

Utilizing Commercial Real-World Data: Truveta

Barnholtz-Sloan Research Team

Jill Barnholtz-Sloan, PhD

Kristin Waite, PhD

Gino Cioffi, MPH

Truveta - Data Overview

Providence

ADVOCATE HEALTH

Trinity Health

Tenet Health

Northwell Health

Advent Health

Baptist Health

Baylor Scott & White Health

BON SECOURS MERCY HEALTH

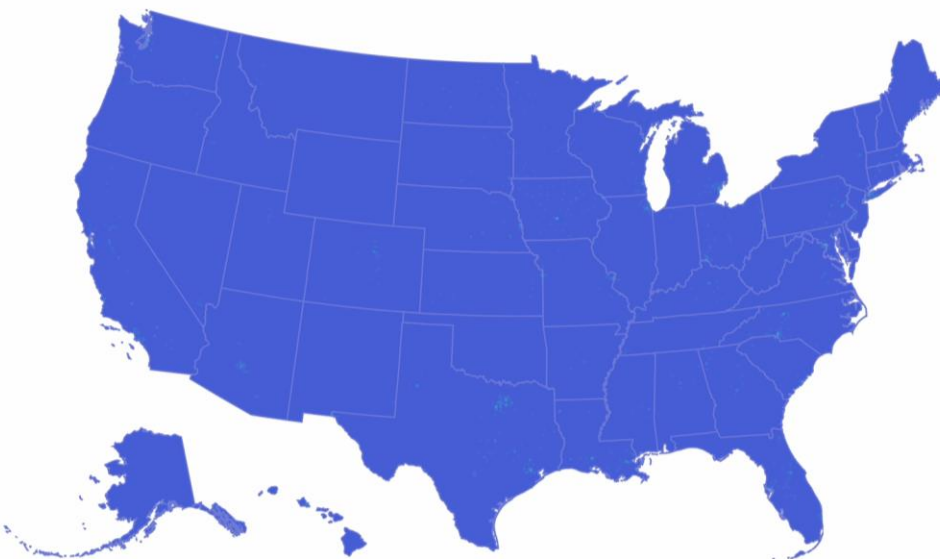
CommonSpirit

HAWAII
PACIFIC
HEALTH

HENRY
FORD
HEALTH

Inova

MEMORIAL
HERMANN



HealthPartners

HONORHEALTH

Lehigh Valley Health Network

MedStar Health

MetroHealth

NOVANT HEALTH

Ochsner Health

Premier Health

Saint Luke's

SANFORD HEALTH

Sentara Health

Texas Health Resources

TriHealth

UnityPoint Health

Virtua Health

WELLSPAN HEALTH

Truveta healthcare members

30 of the largest health
systems in the US

1 out of 3 Americans
represented in Truveta Data

Truveta Data: Complete, timely, and clean regulatory-grade EHR data



120M+

Patients and growing

Semi-structured data

- Diagnosis (SNOMED, ICD)
- Procedure (CPT, HCPCS)
- Medication (RXNORM, NDC)
- Labs (LOINC, UCUM)
- Immunizations (CVX)
- Implanted/explanted Device (UDI)
- Inpatient/outpatient care settings
- Pharmacy
- Mother-child link

Unstructured data

- Clinical notes
- Images
- Genomics

Tokenization and integration with third-party datasets

5B+

Clinical notes

85M+

Imaging studies

1M+

Mother-child pairs

5+

Years longitudinal patient history

200M

Closed Claims Patients

188K+

Unique devices

This data is additionally linked with SDOH, mortality, and claims data

Data from 5B+ clinical notes available for research

Heart Failure - NYHA

Patient Active Problem List Diagnosis • Appendicitis, acute • DM (diabetes mellitus) • Facial cellulitis • Hyponatremia • Essential hypertension • Dilated cardiomyopathy • Pulmonary embolism without acute cor pulmonale • Acute deep vein thrombosis (DVT) of popliteal vein • LBBB (left bundle branch block) • Chronic combined systolic and diastolic CHF, NYHA class 2 and ACA/AHA stage C • Lightheadedness • Suspected sleep apnea • Congestive heart failure • Snoring

History of Present Illness: Information is gathered from patient ... with a past medical history of hypertension, hyperlipidemia, cardiomyopathy, history of left lower extremity DVT, diabetes, history of left bundle branch block, pulmonary embolism, combined systolic and diastolic heart failure with EF of 17% at diagnosis

