TRUVETA

Whitepaper

Truveta Language Model

June 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 100 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.

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 100 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.

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.

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.

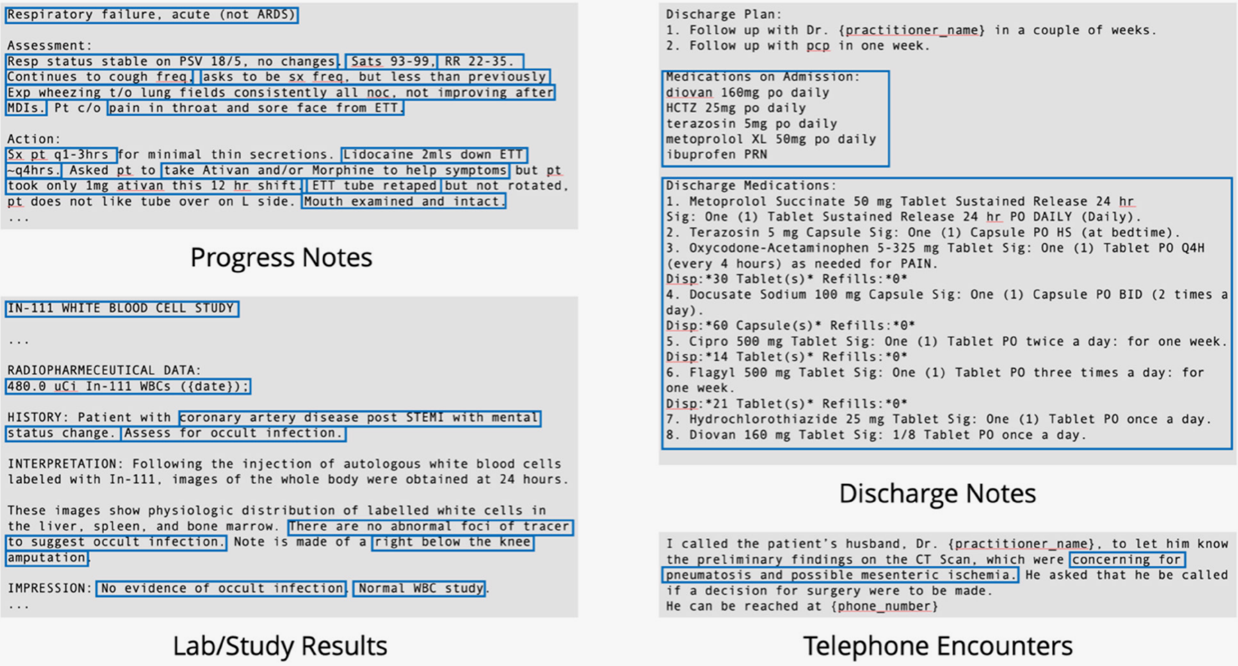
Raw medical record text	Lab Results data after TLM normalization		
	Lab Name (LOINC)	Unit (UCUM)	Value
RBC COUNT, RBC|CBC WITH AUTOMATED DIFF |3.80| M/uL|2.70 |4.90	789-8	10*6/uL	3.80
CBC: 3/9 07:45PM WBC-8.1 RBC-3.89 Hgb-11.7	6690-2	10*3/uL	8.1
	789-8	10*6/uL	3.89
	718-7	g/dL	11.7

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 100 million patients.



```
Respiratory failure, acute (not ARDS)

Assessment:
Resp status stable on PSV 18/5, no changes. Sats 93-99, RR 22-35.
Continues to cough freq, asks to be sx freq, but less than previously
Exp wheezing t/o lung fields consistently all noc, not improving after
MDIs. Pt c/o pain in throat and sore face from ETT.

Action:
Sx pt q1-3hrs for minimal thin secretions. Lidocaine 2mls down ETT
~q4hrs. Asked pt to take Ativan and/or Morphine to help symptoms but pt
took only 1mg ativan this 12 hr shift. ETT tube retaped but not rotated,
pt does not like tube over on L side. Mouth examined and intact.
...
```

