INTEGRATING COMMERCIAL HEALTHCARE DATASETS FOR AEROMEDICAL RISK ANALYSES

Sponsor: Federal Aviation Administration
Dept. No.: P233
P233 B23 01
Contract No.: 693KA8 22 C 00001
Project No.: 100976.10.102.1011.CD4
Outcome No: 4 5.B.1 1
PBWP Reference: AAM Strategic Roadmap 2022 2030
DOI: https://doi.org/10.21949/158556

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

This document was prepared for authorized distribution only. It has not been approved for public release.

McLean, VA

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-23/15	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Integrating Commercial Healthcare Datasets for Aeromedical Risk Analyses	5. Report Date May 31, 2023 6. Performing Organization Code
7. Author(s) I. Lisann[1]; J. O'Connor[2]; S. Roessner[3] All authors from The MITRE Corporation	8. Performing Organization Report No. Product 4-5.B.1-3
9. Performing Organization Name and Address The MITRE Corporation 7515 Colshire Drive McLean, VA 22102	10. Work Unit No. (TRAIS) 11. Contract or Grant No. 693KA8-22-C-00001
12. Sponsoring Agency Name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code

15. Supplementary Notes

Author ORCIDs:
[1] Lisann (0000-0003-1836-1704);
[2] O'Connor (0000-0001-7337-8141)
[3] Roessner (0000-0001-7072-335X)
Technical report DOI: https://doi.org/10.21949/158556

16. Abstract

The Federal Aviation Administration (FAA) Office of Aerospace Medicine requires comprehensive longitudinal healthcare datasets to augment internal data for the purpose of conducting safety risk assessments to update medical standards (i.e. data dirven, risk based decision making). The Federal Aviation Administration (FAA) tasked The MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD), in its Innovation Partner role, to identify commercial healthcare datasets that hold potential value in forecasting medical risk and are suitable for integration into the Aeromedical Data Environment. MITRE CAASD performed a market survey of existing healthcare datasets available commercially or for public use. This market survey led to the identification of over 40 healthcare data sources, many of which contain numerous subordinate sets. An initial set of screening criteria ensured that candidate data sources were sufficiently suitable for modeling objectives; this screening reduced the set to three final candidate data sources. These three data sources were compared using a set of features relevant to risk modeling of aeromedically relevant outcomes by condition. This set of comparison features included their coverage of medical conditions of interest to the FAA, as well as factors impacting integration into the aeromedical data environment.

17. Key Word Aviation, Certification, Medicine, Pilot, Safety	18. Distribution Statement Unlimited

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

.

EXECUTIVE SUMMARY

Background

FAA's Office of Aerospace Medicine conducts Safety Risk Management (SRM) through analysis of pilot medical data. However, medical information from the pilot population is limited in detail, may be biased due to underreporting, and is censored due to indeterminate dropouts. These issues limit analysts' ability to compute likelihoods of aeromedically significant events by medical condition – an important element of risk analysis. Commercially available healthcare datasets may allow analysts to overcome these limitations and better support SRM.

Method

MITRE applied a set of screening criteria to limit the set of potential data sources to those that could support quantitative modeling. A subsequent set of evaluation criteria were used to compare the viable data sources.

Screening Criteria

A market survey revealed more than 40 commercially available healthcare data sources. Sreening criteria (Figure 1) yielded three potential data solutions: Truveta Studio, IQVIA's Ambulatory EMR database linked with their Longitudinal Prescription database and their Medical Claims Database, or Merative's Claims-EMR dataset.

