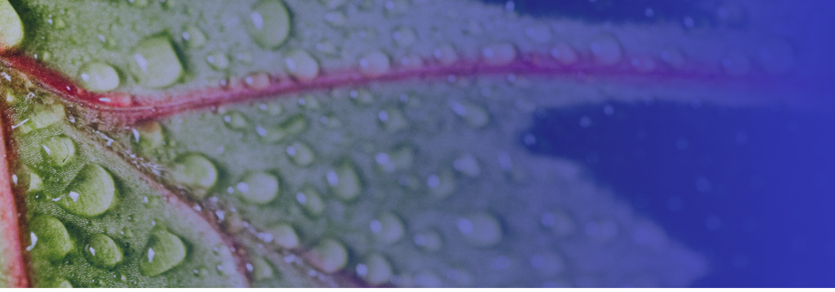
TRUVETA

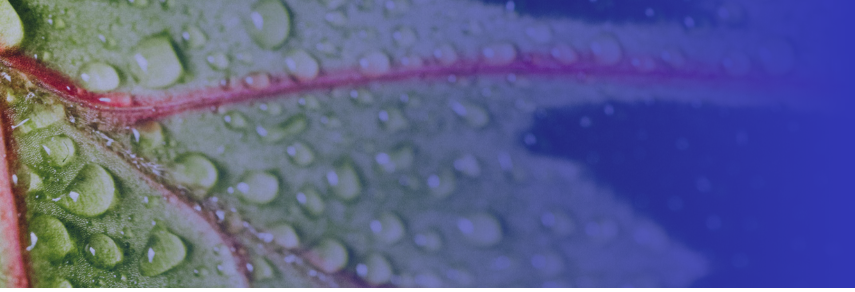
Whitepaper

Our approach to data quality

June 2024



Contents



Raising the bar for healthcare data quality

Access to high-quality data is essential for advancing healthcare research, driving innovation, and improving patient outcomes. Clinical trials and registries have long served as the cornerstone for evidence generation, offering valuable insights into the safety and effectiveness of medical interventions. However, these data sources are not without their limitations. Both trials and registries are time- and labor-intensive and generally result in data that are not generalizable.

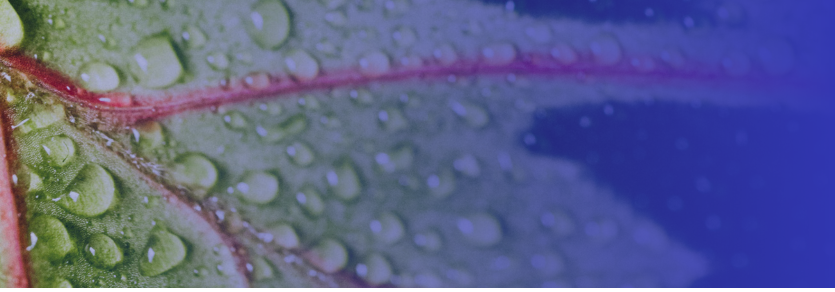
> ***Truveta aims to deliver the most complete, timely, and clean regulatory-grade data representative of a large and diverse population, compared to traditional healthcare data sources.***

Researchers are increasingly looking to real-world data (RWD) captured from healthcare encounters as a promising solution for representative and generalizable data. Unlike data collected in controlled clinical settings, RWD are derived from treatment interactions across varied settings, capturing information from diverse patient populations. This breadth of data holds great potential for overcoming the limitations of traditional data sources and unlocking answers to a broad range of research questions.

However, since RWD are not collected explicitly for research purposes, they are often fragmented, heterogeneous, and subject to measurement and documentation errors, impacting data quality. Researchers must typically navigate the complex task of data cleaning and validation to ensure the reliability and validity of their findings.

Truveta aims to deliver the most complete, timely, and clean regulatory-grade data representative of a large and diverse population, compared to traditional healthcare data sources. Truveta Data includes complete electronic health record (EHR) data for more than 100 million patients, collected from more than 20,000 clinics and 900 hospitals daily. These medical records are then linked with social drivers of health (SDOH), mortality, and claims data for a complete view of patient journeys.

This whitepaper outlines Truveta's approach to data quality, which includes an advanced Quality Management System, industry certifications, and rigorous third-party and customer audits to ensure evidence generated using Truveta Studio is fit for regulatory submission. For information on patient privacy, security, or how the Truveta Language Model is used to clean EHR data for research, download our other whitepapers.