Progress Notes

```
IN-111 WHITE BLOOD CELL STUDY

...

RADIOPHARMECEUTICAL DATA:
480.0 uCi In-111 WBCs ({date});

HISTORY: Patient with coronary artery disease post STEMI with mental
status change. Assess for occult infection.

INTERPRETATION: Following the injection of autologous white blood cells
labeled with In-111, images of the whole body were obtained at 24 hours.

These images show physiologic distribution of labelled white cells in
the liver, spleen, and bone marrow. There are no abnormal foci of tracer
to suggest occult infection. Note is made of a right below the knee
amputation.

IMPRESSION: No evidence of occult infection. Normal WBC study.
...
```

Lab/Study Results

```
Discharge Plan:
1. Follow up with Dr. {practitioner_name} in a couple of weeks.
2. Follow up with pcp in one week.

Medications on Admission:
diovan 160mg po daily
HCTZ 25mg po daily
terazosin 5mg po daily
metoprolol XL 50mg po daily
ibuprofen PRN

Discharge Medications:
1. Metoprolol Succinate 50 mg Tablet Sustained Release 24 hr
Sig: One (1) Tablet Sustained Release 24 hr PO DAILY (Daily).
2. Terazosin 5 mg Capsule Sig: One (1) Capsule PO HS (at bedtime).
3. Oxycodone-Acetaminophen 5-325 mg Tablet Sig: One (1) Tablet PO Q4H
(every 4 hours) as needed for PAIN.
Disp:*30 Tablet(s)* Refills:*0*
4. Docusate Sodium 100 mg Capsule Sig: One (1) Capsule PO BID (2 times a
day).
Disp:*60 Capsule(s)* Refills:*0*
5. Cipro 500 mg Tablet Sig: One (1) Tablet PO twice a day: for one week.
Disp:*14 Tablet(s)* Refills:*0*
6. Flagyl 500 mg Tablet Sig: One (1) Tablet PO three times a day: for
one week.
Disp:*21 Tablet(s)* Refills:*0*
7. Hydrochlorothiazide 25 mg Tablet Sig: One (1) Tablet PO once a day.
8. Diovan 160 mg Tablet Sig: 1/8 Tablet PO once a day.
```

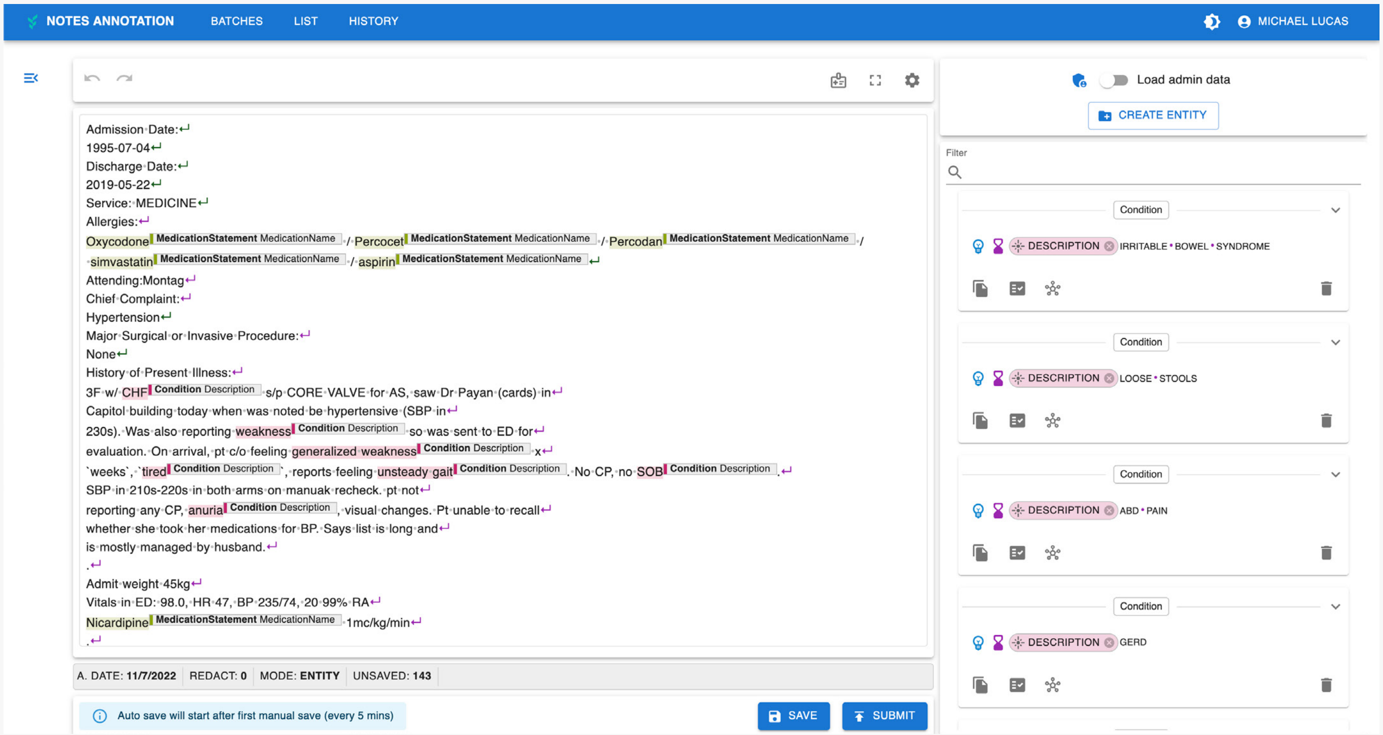
Discharge Notes

```
I called the patient's husband, Dr. {practitioner_name}, to let him know
the preliminary findings on the CT Scan, which were concerning for
pneumatosis and possible mesenteric ischemia. He asked that he be called
if a decision for surgery were to be made.
He can be reached at {phone_number}
```

