TRUVETA

Whitepaper

Truveta Language Model

January 2024

Contents

Introduction

Healthcare organizations generate an immense volume of data, with the average hospital producing roughly 50 petabytes of data a year. However, an estimated 95 percent of this data goes unused – largely because it is fragmented, inaccessible, and unstructured. Unfortunately, the majority of valuable clinical information is contained within unstructured data.

Having access to complete, timely, clean, research-ready data from Electronic Health Records (EHRs) – including concepts from free-text clinical notes – unlocks a tremendous range of opportunities to advance research, innovation, and patient care.

Delivering clean data at scale across disease areas has historically been infeasible due to the time, cost, and scope of expertise required. Advancements in AI have presented a unique opportunity to transform and clean massive streams of healthcare data.

Truveta is a growing collective of more than 30 health systems who provide over 18% of the daily clinical care across the United States. Member health systems provide complete medical records for more than 100 million patients, which are then linked across health systems and augmented with claims, social drivers of health (SDOH), and mortality data to provide a complete, longitudinal view of patient journeys.

Every day, the Truveta Language Model (TLM) cleans these billions of data points to prepare them for research. TLM's healthcare expertise is trained on the largest collection of complete medical records representing the full diversity of the United States. It is the first large-language model specifically designed to empower researchers to study patient care and outcomes. This whitepaper explains TLM and how it works.

For information about our data quality process, see this whitepaper.

The Truveta Language Model (TLM) cleans billions of clinical data points to prepare them for research. TLM's healthcare expertise is trained on the largest collection of complete medical records representing the full diversity of the United States.

Truveta Language Model

As healthcare considers the potential of AI and real-world data, the opportunities and potential consequences are real. General large language models understand language but are inaccurate within the medical domain due to being trained on the public Internet, which contains no real medical records. In contrast, TLM fine tunes open large language models with additional training on Truveta Data to achieve above 90% precision and recall across clinical domains.

TLM fine tunes open large language models with additional training on Truveta Data to achieve above 90% precision and recall across clinical domains.

TLM can normalize 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. Having access to both semi-structured and unstructured data is essential for powering critical research, given that notes contain an estimated 60-80% of clinical data relevant to research questions. Specifically, notes contain information about family history, disease staging, adverse events, symptoms, reasons for a medication change, interpretations of findings, recommendations for follow-up, and other clinical context. These pieces of information may offer researchers access to critical measures of interest or help contextualize other data points.

