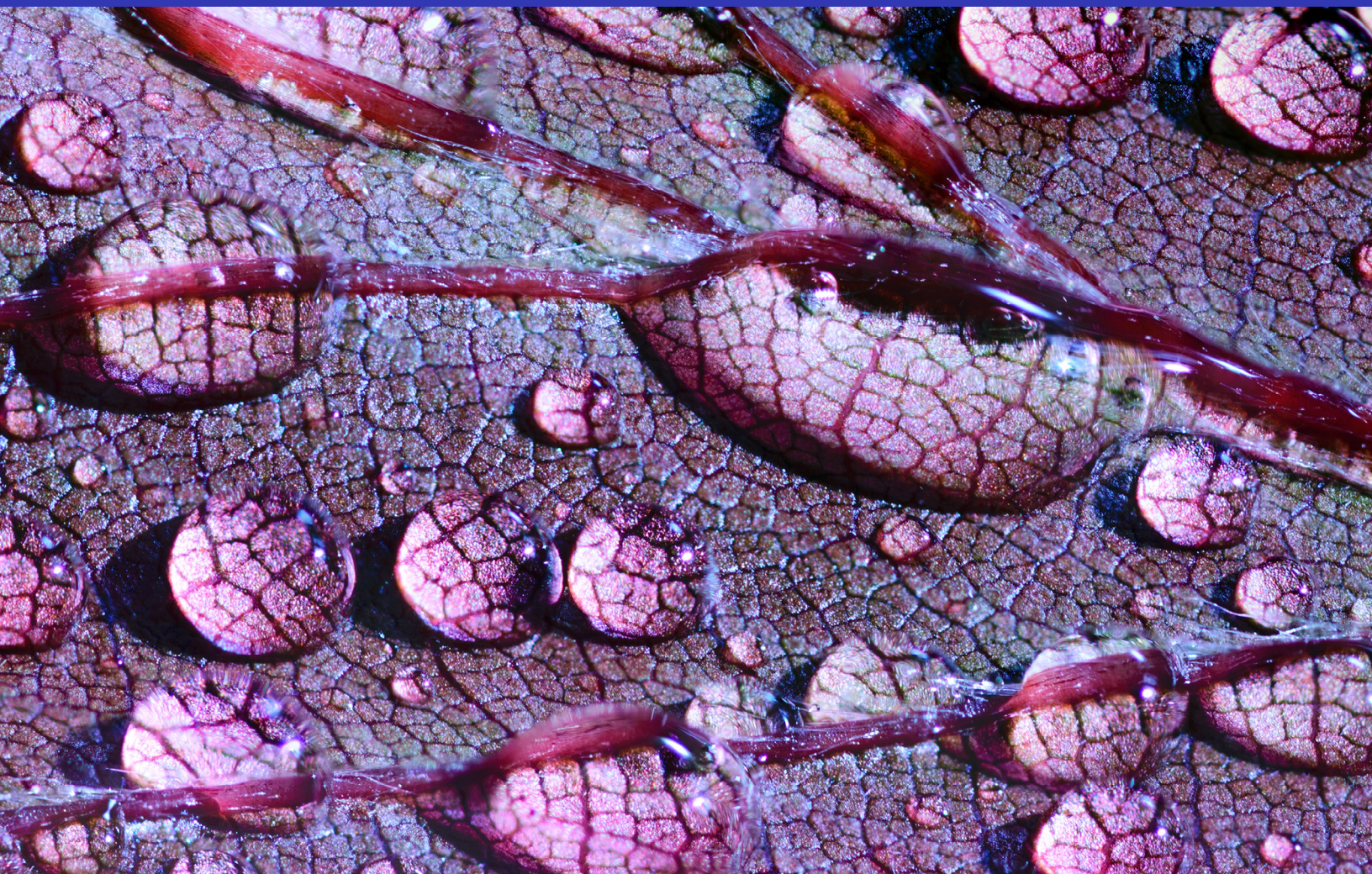


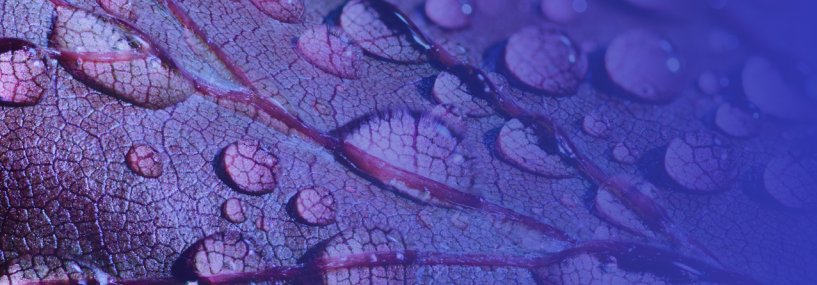


Whitepaper

# Our approach to analytics: Truveta Studio

January 2024





## Contents

<b>Introduction</b> .....	3
<b>The process of scientific inquiry</b> .....	4
<b>Organize hypotheses into studies</b> .....	5
<b>Gather the data</b> .....	6
<b>Build precise inclusion and exclusion criteria</b> .....	6
<b>Truveta Prose: a language for transparent and accurate data definitions</b> .....	7
<b>Never start from scratch with Truveta Library</b> .....	9
<b>Immediate access to row-level data through snapshots</b> .....	11
<b>Analyze the data</b> .....	12
<b>Perform powerful statistics</b> .....	12
<b>Publish daily updating dashboards</b> .....	13
<b>Train AI models</b> .....	13
<b>Publish the study</b> .....	14
<b>Conclusion</b> .....	14



## Introduction

The pace of new therapy adoption has been limited by the ability to produce data that the medical community trusts. Clinical researchers have historically had difficulty accessing and cleaning data, wrestling with hard-to-use tools, and developing and maintaining their own bespoke analytics code. Collaborative efforts required manual oversight of access permissions, necessitating meticulous version tracking for any alterations made to the data, analysis code, and any study content by collaborators. Collectively, these hurdles have slowed the pace of learning in healthcare, limiting advancements in care delivery. Truveta can help solve these challenges.