Telephone Encounters

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").



Admission Date:
1995-07-04
Discharge Date:
2019-05-22
Service: MEDICINE
Allergies:
Oxycodone MedicationStatement MedicationName / Percocet MedicationStatement MedicationName / Percodan MedicationStatement MedicationName /
simvastatin MedicationStatement MedicationName / aspirin MedicationStatement MedicationName
Attending:Montag
Chief Complaint:
Hypertension
Major Surgical or Invasive Procedure:
None
History of Present Illness:
3F w/ CHF Condition Description s/p CORE VALVE for AS, saw Dr Payan (cards) in
Capitol building today when was noted be hypertensive (SBP in
230s). Was also reporting weakness Condition Description so was sent to ED for
evaluation. On arrival, pt c/o feeling generalized weakness Condition Description x
`weeks`, `tired Condition Description `, reports feeling unsteady gait Condition Description . No CP, no SOB Condition Description .
SBP in 210s-220s in both arms on manuak recheck. pt not
reporting any CP, anuria Condition Description , visual changes. Pt unable to recall
whether she took her medications for BP. Says list is long and
is mostly managed by husband.
.
Admit weight 45kg
Vitals in ED: 98.0, HR 47, BP 235/74, 20 99% RA
Nicardipine MedicationStatement MedicationName 1mc/kg/min
.

A. DATE: **11/7/2022** | REDACT: **0** | MODE: **ENTITY** | UNSAVED: **143**

Auto save will start after first manual save (every 5 mins)

