TRUVETA

Truveta Language Model

December 2024

Contents

Introduction

Healthcare organizations generate an immense volume of data, with the average hospital producing roughly 50 petabytes of data a year. This includes semi-structured electronic health record (EHR) data, unstructured clinical notes, medical imaging studies, and genomic data. However, an estimated 95 percent of this data goes unused – largely because it is fragmented, inaccessible, and unstructured.

Artificial intelligence (AI) presents an opportunity to unlock the insights hidden within healthcare data in a timely, efficient, and scalable way and transform care delivery. By cleaning massive daily streams of clinical healthcare data from across the country, AI can facilitate access to research-ready clinical data at scale. This, in turn, can lead to breakthroughs that accelerate therapy adoption and improve patient care.

The cornerstone of Truveta's AI is the Truveta Language Model (TLM), a large-language, multi-modal AI model used to clean billions of daily EHR data points for scientifically rigorous research. TLM fine-tunes open large language models with additional training on de-identified EHR data from more than 120 million patients, including 8.4 billion diagnoses, 4.1 billion encounters, and 4 billion medication orders. TLM's healthcare expertise is trained on the largest collection of complete medical records representing the full diversity of the United States.

This specialized training on electronic health records data is one of the core features that sets TLM apart from general large language models, which understand language but are inaccurate within the medical domain due to being trained on the public Internet. TLM's specialized training on healthcare data is critical for ensuring the clinical validity of the content being normalized. Additionally, our rigorous data quality standards ensure regulatory-grade evidence of data accuracy.

This whitepaper explains how Truveta's clinical expert-led AI handles the ingestion and cleaning of healthcare data to ensure high-quality inputs for research. For detailed information about our approach to data quality or patient privacy, see our other whitepapers.

Truveta is a growing collective of 30 health systems committed to Saving Lives with Data. Member health systems provide complete EHR data for more than 120 million patients, which are then linked with social drivers of health (SDOH), mortality, and claims data to provide a complete, longitudinal view of patient journeys.

Cleaning EHR data using the Truveta Language Model

TLM can clean all types of EHR data, whether semi-structured data such as lab tests or diagnoses, or unstructured data such as the contents of clinical notes or imaging reports. This process is complex, as most healthcare information documented in the EHR is not standardized. There are millions of ways clinicians, hospitals, and health systems express observations, diagnoses, medication plans, and other clinical concepts. For example, a clinician might document COVID-19 as "acute COVID -19," "COVID," "COVID-19," "COVID infection," or "COVID19 _ acute infection" and Ibuprofen usage as "600mg Ibuprofen" or "Ibuprofen 600mg tablets by mouth." Before TLM, this variability presented a very expensive data cleaning challenge.

With different types of data, TLM learns how to normalize raw medical text to the most appropriate medical information ontology:

Concept Type	Ontology
Diagnoses	SNOMED, ICD
Lab Tests	LOINC, UCUM
Drugs	RxNorm, NDC
Devices	GUDID
Procedures	CPT, HCPCS, ICD10PCS
Vital signs and observations	LOINC, SNOMED
Immunizations	CVX
Genomics	HGNC
Site of care	CMS Place of Service
Provider	NPPES NPI Registry

Fig 1. TLM maps clinical concepts to standard medical ontologies.

The below figure offers an example of TLM's data cleaning process applied to lab test results. Here, TLM structured two sets of lab test results into four rows of the LabResult table within the Truveta Data Model (TDM). Each test is mapped to a standard medical ontology with standard units of measurement.

Raw medical record text	Lab Results data after TLM normalization		
	Lab Name (LOINC)	Unit (UCUM)	Value
RBC COUNT, RBC|CBC WITH AUTOMATED DIFF |3.80| M/uL|2.70 |4.90	789-8	10*6/uL	3.80
CBC: 3/9 07:45PM WBC-8.1 RBC-3.89 Hgb-11.7	6690-2	10*3/uL	8.1
	789-8	10*6/uL	3.89
	718-7	g/dL	11.7

Fig 2. Example of TLM mapping lab results to the appropriate standard medical ontology.