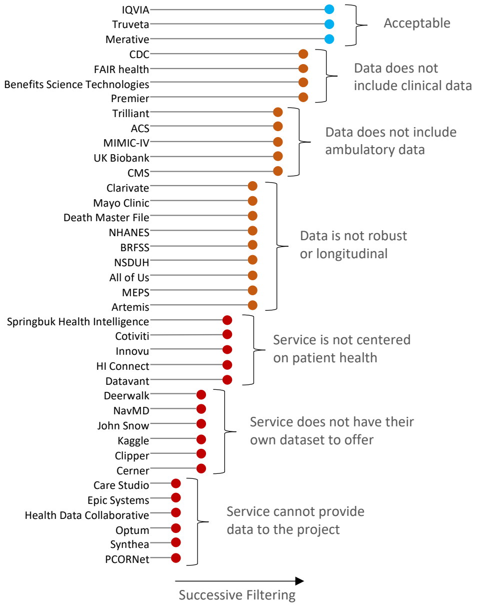
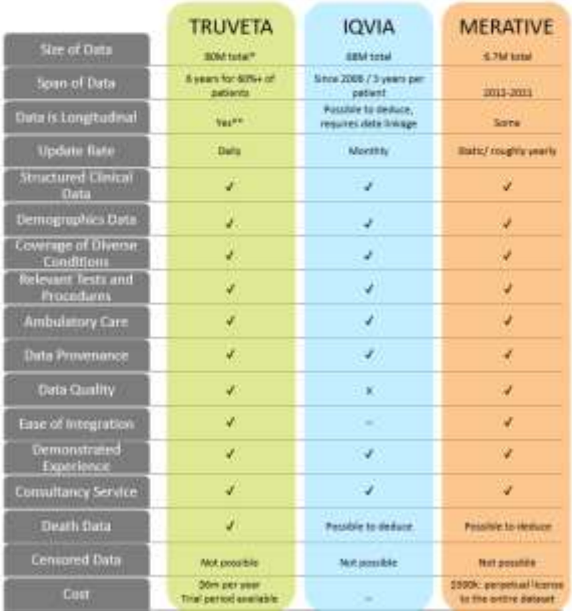

IQVIA
Truveta
Merative
Acceptable
CDC
FAIR health
Benefits Science Technologies
Premier
Data does not include clinical data
Trilliant
ACS
MIMIC-IV
UK Biobank
CMS
Data does not include ambulatory data
Clarivate
Mayo Clinic
Death Master File
NHANES
BRFSS
NSDUH
All of Us
MEPS
Artemis
Data is not robust or longitudinal
Springbuk Health Intelligence
Cotiviti
Innovu
HI Connect
Datavant
Service is not centered on patient health
Deerwalk
NavMD
John Snow
Kaggle
Clipper
Cerner
Service does not have their own dataset to offer
Care Studio
Epic Systems
Health Data Collaborative
Optum
Synthea
PCORNet
Service cannot provide data to the project
Successive Filtering

Figure 1. Market Survey Screening Criteria



	TRUVETA	IQVIA	MERATIVE
Size of Data	80M total*	68M total	6.7M total
Span of Data	8 years for 60%+ of patients	Since 2006 / 3 years per patient	2013-2021
Data is Longitudinal	Yes**	Possible to deduce, requires data linkage	Some
Update Rate	Daily	Monthly	Static/ roughly yearly
Structured Clinical Data	✓	✓	✓
Demographics Data	✓	✓	✓
Coverage of Diverse Conditions	✓	✓	✓
Relevant Tests and Procedures	✓	✓	✓
Ambulatory Care	✓	✓	✓
Data Provenance	✓	✓	✓
Data Quality	✓	x	✓
Ease of Integration	✓	–	✓
Demonstrated Experience	✓	✓	✓
Consultancy Service	✓	✓	✓
Death Data	✓	Possible to deduce	Possible to deduce
Censored Data	Not possible	Not possible	Not possible
Cost	$6m per year Trial period available	–	$300k: perpetual license to the entire dataset

Figure 2. Feature comparison table for three data solutions that satisfy key criteria.

Evaluation Criteria.

MITRE identified eight medical conditions and their relevant tests and procedural information believed to be most conducive to medical risk forecasting. These conditions were selected as they were found to have a reasonable likelihood of acutely precluding pilots from flight duties and became part of our feature chart (Figure 2) to compare datasets. The remaining criteria used to compare the data were developed in collaboration with subject matter experts in data science and artificial intelligence to ensure that characteristics were relevant to data modeling and risk forecasting.

Conclusions

Truveta is the most feature rich and is expected to hold greater than 130 million patient records by the end of 2023. Because Truveta encompass numerous health systems, patient histories are likely to be continuous even if they cross platforms.

Access to Truveta Studio for three years and the ability to export seven medical conditions into a separate cloud environment is roughly $6M per year (subject to change); in comparison, Merative's CED costs $300 thousand. At the time of this study, pricing information for IQVIA was not available.

Leveraging commercially available healthcare datasets could allow the FAA's Office of Aerospace Medicine to expand its data-driven, risk-based, systems approach to Safety Risk Management.