Truveta offers the most complete, timely, and clean electronic health record (EHR) data enabled by a growing health system collective that together provide more than 18% of all daily clinical care in the US. Data is linked across health systems and integrated with claims, social drivers of health (SDOH), and mortality data for a complete view of patient journeys. Truveta Data is normalized, de-identified, and updated daily.

[Truveta Data](#) is available in [Truveta Studio](#), empowering researchers to build precise and representative populations in seconds, study large populations through flexible and powerful analytics, and collaborate in real-time. Researchers that prefer their own environment can also extract Truveta Data into SAS, Excel, or use other tools such as Microsoft Power BI.

This whitepaper provides an overview of Truveta Studio and how it can accelerate research.

---

*Truveta offers the most complete, timely, and clean electronic health record (EHR) data enabled by a growing health system collective that together provide more than 18% of all daily clinical care in the US.*

## The process of scientific inquiry

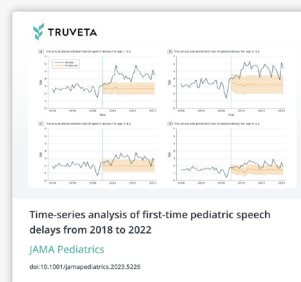
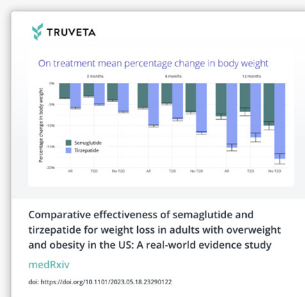
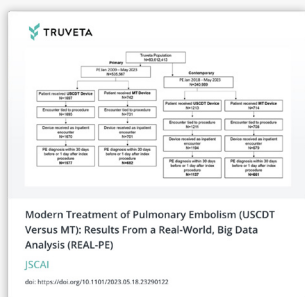
Truveta Studio is designed to support the ongoing process of scientific inquiry, taking inspiration from this [UC Berkeley paper](#) describing how trusted knowledge of the natural world is built through an iterative learning process. What follows is Truveta’s interpretation of this process, brought to life via Truveta Studio.