SAVE SUBMIT

Condition
DESCRIPTION IRRITABLE • BOWEL • SYNDROME

Condition
DESCRIPTION LOOSE • STOOLS

Condition
DESCRIPTION ABD • PAIN

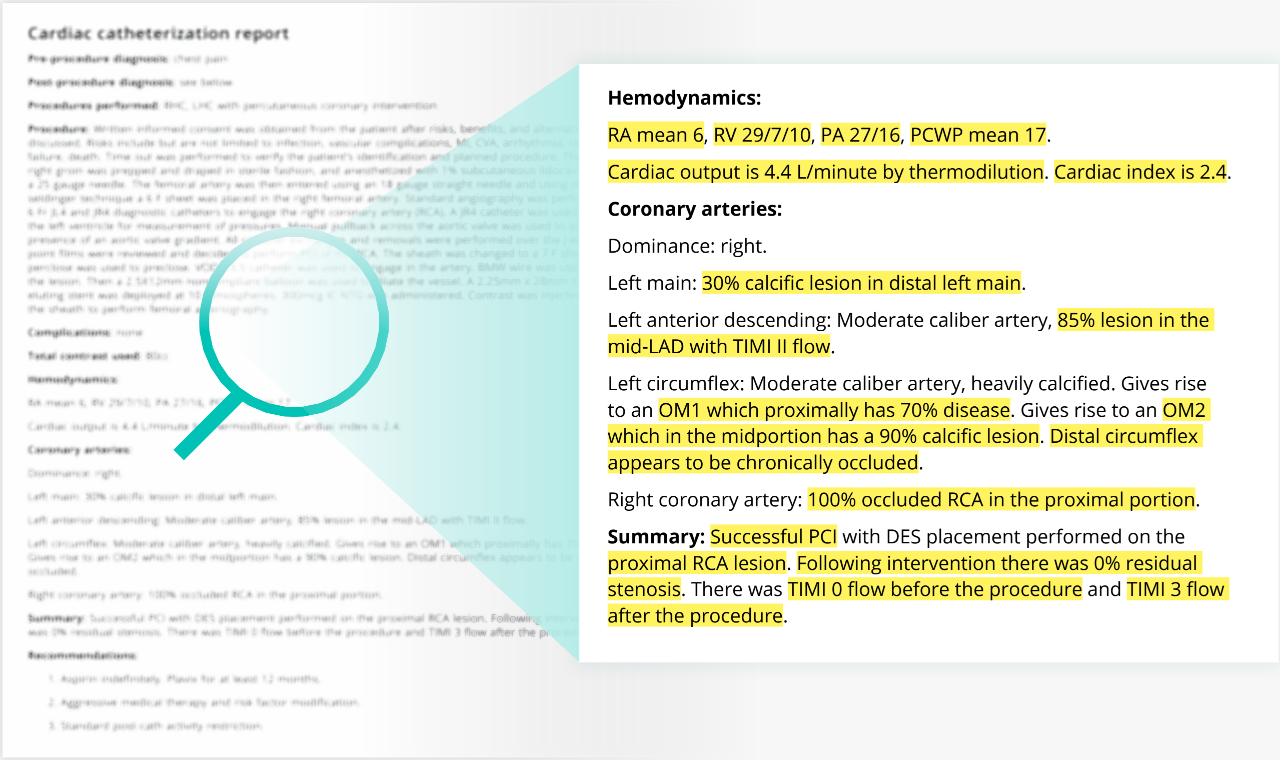
Condition
DESCRIPTION GERD

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 further fine-tuned for accuracy within specific domains of study. Customers have benefitted from the breadth and depth of Truveta's clinical notes for many conditions, including colon cancer, migraines, seizures, NASH, heart failure, vessel disease, hypercholesterolemia, rare diseases, and others.

For example, Truveta has structured 69 distinct clinical measurements from 3.4 million echocardiogram reports, empowering researchers to conduct innovative cardiovascular studies with deep understanding of heart structure and function for patients of interest. From essential metrics like ejection fraction to specialized measures such as tricuspid annular systolic velocity (TASV), the availability of these measures at scale enables researchers to conduct more comprehensive cardiovascular studies.

Similarly, researchers can access insights from more than 700,000 cardiac catheterization reports from more than 440,000 patients, resulting in 4.7 million distinct catheterization measurements available for study. This includes measurements from right heart catheterizations such as hemodynamic measurements, as well as measurements from left heart catheterizations such as amount and location of vessel disease, lesion complexity, thrombolysis in myocardial infarction (TIMI) grades, and whether an intervention was performed.



Cardiac catheterization report

Hemodynamics:

RA mean 6, RV 29/7/10, PA 27/16, PCWP mean 17.

Cardiac output is 4.4 L/minute by thermodilution. Cardiac index is 2.4.

