Protege

CLERK by Protege

Care Lifecycle Events and Records Knowledgebase (CLERK)

A Curated EHR × Claims Data Product for Training AI Models for Healthcare Administration Tasks

CLERK BY PROTEGE

Table of Contents

Introduction

AI applications are capable of transforming every segment of healthcare, but deployment at scale remains a persistent challenge. A recent global survey found that only 1 in 4 companies focused on AI were able to successfully implement an AI model in production over a 12-month period, with another reporting that 21% of generative AI initiatives failed to scale due to computational costs alone. However, while institutional implementation can be slow, many physicians are independently adopting AI tools in their daily practice: a recent AMA survey found that 2 in 3 physicians are now using health AI, representing a 78% increase from 2023 (AMA, 2024).

Administrative costs in healthcare are widely recognized as a key factor distinguishing the United States' healthcare spending, which is approximately double that of other developed nations. Research has shown that U.S. healthcare requires twice as many administrators per patient encounter compared to Canada, highlighting significant inefficiencies. Consequently, automation using AI presents substantial opportunities to enhance productivity and reduce unnecessary spending within the U.S. healthcare system.

Healthcare administration tasks include billing, coding, prior authorization, and claims management, all of which are well-suited tasks for AI. These tasks demand intricate analysis of both clinical data (from electronic health records) and financial data (claims information) to effectively support administrative workflows. Research estimates that approximately $265 billion (representing about 7% of total $4 trillion in healthcare expenditure) could be saved through targeted interventions in administrative efficiency (Health Affairs 2022, JAMA 2019).

The magnitude of these savings raises a second question: If AI in healthcare can be so valuable, what is inhibiting its deployment in medical coding and billing at a higher velocity? We believe the limitation is data. Today, many AI builders leverage off-the-shelf large language models (LLMs)—which, as a recent NEJM study highlighted, are not yet optimized for medical coding—and adapt them using customer-specific data to create specialized administrative AI products. The typical product development lifecycle involves a builder leveraging customer-specific data for training, leading to:

Long deployment times
Customer data must be ingested, a bespoke model is created, and then deployed.

Data sharing security and privacy concerns
Transfer of healthcare data between customer and AI builder carries inherent risks.

Insufficient data volumes
Individual customers typically lack the number of varied observations for robust model adaptation.

Expensive, bespoke model development
Compute costs are high when training is done on a per-customer basis, and products also do not scale across customers.

Access to large, well-curated, diverse datasets for training LLMs in healthcare would alleviate some of these burdens, but they are difficult to acquire. Curation of such a data set requires connecting two distinct and complex data types: **Electronic Health Records (EHR)** and **claims data**. However, several challenges persist:

ⓘ FRAGMENTATION OF EHR DATA

EHR records are scattered across diverse settings—urban and rural, large hospital systems, community hospitals, and independent clinics—making it challenging to aggregate a database of the approximately 1 million healthcare providers in the U.S.

ⓘ ANCHORING COMPLEXITY WITH OPEN CLAIMS

Effective pairing of an EHR encounter with its respective claim requires open claims (provider-centric data) rather than closed claims (patient-centric historical data), as open claims comprehensively capture all encounters from the provider's perspective, enabling accurate alignment with corresponding EHR records.

ⓘ RAPIDLY EVOLVING RULES AND NEED FOR COMPREHENSIVE DATA

Insurance rules and value based purchasing billing standards evolve quickly, necessitating continuous data refreshes.

ⓘ DATA VOLUME AND REPRESENTATIVENESS

Encounters that follow the quickest path to implementation (e.g., outpatient surgical procedures) often occur in smaller, independent clinics, limiting data volume available from any single care site.

These challenges have significantly lengthened the training-to-deployment cycle and impeded the availability of ready-to-use applications.

CLERK: Protege's Solution

CLERK is designed to eliminate these barriers by offering AI developers access to a **highly curated, encounter-linked EHR and claims dataset**, purpose-built for scalable training of AI models in healthcare administration. Below, we describe how we build our product along with a how we validated it on a subset of data in our catalog.