Extracting concepts from clinical notes

In addition to semi-structured EHR data as shown above, TLM can normalize concepts contained in unstructured clinical notes, which offer critical information for understanding longitudinal patient journeys. In fact, notes contain nearly 80% of clinical data relevant to research, such as information about family history, disease staging, adverse events, symptom severity, reasons for a medication change, interpretations of findings, recommendations for follow-up, and other clinical context.

Truveta receives all clinical notes written during a patient's care, including progress and procedure notes; nursing evaluations; pathology and diagnostic imaging studies; laboratory reports; consulting, referral, history, physical, and plan of care notes; evaluation and plan notes; surgical operation notes; discharge summaries; and more. These pieces of information may offer researchers access to critical measures of interest or help contextualize other data points. Truveta Data today includes more than 5 billion notes from more than 120 million patients.

Fig 3. TLM extracts critical data points from clinical notes, such as disease staging, adverse events, and medication rationale changes.

To clean the clinical concepts contained in notes and add them to TDM, Truveta must first identify and extract those concepts. Using a custom tool, Truveta's clinical expert annotation team labels raw clinical terms associated with a given concept (e.g., ejection fraction, seizure frequency, migraine severity) to train and evaluate TLM with a focus on clinical accuracy. This annotation process is complex and nuanced, and accounts for the immense variation of human language, including misspellings, abbreviations, negation (e.g., "patient denies feeling fatigued"), hypotheticals/conditionals (e.g., "Will consider starting low-dose glipizide if A1C still grossly elevated"), and family history (e.g., "Family Hx: Mother: Diabetes, Father/son: bipolar disorder").
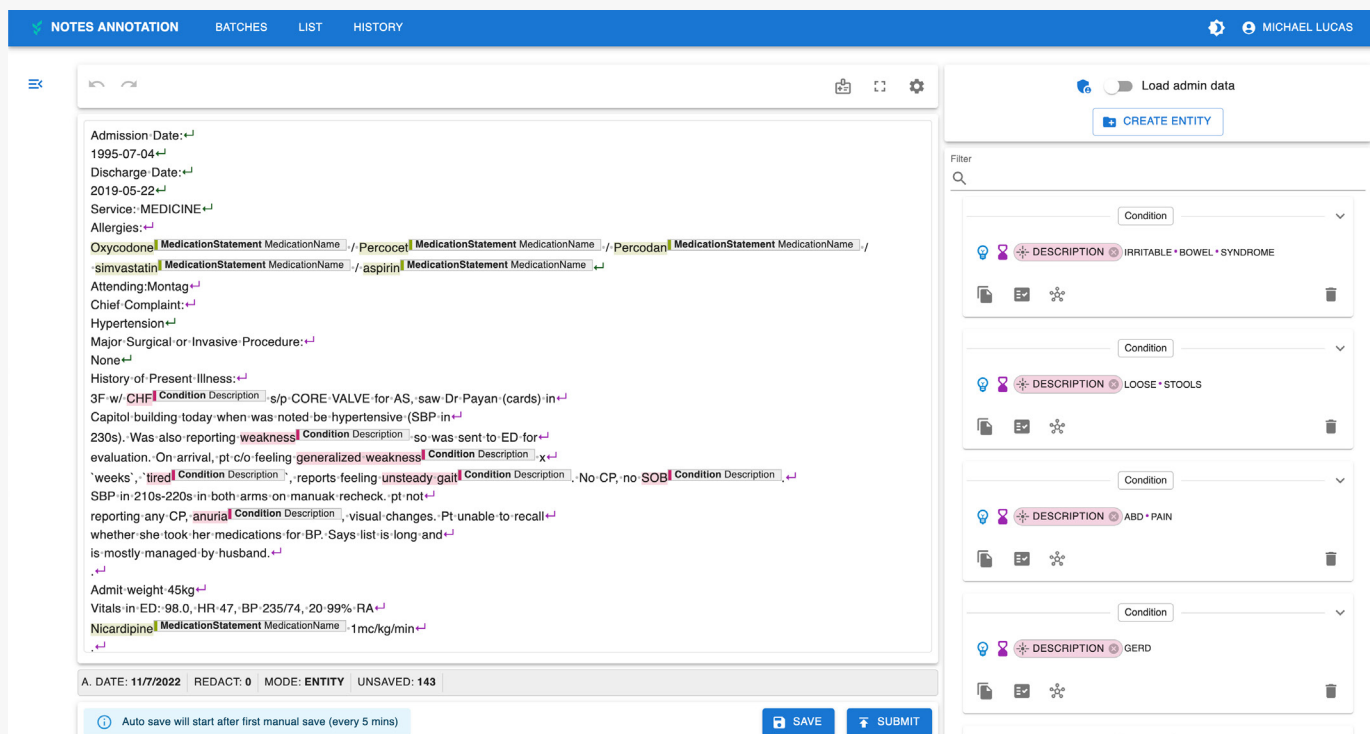
Fig 4. Custom tool Truveta annotators use to label unstructured medical record data for normalization.