Coronary arteries:

Dominance: right.

Left main: 30% calcific lesion in distal left main.

Left anterior descending: Moderate caliber artery, 85% lesion in the mid-LAD with TIMI II flow.

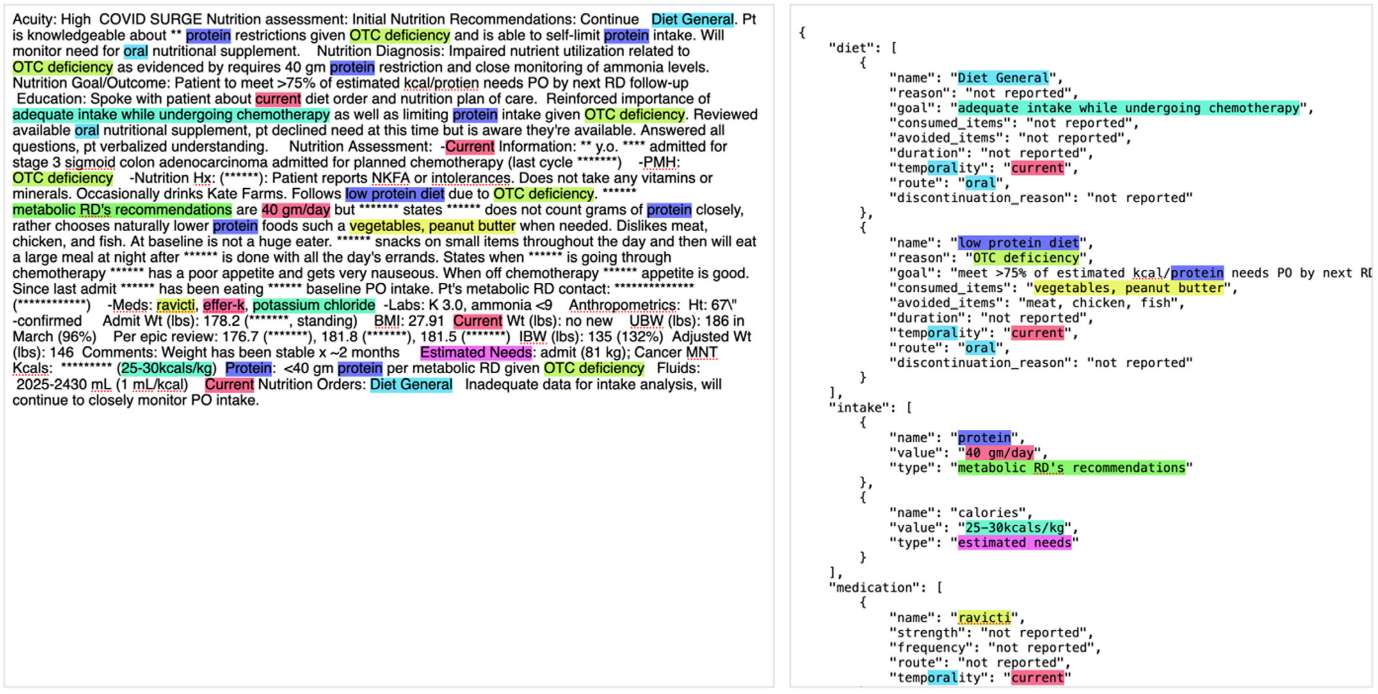
Left circumflex: Moderate caliber artery, heavily calcified. Gives rise to an OM1 which proximally has 70% disease. Gives rise to an OM2 which in the midportion has a 90% calcific lesion. Distal circumflex appears to be chronically occluded.

Right coronary artery: 100% occluded RCA in the proximal portion.

Summary: Successful PCI with DES placement performed on the proximal RCA lesion. Following intervention there was 0% residual stenosis. There was TIMI 0 flow before the procedure and TIMI 3 flow after the procedure.

Fig 5. Example data extracted from a cardiac catheterization report.