Truveta Studio empowers researchers to naturally move between the various steps in the process, from an original hypothesis, to organizing hypotheses into studies, gathering and analyzing data, soliciting feedback, and then publishing or submitting for regulatory review. Additionally, researchers can train AI models to augment traditional analytics. The scientific process often isn’t linear, and Studio empowers researchers to iterate and easily move between steps in the process as they continue to discover and learn.

Truveta Studio is already being used by leading health systems such as Providence, Baylor Scott and White, and Novant Health; and innovative life science companies such as Boston Scientific, Boehringer Ingelheim, Reprise Cardiovascular, and SK Life Science, to monitor safety, demonstrate real-world comparative effectiveness, advance discovery, [and more](#).

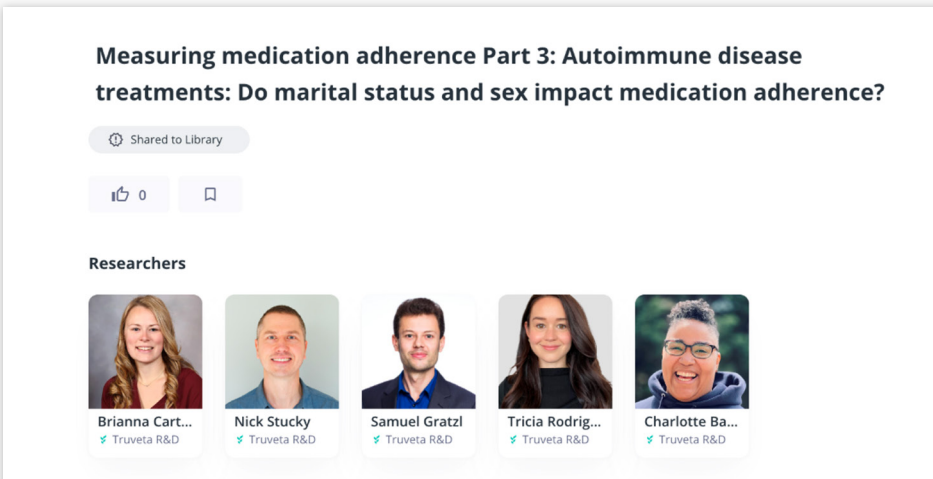
You can see examples of published research [here](#).



## Organize hypotheses into studies

All research starts with a research interest, which can lead to a hypothesis or exploratory research. Often, hypotheses stem from external events like public policy questions, new therapy development, or ideas for improving patient care. Within Truveta Studio, these hypotheses are organized within a Truveta study.

Researchers can seamlessly collaborate on study development with others. Often these researchers belong to the same organization, but there are instances when researchers across organizations collaborate on a single study. An example might be a medical device company conducting an effectiveness study jointly with one or more healthcare institutions, or a life sciences company conducting clinical trials at trial sites across various health systems. Truveta Studio enables such collaborative projects by allowing researchers to invite other researchers to the Truveta study in which they have organized their hypotheses. The researchers can be within their organization, guest researchers from another organization, or even independent researchers with specific expertise required for the study. Truveta Studio enables researchers to do this with ease.








**Measuring medication adherence Part 3: Autoimmune disease treatments: Do marital status and sex impact medication adherence?**

Shared to Library

0

**Researchers**

-   
Brianna Cart...  
Truveta R&D
-   
Nick Stucky  
Truveta R&D
-   
Samuel Gratzl  
Truveta R&D
-   
Tricia Rodrig...  
Truveta R&D
-   
Charlotte Ba...  
Truveta R&D

*An example hypothesis and related study summary in Truveta Studio.*



## Gather the data

The next, and historically most onerous, step in the research process is obtaining the right data to test the hypothesis and complete the scientific study. Today, researchers struggle to access and collect the relevant data needed for their study. EHR data are stored in numerous formats and often across siloed databases. Even within a single EHR vendor there are substantial differences in implementation, complicating research investigations that leverage data from multiple health systems. In fact, the bulk of a clinical researcher's study time is dedicated to data acquisition and cleaning, with approximately [80%](#) of the investigator's time on a study being attributed to data cleaning alone.