TLM is designed for broad applicability, enabling extraction of clinical concepts relevant to both common and rare diseases or clinical scenarios. TLM can also be fine-tuned for accuracy within specific research domains, making it a versatile tool for a wide range of therapeutic areas, including cardiovascular, metabolic diseases, oncology, neurology, hepatology, pulmonology, rare diseases, and others. Researchers can access a continually expanding library of clinical concepts to enable more comprehensive research.

For example, Truveta has structured 69 distinct clinical measurements from 3.4 million echocardiogram reports, providing researchers with deep insights into heart structure and function. Key metrics like ejection fraction and specialized measures such as tricuspid annular systolic velocity (TASV) empower researchers to conduct more comprehensive cardiovascular studies.

Beyond cardiovascular disease, TLM delivers detailed insights into treatment patterns, disease progression, and patient outcomes across other therapeutic areas. In metabolic research, for instance, TLM enables analysis of GLP-1 therapy status and reasons for discontinuation. Oncology researchers benefit from access to ECOG Performance Status scores, cancer staging details, and pathology findings, while pulmonology studies are enriched by structured data from pulmonary function tests. In neurology, TLM captures data on migraine and seizure frequency, severity, symptoms, and treatment responses, enabling a deeper understanding of these conditions.

Examples of extracted clinical concepts

Cardiovascular
- Tricuspid valve regurgitation
- Echocardiograms: Quantitative results
- Cardiac cath reports: Hemodynamic measurements, CAD concepts
- NYHA and KCCQ scores
- QRS duration
- Vessel disease

Metabolics
- GLP-1 status and discontinuation reason

Oncology
- Hepatocellular carcinoma: ECOG Performance status scores, Child-Pugh scores, and Barcelona Clinic Liver Cancer stage
- Colon cancer: Staging, family history, pathology report findings

Neurology
- Seizure frequency
- Migraine frequency & severity
- Migraine treatment response
- Migraine symptoms & triggers
- Migraine treatment (triptans) status and discontinuation reason

Hepatology
- Fibrosis stage, steatosis, and hepatocyte ballooning

Pulmonology
- Pulmonary function test results

Rare disease
- HoFH: Confirmation of diagnosis
- OTC Deficiency: symptoms, disease progression, & dietary intake

Example cardiovascular concepts extracted from notes

Sampling of normalized echocardiogram data in Truveta Data

Clinical measure	Measure frequency	Patients with 3+ instances
Left ventricular ejection fraction (LVEF)	6.8M	861K
Aortic valve orifice area	1.3M	170K
Aortic valve mean pressure gradient	1.3M	124K
Aortic valve maximum velocity	1.1M	94K
Tricuspid annular lane systolic excursion (TAPSE)	769K	54K