TLM is also designed to extract data relevant to more niche research areas. Figure 6 shows extraction of dietary information, which is critical for the treatment and management of a rare genetic disorder called Ornithine Transcarbamylase (OTC) deficiency, but rarely captured in structured data. The left-hand side of the figure shows how information such as protein and caloric intake, specific foods consumed, and route of protein intake can be extracted from a free-text clinical note.



Acuity: High COVID SURGE Nutrition assessment: Initial Nutrition Recommendations: Continue Diet General. Pt is knowledgeable about ** protein restrictions given OTC deficiency and is able to self-limit protein intake. Will monitor need for oral nutritional supplement. Nutrition Diagnosis: Impaired nutrient utilization related to OTC deficiency as evidenced by requires 40 gm protein restriction and close monitoring of ammonia levels. Nutrition Goal/Outcome: Patient to meet >75% of estimated kcal/protien needs PO by next RD follow-up Education: Spoke with patient about current diet order and nutrition plan of care. Reinforced importance of adequate intake while undergoing chemotherapy as well as limiting protein intake given OTC deficiency. Reviewed available oral nutritional supplement, pt declined need at this time but is aware they're available. Answered all questions, pt verbalized understanding. Nutrition Assessment: -Current Information: ** y.o. **** admitted for stage 3 sigmoid colon adenocarcinoma admitted for planned chemotherapy (last cycle *******) -PMH: OTC deficiency -Nutrition Hx: (******): Patient reports NKFA or intolerances. Does not take any vitamins or minerals. Occasionally drinks Kate Farms. Follows low protein diet due to OTC deficiency. ****** metabolic RD's recommendations are 40 gm/day but ******* states ****** does not count grams of protein closely, rather chooses naturally lower protein foods such a vegetables, peanut butter when needed. Dislikes meat, chicken, and fish. At baseline is not a huge eater. ****** snacks on small items throughout the day and then will eat a large meal at night after ****** is done with all the day's errands. States when ****** is going through chemotherapy ****** has a poor appetite and gets very nauseous. When off chemotherapy ****** appetite is good. Since last admit ****** has been eating ****** baseline PO intake. Pt's metabolic RD contact: ************** (************) -Meds: ravicti, effer-k, potassium chloride -Labs: K 3.0, ammonia <9 Anthropometrics: Ht: 67\" -confirmed Admit Wt (lbs): 178.2 (*******, standing) BMI: 27.91 Current Wt (lbs): no new UBW (lbs): 186 in March (96%) Per epic review: 176.7 (*******), 181.8 (*******), 181.5 (*******) IBW (lbs): 135 (132%) Adjusted Wt (lbs): 146 Comments: Weight has been stable x ~2 months Estimated Needs: admit (81 kg); Cancer MNT Kcals: ********* (25-30kcals/kg) Protein: <40 gm protein per metabolic RD given OTC deficiency Fluids: 2025-2430 mL (1 mL/kcal) Current Nutrition Orders: Diet General Inadequate data for intake analysis, will continue to closely monitor PO intake.

```
{
    "diet": [
        {
            "name": "Diet General",
            "reason": "not reported",
            "goal": "adequate intake while undergoing chemotherapy",
            "consumed_items": "not reported",
            "avoided_items": "not reported",
            "duration": "not reported",
            "temporality": "current",
            "route": "oral",
            "discontinuation_reason": "not reported"
        },
        {
            "name": "low protein diet",
            "reason": "OTC deficiency",
            "goal": "meet >75% of estimated kcal/protein needs PO by next RD
            "consumed_items": "vegetables, peanut butter",
            "avoided_items": "meat, chicken, fish",
            "duration": "not reported",
            "temporality": "current",
            "route": "oral",
            "discontinuation_reason": "not reported"
        }
    ],
    "intake": [
        {
            "name": "protein",
            "value": "40 gm/day",
            "type": "metabolic RD's recommendations"
        },
        {
            "name": "calories",
            "value": "25-30kcals/kg",
            "type": "estimated needs"
        }
    ],
    "medication": [
        {
            "name": "ravicti",
            "strength": "not reported",
            "frequency": "not reported",
            "route": "not reported",
            "temporality": "current"
```

Fig 6. Visualization of dietary information extracted by TLM for a patient with OTC deficiency.