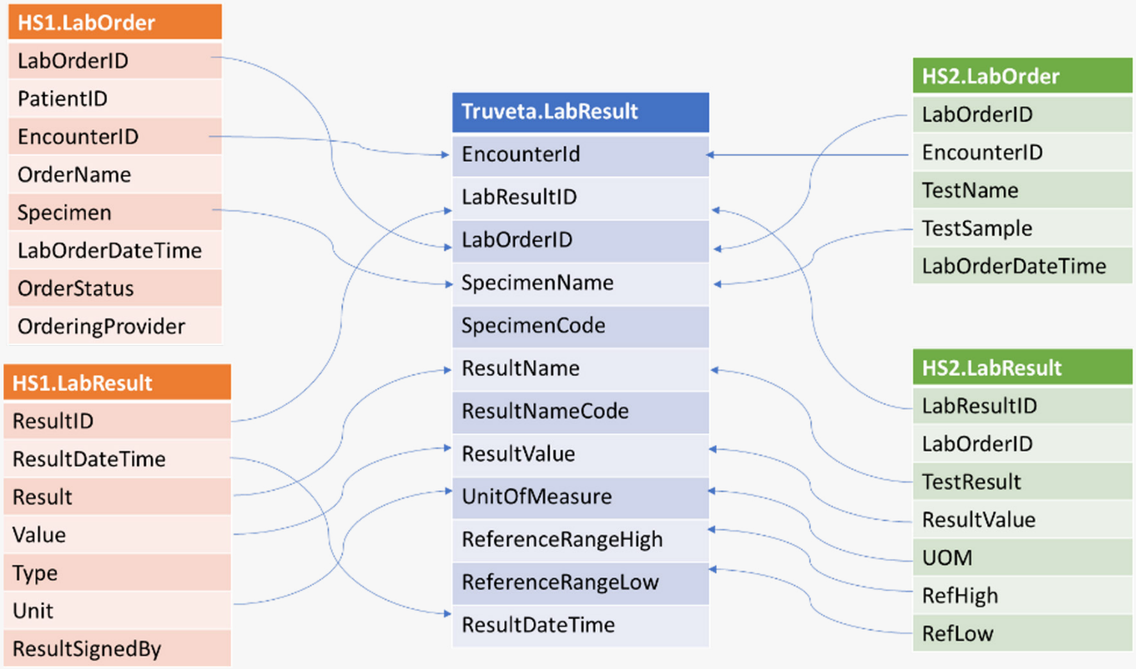
Ingesting and cleaning data from member health systems

EHR data is notoriously ‘messy’, with crucial information contained in semi-structured categorical strings that must be standardized into structured data. Unstructured clinical notes, found in progress notes, discharge reports, lab reports, and transcribed telephone encounters, for example, contain details such as symptoms, clinical status measures or scores, and reasons for medication changes. These data require cleaning to be ready for analytics, which includes normalizing values to standard ontologies, standardizing units of measure, removing erroneous values/records, and de-identifying data to protect patient privacy.

Truveta is a growing collective of 30 health systems that provide complete EHR data for inclusion in Truveta Data. Raw patient data from each member health system are sent daily to a secure, cloud-based environment called a Truveta Embassy. Direct connections with member health systems create a rapid feedback loop for improving data quality, with members receiving feedback via data health dashboards. This loop is strengthened as member health systems use Truveta Data and Truveta Studio for research and analytics at their organizations.

Incoming data are processed and assessed continually for quality. Next, these data, which come in enormously heterogeneous formats and data schema, are transformed to a single data model called the Truveta Data Model (TDM). We refer to this process of data model unification as syntactic normalization. We receive hundreds of disparate tables from health systems across EHR vendor schemas. Our process of syntactic normalization requires meticulous alignment of all these data to dozens of tables in TDM.

The example below illustrates how laboratory result data from two health systems would be integrated. In some instances, we see identical source fields (e.g., LabOrderID) that are mapped directly to target fields in our model. In other cases, we define a new field (e.g., SpecimenName) that covers several similarly named fields. We may also remove fields from our model (e.g., ResultSignedBy) that may be redundant or have no value for clinical research or add new fields (e.g., ResultNameCode) to further clarify our model.



