Supplementary Online Content

Riskin DJ, Monda KL, Gagne JJ, et al. Implementing accuracy, completeness, and traceability for data reliability. *JAMA Netw Open.* 2025;8(3):e250128. doi:10.1001/jamanetworkopen.2025.0128

This supplementary material has been provided by the authors to give readers additional information about their work.

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eTable 1. Definitions of Accuracy, Completeness, and Traceability From the US Food and Drug Administration Draft Guidance for Real-World Data[1]

Metric	**Definition**
Accuracy	Closeness of agreement between the measured value and the true value of what is intended to be measured.
Completeness	The "presence of the necessary data"
Traceability	Permits an understanding of the relationships between the analysis, analysis datasets, tabulation datasets, and source data.

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eTable 2. Source of Truth for Evaluating Reliability and Rationale for its Selection by Data Type

Data type	Source of truth	Rationale
Clinical characteristic[a]	EHR unstructured data	Documented at point of care
Lab value	EHR structured data	Most reliable storage location
Medication	Pharmacy claim	Reflects medication receipt
Procedure	EHR structured data	Documented at point of care

[a] Clinical characteristics include conditions, comorbidities, symptoms and findings

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eTable 3. Code Sets Used to Define Select Features Across Variable Types

Feature	**Codes**
Severe asthma	ICD10CM: J45.5, J45.50, J45.51, J45.52 ICD9CM: No sufficient granular codes available SNOMED: 370221004, 426656000, 707447008, 707513007, 707979007, 124991000119109, 10675391000119101, 10675431000119106, 10675471000119109, 10675551000119104, 10675591000119109, 10675631000119109, 10675671000119107, 10675711000119106, 10675751000119107
Nasal polyp	ICD10CM: J33.0, J33.1, J33.8, J33.9 ICD9CM: 471.0, 471.1, 471.8, 471.9 SNOMED: 52756005, 141390005, 155520002, 164193006, 195752006, 195754007, 195762004, 233686006, 373605001, 736499003, 736500007, 736591003, 736635001, 736636000, 41931000119102
Asthma exacerbation	ICD10CM: J45.21, J45.22, J45.31, J45.32, J45.41, J45.42, J45.51, J45.52, J45.901, J45.902 ICD9CM: 493.01, 493.02, 493.11, 493.12, 493.21, 493.22, 493.91, 493.92 SNOMED: 57546000, 281239006, 304527002, 425969006, 707445000, 707446004, 707447008, 707979007, 708038006, 708090002, 708093000, 708094006, 733858005
Pulmonary function test	CPT: 94010, 94060 SNOMED: 3862003, 15435001, 22238008, 23426006, 28101009, 29893006, 30756005, 33875008, 36150003, 36421003, 37701007, 41187008, 44236004, 46857000, 48806009, 54831007, 61646000, 66904006, 67226009, 69706004, 70115007, 74252006, 76572000, 87529006, 91225006, 127783003, 127802000, 127803005, 165011001, 171255006, 252483008, 252484002, 252485001, 252486000, 252487009, 252488004, 252489007, 252493001, 252494007, 252520007, 258058009, 314979004, 426345008, 438575002, 439321001, 440011007, 440209007, 440225002, 440370007, 440371006, 444642008, 446461001