PROBLEM STATEMENT

FAA's Office of Aviation Safety uses a Safety Management System (SMS) as directed by Order 8000.369. In accordance with Order 8040.4, the Office of Aerospace Medicine conducts Safety Risk Management (SRM) through analysis of pilot medical data and promulgates safety controls through updates to OneGuide. However, these data are limited in detail, may be biased due to underreporting, and are censored (truncated) due to indeterminate dropouts. These issues limit risk analysts' ability to compute likelihoods of aeromedically significant events by medical condition.

PROPOSED SOLUTION

Commercially available healthcare datasets may allow analysts to overcome these limitations. Leveraging such datasets could allow the Office of Aerospace Medicine to expand its data-driven, risk-based, systems approach to Safety Risk Management. This task supports the "analyze safety risks" step of SRM and positions the FAA to more effectively respond to system changes and manage regulatory risk through OneGuide.

METHOD

MITRE identified an initial set of datasets and data services with the help of Subject Matter Experts (SMEs) familiar with healthcare datasets and the types of attributes they provide. Additional research from rating organizations was used those to identify others. Independent research was conducted separately by custom research experts to provide another source of datasets. We evaluated those datasets against a common set of characteristics, allowing us to compare them and determine their potential value to the Office of Aerospace Medicine. This set of characteristics included their coverage of medical conditions of interest to the FAA, as well as the relevant tests and procedural information used to monitor them. We also considered factors impacting the ease of integration into the aeromedical data environment.

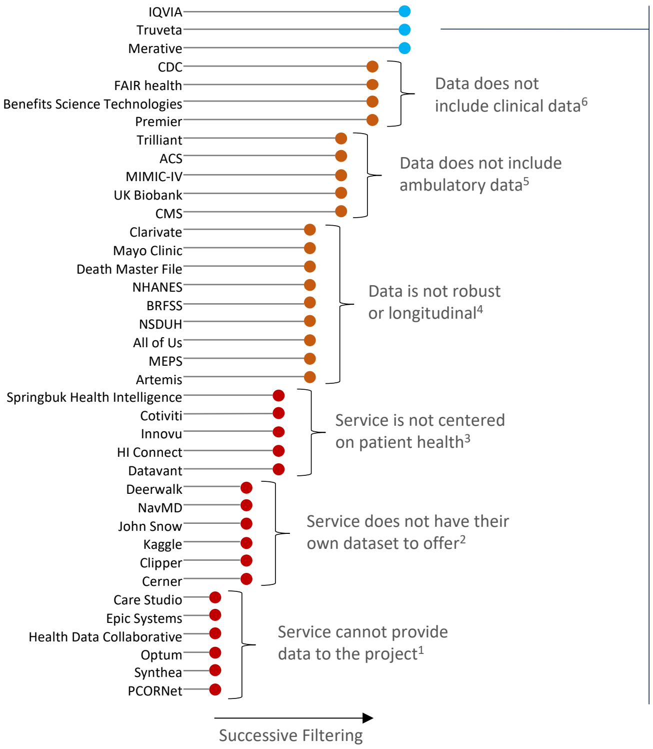
IDENTIFICATION AND EVALUATION OF DATA SOURCES

Figure 3 is a graphical representation of the data sources that the MITRE team considered. Data sources are grouped and categorized by their evaluation and are sorted, by group, in terms of MITRE's evaluation of their viability for the project.

MITRE identified an initial set of datasets with the help of SMEs familiar with healthcare datasets and the types of attributes they provide, rating organizations, and custom research experts. The data requirements in this screening were developed in collaboration with the FAA and SMEs in medicine and health informatics.

Though data sources listed closer to the top satisfy more requirements than do those towards the bottom, we believe that only the top three data sources listed would ultimately meet the FAA's needs. For the other data sources, the most significant requirements shortfalls were severe enough to discount the data source. Data sources were sometimes unable to meet the data requirements for multiple reasons. In those instances, we placed them in the most applicable grouping.

Many of these data sources contain numerous datasets. The graph is purposefully organized by the overarching data source rather than the individual dataset to allow for greater readability.



Truveta offers a platform called Truveta Studio that integrates and normalizes data from over 30 different health system members into a cloud environment. We recommend a subscription to Truveta Studio for access to all their data.

IQVIA has multiple data sets that can be linked together. We recommend linking their Ambulatory EMR (AEMR) database, Longitudinal Prescription (LRx) database, and Medical Claims (Dx) database to provide a more comprehensive image of patient health.