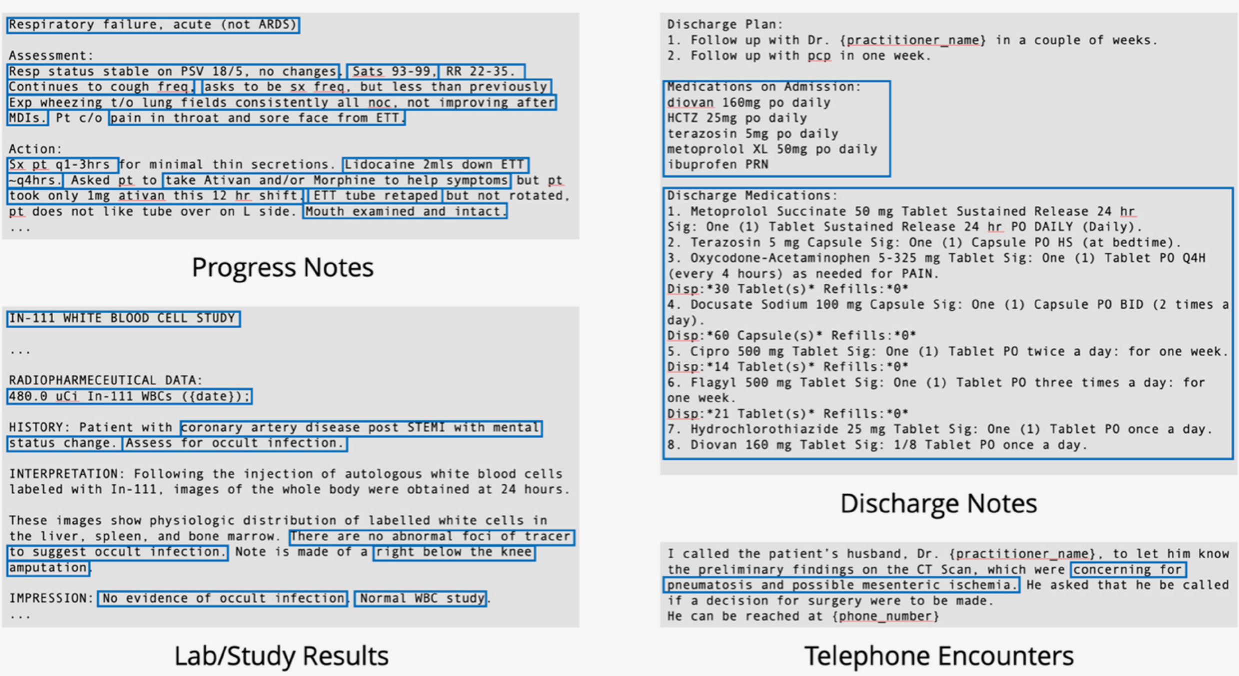

Fig 1. TLM extracts critical data points not available within claims data, such as disease staging, adverse events, and medication rationale changes from clinical notes.

The normalization 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, and medication plans, for instance. "Acute COVID-19," "COVID," "COVID-19," "COVID infection," and "COVID19 _ acute infection" all refer to the same disease process, and "600mg Ibuprofen" and "Ibuprofen 600mg tablets by mouth" are the same medical products. Before TLM, this unstructured data presented a very expensive data cleaning challenge for analytics.

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 2. How 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 LabResults 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 3. Example of TLM mapping lab results to the appropriate standard medical ontology.

Training Truveta Language Model on clinical concepts

TLM is trained on data from Truveta's health system members, currently representing more than 100 million patient journeys, including 8.4 billion diagnoses, 4.1 billion encounters, and 4 billion medication orders.

Using this data, Truveta's clinical expert annotation team labels thousands of raw clinical terms, including misspellings and abbreviations, to train and evaluate TLM with a focus on clinical accuracy. This annotation process is complex and nuanced. Sometimes even experts disagree on the best normalization approach, which is why all terms are assessed by multiple experts. In the event of disagreement, those experts discuss and reach alignment.

Clinical experts label concepts, build consensus, and review low confidence TLM results using a custom tool designed to continuously improve clinical accuracy of Truveta Data over time.

After running TLM, each concept receives a statistical "confidence score". Low confidences results are reviewed to create additional training data for the model.

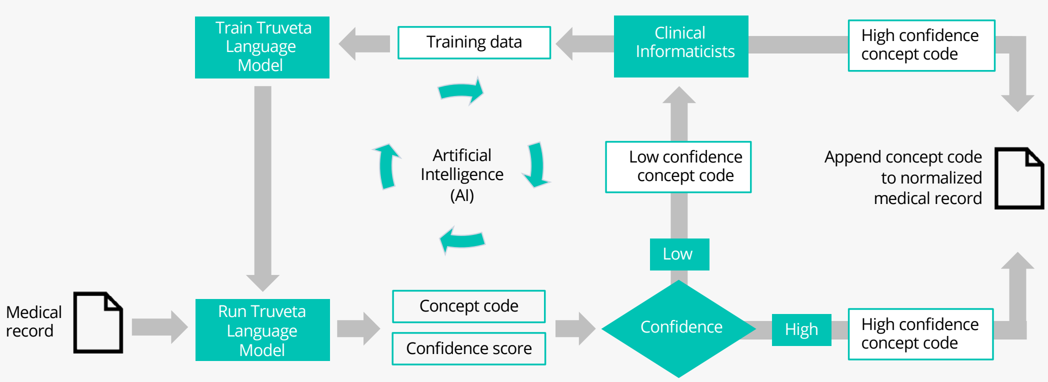

Fig 4. Depiction of the iterative model training process.

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 to start normalizing data.