Researchers in Truveta Studio have immediate access to all Truveta Data. Because Truveta Data is a clean dataset, researchers can focus their efforts on selecting the right inclusion/exclusion criteria, refining their methodology, and doing great science. To learn more about how Truveta Data is the most complete, timely, and clean data, please read our [data quality whitepaper](#).

## Build precise inclusion and exclusion criteria

Inclusion and exclusion criteria are core to all research, ensuring hypotheses are evaluated against an accurate patient population germane to answering the research question. With respect to real-world data (RWD), these criteria are generally defined through a combination of ontology codes (e.g., ICD-10-CM, LOINC, RxNorm) signifying the clinical concept in tandem with complex logic over the clinical events in a patient's longitudinal health record.

For example, a researcher may wish to study a population with type 2 diabetes mellitus, applying specific inclusion criteria that require two separate measurements of hemoglobin A1c above a particular threshold, and a diagnosis of type 2 diabetes mellitus at least one encounter. This researcher may also impose an exclusion criterion of type 1 diabetes mellitus diagnosis. Beyond inclusion/exclusion, researchers will typically define covariates or other variables integral to the research.

*Researchers in Truveta Studio have immediate access to all Truveta Data. Because Truveta Data is a clean dataset, researchers can focus their efforts on selecting the right inclusion/exclusion criteria, refining their methodology, and doing great science.*



We refer to these definitions for clinical criteria as data definitions (also known as computable phenotypes in certain RWD research communities). Despite the profound importance of data definitions in research methodology and study outcomes, the underlying logic of many research studies is often opaque, described only cursorily within methods sections of journals or hidden in difficult-to-digest and impossible-to-reproduce supplemental documents. Truveta elevates the stature of data definitions to where they belong in the research process; data definitions in Studio are fully transparent, reusable, and shareable either with specific research collaborators or the Truveta research community.

## Truveta Prose: a language for transparent and accurate data definitions

In traditional research, data definitions are stored in legacy formats, such as Excel documents, and implemented through hard-to-understand code in traditional analytic languages, such as SQL, R, STATA, and SAS. Computing these definitions over large-scale data (such as Truveta Data, which includes over 100M+ patient lives and growing) takes days at best and weeks at worst. Code snippets can be shared among researchers, but it is a clunky experience and lacks the transparency and reproducibility researchers seek when collaborating.

In Truveta Studio, a researcher can compute even the most complex data definitions over the entirety of Truveta Data in seconds. Data definitions can be shared and reused readily by researchers. **To bring this power to researchers, we knew we couldn't rely on existing languages or legacy technologies; we needed to build a new language, which we call Truveta Prose.** Truveta Prose is a query language written specifically for clinical research. The language has full range of expressing any clinical idea, which can then be leveraged for defining inclusion/exclusion criteria or as variables needed elsewhere in the study.

*Truveta elevates the stature of data definitions to where they belong in the research process; data definitions in Studio are fully transparent, reusable, and shareable either with specific research collaborators or the Truveta research community.*







Because Prose was purpose-built for clinical research, the language excels for authoring and executing time-based queries (i.e., where temporality of events on the patient timeline must be calculated and compared). These types of queries are typically laborious to author in other languages and difficult to execute across massive datasets. Prose also enables researchers to build queries over very sensitive data, while protecting patient privacy with full regulatory compliance. Researchers can leverage very specific temporal, geographic, or demographic criteria without having exposure to row-level, identifiable data. To learn more about our commitment to privacy, see our [approach to privacy whitepaper](#).

## Never start from scratch with Truveta Library

Truveta Library is the largest and most complete collection of data definitions purpose-built for clinical research. Built by contributions from the Truveta clinical informatics team as well as researchers in the Truveta community, Library is an ever-growing repository ready for immediate use by researchers in Studio. Applying inclusion and exclusion criteria can be accomplished in minutes by using definitions from the Library. The below example highlights how a researcher can build a population of patients with diabetes and renal transplant within two years of onset of diabetes. In this example, the researcher uses the diabetes and renal transplant definitions from Library, and then builds upon these definitions to apply very precise inclusion criteria for the population.