ECHO Complete Result Date:

1. Mildly dilated left ventricle with normal thickness, LVEDD 5.9 cm global hypokinesis with mildly reduced calculated LVEF 42% which on visual comparison is similar to the images
2. The RV is not seen well but the size appears normal with normal lateral annular motion and TAPSE suggestive of normal RV SF
3. Normal biatrial sizes. Intracardiac lead seen intermittently in the right atrium
4. The mitral and aortic valves are seen adequately without echodensities suggestive of vegetation. The tricuspid valve is not seen completely but there are no large echodensities visualized. There is trace tricuspid and mitral valve regurgitations
5. The aortic root and feels normal
6. There is a trace pericardial effusion
7. Normal CVP, normal left-sided filling pressure
8. Sinus 70s ECHO Transesophageal (TEE)

Patient Id	Note Date	Clinical Measure	Value
PT1234	2023-01-22	NYHA Class	2
....
....

Patient Id	Note Date	Clinical Measure	Value
PT1234	2023-01-22	LVEF	42%
....
....

Examples of concepts extracted from notes:

- Ejection fraction
- Seizure frequency
- Cancer stage
- Heart failure stage and symptoms
- Cardiac catheterization measures
- Metabolic disease progression

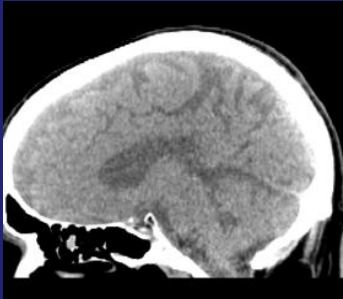
More than 1M matched mother-child pairs



Longitudinal EHR data for more than 1 million mother-child pairs, enabling research from pre-pregnancy through the first 5 years of the child's life