HS1.LabOrder
- LabOrderID
- PatientID
- EncounterID
- OrderName
- Specimen
- LabOrderDateTime
- OrderStatus
- OrderingProvider

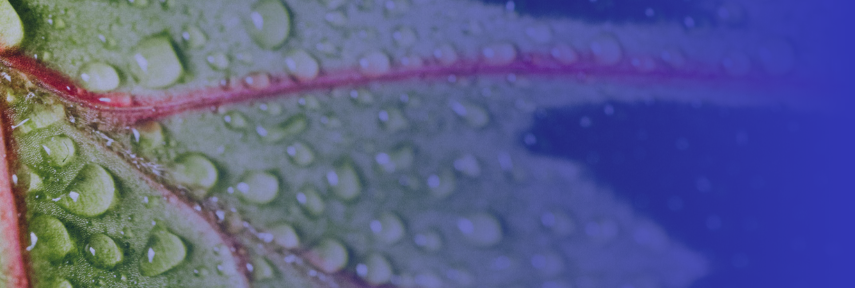
HS1.LabResult
- ResultID
- ResultDateTime
- Result
- Value
- Type
- Unit
- ResultSignedBy

Truveta.LabResult
- EncounterId
- LabResultID
- LabOrderID
- SpecimenName
- SpecimenCode
- ResultName
- ResultNameCode
- ResultValue
- UnitOfMeasure
- ReferenceRangeHigh
- ReferenceRangeLow
- ResultDateTime

HS2.LabOrder
- LabOrderID
- EncounterID
- TestName
- TestSample
- LabOrderDateTime

HS2.LabResult
- LabResultID
- LabOrderID
- TestResult
- ResultValue
- UOM
- RefHigh
- RefLow

Example of syntactic normalization process, which involves mapping similar fields in records from two health system members to one common data model, Truveta Data Model (TDM).



Establishing data quality standards

After ingesting and cleaning data provided by member health systems, we continuously measure data quality to ensure the usefulness and trustworthiness of the data for research. Our robust approach to assessing data quality aligns to the following industry-standard categories and RWD regulatory submission standards:

- **Representativeness** measures how the patient diversity of Truveta Data compares to the overall US population.
- **Completeness** measures whether all expected data fields and values are present in the linked, longitudinal patient record across EHR, notes, images, genomes, claims, SDOH, and mortality data.
- **Timeliness** measures how quickly data is delivered to Truveta and made available for research following its origination in point-of-care data capture systems.
- **Cleanliness** measures whether the data are accurate and plausible, and thus usable for research analytics. Creating a foundation for clean data requires three unique processes:
    - Semantic normalization, which measures how data is translated from source strings to target ontologies (e.g., SNOMED, LOINC).
    - Value and unit of measure normalization, which involves normalizing values to research-ready concepts and standardizing the units of all measurements.
    - Clinical validity, which measures whether values, disease prevalence, and other metrics meet clinical expectations.

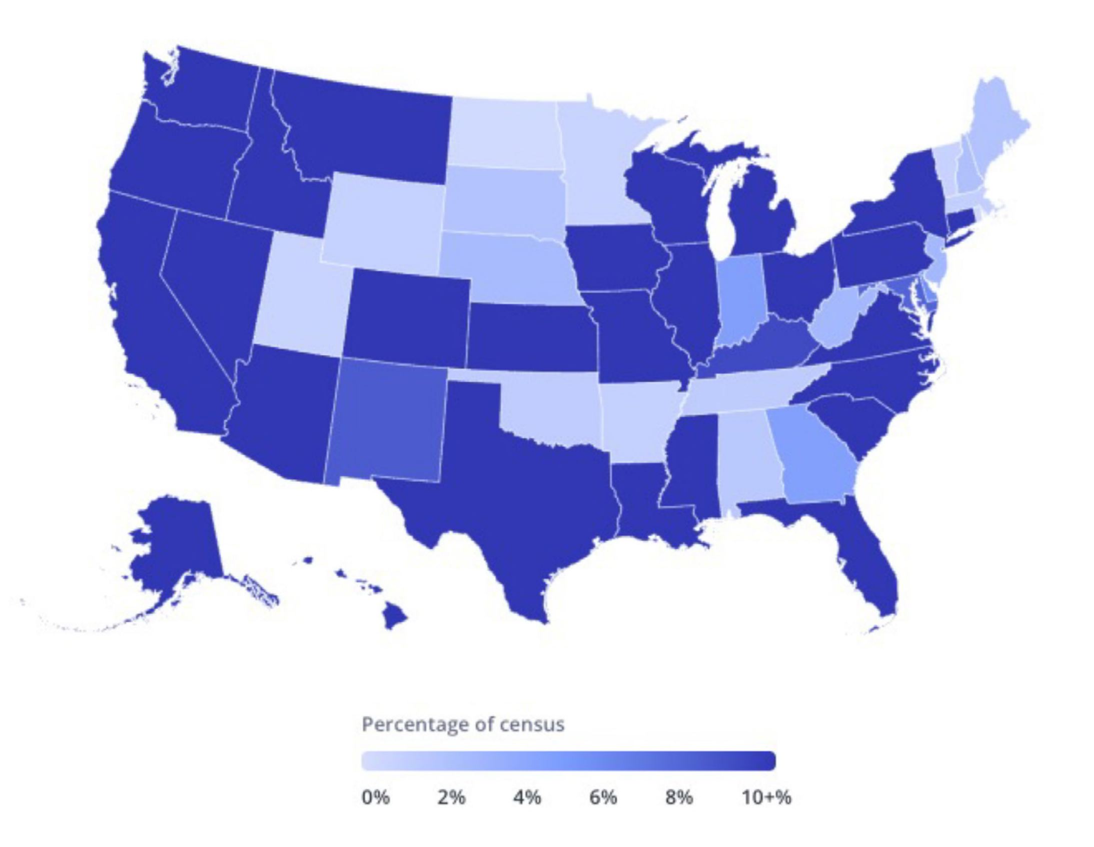
To ensure transparency into the quality of our data, our national dataset and any patient population being studied is accompanied by a “population datasheet.” This data summary provides counts and high-level metrics related to the underlying data, insight into the representativeness of the data benchmarked against the US census, and clinical validity metrics of the study population compared to all Truveta Data.