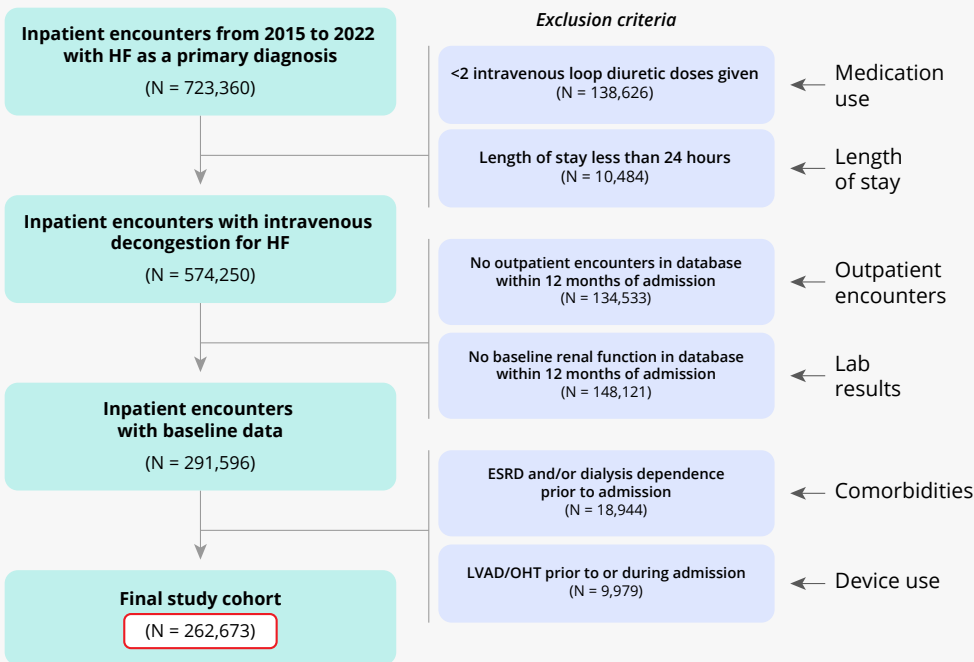
### Diabetes + Renal Transplant (2 years)

Population

```
defDiabetes = import "https://library.truveta.com/o/truveta-research/d/diabetes"
defRenalTransplant = import "https://library.truveta.com/o/truveta-research/d/renal-transplant"
transplant2years = sequence(defDiabetes d, defRenalTransplant r)
{
  | r.RecordedDateTime > d.OnSetDateTime and r.RecordedDateTime <= d.OnSetDateTime + 2years
}
population defDiabetes and defRenalTransplant
```

By leveraging data definitions as composable building blocks, researchers can apply the precise inclusion/exclusion criteria needed to power their research. The following example showcases the clinical depth possible through data definitions written in Prose and published to Library for widespread research use:

The clinical depth in Truveta Data supports populations with highly specific I/E criteria for any disease, drug, or device



An example of inclusion/exclusion criteria for Truveta Data, leading to final study population.