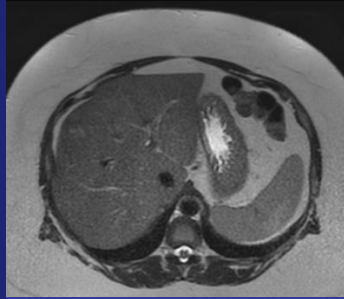
Images

Largest collection of medical images integrated with EHR data



16.5M

Computed
tomography



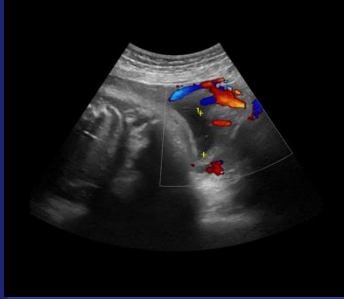
5.2M

Magnetic
resonance
imaging



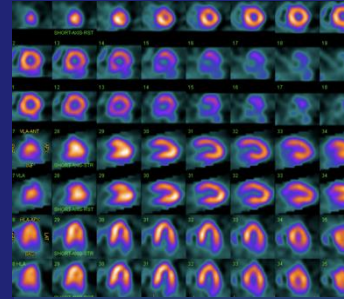
38.7M

Digital
x-ray



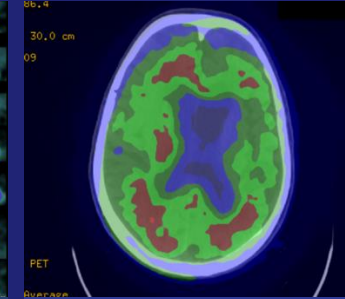
13.9M

Ultrasound



500K

Nuclear
medicine



228K

PET
imaging



10.1M

Mammography



More than 85 million imaging studies searchable by modality and protocol

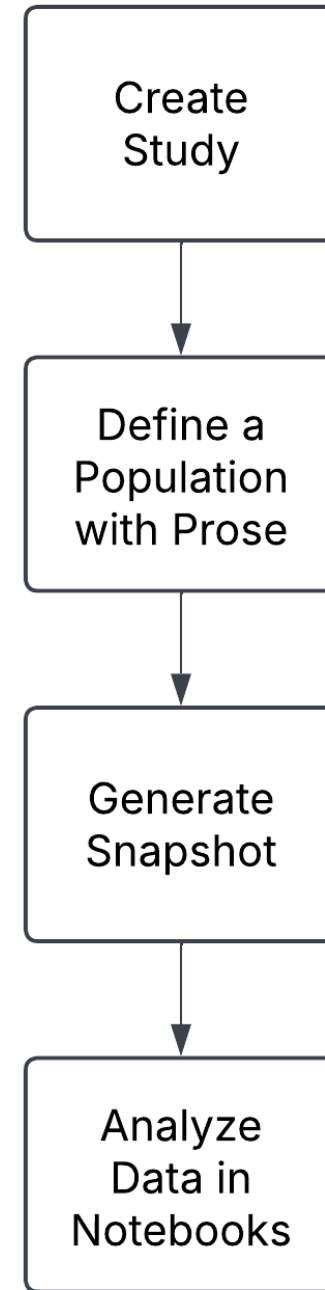
Types of Studies Reported

- Treatment analysis
- Mother - Child studies (pregnancy to 5 yr old)
- Adherence studies
- COVID analysis
- Prescription trends and monitoring
- Testing/screening trends
- Injury analysis
- Office, ER visit analysis
- Monitoring reports