Sign up for a free account to view examples of complete studies in Truveta Studio. >>

Representativeness

Clinical research demands representative data where findings can generalize to a national population. We measure representativeness by benchmarking geographic, race, ethnicity, gender, and age diversity against the US Census.

We benchmark representativeness at a state level based on whether we have data coverage for at least 10% of the state's population of covered lives. This benchmark is set by the Center for Medicare and Medicaid Services (CMS) under their qualified entity framework. As of publication, Truveta has data from 34 states that meets this benchmark.



Truveta Data representative coverage as of May 2024.

Completeness

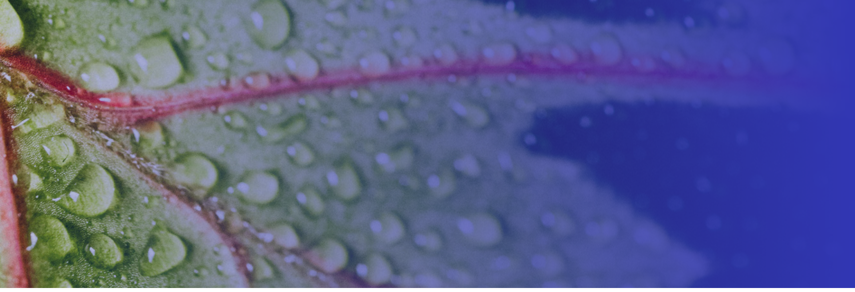
Truveta views completeness as a measure of whether the data reflect full longitudinal patient history. EHR data provide an extremely granular understanding of the clinical care a patient receives, including semi-structured data (e.g., medication administrations, conditions, procedures, encounters) and unstructured data (e.g., clinical notes, images, genomes). Given that patients may be seen across multiple health institutions over the course of their life, Truveta invests heavily in technology to link the same patient across multiple health systems to maintain a complete longitudinal view. This is made possible through TruvetaID, a privacy-preserving tokenization process applied to all patients in Truveta Data, enabling high-precision patient linkage both within and across health systems.

A limitation of the EHR is that this granular information typically only pertains to care the patient receives within that respective system. Truveta's member network is substantive but does not capture 100% of care. Therefore, Truveta builds partnerships to obtain patient claims data to ensure all clinical care is captured and that patient journeys are as complete as possible.

We also acknowledge the EHR is not optimized for mortality data (with two-thirds of deaths occurring outside of clinical settings) or for social drivers of health (SDOH) data, much of which is not captured consistently by providers. Truveta has partnered with LexisNexis to link highly accurate and up-to-date mortality and SDOH data to the patient record. Our patient-level SDOH data includes more than 45 attributes including education, income, housing stability, and social support, which enable researchers to study the impact of lifestyle and social factors on clinical care and outcomes.