Merative has multiple datasets that meet specific needs. Though their data cannot be linked together as easily as IQVIA's, they offer their Claims-EMR Dataset (CED) which links some of their data from their Explorys EMR set with some of their MarketScan administrative claims data. We recommend purchasing their CED. Having the Explorys and MarketScan datasets on their own does not allow them to be linked.

Figure 3. Market Survey Data Source Filtering Criteria

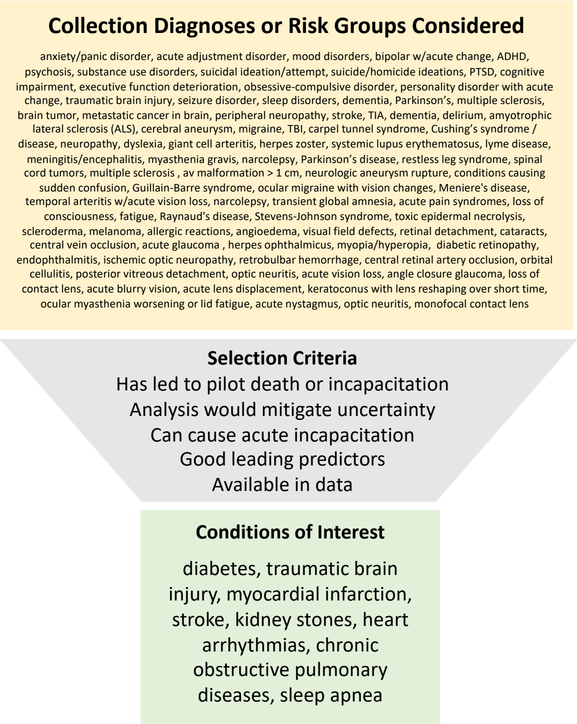
[1] These data sources do not possess any data that the FAA can use, due to legal reasons or still being new.
[2] These data sources do not have their own data sets. They either have services to help analyze healthcare data or they act as a connector or consolidator between multiple data sets.
[3] Most of the data sources have reducing healthcare costs or improving business as their mission statement, not improving patient outcomes.
[4] The data provided by these sources are either missing too many fields and/or medical conditions, or most of their data only captures a snapshot in time and is not longitudinal.
[5] This data is focused on claims information.
[6] Though the data has clinical data, it does not include robust outpatient (ambulatory) data. This data is necessary as it includes patient visits with vital signs and diagnostic tests like lab results. Having only hospital data would skew the population to one that is already sick.

EVALUATION DETAILS

Medical Condition Evaluation

The MITRE team considered relevant medical conditions while evaluating commercially available data. This ensured that analysis of the data leads to impactful insight in medical risk forecasting.

We identified eight medical conditions that we believed would be most conducive to medical risk forecasting. These conditions were taken from a list of over one hundred medical conditions from an aeromedical risk characterization workshop in Februrary 2022 moderated by the Aerospace Medical Research Division (AAM-600). These conditions were found to have a reasonable likelihood of acutely precluding pilots from flight duties.



Collection Diagnoses or Risk Groups Considered

anxiety/panic disorder, acute adjustment disorder, mood disorders, bipolar w/acute change, ADHD, psychosis, substance use disorders, suicidal ideation/attempt, suicide/homicide ideations, PTSD, cognitive impairment, executive function deterioration, obsessive-compulsive disorder, personality disorder with acute change, traumatic brain injury, seizure disorder, sleep disorders, dementia, Parkinson's, multiple sclerosis, brain tumor, metastatic cancer in brain, peripheral neuropathy, stroke, TIA, dementia, delirium, amyotrophic lateral sclerosis (ALS), cerebral aneurysm, migraine, TBI, carpel tunnel syndrome, Cushing's syndrome / disease, neuropathy, dyslexia, giant cell arteritis, herpes zoster, systemic lupus erythematosus, lyme disease, meningitis/encephalitis, myasthenia gravis, narcolepsy, Parkinson's disease, restless leg syndrome, spinal cord tumors, multiple sclerosis , av malformation > 1 cm, neurologic aneurysm rupture, conditions causing sudden confusion, Guillain-Barre syndrome, ocular migraine with vision changes, Meniere's disease, temporal arteritis w/acute vision loss, narcolepsy, transient global amnesia, acute pain syndromes, loss of consciousness, fatigue, Raynaud's disease, Stevens-Johnson syndrome, toxic epidermal necrolysis, scleroderma, melanoma, allergic reactions, angioedema, visual field defects, retinal detachment, cataracts, central vein occlusion, acute glaucoma , herpes ophthalmicus, myopia/hyperopia, diabetic retinopathy, endophthalmitis, ischemic optic neuropathy, retrobulbar hemorrhage, central retinal artery occlusion, orbital cellulitis, posterior vitreous detachment, optic neuritis, acute vision loss, angle closure glaucoma, loss of contact lens, acute blurry vision, acute lens displacement, keratoconus with lens reshaping over short time, ocular myasthenia worsening or lid fatigue, acute nystagmus, optic neuritis, monofocal contact lens