Sampling of normalized cardiac catheterization data in Truveta Data

Clinical measure	Measure frequency	Patients with 3+ instances
Stenosis (coronary artery disease)	3M	281K
Right atrial pressure	122K	6.9K
Pulmonary capillary wedge pressure	85K	6.5K
Cardiac output	97K	8.5K
Cardiac index	82K	6.8K

Fig 5. Examples of clinically relevant concepts extracted from notes.

TLM also extracts data that transcend specific disease states, capturing concepts relevant across therapeutic areas and research needs. For example, nutrition data is a critical yet often overlooked component of health that is rarely captured in structured formats. Figure 6 illustrates the extraction of dietary information, which is essential for managing conditions like Ornithine Transcarbamylase (OTC) deficiency. The left-hand side of the figure highlights how details such as protein and caloric intake, specific foods consumed, and the route of protein intake can be extracted from free-text clinical notes, providing actionable insights for both niche and broad applications.
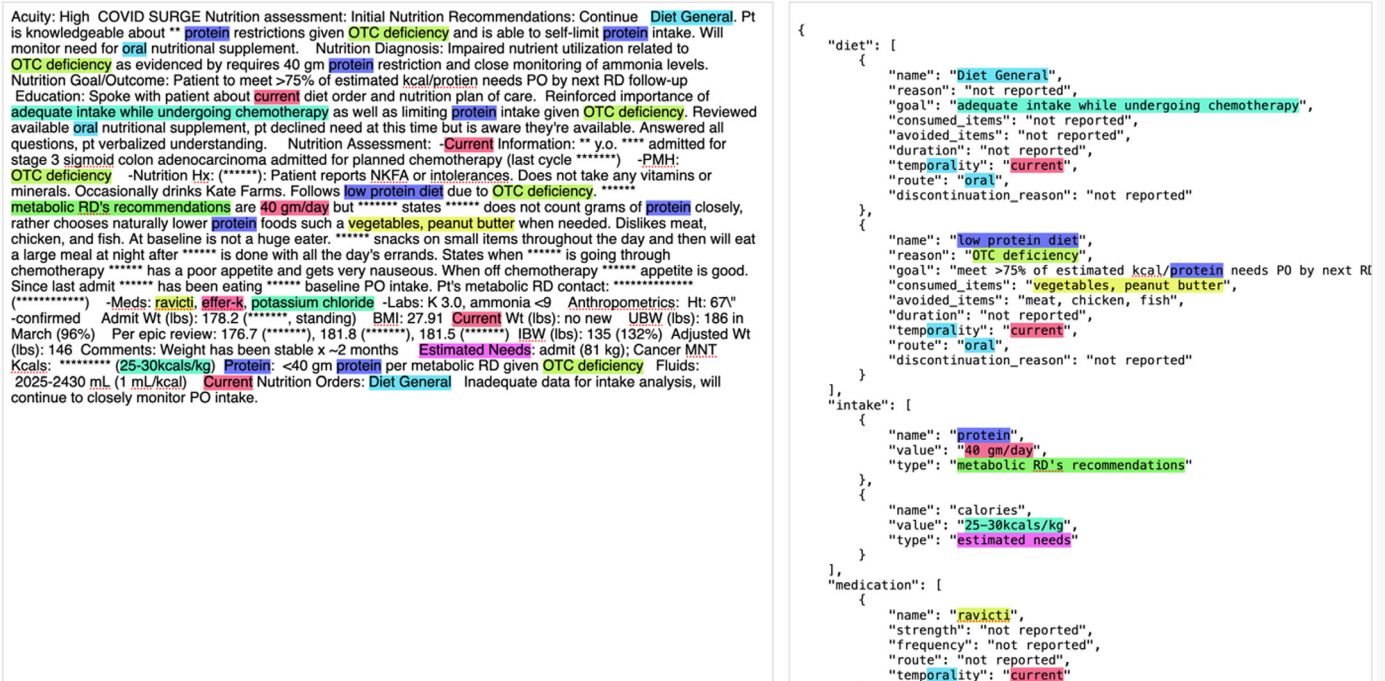
Fig 6. Visualization of dietary information extracted by TLM for a patient with OTC deficiency.