After extraction, these raw strings will be cleaned to target ontology codes. Each record will then be transformed to an observation record in TDM with a source provenance indicating that the information was extracted from clinical notes. This type of extraction is relevant to any disease.

The power of TLM lies in its AI-driven approach, which eliminates the need for human-based extraction. TLM enables the consistent extraction of both common and nuanced data at scale and with equal determination. Ready access to data 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.

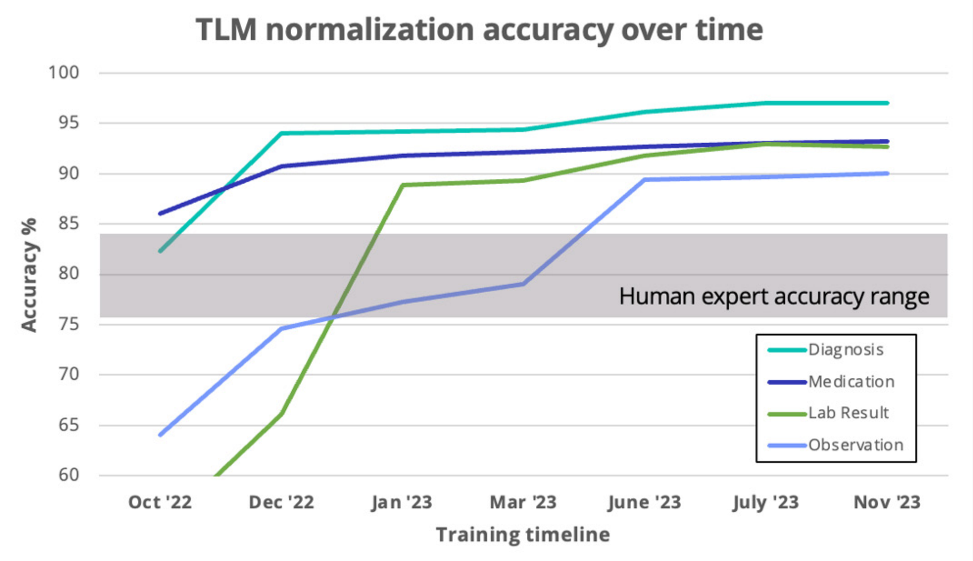


Fig 7. TLM normalization capabilities as compared to human experts.

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.

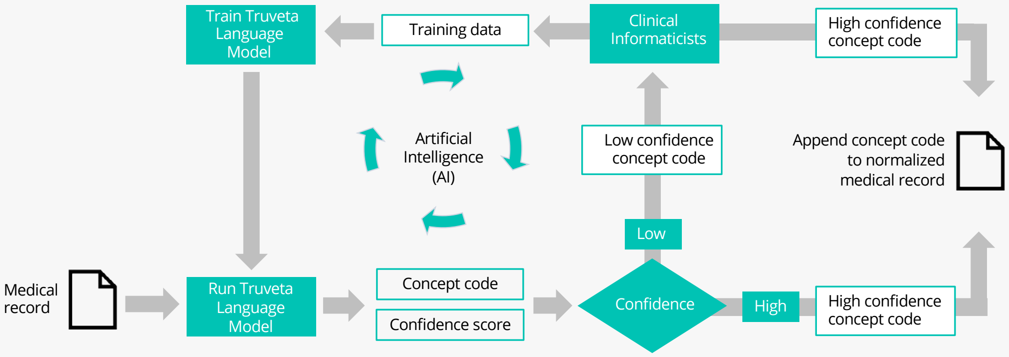

Train Truveta Language Model
Training data
Clinical Informaticists
High confidence concept code
Artificial Intelligence (AI)
Low confidence concept code
Append concept code to normalized medical record
Low
Medical record
Run Truveta Language Model
Concept code
Confidence score
Confidence
High
High confidence concept code

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 quality 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

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

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

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

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		
Ejection Fraction	#Patients	#Notes	Precision	Recall	Accuracy
	100	100	1.00	0.91	0.91