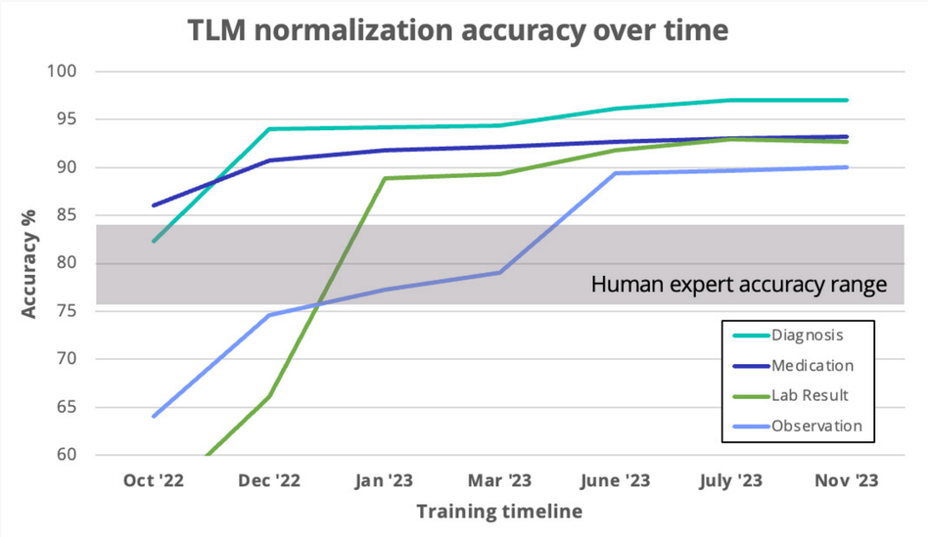

Fig 5. TLM normalization capabilities as compared to human experts.

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.

Training Truveta Language Model on clinical notes

TLM not only identifies and normalizes clinical concepts, but also extracts those concepts from clinical notes. This extraction accounts for nuances such as negation (e.g., “patient denies feeling fatigued”), hypotheticals/ conditionals (e.g., “Will consider starting low-dose glypizide if A1C still grossly elevated”), and family history (e.g., “Family Hx: Mother: Diabetes, Father/son: bipolar disorder”).