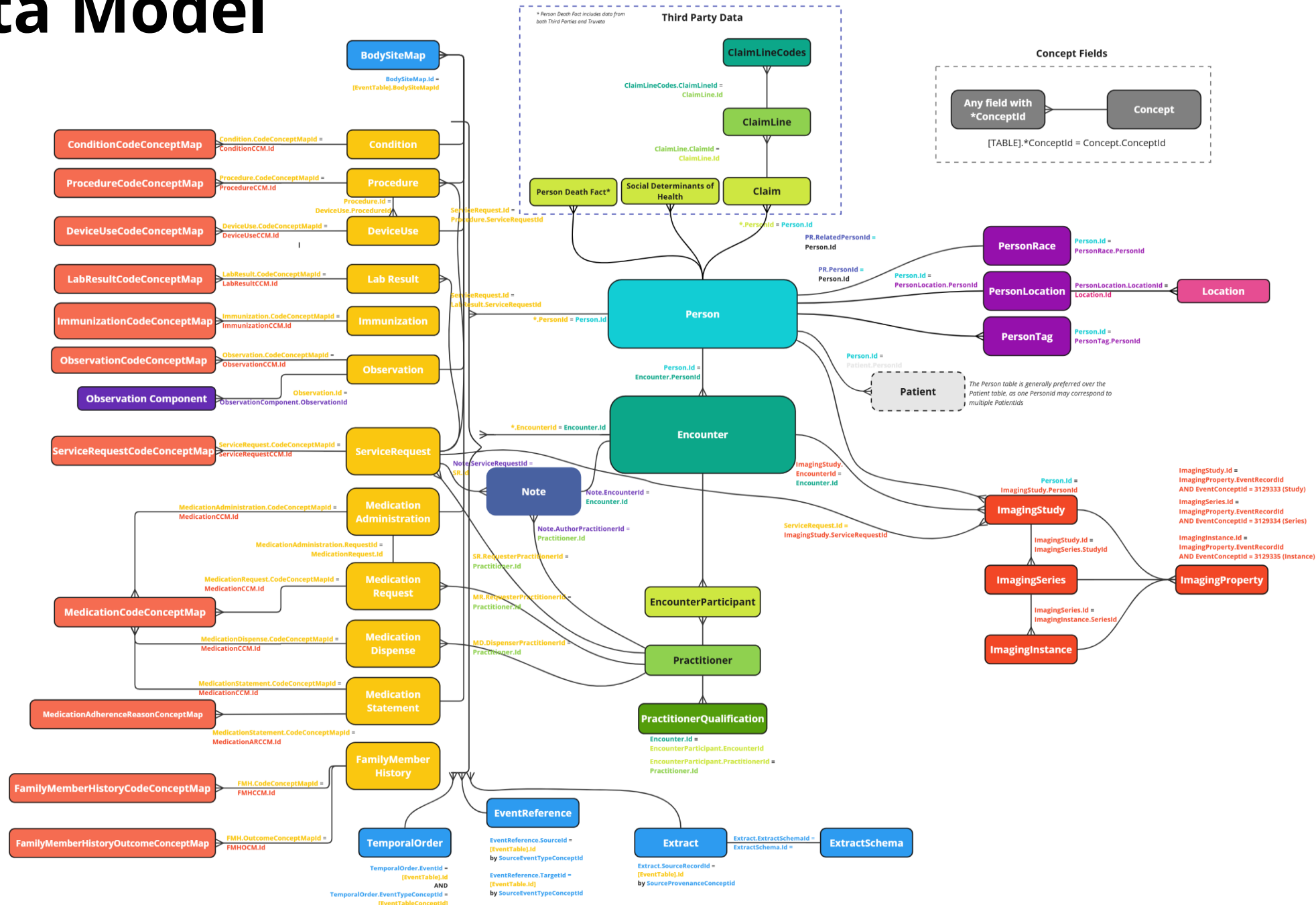
Truveta - User Overview

Truveta Studio

- Web interface for engaging with data
 - All cloud-based, accessed through browser
- “Studies” contain data definitions, snapshots, and notebooks for a given project



Truveta Data Model



Creating Data Definitions

- Prose
 - Truveta query language
- Event and Codesets
 - Provides broad overview of conditions, procedures, or treatments
- Code identification and mapping tools
 - Truveta Studio Concept Search tool
 - [ATHENA](#)

The screenshot displays the Truveta Studio interface for creating data definitions. The sidebar on the left lists various studies, with 'GBM SNOMED_ICDO...' selected. The main editor area shows a code definition for 'GBM' with a 'Definition estimate' of 16K. The code defines a research base and filters for glioblastoma events. The 'Demographics' panel on the right shows the distribution of sex and race for the defined population.

GBM Definition estimate 16K

```
1 defResearchBase = include "https://library.
2   truveta.com/o/truveta/d/research-base"
3
4 Edit Codeset
5 codes = codeset(
6   "SNOMEDCT",
7   selfAndDescendants,
8   "35262004"Gliosarcoma,
9   "393563007"Glioblastoma multiforme,
10  "44529004"Giant cell glioblastoma,
11  "63634009"Glioblastoma
12 )
13 events = filter(defResearchBase.
14   ConditionSet c) {
15   c.CodeConceptId matches { Id in codes }
16 }
17 export events
```

Demographics

Sex

Male	55.9%
Female	38.8%
No Information	5.3%

Race

White	73%
Unknown	20%
Black or African American	3.2%
Other Race	2.2%
Asian	1.3%
American Indian or Alaska Native	0.2%

Defining a Study Population

- Use Prose Definitions to define population
- Generate "Snapshot"
 - Point-in-time collection of records
 - To be analyzed in notebook environment
- Select Tables to be included
 - Remove unnecessary tables to reduce cloud costs

The screenshot displays the TRUVETA web application interface. The top navigation bar includes the TRUVETA logo, a search bar, and a user profile icon. The main content area is titled "Population" and shows a "Population estimate" of 15K. The left sidebar lists various studies and populations, with "GBM Sex Differences nc1" selected. The central pane displays a Prose definition for the population, which includes imports for various tables and a definition for the index event. The right sidebar shows demographic data, including a bar chart for Age and horizontal bar charts for Sex and Race.

Population

Population estimate 15K

```
1 // #region imports
2 defResearchBase = include "https://library.truveta.com/o/truveta/d/
  research-base"
3 defGbmSnomedGcTest = import "/definitions/gbm-snomed-gc-test"
4 defGbmSnomedIcdo3GcTest = import "/definitions/gbm-snomed-icdo-3-gc-test"
5 defGbmSurgicalResection = import "/definitions/gbm-surgical-resection"
6 defGbmTemozolomide = import "/definitions/temozolomide-gbm-total-count"
7 defGbmRadiationTherapy = import "/definitions/gbm-radiation-therapy"
8 defGeneralBrainTumorDiagnosisCodes = import "/definitions/
  general-brain-tumor-diagnosis-codes"
9 // #endregion
10
11 // define index event
```

Demographics

Age

Age Group	Percentage
5	0%
15	0%
25	0%
35	0%
45	0%
55	0%
65	0%
75	0%

Sex

Sex	Percentage
Male	55.9%
Female	39.2%
No Information	5%

Race

Race	Percentage
White	73%

Good news!
No problems detected in your Prose

Accessing Data

- Spark-based integrated notebook environment
 - PySpark (Python)
 - SparkR (R)
- Data Wrangling
 - SQL
 - Koalas (Python)
 - Sparklyr (R)