The power of TLM lies in its AI-driven approach, eliminating the need for human-based extraction. By enabling the consistent extraction of both common and nuanced data at scale, TLM supports the continuous expansion of our clinical concept library.

Ready access to concepts from notes can confirm the accuracy of data points gathered from semi-structured EHR data, elevating confidence in clinical truth. It also enables researchers to pair insights extracted from notes with other critical data points to study the complete patient journey and explore treatment effectiveness, disease risk factors, patient subgroups, and more.

Ensuring clinical accuracy through continuous quality assessment

The goal of TLM is to exceed the accuracy of clinical experts reviewing medical records. When the model achieves greater accuracy than clinical experts in a particular healthcare domain (e.g., clinical observations, lab results), the model is deployed into Truveta Embassies. These embassies are secure, cloud-based environments where health systems send their raw medical records data for normalization.
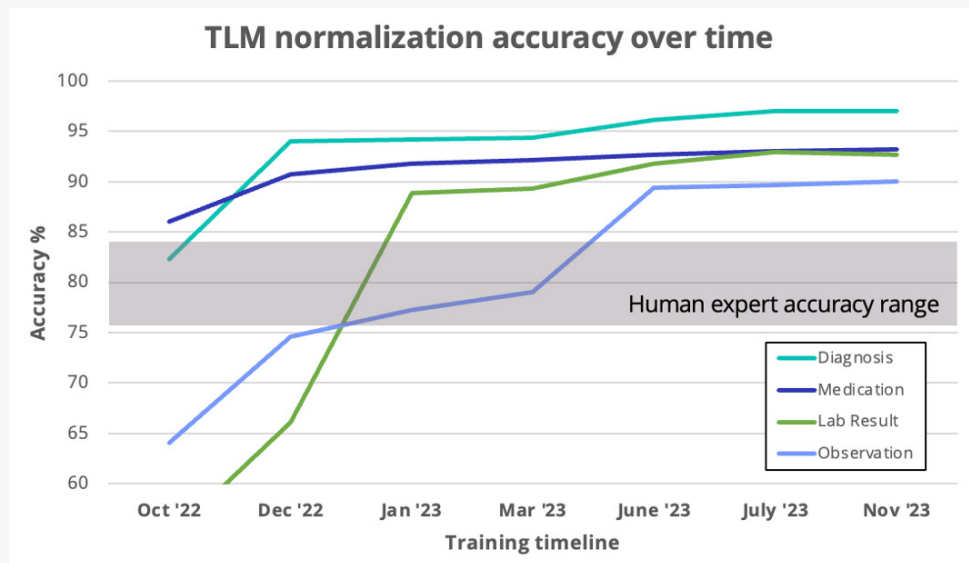
Fig 7. TLM normalization capabilities as compared to human experts.

TLM is currently achieving high accuracy on diagnoses, medications, lab results, lab values, clinical observations, and more. TLM's accuracy improves over time with ongoing training but already today outperforms state-of-the-art approaches, including GPT-4, LogMap, AML, BERTMap, and the latest ontology matching frameworks from the Ontology Alignment Evaluation Initiative. You can read more about the underlying AI here.

If TLM performance on any concept falls below the human expert range, TLM stops processing that concept and our AI team commences additional annotation and/ or model training to improve performance. This iterative process ensures that the models become more robust and capable of handling a larger scale of incoming data without requiring a proportional increase in human intervention.