Truveta also benefits from broad investments in healthcare data interoperability in which health systems share data for patients who receive care outside of the health system. Data from health system interoperability efforts are integrated into a patient's longitudinal record in Truveta Data. For example, Truveta receives discharge summaries from non-Truveta providers that capture services administered outside of the Truveta network.

In addition to ensuring completeness of core clinical data domains in a patient's record (e.g., conditions, medications, SDOH, mortality data), we also measure and ensure quality within each individual record. For example, a medication administration record should contain (among other items) the medication and dose administered, a linked encounter during which the administration took place, and a timestamp for when the administration began. Missing one or more of these fields in a medication administration record renders it less usable or even unusable for research. Truveta develops and operates metrics across all tables in the TDM to assure this finer dimension of completeness is maintained for data records. Completeness gaps in data are fed back to member health systems so they can improve and complete the data feeds that arrive to Truveta.



Timeliness

Truveta aims to deliver the most up-to-date data available for healthcare research, with a goal of providing access to data within one day of care provision. Health systems contribute daily data feeds with up-to-date encounter, condition, medication, and other data for immediate ingestion – regardless of whether the encounter has been closed or coded. This approach ensures that researchers have access to timely data, enabling immediate insight into patient care pathways and outcomes.

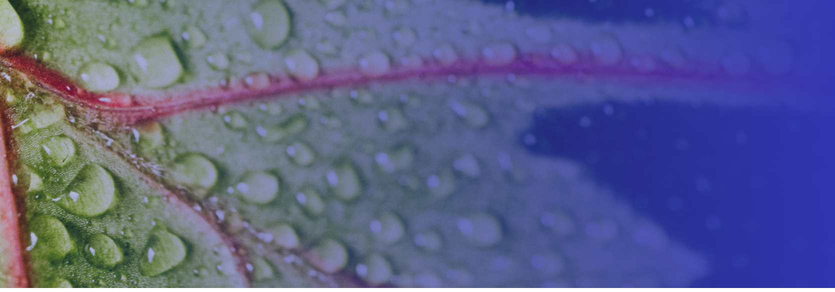
Cleanliness

We define data cleanliness as a researcher having access to accurate, consistent, error-free, and research-ready data. We believe there are three critical components to bringing clean data to the researcher: semantic normalization, unit of measure normalization, and clinical validity.

Semantic normalization

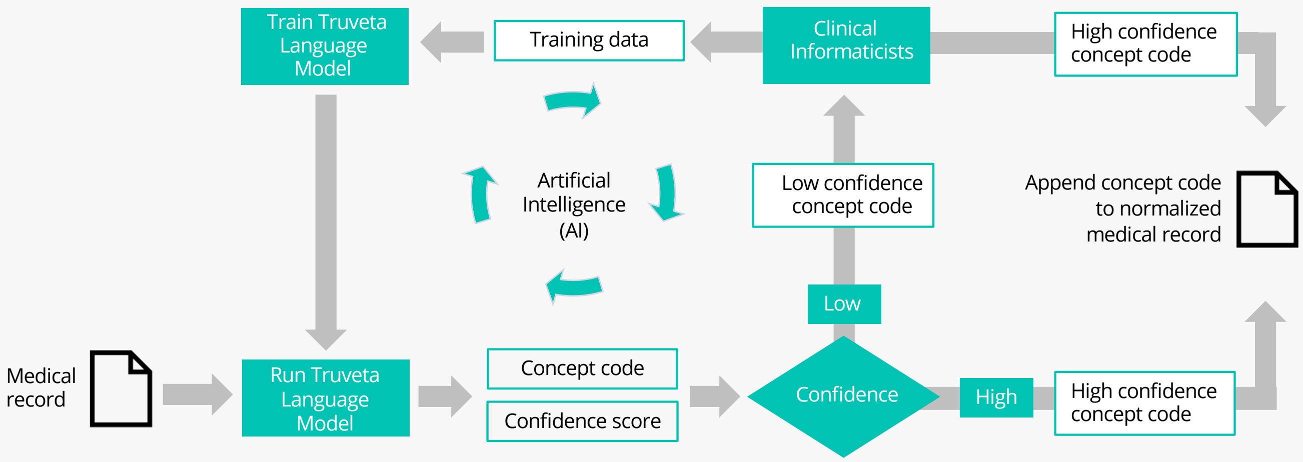
Data from health system members is heterogeneous, arriving as semi-structured data containing many different strings and/or codes for the same clinical concept. In this context, a concept refers to a specific clinical idea or piece of information. For example, one system may label a condition as “acute renal insufficiency,” whereas another will label the same condition as “AKI.” Both text-strings are intended to express the same concept, but such inconsistencies in documentation make it challenging to meaningfully work across medical records to perform research.