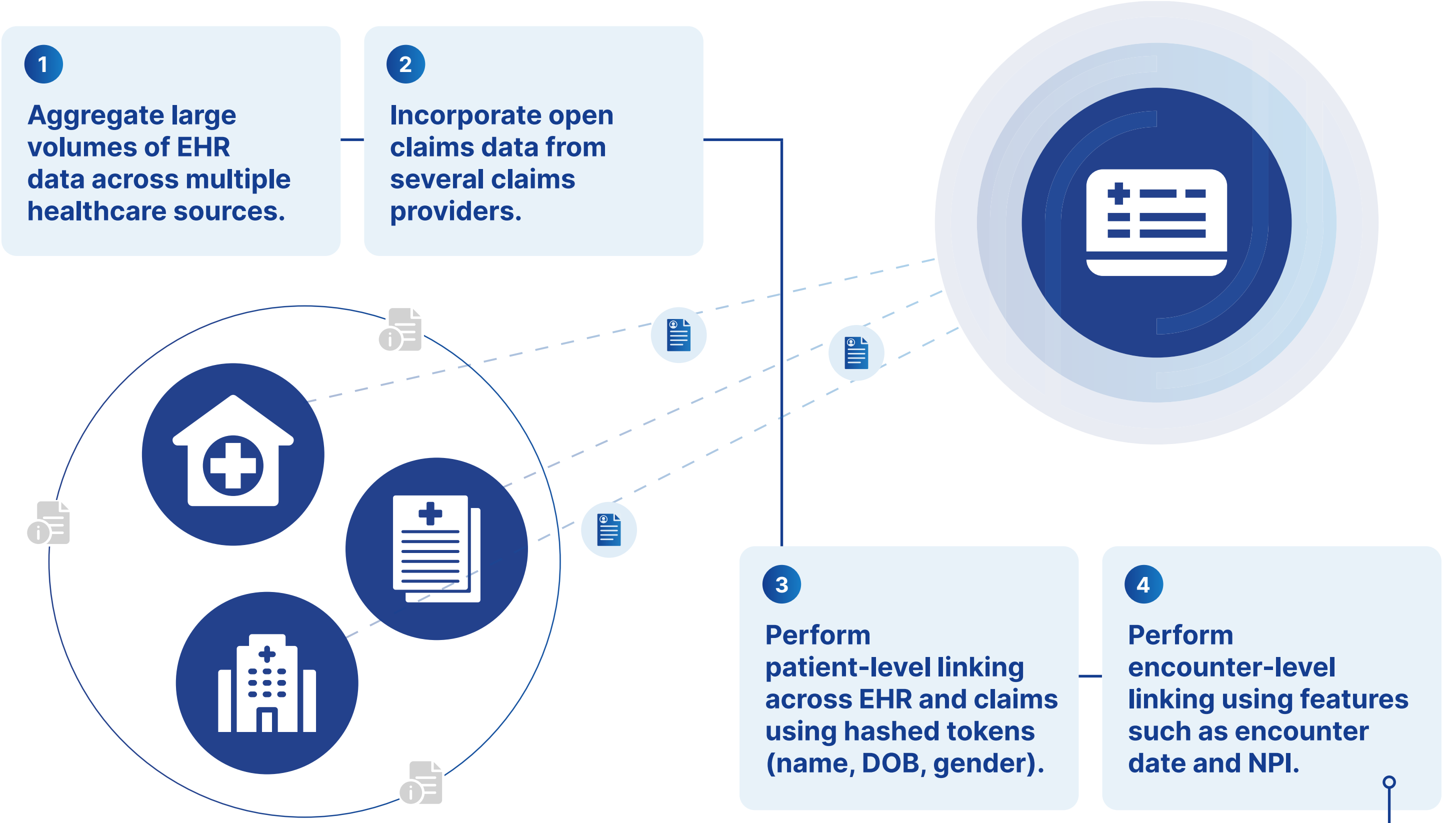
Our solution consists of two core components:

1

Encounter-Level Linked EHR/Claims Dataset

Protege aggregates and meticulously links EHR and claims data at the **encounter level,** achieving the depth and breadth necessary for training robust AI models. Below, we describe our validation process using a subset of data in our catalog. At scale, our full set of encounter-level linked EHR x claims data covers tens of millions of encounters.

KEY STEPS



1 Aggregate large volumes of EHR data across multiple healthcare sources.

2 Incorporate open claims data from several claims providers.

3 Perform patient-level linking across EHR and claims using hashed tokens (name, DOB, gender).

4 Perform encounter-level linking using features such as encounter date and NPI.

VALIDATION OF ENCOUNTER-LEVEL LINKING

To validate our methodology for encounter-level linking, we leveraged two very large datasets of EHR and open claims, both aggregated from multiple sources. Our validation process then proceeded as follows:

1) Subsample 15M patients who appear in both EHR; linked deterministically across data sources. Across these 15M patients, 752M claims were identified.

2) Identify EHR x claims linked encounters by connecting on NPI from both claims and the EHR system and the encounter date. Note that we implement a date-offset of one day. Across all 752M claims, we identified 30M claims x EHR linked encounters.

> At this stage, we created a **analysis arm** and a **falsification arm** of our cohort. The validation arm linked encounters using the methodology described above, and the falsification arm linked encounters using a randomly shuffled demographic token.

3) Validate encounter-level linkage by measuring the share of linked encounters where the ICD codes in the claim and EHR match, and comparing this match rate between the analysis arm and the falsification arm. We found an 80% concordance rate of ICD codes between EHR and claims encounters in the analysis arm, whereas ICD concordance was consistently less than 2% in the falsification arm. This process demonstrates that our method of connecting encounters results in a high volume of true matches between EHR and claims sources. The table below shows three examples of ICD concordance at the encounter level.