Fig 8. Depiction of the iterative model training process.

The quality review process for notes involves evaluating both extraction and normalization to ensure that previously unstructured text strings map accurately into the TDM. As the AI model for these tasks undergoes training, the team evaluates its performance on precision and recall against an evaluation set produced by trained clinical experts. Truveta then performs additional scenario-based q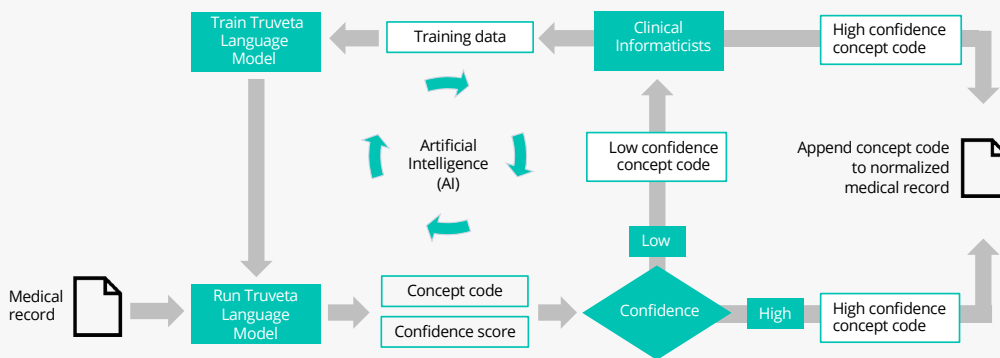uality validations to ensure accurate normalization and effectiveness in fulfilling the target research use case(s). Once a model meets or exceeds the performance of a trained human annotator and is deployed to production, the AI team continually assesses quality to ensure optimal performance across variations in clinical documentation, especially from newly onboarded health system members.

Our data quality goal is to provide the transparency and rigor required to be trusted by regulators. Thus, each clinical concept extracted from notes is accompanied by documentation on the concept definition, modeling methods, and the accuracy of TLM's extraction. This documentation may be embedded in a methods section in a manuscript or submitted to regulators. An example of this regulatory-grade documentation can be found in the appendix.

Further, Truveta has invested in a robust Quality Management System (QMS), which includes policies, procedures, and software controls to uphold the most stringent data quality standards. For studies marked as intended for regulatory submission, Truveta generates additional logging detail affirming successful system performance during the study. See our data quality whitepaper for more details.

Empowering researchers with trustworthy and complete data

TLM is a profound innovation for making healthcare data trustworthy and complete for scientifically rigorous research. Truveta is empowering life science, academic research, government, and healthcare organizations to achieve our shared mission of Saving Lives with Data.

We look forward to the development of industry models that seamlessly integrate with foundational large language models, unlocking the full potential of AI to improve human health – and operationalizing them at massive scale.

To learn more about Truveta, please visit the Truveta website, follow us on LinkedIn, or contact us at info@Truveta.com.

Appendix: Regulatory evidence report

Ejection Fraction Quality Evidence Report
Created: 2024-03-08

Background

Data was extracted from echocardiogram reports. This data was then transformed to Truveta Data Model. The following parameters were in place during extraction of the data:

1. Ejection fraction quantitative measurements obtained during an echocardiogram and recorded in that echocardiogram report were annotated as observations, along with the associated values. Any dates or time periods associated with those observations were captured as well.

2. If available, the datetime that the study (echocardiogram) was performed was annotated in each note and transformed to the EffectiveDateTime for all observation records of measurements reported in that note. The datetime the note was signed was annotated in each note (if available) and transformed to the RecordedDateTime for all observation records of measurements reported in that note. If different datetimes were available in the note that better represented the EffectiveDateTime or RecordedDateTime of any measurement(s) than the study datetime or signed datetime of the note, those datetimes were captured and transformed to the EffectiveDateTime or RecordedDateTime fields, as appropriate.