Selection Criteria

Has led to pilot death or incapacitation
Analysis would mitigate uncertainty
Can cause acute incapacitation
Good leading predictors
Available in data

Conditions of Interest

diabetes, traumatic brain injury, myocardial infarction, stroke, kidney stones, heart arrhythmias, chronic obstructive pulmonary diseases, sleep apnea

Figure 4. Criteria applied when considering medical conditions of interest.

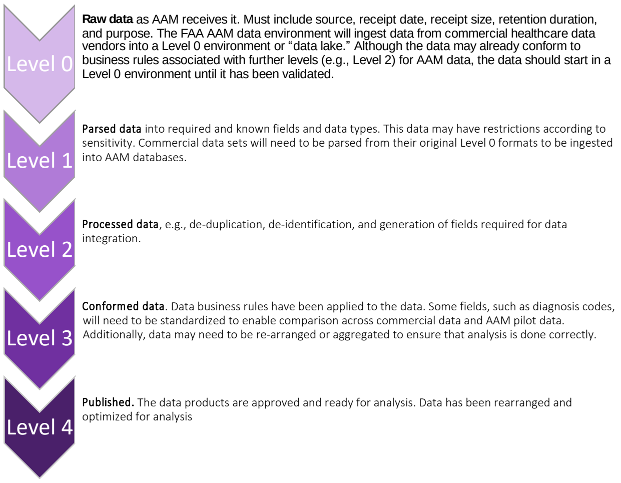
Figure 4 displays the criteria applied to narrow this long list down to eight. Special consideration was given to conditions that can cause acute incapacitation. The following eight conditions were determined as meeting the criteria after conducting literature reviews and consulting with physicians and electronic health records SMEs. These conditions can be used to generate cohorts from the acquired commercial datasets to compute likelihoods of aeromedically significant events by medical condition when conducting risk analysis during SRM.

For detailed information on the eight identified conditions and their leading clinical predictors for artifical intelligene /machine learning use, consult "Medical Conditions Analysis for Electronic Healthcare Records" (https://doi.org/10.21949/1528558). These excel sheets are organized by medical condition and contain the relevant ICD-9, ICD-10, SNOMED, and LOINC codes for each leading predictor, as well as the normal ranges the data should fall within.

We were also interested in the role of pilots' mental health conditions in performance of duties, and the leading indicators for such conditions. However, identifying reliable predictors in healthcare datasets continues to be a challenge. We recommend using these data to study the efficacy of medications used to treat mental health conditions.

Integrating Commercial Data

As a data environment is developed, the commercial data will need to be integrated into the system so that it can be used to perform analyses. These levels provide a systematic approach to organize and cleanse the dataset chosen by the FAA.



Raw data as AAM receives it. Must include source, receipt date, receipt size, retention duration, and purpose. The FAA AAM data environment will ingest data from commercial healthcare data vendors into a Level 0 environment or "data lake." Although the data may already conform to business rules associated with further levels (e.g., Level 2) for AAM data, the data should start in a Level 0 environment until it has been validated.

Parsed data into required and known fields and data types. This data may have restrictions according to sensitivity. Commercial data sets will need to be parsed from their original Level 0 formats to be ingested into AAM databases.

Processed data, e.g., de-duplication, de-identification, and generation of fields required for data integration.

Conformed data. Data business rules have been applied to the data. Some fields, such as diagnosis codes, will need to be standardized to enable comparison across commercial data and AAM pilot data. Additionally, data may need to be re-arranged or aggregated to ensure that analysis is done correctly.