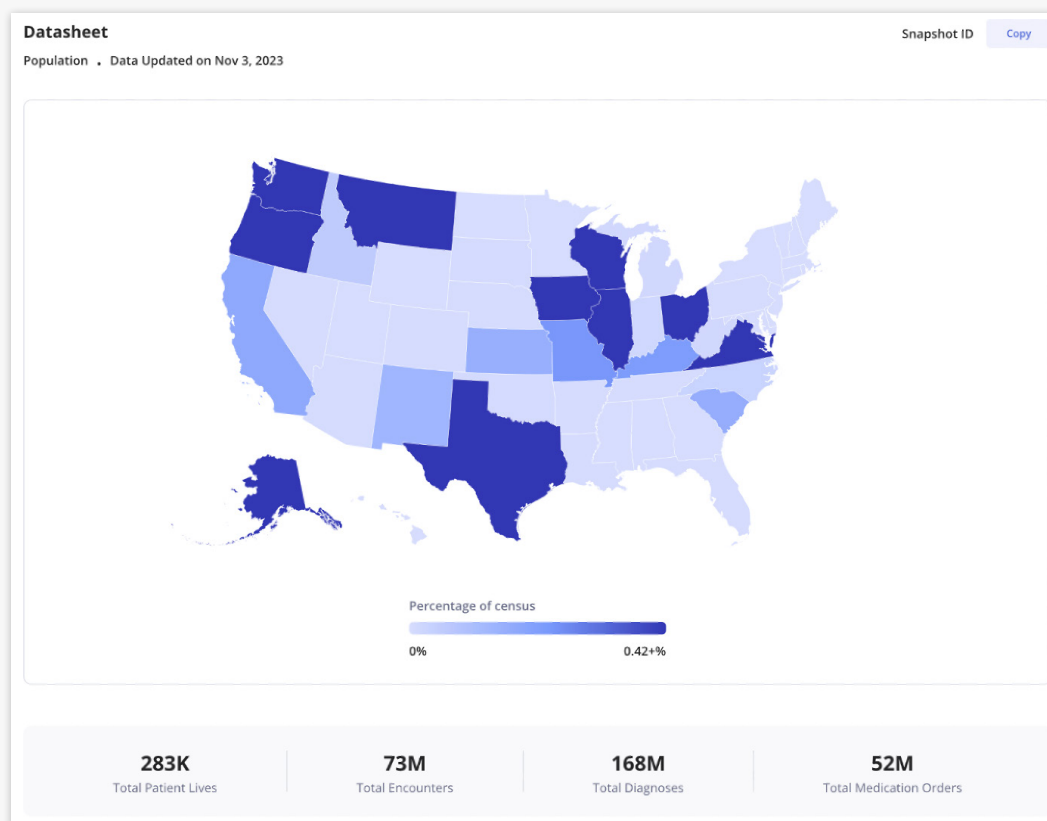
Truveta Library extends beyond data definitions; researchers can share full inclusion/exclusion criteria (which generally leverage numerous data definitions) and can even share complete research studies. **Truveta empowers the acceleration of the research process, ensuring researchers can readily build from and grow community knowledge.**

## Immediate access to row-level data through snapshots

Once inclusion and exclusion criteria are defined and the researcher has confidence in their population, a researcher needs row-level data to perform their analyses, whether a simple descriptive analysis for a rare disease or a rigorous comparative effectiveness analysis evaluating outcomes for two therapies. To obtain the row-level data needed, Studio researchers take a population snapshot. **Snapshots are immutable, complete, row-level datasets for the population of interest, containing all relevant Truveta Data.** These data are inclusive of deidentified, row-level data across all semi-structured EHR data (encounters, medication administrations/requests/dispenses, devices, laboratory results), unstructured EHR data with concepts extracted from clinical notes, as well as linked claims, SDOH, and mortality data.

Because Truveta Data is ever-growing and ever-changing, snapshots over the same population separated by any amount of time will always be different. **Snapshot immutability enables full confidence by researchers that the data behind their analyses remains constant over time, the full analyses can be readily reproduced, and the dataset will be stable for publication and/or regulatory submission.**

Each snapshot generates a population data sheet that conveys the representativeness, completeness, timeliness, and cleanliness of the underlying data.



*A datasheet built from a population snapshot, conveying representativeness, timeliness, completeness, and cleanliness of the underlying data to the researcher.*



## Analyze the data

Truveta Studio enables a powerful analytics experience for researchers to perform rigorous scientific analyses over snapshots generated for precision populations. Studio empowers analysis of these typically massive populations using pre-installed notebooks with full support for R and Python, pre-loaded with the latest medical statistics, AI, and visualization libraries (including pandas, NumPy, Matplotlib, SciPy, Tidyverse, Arrow, and dplyr). If researchers prefer, they can also extract data into their preferred environment, including SAS and Excel, and use other analytics tools like Microsoft Power BI.