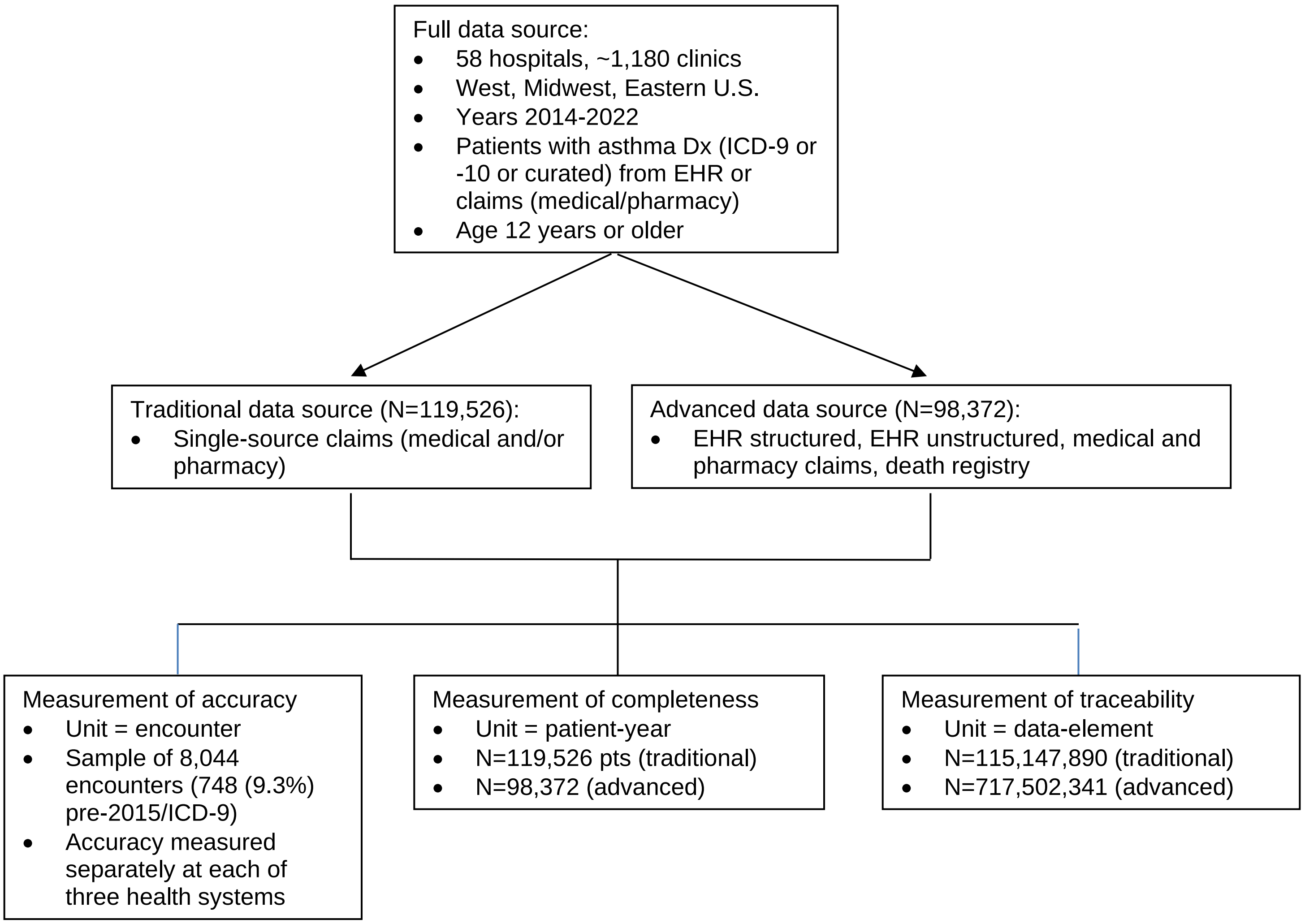
Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eTable 4. Data Source Contributions Observed and Applied Weighting Model Used to Calculate Completeness

Data source	**Average observed # of data elements per patient-year using traditional approach**	**Average observed # of data elements per patient-year using advanced approach**	**Weight used to calculate completeness**
Medical claim	12,286	14,874	25%
Pharmacy claim	1,831	1,987	25%
EHR structured data	0	16,412	25%
EHR unstructured data	0	193,168	25%

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eFigure 1. Dataset Design Schema



Full data source:
- 58 hospitals, ~1,180 clinics
- West, Midwest, Eastern U.S.
- Years 2014-2022
- Patients with asthma Dx (ICD-9 or -10 or curated) from EHR or claims (medical/pharmacy)
- Age 12 years or older

Traditional data source (N=119,526):
- Single-source claims (medical and/or pharmacy)

Advanced data source (N=98,372):
- EHR structured, EHR unstructured, medical and pharmacy claims, death registry

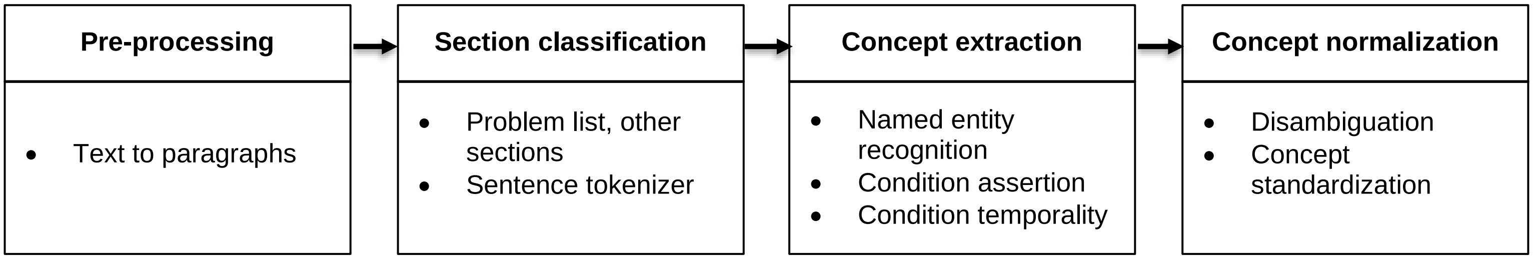
Measurement of accuracy
- Unit = encounter
- Sample of 8,044 encounters (748 (9.3%) pre-2015/ICD-9)
- Accuracy measured separately at each of three health systems