Published. The data products are approved and ready for analysis. Data has been rearranged and optimized for analysis

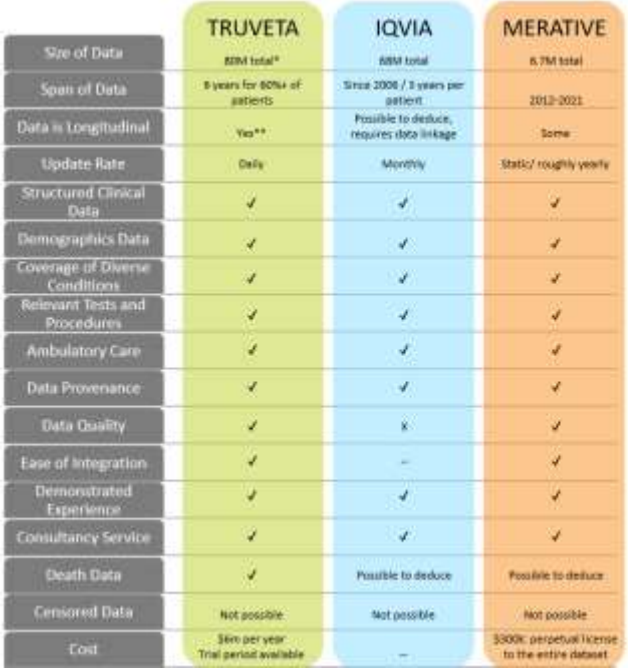
Figure 5. Levels of data integration in the context of the FAA integrating commercial healthcare data

The aeromedical data environment is still undergoing development, and we were unable to obtain a free trial to experiment with integration. However, these levels provide a systematic approach to doing so.

CONCLUSIONS

Evaluation Criteria

The criteria we used to compare the data in Figure 2 were developed in collaboration with SMEs in data science and artificial intelligence to ensure that we were considering characteristics relevant to data modeling and risk forecasting. Three vendors offer data sets that satisfy these criteria.



	TRUVETA	IQVIA	MERATIVE
Size of Data	80M total*	68M total	6.7M total
Span of Data	8 years for 80%+ of patients	Since 2006 / 3 years per patient	2012-2021
Data is Longitudinal	Yes**	Possible to deduce, requires data linkage	Some
Update Rate	Daily	Monthly	Static/ roughly yearly
Structured Clinical Data	✓	✓	✓
Demographics Data	✓	✓	✓
Coverage of Diverse Conditions	✓	✓	✓
Relevant Tests and Procedures	✓	✓	✓
Ambulatory Care	✓	✓	✓
Data Provenance	✓	✓	✓
Data Quality	✓	x	✓
Ease of Integration	✓	–	✓
Demonstrated Experience	✓	✓	✓
Consultancy Service	✓	✓	✓
Death Data	✓	Possible to deduce	Possible to deduce
Censored Data	Not possible	Not possible	Not possible
Cost	$6m per year Trial period available	–	$300K: perpetual license to the entire dataset

Figure 6. Feature comparison table for three data solutions that satisfy key criteria.

Data Recommendations

MITRE recommends the use of three data sources, in order of preference:

1. Truveta
2. IQVIA
3. Merative

Truveta's platform is expected to have greater than 130 million patient records by the end of 2023 and to continue growing beyond then. Because Truveta encompass so many health systems, even if patients change health providers, they are likely to still be captured in their data platform.

IQVIA does not offer the same breadth in ambulatory data as Truveta, but it does provide the option to link its ambulatory dataset to other datasets to provide greater longitudinal information. For example, its prescription database captures most of the American population. Therefore, even if a patient leaves the ambulatory dataset, that patient information can be linked to other datasets to track that patient's journey.

Merative's CED has fewer than seven million patient records. That number drops to fewer than two million patients that are in the dataset for at least two years, and fewer than half a million with data over five years. This limitation reduces modeling accuracy and makes it more likely that patients will leave the dataset if they switch healthcare providers.

Truveta is also the only vendor we interviewed that has mortality information. In other data sources, patient death must be inferred.

Truveta's most significant constraint is its cost. Access to their studio for three years and the ability to export seven medical conditions into a separate cloud environment is slightly over six million dollars per year, subject to negotiation.[1] By contrast, Merative's CED would cost three hundred thousand dollars. At this time, IQVIA was unable to provide pricing information.

[1] Truveta pricing may need increase to accommodate additional medical conditions for data