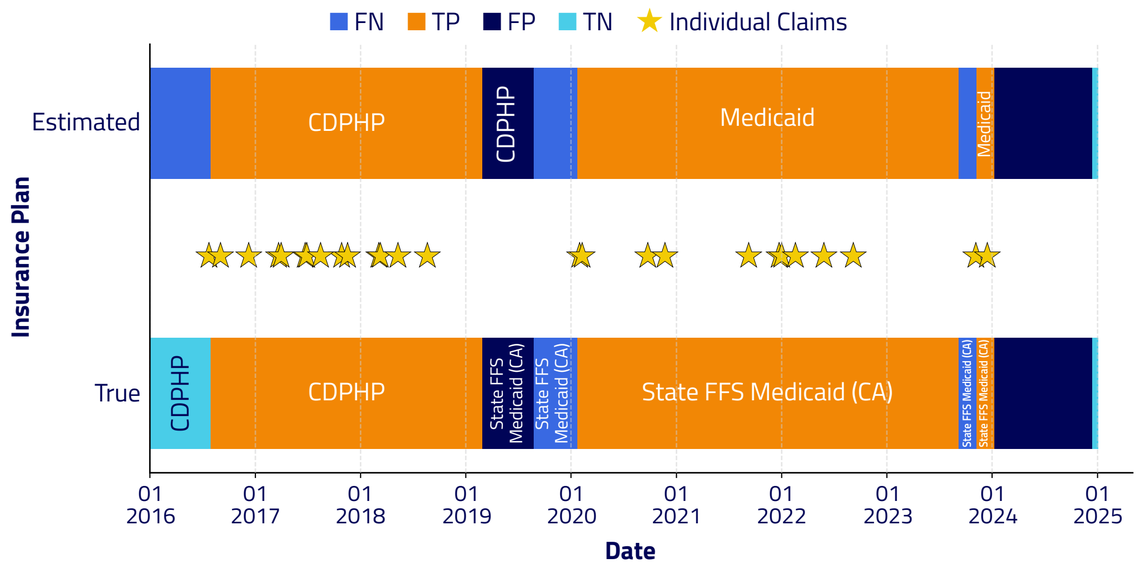

Figure 2. Patient-level insurance timelines (truth vs. KPI estimate). Horizontal bars show contiguous insurance enrollment spans over calendar time and by source of truth vs. model output. Bar colors encode match quality: TP (orange), FP (dark blue), FN (blue), TN (light blue). Yellow stars mark individual claim service/fill dates positioned between the paired rows for each claim type. The visualization summarizes where and when enrollments align or disagree, from which a weighted accuracy of 0.71 is computed.

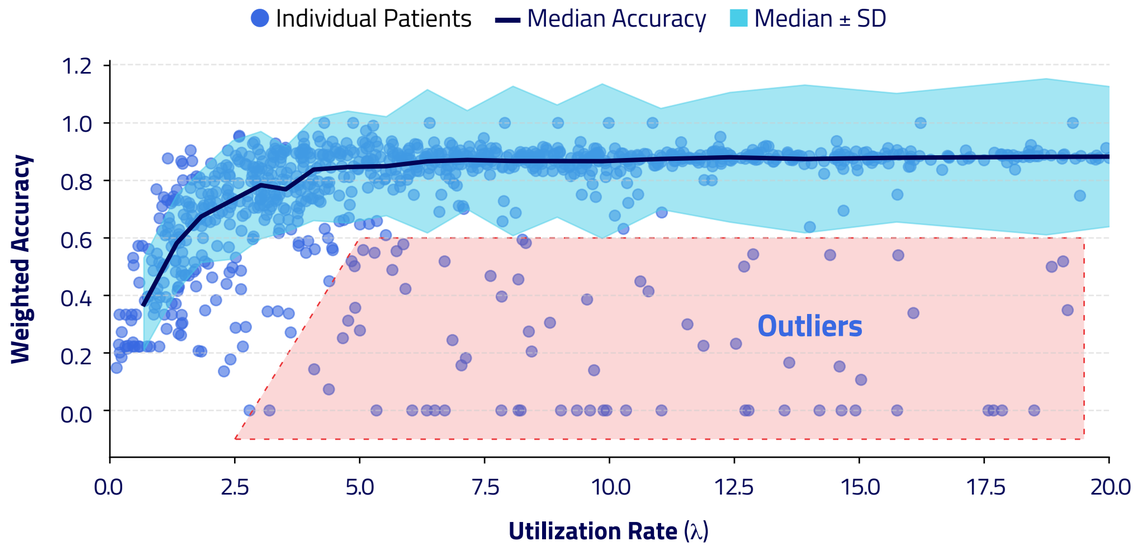

Figure 3. Weighted accuracy versus utilization rate. Each point represents a synthetic patient showing the relationship between healthcare utilization rate (λ) and KPI's weighted accuracy. The solid line traces the median accuracy across equal-sized λ bins, with the shaded region denoting ±1 standard deviation. The highlighted red polygon marks an outlier region (λ≈2.5–20) where model accuracy exhibits elevated variability among high-utilization patients.

		Predictions	
		Positive	Negative
Actual	Positive	2.3M (days)	513K (days)
	Negative	298K (days)	100K (days)

Table 1. Confusion matrix of insurance enrollment accuracy (weighted by days): Each cell represents the total number of patient-days classified into TP, FP, FN, and TN, based on alignment between the model-estimated and true insurance coverage spans. Day-level weighting accounts for the temporal duration of correct and incorrect classifications.

Discussion & Conclusion

- This work introduces a scalable, principled, and domain-aware framework for end-to-end testing of the robustness of healthcare data products, ensuring industry-leading data quality
- This approach is relevant for evaluating data used in clinical and research contexts
- Our synthetic claims simulate real-world complexity with known ground truth, providing a robust benchmark for evaluating the KPI data product
- This work highlights the value of synthetic data as a reliable proxy for real-world data in scenarios where ground truth is paramount

References

1. Ganguli D et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv reprint. 2022. arXiv:2209.07858
2. Walonoski J et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *JAMIA*. 2018;25(3):230-238.

The authors wish to thank Kim Jacoby and Alona May for their input in preparing this work.

This work is sponsored by Komodo Health. The authors are employees of Komodo Health and have shares or options.

Scan here to download poster or inquire for more info.

komodo®