3. All observation records of measurements extracted from notes were assigned a SourceProvenanceConceptId of 3056721 ("note").

4. The names of the measurements that were extracted from notes were normalized to LOINC concepts.

5. The values and units of measurements were normalized to the appropriate value fields in the observation table.

Assumptions

To best capture the annotated data from notes in Truveta Data Model, the following assumptions were made during normalization:

1. Ejection fraction terms without a specified laterality in the note were assumed to be left ventricular ejection fraction.

2. Many LOINC codes specify an ultrasound method (2D vs M-mode, for example) or calculation method. If the method by which a measurement was obtained was not specified in the echo report, the terms were mapped to a LOINC code if the normalization team could verify that the method (such as 2D) is the way a specific measurement must be obtained and thus the only option. If this could not be confirmed through published echocardiography guidelines or other reputable sources, the measurement names were mapped to null flavor codes or more generalized, non-measurement descriptor codes, typically SNOMED.

Examples

Note text	TDM representation (observation element)
Visit date: 3/1/2023 Ejection fraction is 60%. Prior ejection fraction on 2/1/2023 was 30-35%. Signed by MD on 3/3/2023.	**CodeConceptId**: 789242 (left ventricular ejection fraction by US) ValueNumeric: 60 **ValueUOM**: 1189997 (%) **ObservationCategory**: 1065645 (imaging) EffectiveDateTime: 3/1/2023 RecordedDateTime: 3/3/2023 SourceProvenanceConceptId: 3056721 (notes)
	CodeConceptId: 789242 (left ventricular ejection fraction by US) ValueRangeLow: 30 ValueRangeHigh: 35 **ValueUOM**: 1189997 (%) **ObservationCategory**: 1065645 (imaging) EffectiveDateTime: 2/1/2023 RecordedDateTime: 3/3/2023 SourceProvenanceConceptId: 3056721 (notes)

Machine learning (ML) model training

Measurements were extracted from echocardiogram reports using machine learning models. The notes were processed using two machine learning models to create structured data. The first is called the clinical extraction model. This model employs natural language processing (NLP) techniques to extract clinical observations. For example: an observation name of *left ventricular ejection fraction* with a corresponding value of *65%*. The clinical measurement names and values are detected by a technique called named entity recognition (NER), and their associations are detected by a technique called relation extraction (RE). The model jointly learns both NER and RE.

After term extraction is complete, a second model called Automapper at Truveta predicts the mappings of the observation names to standard concepts, primarily LOINC codes as described above. Each prediction has a confidence of high, medium, or low confidence. High confidence mappings mean the model is 99% confident in the prediction. Only high confidence mappings from Automapper were used. Human experts then mapped the most prevalent medium and low confidence terms to ensure mapping completeness and to continue to train the model for future updates.

Both models were trained on data extracted from 500 echocardiogram reports by human annotation. At least 50% of these notes were double annotated (annotated by two human experts) to ensure data quality. Terms from these same notes were normalized to standard ontologies (LOINC and SNOMED) by a separate team of human experts.

Clinical concept extraction pipeline

Concepts are extracted from clinical notes in two major steps: extraction and mapping. The data from these subsequent processes is ingested into the Truveta Data Model (TDM) as structured, normalized data. Evaluation of the extraction pipeline is done by comparing the output data in TDM to the ground truth as annotated by a human expert. In this way, the entire system of the clinical extraction pipeline is evaluated wholistically, rather than in a piece meal fashion.

Quality Assurance evidence report for Ejection Fraction

This section reports the performance of the Notes Extraction Pipeline for this scenario. We establish the validation criteria for this scenario according to the definitions and validation methodologies described below. Multiple models, as well as a rules-based engine and our data processing pipeline, impact the final data made available to the customer. Therefore, the quality assurance (QA) evidence report seeks to validate the entire system of clinical notes extraction end to end, rather than individual model(s).