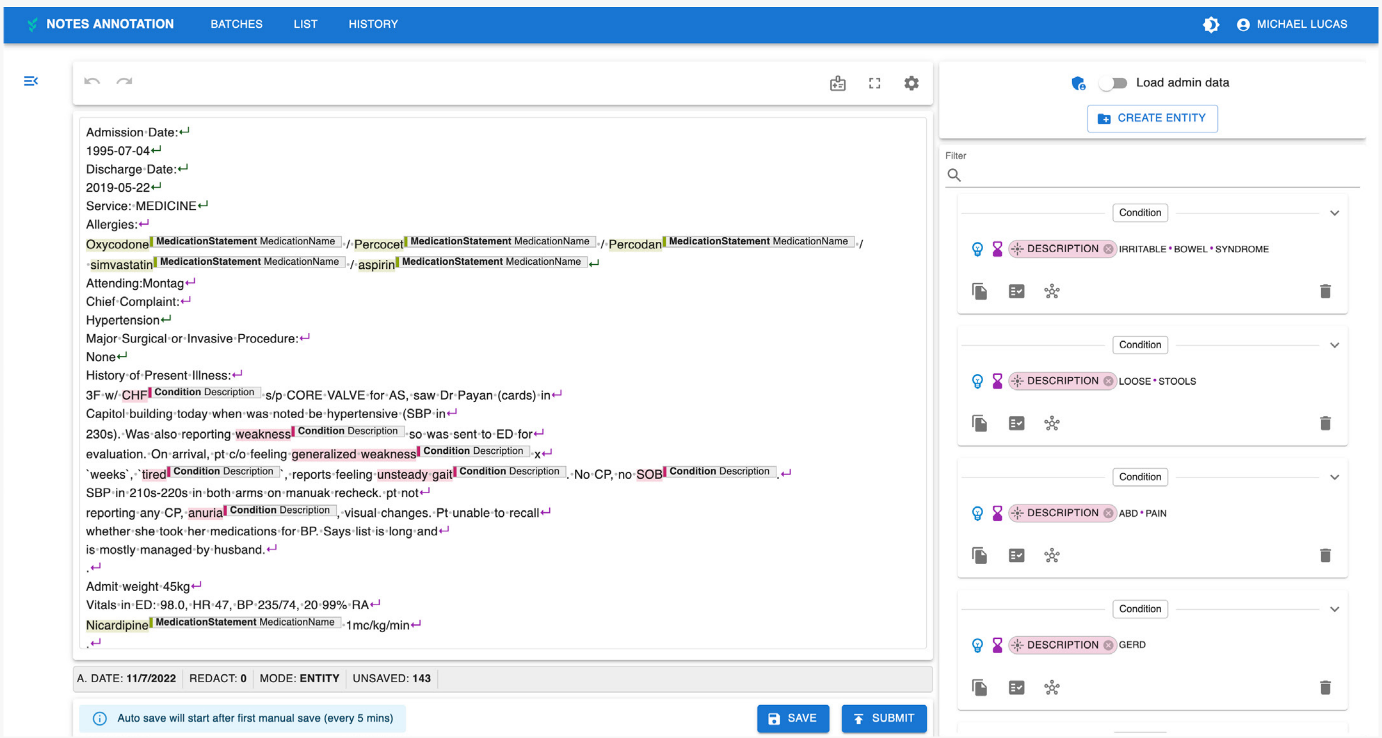

Fig 6. Custom tool Truveta annotators use to label unstructured medical record data for normalization.

In the example below, a clinician documented the explicit absence of symptoms like coughing and dyspnea, as well as updates to the patient's medication regimen. These pieces of information were deemed relevant enough to be documented by a clinician, but cannot be analyzed for research without being extracted, properly normalized, and negated in the structured data. The power of TLM is providing this functionality at scale while exceeding the quality of human experts.

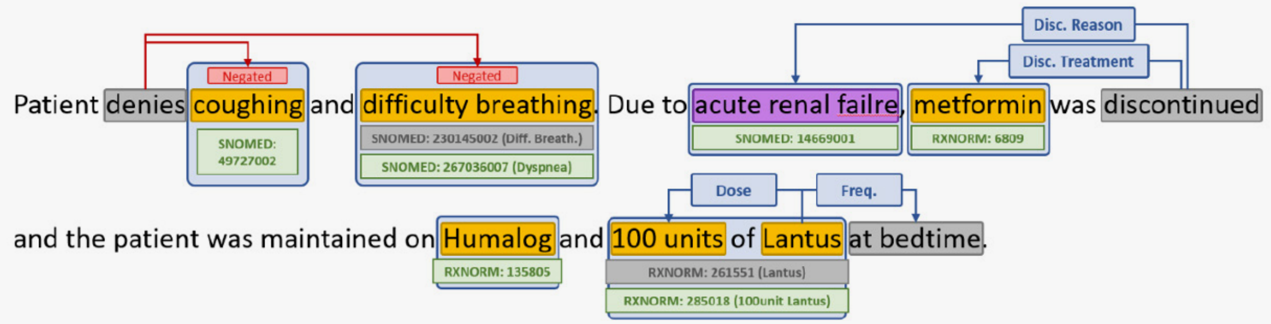

Fig 7. An illustration of many of the tasks TLM executes: detecting clinical concepts, normalizing concepts to target ontologies, linking related concepts, and performing context-based negation.

Accessing clinical concepts from patient notes is especially important for advancing research on rare diseases, where much clinical nuance around diagnosis and treatment is only captured in clinical notes. However, until recently, notes extraction has typically focused on high-prevalence, well-documented disease areas, given the effort involved. In the absence of clinical notes data, researchers studying rare diseases have typically had to acquire and combine data from many sources. This approach is highly manual and not easily replicable. TLM is designed for broad applicability and provides expert-level extractions of both common and rare diseases. TLM can also be further fine-tuned for accuracy within specific domains of study.