Study report | Mortality_Final

Attach to: sp34cpuXme41 | Language: SparkR (R) | Clear outputs | Variable explorer | Outline | Save | X | Play | Download | Checkmark | Refresh

● Not connected | Selected Cell 19 of 101 cells

Creating tables

1. Create a dataframe called "condition_encounter" selecting patients with absolute year for 1st diagnosis of GBM between 2016-2023

```
1 sql<-
2 "select Condition.EncounterId, Condition.PersonId, SearchResult_first_dx.dt, Encounter.StartDateTime, Encounter.TypeConceptId, Encounter.Cl
3 from SearchResult_first_dx
4 INNER JOIN Condition on Condition.PersonId = SearchResult_first_dx.PersonId
5 INNER JOIN Encounter on Encounter.Id = Condition.EncounterId
6 INNER JOIN ConditionCodeConceptMap on Condition.CodeConceptMapId = ConditionCodeConceptMap.Id
7 INNER JOIN Concept on Concept.ConceptId = ConditionCodeConceptMap.CodeConceptId
8 WHERE ConceptCode IN
9 ('393563007',
10 '63634009',
11 '276828006',
12 '1163375002',
13 '684931000119100',
14 '684911000119105') AND absolute_AdministeredDateTime in (2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023)
15
16 condition_encounter <- load_sql_table(con, snapshot, sql, view_name='tbl_condition_encounter', cache='TRUE', output_mode='sparklyr')
17 display_df(condition_encounter, 5)
18
```

[11] ✓ - Command executed in 53 sec 475 ms on 6:52:55 PM, 7/25/24 | SparkR (R)

Created view: `tbl_condition_encounter`
cached rows:452789

Index	EncounterId	PersonId	dt	StartDateTime	TypeConceptId	ClassCon
1	16859e59-cd37-c1f1-e6c6-a8b04ab...	b93ad37f-4308-5d8a-36f6-fce4d8cf...	2015-02-21 10:42:00	2015-05-16 11:23:00	3059272	1065217
2	16859e59-cd37-c1f1-e6c6-a8b04ab...	b93ad37f-4308-5d8a-36f6-fce4d8cf...	2015-02-21 10:42:00	2015-05-16 11:23:00	3059272	1065217
3	5dace848-0a32-4eab-87c8-d58fb36...	68bc8423-a7fb-ef12-cab3-92edbc...	2020-12-25 00:00:00	2021-01-12 00:00:00	3059269	2649591
4	5dace848-0a32-4eab-87c8-d58fb36...	68bc8423-a7fb-ef12-cab3-92edbc...	2020-12-25 00:00:00	2021-01-12 00:00:00	3059269	2649591
5	7d567d8c-fa...	2016-10-29 15:14:16	2649591	2649591

PySpark



Koalas

APACHE Spark R



Analyzing Data & Exporting Results

- Analytical dataset can be analyzed using regular R/Python syntax
- Exporting Results
 - Can export figures (png, jpeg, etc.) and tables (html)
 - Can **not** export data files

2_analyze_GC_CU...PopulationSnapshots - Popu...

Attach to
sp34cpuXme41

Language
SparkR (R)

Save

Not connected

Selected Cell 17 of 49 cells

Surgery

M4

```
1 df3 %>% select(Sex, AgeGroup, Race,
2   CollegePastPresent,
3   ProspectEstimatedIncomeRange,
4   AddressChangeCountLast12Months, CurrAddrLenOfRes,
5   HouseholdMembersCount,
6   DistanceToCloseTies,
7   HouseholdMotorizedPropertyRegistrationsCount,
8   HealthcareCostRiskCategory, MedicationNonAdherenceRiskCategory, HealthManagementNonMotivationRiskCategory,
9   Surg_Delay) %>% filter(!is.na(Surg_Delay)) %>% #na.omit %>%
10  tbl_summary(by = Surg_Delay,
11             label = list(
12               AgeGroup ~ "Age Group",
13               CollegePastPresent ~ "Attended College",
14               ProspectEstimatedIncomeRange ~ "Estimated Yearly Income",
15
16               AddressChangeCountLast12Months ~ "Address Change in Last 12 Months",
17
18               HouseholdMembersCount ~ "Only Member of Household",
19               DistanceToCloseTies ~ "Distance to Close Ties",
20               HouseholdMotorizedPropertyRegistrationsCount ~ "Motorized Vehicle Registered to Current Address",
21               HealthcareCostRiskCategory ~ "Healthcare Cost Risk Category",
22               MedicationNonAdherenceRiskCategory ~ "Medication Non-Adherence Risk Category",
23               HealthManagementNonMotivationRiskCategory ~ "Health Management Non-Motivation Risk Category"
24             ),
25             type = all_dichotomous() ~ "categorical",
26             missing = "no") %>% add_p() %>% bold_p() %>%
27  add_n(statistic = "{n_miss} ({p_miss}%)" %>%
28  modify_header(n = "**Missing**") %>%
29  bold_labels() %>% as_gt() %>%
30  gt::tab_header(title = gt::html("Table 1. Descriptive statistics of demographic and SDOH factors,<br>
31  stratified by surgical resection delay status (Truveta, 2018-2024)",
32  subtitle = "Delay defined as surgical resection occurring 10 days post GBM diagnosis") %>%
33  gt::as_raw_html() %>% as.character() %>% displayHTML())
```