Patient cohort selection

For the purposes of the sample validation report, all patients with an echocardiogram report were included in the patient cohort and 100 patients were randomly selected for validation. For the sample report, data from only one contributing member health system was considered. The full validation report will include a stratified sample of data from all contributing health systems.

Note selection

Note type: echocardiogram or echo
Note status: signed or addendum

From the patient cohort above, we selected a random sample of notes based on the following protocol:

Using a randomizing algorithm, 100 random patients were selected from the above cohort. If a patient has more than 1 qualifying note, an additional randomization was applied to select 1 note per patient. This resulted in an evaluation set of 100 randomly selected notes for 100 randomly selected patients from the cohort of interest.

Creation of evaluation set

1. Notes were annotated by expert clinical terminologists to establish a ground truth.

2. Terms were mapped to the standard ontology by expert clinical terminologists.

3. Normalization of values by rules-based engine was reviewed by expert clinical terminologists for accuracy.

The resulting data serves as the ground truth for the scenario. The machine learning development process did not train on this QA test set.

Evaluation methodology

- Record level data from the evaluation set was tabulated as shown below in the sample. Specifically, column (A) contains the concept to be extracted, column (B) contains the expected value, and column (C) contains the expected standard unit of measure (UOM).

- The clinical concept extraction pipeline was used to extract normalized concepts from the evaluation set of notes (columns D-F).

- The concepts in columns (D-F) were evaluated into "Equivalent", "Wide", "Narrow", "Missing", and "Wrong" categories by multiple experts and their combined judgment was entered in column (G).

    ◦ Equivalent: All columns are populated correctly

    ◦ Wider: All columns are populated but the concept of interest is mapped to a less specific (but still correct) LOINC or SNOMED code than in the evaluation set.

    ◦ Narrow: The concept of interest is mapped to a more specific (and therefore incorrect) LOINC or SNOMED code than in the evaluation set, such as one implying a method that is not specified in the note.

    ◦ Missing: The clinical concept extraction pipeline did not return a record, or the record is missing a field critical to the utility of the data (in this case, the normalized value).

    ◦ Wrong: The clinical concept extraction pipeline extracted text outside of the concepts of interest, or extracted the concept an incomplete fashion so that it could not be normalized to a valid standard concept.

- The mappings were then evaluated as true positive (TP), false positive (FP), and false negative (FN) by the experts in a scenario dependent manner and populated in column (H). Specifically

    ◦ Wider, equivalent: TP

    ◦ Missing: FN

    ◦ Narrow, Wrong: FP

- Finally, the record level evaluations were rolled up to a note level evaluation in the following manner:

    ◦ False positives are evaluated. If the false positive in some way changes the interpretation of the note (for example, indicates a conflicting EF value), the false positive is included in the overall evaluation score. In this case, no false positives were found which impact the overall interpretation of the data.

    ◦ False negatives are evaluated. If the false negative changes the interpretation of the note, the false negative is included in the overall evaluation score. In the majority of cases, the notes have 2 or more mentions of the same ejection fraction result and the model captured at least 1, so the false negative did not impact overall quality.

    ◦ Remaining notes are counted as true positives.

Metrics definition

$$Precision = \frac{Total\ TP}{Total\ TP + Total\ FP}$$

$$Recall = \frac{Total\ TP}{Total\ TP + Total\ FN}$$

$$Accuracy = \frac{Total\ TP}{Total\ TP + Total\ FN + Total\ FP}$$

Note: When calculating accuracy, we do not consider true negatives (TN) since most of the spans are not labeled and would lead to an inflated accuracy close to 100%.

Results summary: Note level extraction quality

Clinical scenario	Evaluation set statistics		QA metrics		
	#Patients	#Notes	Precision	Recall	Accuracy
Ejection Fraction	100	100	1.00	0.91	0.91