Truveta receives hundreds of millions of heterogeneous concepts from health systems that would be impossible for a researcher to parse through and categorize on their own. Through semantic normalization, Truveta translates these concepts to target medical ontologies (e.g., SNOMED-CT, LOINC, RxNorm), ensuring their immediate usability for research. Furthermore, Truveta normalizes to concepts within these ontologies that are more specific than how concepts are coded by the health system. For example, the SNOMED-CT ontology confers much more clinical depth for diagnostic codes than ICD-10-CM, the usual ontology used for claims data.

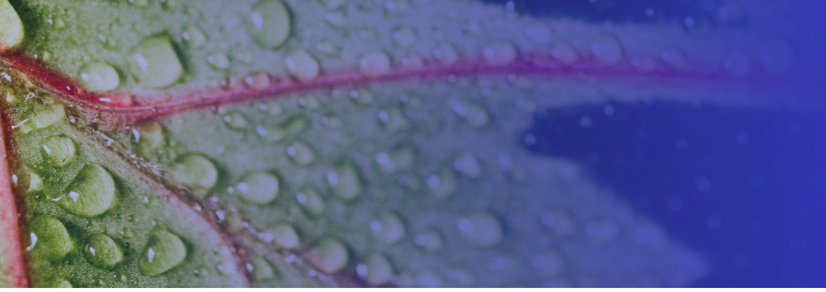


Traditionally, semantic normalization takes place through simple rule-based engines or manually by human annotators. These legacy technologies do not perform at the scale of Truveta Data. To accurately normalize data in real-time, we created the Truveta Language Model (TLM), a large-language, multi-modal AI model used to clean billions of data points for health research. TLM fine-tunes open large language models with additional training on de-identified EHR data from more than 100 million patients, enabling highly accurate semantic normalization without the commercial bias present in claims data.

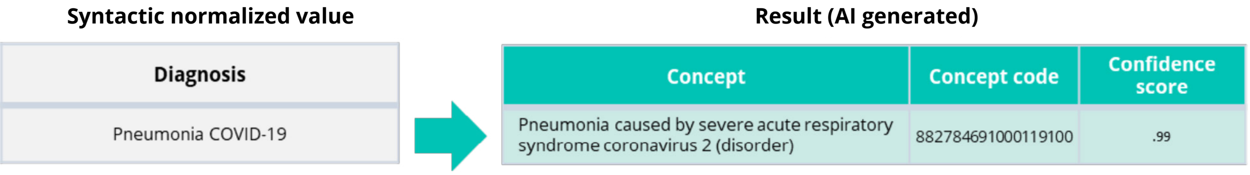
The diagram below shows how TLM is used to perform expert-led, AI-driven semantic normalization. For each incoming medical record, our system employs AI to associate each term contained in the record with standard terms or concepts within industry-standard ontologies. Each match is assigned a confidence score, reflecting the model's assessment of how closely the concept aligns with the term in the record.


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

Truveta's AI model leverages the expertise of clinical informaticists to resolve low-confidence concept mappings and generate training data to continuously improve the models.



In the example below, TLM has mapped one diagnosis to one standard SNOMED CT concept and provided an associated confidence score. A high-confidence score suggests a strong alignment between the SNOMED concept and the terms in the record. Concept codes with high-confidence scores are then added to the record and de-identified before being transferred to Truveta Studio.



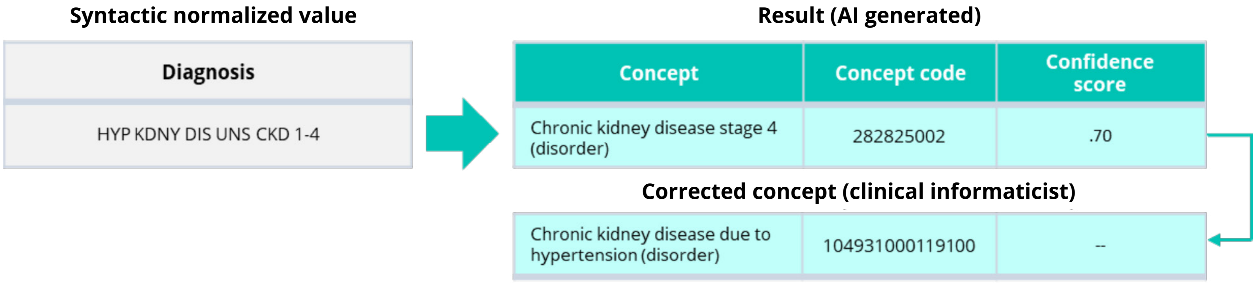
Syntactic normalized value

Diagnosis
Pneumonia COVID-19

Result (AI generated)