[13]

✓ - Command executed in 9 sec 607 ms on 11:16:37 AM, 10/18/24

SparkR (R)

Table 1. Descriptive statistics of demographic and SDOH factors, stratified by surgical resection delay status (Truveta, 2018-2024)

Delay defined as surgical resection occurring 10 days post GBM diagnosis

Characteristic	Missing	Delay, N = 789 ¹	No Delay, N = 2,731 ¹	p-value ²
Sex	22 (0.6%)			>0.9
Female		318 (41%)	1,101 (41%)	
Male		465 (59%)	1,614 (59%)	
Age Group	0 (0%)			<0.001
18-44		135 (17%)	221 (8.1%)	
45-64		357 (45%)	1,090 (40%)	
65+		297 (38%)	1,420 (52%)	
Race	475 (13%)			0.3

Future Applications

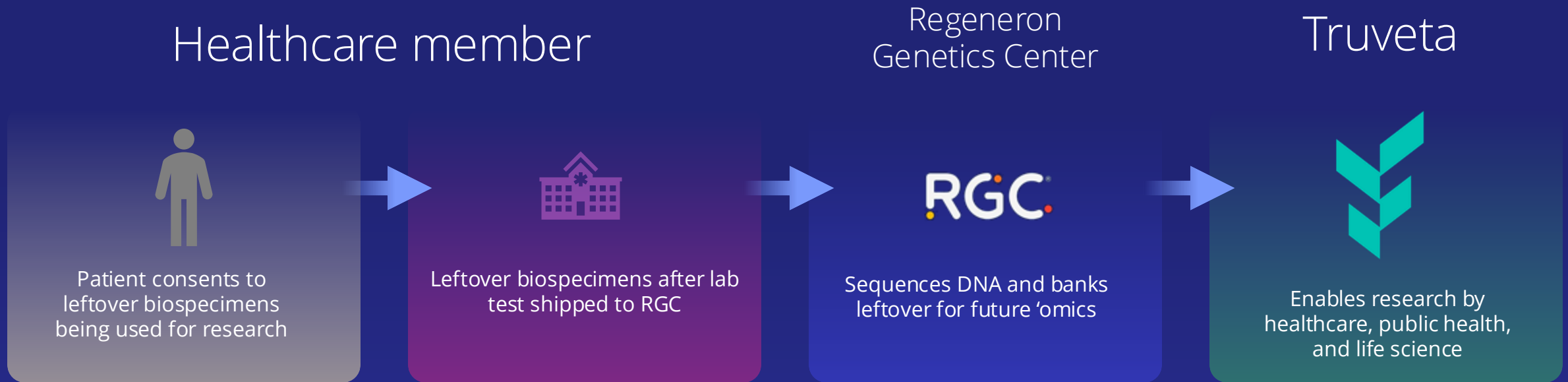


Truveta Genome Project

Creating the largest and most diverse database of genotypic and phenotypic information ever assembled

- Delivering genetic data on 10s of millions, over 10x the scale of previous endeavors
- Creating the world's largest genetic biobank to support future multi-omics
- Enabling drug discovery and optimized clinical trials
- Transforming how diseases are prevented, diagnosed, and cured

Truveta Genome Project: Learning the biology of disease



Questions

- For feasibility questions, please reach out to Kasie Bailey at Truveta
 - Email: kasieb@truveta.com