## Perform powerful statistics

Truveta Studio leverages Apache Spark as its underlying analytics engine. Snapshots generated for the study population can be loaded into an integrated notebook. Spark handles big data workloads to be distributed in parallel across a cluster of nodes in a high-performing manner. This significantly speeds up iterative data exploration and analysis. Here are just a few examples of the research techniques that Truveta researchers are using inside notebooks on Truveta Studio:

- **Comparing treatment effectiveness for medications or devices of interest.** Truveta Studio allows researchers to perform a variety of analyses when comparing medications or other clinical exposures. Researchers can build balanced cohorts through a variety of patient matching algorithms (e.g., propensity score matching) to better evaluate clinical outcomes. For instance, a researcher evaluating post-market surveillance for a new anticoagulant could compare severe bleeding outcomes to a matched cohort of patients on more established anticoagulants. Building on this, the researcher could use a variety of statistical models (e.g., logistic regression, Cox regression) to evaluate covariates of interest in the outcome.
- **Investigating disease risk with linear regression.** Researchers can use linear regression, a technique for describing the relationship between a continuous response variable and one or more predictors, to estimate the association between a biomarker such as bilirubin and diabetes, BMI, leukemia, or other patient health descriptors. Other commonly used, generalized linear models can be run in a notebook including logistic regression and Poisson regression. For example, Truveta researchers can use logistic regression on Truveta Data to measure the odds of COVID-19 infection for patients with kidney disease versus those without.

- Estimating survival outcomes to understand treatment effectiveness.**  
 Truveta Studio provides a comprehensive suite of time-to-event analytic methods, including well-established techniques like Kaplan-Meier analysis and Cox Proportional Hazards models. These methods enable researchers to assess and compare survival probabilities, identify significant prognostic factors, and evaluate the impact of various treatments or interventions on survival outcomes. Researchers in Truveta Studio can derive valuable insights that contribute to evidence-based decision-making.

Templates are available to help researchers get started with these analyses.

## Publish daily updating dashboards

Population snapshots can also be scheduled to refresh daily, weekly, or monthly, with the latest data automatically streamed into dashboards using PowerBI integrated within Truveta Studio.



*A real-time dashboard showing top COVID-19 therapies breakthrough trends with daily updated Truveta Data.*

## Train AI models

Truveta Studio enables customers to develop, deploy, use, and share powerful AI models on top of Truveta Data, their own proprietary datasets, and rich public datasets such as Drug Central, PubMed, MeSH, Reactome, and DisGeNET. To build or fine-tune AI models, Truveta Studio provides a robust AI infrastructure, which includes generic toolsets including Kubeflow and PyTorch/TF, as well as specific generative frameworks like Nemo and Hugging Face Transformers to support state-of-the-art generative AI.



## Publish the study

Once the data is analyzed, it is easy for a researcher to publish their study as a PDF to submit to a peer-reviewed journal, or to share online with other users of Truveta Studio.

Researchers can also choose to submit their research for regulatory review. Truveta Studio enables this process by automatically compiling regulatory evidence, metadata, methodology documentation, quality management summarization, audit trails, and all underlying definitions and code for research. These compiled resources are made available to the researcher in an easy-to-export format that is ready for submission to regulatory bodies along with research findings.

## Conclusion

Truveta Studio enables scientifically rigorous research with powerful analytics. As we consider the scientific process, Truveta Studio contains all the tools needed to create and share scientific studies, while empowering researchers to accelerate research.

Truveta Studio provides healthcare, life sciences, and government organizations with a powerful analytics solution to support a broad range of research, including:

- **Safety monitoring:** Fulfill post-market safety requirements more efficiently with regulatory-grade data.
- **Comparative effectiveness:** Establish product differentiation and increase market adoption with timely patient-level outcomes and SDOH data.
- **Label expansion:** Identify emerging opportunities to expand product usage and improve care.
- **Clinical trials:** Optimize clinical trial protocols through dynamic testing of inclusion and exclusion criteria and accelerate trial timelines by identifying and enrolling the right patients faster.
- **Discovery:** Prioritize new targets, identify new indications, or train AI models.
- **Closing care gaps:** Collaborate with health systems to advance adoption of insights into patient care.

As with all parts of Truveta, Truveta Studio benefits from continuous improvement. Our Truveta Library is expanding each day — from new reference studies to increasingly more advanced Truveta definitions covering a diverse set of clinical subjects, including rare diseases. Contributions and feedback by our Truveta community provide ongoing validation of these definitions and studies.

To learn more about Truveta, please visit our [Truveta website](#), follow us on [LinkedIn](#), or contact us at [info@Truveta.com](mailto:info@Truveta.com).