Concept	Concept code	Confidence score
Pneumonia caused by severe acute respiratory syndrome coronavirus 2 (disorder)	882784691000119100	.99

In the next example, TLM has mapped one diagnosis to a concept with a low-confidence score. In this case, the record is sent to our team of clinical informaticists for review and comparison. The concept code is updated with a more appropriate concept, added to the record, and sent back into our pipeline.



Syntactic normalized value

Diagnosis
HYP KDNY DIS UNS CKD 1-4

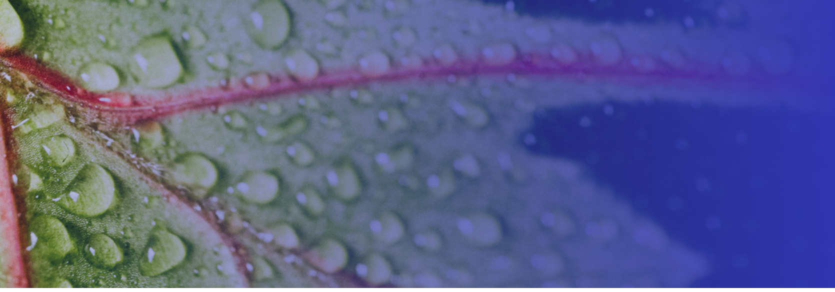
Result (AI generated)

Concept	Concept code	Confidence score
Chronic kidney disease stage 4 (disorder)	282825002	.70

Corrected concept (clinical informaticist)

Concept	Concept code	Confidence score
Chronic kidney disease due to hypertension (disorder)	104931000119100	--

The updated codes enhance TLM through ongoing model training. Truveta retains a comprehensive record of codes to preserve iterative improvements for research. We also use transfer learning, an advanced AI technique that leverages knowledge gained from solving one problem to address a related one. By combining the power of AI, expert-led training, and an intense focus on quality, our system tackles millions of normalization challenges every day, confidently mapping medical concepts and delivering clean data to the research community.



Value normalization and unit of measure standardization

Truveta's unit of measure value normalization and unit of measure standardization also ensure data consistency. For example, weight data received from health systems may arrive in a variety of units, such as grams, kilograms, ounces, or pounds. Occasionally, erroneous units (like seconds) may be attributed to a weight record. Truveta not only normalizes all source strings indicative of weight measurements to the proper target ontology, but also standardizes the values to align with the standard unit of measure for the weight record. This process ensures that researchers have a consistent experience working with the data and can readily use it without laborious cleaning.

Clinical validity

To achieve our goal of delivering research-ready data, we build processes to measure the clinical validity of the data. It is common for each research group to have their own processes for confirming data validity, asking questions such as:

- Do patients with hypertension have blood pressure measurements around the time of their blood pressure diagnoses?
- Do prevalence and incidence rates of rheumatoid arthritis align with rates published by the CDC?
- Are the ranges of hemoglobin A1c as I would expect (between 4-20% with a right-skewed distribution)?
- Do times of medication administration always follow the ordering provider's request for that medication?

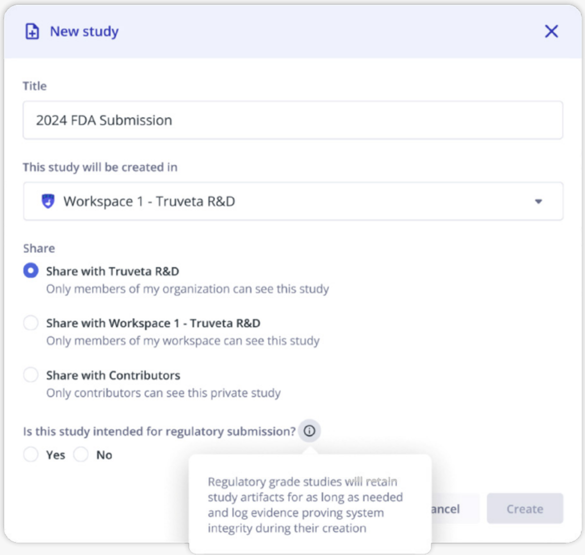
Truveta operationalizes these types of questions into a large and growing library of metrics that run consistently over all Truveta Data, outlining consistencies and inconsistencies that are immediately actionable. Clinical validity issues are reviewed by experts to improve data processing or are fed back to health system members so they can improve their data feeds. Metrics are provided to customers through the population data sheet, providing transparency and confidence in the population the researcher will study.