Figures 8 and 9 below illustrate how TLM extracts key research data from clinical notes for a rare genetic disorder called Ornithine Transcarbamylase Deficiency (OTC Deficiency). Since OTC Deficiency affects only 1 in 14,000 to 17,000 people, researchers interested in this disease typically face challenges with data availability and consistency. Most of the data relevant to OTC Deficiency research is not available in traditional data sources such as claims. Instead, it is frequently communicated between practitioners via clinical notes.

Figure 8 shows a more classical use of clinical Natural Language Processing (NLP) applied to a note for OTC Deficiency: extraction of general conditions and symptoms. The right-hand side shows a list of current non-negated conditions the patient is experiencing. After extraction, these raw strings will be normalized to target ontology codes. Each record will then be transformed to an Observation record in the TDM with a source provenance indicating that the information was extracted from clinical notes. This type of extraction is relevant to any disease.

DISCHARGE SUMMARY Patient: ************************ DOB: ******** MRN: *********** Date of Admission: ******** Date of Discharge ******** Primary Care Provider: *****************, MD Admitting Provider: ********************, MD Discharging Provider: *****************, MD Discharge Diagnoses Principal Problem: Altered mental status, unspecified altered mental status type Active Problems: Hyperammonemia Ornithine transcarbamylase deficiency Hypertension Resolved Problems: * No resolved hospital problems. * Consultants None Issues to be Addressed at Follow-Up - Continued care at ******** Reason for Hospital Admission Please refer to the H&P for full details. Briefly, this is a ** y.o. ****** with history of ornithine transcarbamylase deficiency with h/o minimal compliance, followed by medical genetic metabolic team at ****** who presented with AMS in the setting of several days of not feeling well and not taking medications. Severely agitated and combative despite droperidol, versed, and zyprexa, and was subsequently intubated. Findings significant for ammonia 248, acute liver failure, and procalcitonin 6. CTH with no evidence of cerebral edema. ****** was given 200 mg/kg of arginine at 146 mL/hr over 90 minutes as well as D10W infusion. ****** is being transferred to **** for continued management. Hospital Course by Problem Ornithine transcarbamylase deficiency Followed by Dr ******* ******* and Dr **** in *********. Hx of inconsistent use of meds. Per EMR review, home arginine (L-arginine) powder 5g in liquid tid, levocarnitine (carnitor) 330 mg oral tablet. Unable to get accurate med rec from ********. General principles of management: Rehydrate and maintain good urine output without overhydration. Remove nitrogen (ammonia) from the body using medications and/or hemodialysis (HD when ammonia persistently > 350-400, or if rapidly \nincreasing, or acute hyperammonemia that is resistant to initial drug therapy). Initial ammonia 248. ****** was given 200 mg/kg of arginine at 146 mL/hr over 90 minutes as well as D10W infusion. Normal blood glucose. CT head with no evidence of cerebral edema. - Monitor ammonia every 4 hours - D10W infusion - Arginine infusion - ****** to start sodium phenylbutyrate - No current indication for HD - Monitor CMP Acute metabolic encephalopathy In the setting of non-adherence to OTC deficiency medications as above. Ammonia 248. No other electrolyte abnormalities, no hypoglycemia. Not hypoxic. CTH normal. Cannot fully r/o infectious component as procalcitonin elevated 6, though normal WBC, afebrile. Potential pulmonary source based on CXR with LLL opacification - c/f atelectasis vs consolidation. UA bland. - Management as above ADHD Per EMR Ritalin 5 mg. ****** reports Adderall, though not taking. Hypertension triamterene-hydrochlorothiazide 37.5-25 mg Oral tablet. Per ********, not taking. Hypertensive - prn PO hydralazine for SBP > 180 Mechanically ventilated Due to severe agitation refractory to medical therapy. Sedated with propofol and precedex. - Wean when appropriate Acute liver failure INR 1.3, encephalopathic, elevated transaminases AST 300, ALT 2300. Tbili 3. Suspect consequence of non-adherence to medications for OTC deficiency. APAP, EtOH wnl. - Acute hepatitis panel pending - Management as above Sepsis of unknown etiology Elevated procalcitonin Abnormal CXR Procal 6.19 in ED. Normal WBC. Potential pulmonary based on CXR with LLL opacification - c/f atelectasis vs consolidation. UA bland. Note: Has PCN allergy. Given IV 2g ceftriaxone and started on IV vancomycin. - Management per ****** Hyperthyroidism TSH 0.04 with free T4 1.25. Pending Labs and Imaging Order Current Status Hepatitis Panel, Acute Collected (********) Discharge \nMedications New Medications Details arginine 100 mg/mL injection Inject 219 mLs into the vein continuous. aka: R-GENE chlorhexidine 0.12% solution Take 15 mLs by mouth 2 times daily. aka: PERIDEX dexmedetomidine in saline 4 mcg/mL infusion Inject 0-164.1 mcg/hr into the vein continuous. aka: PRECEDEX propofol infusion 10 mg/mL infusion Inject 0-7,658 mcg/min into the vein continuous. aka: DIPRIVAN vancomycin 1,250 mg in sodium chloride 0.9% 500 mL IVPB Inject 1,250 mg into the vein once for 1 dose. Indications: Sepsis of Unknown Etiology Unchanged Medications Details Arginine Powd \"L-Arginine Powder\" Amino Acid Powder. You might also be taking other medications not listed above. If you have questions about any of your other medications, talk to the person who prescribed them or your Primary Care Provider. Discontinued Medications methylphenidate 20 MG tablet aka: RITALIN Condition at Discharge Stable Current Vital Signs Temp: 36.2 \u00b0C (97.2 \u00b0F), Pulse: 112, Resp: 30, BP: 136/62, SpO2: 100 % on 40% fiO2 mechanical ventilation Last Weight Weight: 105.9 kg (233 lb 7.5 oz) (********) Disposition ******** Current Support Family Services Ordered None Follow-up Information No follow-up provider specified Close outpatient follow-up recommended: YES Complex case management needed: Unknown Code Status


```
{
    "visit_date": "not reported",
    "signature_date": "REDACTED [***************]",
    "symptoms_diagnoses": [
        {
            "name": "Altered mental status",
            "date": "not reported",
            "temporality": "current",
            "status": "positive"
        },
        {
            "name": "Hyperammonemia",
            "date": "not reported",
            "temporality": "current",
            "status": "positive"
        },
        {
            "name": "Ornithine transcarbamylase deficiency",
            "date": "not reported",
            "temporality": "current",
            "status": "positive"
        },
        {
            "name": "Hypertension",
            "date": "not reported",
            "temporality": "current",
            "status": "positive"
        },
        {
            "name": "severely agitated",
            "date": "not reported",
            "temporality": "current",
            "status": "positive"
        },
        {
            "name": "combative",
            "date": "not reported",
            "temporality": "current",
            "status": "positive"
        },
        {
            "name": "inconsistent use of meds",
            "date": "not reported",
            "temporality": "historical",
```