ICD Comparison Samples to Illustrate Negatives and Positives

ID	NPI (EHR)	NPI (CLAIMS)	Date (EHR)	Date (CLAIMS)	ICD (EHR)	ICD (CLAIMS)	Match?
1	112212	—	2023-02-01	2023-02-01	I10	Z00	No
2	996425	996425	2025-01-12	2025-01-12	E66	T84 - device comp.	No
3	643901	643901	2023-06-04	2023-06-04	R30, N39, Z78	R30 - urinary symp.	Yes

2

Specialty-Specific Data Curation

Many AI builders are interested in adapting their models for a specific clinical specialty, or for a specific healthcare administration use case. The second and final step of curating the CLERK training data set involves the following components:

Curating for specialty coverage

A critical element of a training dataset for administrative tasks such as billing and coding is observation volume across all specialty-relevant CPT codes. Generalizable models require many EHR observations of the same CPT code in order to learn variations in clinical documentation. Across the subset 30M encounters described in the validation study above many highly uncommon codes were observed over 100 times. Below, we show three examples of CPT codes specific to cardiology that vary in how commonly they are observed in the real world, and show how commonly they were observed in our subset of 30M encounters.

Many highly specific CPT Codes are observed at least **100 times**

EXAMPLE

75822

Venography, extremity, unilateral, radiological supervision and interpretation

Many distinct CPT Codes are observed at least **1000 times**

EXAMPLE

33249

Insertion or replacement of permanent implantable defibrillator system

Many common CPT Codes are observed at least **100,000 times**

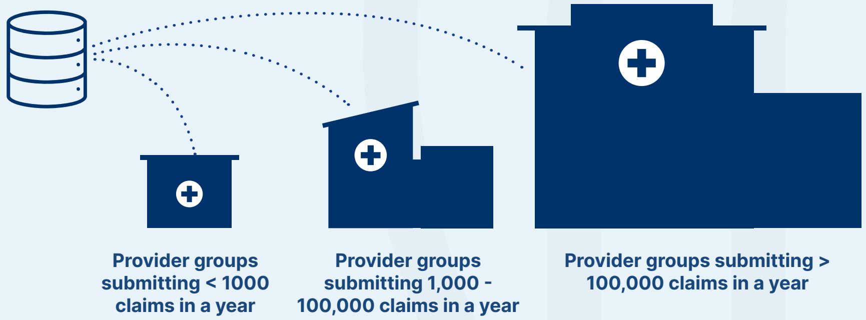
EXAMPLE

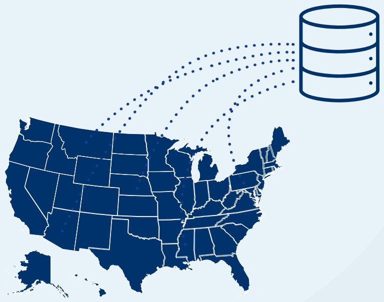
93000

Routine electrocardiogram with at least 12 leads

Ensure generalizability across clinical settings

Clinical documentation can vary widely across the U.S., influenced by both geographic differences and individual provider practices. The data leveraged to build the CLERK product, contains records from nearly all US states, and provider practices ranging from very small (submitting less than 1,000 per year) to extremely large (submitting more than 100,000 claims in a year).


Provider groups submitting < 1000 claims in a year
Provider groups submitting 1,000 - 100,000 claims in a year
Provider groups submitting > 100,000 claims in a year



Bias mitigation

As datasets become more curated, they often become biased—a direct reflection of the classic **precision-recall tradeoff.** Higher precision in matching claims to EHR typically narrows the dataset to a specific subgroup of claims. Bias emerges if this subgroup is correlated with claim success or denial.

Claims are denied for many reasons: improper billing is just one, while others, like lack of beneficiary enrollment (particularly common in Medicaid), are equally important. To build effective AI automation that boosts successful claim processing, **models must be trained on claims with diverse denials beyond coding errors**. It's crucial to evaluate whether high-precision matching disproportionately captures only straightforward or uncomplicated insurance scenarios. One bias mitigation exercise we consider directly addresses this challenge.

Second, we create a broader "**bias summary index**" by modeling the probability that a claim successfully matches an encounter. Specifically, we regress an indicator for successful matching on various claim characteristics—including demographic data, healthcare costs, disease severity, and insurance type—capturing both linear and nonlinear effects across potentially hundreds of features. Characteristics significantly associated with a lower probability of matching reveal systematic differences, helping quantify the bias between matched and unmatched claims.

These quality controls ensure that AI models trained on CLERK data perform reliably across diverse clinical and administrative environments.

Conclusion

CLERK unlocks the potential for AI builders to move beyond bespoke, slow-to-deploy models by providing **scalable, generalizable training datasets.** This in turn enables:

- **Faster and more secure product development cycles**
- **More cost-effective product development**
- **Increased ability to serve multiple customers with off-the-shelf models**

We believe CLERK is a foundational product for the next wave of healthcare administrative AI innovation, empowering faster deployment, better generalization, and ultimately, better healthcare outcomes.

We are continuously expanding our data curation techniques and exploring additional strategies for enhancing encounter-level linkage fidelity and reducing bias—stay tuned for future updates to the CLERK product suite.

Protege

withprotege.ai