Ensuring data fit for regulatory submissions

Truveta upholds the highest standards in building quality systems to gain researchers' trust and ensure data support for regulatory submissions. We have established a robust quality management system (QMS) with documented procedures, roles, and responsibilities to ensure data integrity and continuous improvement. Each component of our QMS adheres to standard operating procedures (SOPs) designed to identify risks and establish purpose-built processes and software controls. Continuous monitoring and evidence logging ensure compliance, enabling customers to demonstrate artifact integrity in regulatory submissions. We've also made significant investments in hundreds of product and process enhancements, more than 70 controls aligned with FDA guidance, and more than 30 SOPs addressing regulatory requirements.

Our security and privacy standards are also third-party audited through industry certifications, including SOC 2 Type 2 examination and ISO certifications (ISO 27001, 27701, and 27018), ensuring adherence to global standards for data protection. We are actively pursuing ISO 9001 certification to attest that our data processing system meets the highest bar for reliability, and we have been audited by industry experts with extensive FDA experience, as well as by top biopharmaceutical customers. These audits further validate our regulatory-grade capabilities and ensure a submission with Truveta Data contains all needed elements to comply with FDA recommendations, as published in guidance from the FDA's Center for Biologics Evaluation and Research (CBER) and Center for Drug Evaluation and Research (CDER).

We have also invested in analytics workflows within Truveta Studio that support regulatory-grade research. Researchers can designate studies, snapshots, and notebooks as intended for regulatory submission, which yields verifiable evidence of data quality and system integrity at the time of analysis and ensures reproducibility of findings. With secure storage and versioning of all data, metadata, analytic code, and evidence, researchers can conduct studies confidently, knowing that evidence is readily retrievable for regulatory audits upon request.



New study

Title

2024 FDA Submission

This study will be created in

Workspace 1 - Truveta R&D

Share

● Share with Truveta R&D
Only members of my organization can see this study

○ Share with Workspace 1 - Truveta R&D
Only members of my workspace can see this study

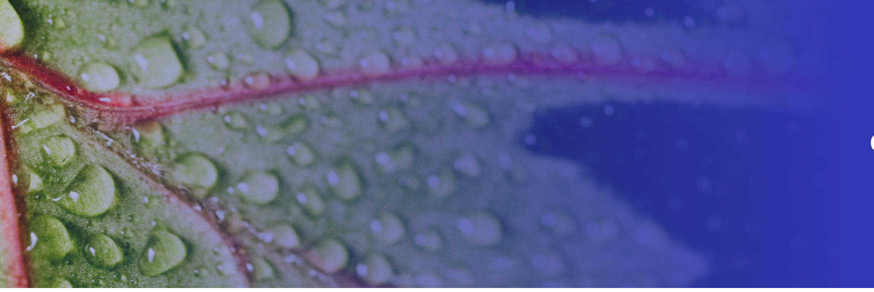
○ Share with Contributors
Only contributors can see this private study

Is this study intended for regulatory submission?

○ Yes ○ No

Regulatory grade studies will retain study artifacts for as long as needed and log evidence proving system integrity during their creation

Researchers can easily designate studies for regulatory submission within Truveta Studio.



Fostering continuous improvement

Continuous data quality improvement is a top priority. We provide routine feedback to health system members to encourage iterative improvement in the data sent to Truveta. Our data monitoring and validation processes focus on completeness, cleanliness, and timeliness, evolving with our understanding of researcher needs. Our machine learning technologies are designed to grow just as quickly as our dataset and research use cases, ensuring consistent data quality.

At the same time, our de-identification and security systems and processes are continuously evolving to outpace security threats and to adhere to the highest privacy standards. We maximize research utility by thoughtfully and precisely redacting data and facilitating research on granular geolocation, absolute time, mother-child linkage, and full SDOH parameters.

Together, these initiatives underscore mission of Saving Lives with Data — bringing the most complete, clean, and timely data to researchers to accelerate adoption of new therapies and advance patient care.

To download other whitepapers on AI, data security, patient privacy, or data analytics, please visit our website. And to learn more, follow us on LinkedIn, or contact us at info@Truveta.com.