Fig 8. Visualization of disease characteristics and symptoms extracted by TLM for a patient with OTC Deficiency.

TLM is also designed to extract nuanced data relevant to more niche research areas. Figure 9 shows a much less common use of clinical NLP – extraction of dietary information, which is very important to the treatment and management of OTC Deficiency but rarely captured in structured data. The power of TLM is the ability to consistently extract both common and nuanced data for clinical researchers at scale and with equal determination.

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 R
            "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 9. Visualization of dietary information extracted by TLM for a patient with OTC deficiency.

Collectively, these processes offer researchers unprecedented access to clinical insights previously hidden in free-text notes, for both high-prevalence conditions and rare diseases. Truveta Data today includes more than 5 billion notes from more than 30 health systems.

Operationalizing Truveta Language Model at scale

With each extracted clinical concept, there is a distinct data pipeline which is managed across billions of data points every day:

1. Measuring performance of concept extraction of notes, ensuring accuracy and coverage exceed human expert range.
2. Measuring performance of normalization of extracted concept strings identified in (1), ensuring precision and recall exceed human expert range

Whenever 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.

Our data quality goal is to provide the transparency 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.

Conclusion

TLM is a profound innovation for making healthcare data trustworthy and useful for analytics. With TLM, Truveta's community of life science, government, and healthcare organizations are studying complete, timely, and clean data to achieve our 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 Truveta website, follow on LinkedIn, or contact at info@Truveta.com.