Measurement of completeness
- Unit = patient-year
- N=119,526 pts (traditional)
- N=98,372 (advanced)

Measurement of traceability
- Unit = data-element
- N=115,147,890 (traditional)
- N=717,502,341 (advanced)

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eFigure 2. High-Level Artificial Intelligence Pipeline



Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eAppendix 1. Description of Underlying Datasets

Data was sourced from 58 hospitals and more than 1,180 associated outpatient clinics in the West, Midwest, and Eastern U.S. across academic and community practices spanning three health systems and incorporating Epic Systems and Cerner EHRs. The study spanned 2014 through 2022. Eligibility criteria included patients 12 years old or greater with a diagnosis of asthma based on physician documentation within clinical or administrative data. For a patient to be included, they also needed to meet requirements for one or more analysis approaches. For example, a patient for whom there were no claims data or evidence of enrollment could not be included in the traditional approach. A patient for whom there were no clinical data could not be included in the advanced approach.

From the same population, more patients were identified within the traditional approach than the advanced approach. The reason is that the advanced approach required presence of more data sources which were not present for all patients. Out of 120,616 patients that met minimum data requirements, there were 119,526 asthma patients identified in the traditional approach and 98,372 patients in the advanced approach.

Accuracy, completeness, and traceability each required slightly different subsets of data for computation.

Accuracy required tedious manual annotation which could not be completed on the thousands of patients and hundreds of thousands of encounters available. Thus, a subset of 8,044 encounters was selected for manual annotation to create an reference standard to test accuracy. The subset was oversampled for presence of an asthma treatment, such as a medication, to discover encounters related to asthma. Furthermore, accuracy determination required matched claims encounters and EHR visits since many clinical concepts vary over time and accuracy testing must be performed at a point in time. Matching claims data and EHR data may create error due to frequent variability in documented date and the potential for mismatch. Thus, to reduce error, the traditional dataset for accuracy determination was based on EHR structured data and EHR open claims. This was slightly broader than the traditional dataset used elsewhere, which included only closed claims data. Each encounter required presence of both structured and unstructured data. Of the encounters that met these criteria, selection for accuracy determination was random. Accuracy was separately tested at each health system. Given local optimization, no significant difference was noted in accuracy between health systems. Of the 8,044 encounters, 748 encounters (9.3%) were pre-2015 and used ICD-9 as diagnosis codes. No significant difference was noted in accuracy between 2014 which was pre-ICD-10 implementation and 2015 and beyond which was post-ICD-10 implementation.

Completeness required different calculations than accuracy. The assessment of various data sources was made on a patient-year basis. Thus, the traditional dataset was the complete dataset of patients with only claims data included. The advanced data set was the complete dataset of patients with all EHR and claims data included. There were 119,526 patients in the traditional dataset and 98,372 in the advanced data set. The denominator was all patient years from index date of first asthma mention for which any data were present. For example, in the advanced approach, a patient may have first mention of asthma in 2018 and may have three years of EHR data or claims enrollment. In this case, completeness for that patient would be measured based on three patient-years. Patients within the traditional and advanced approaches rarely had all data for all years. Thus, the preferred approach for calculating completeness is to measure completeness of data sources for the years in which data are present and to separately report on longitudinality of the data set as measured by average years of data per patient.

Traceability required an assessment of all data elements against a source of truth. The underlying datasets were the same as those used in the completeness analysis, but the computation was performed on a per-data-element basis. In the traditional dataset, there were 115,147,890 clinical data elements of which 13,188,942 were traceable. In the advanced dataset, there were 716,502,341 clinical data elements of which 553,776,479 were traceable. There were untraceable elements in both datasets due primarily to lack of support, lack of data availability, and documentation or linkage errors. As an example of lack of data availability, tracing a medication may have been impossible in the advanced data set due to lack of pharmacy claim availability in the year the EHR data were available. As another example, tracing a clinical characteristic may have been impossible in the traditional data due to lack of clinical data in that set. Traceability was aimed at discovering systematic error and did not attempt to address data relevance questions such as the importance of a prescription that was not dispensed.

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eAppendix 2. Description of Technology

The Verantos Evidence Platform utilizes artificial intelligence to perform information extraction from clinical documents. The pipeline is composed of four layers as shown in **eFigure 2**.

Preprocessing
A clinical document can be lengthy and variable in formatting. To ensure consistent processing by a deep learning model the original content must be segmented. The preprocessing layer uses a combination of formatting and syntactic cues to split a document into paragraph-sized segments, which also frequently coincide with the boundaries of clinical sections.

Section classification
Section information can be helpful for locating specific medical concepts such as medications and aid in disambiguating abbreviations and acronyms. A transformer model is trained on annotated examples to classify each segment into one of the predefined categories such as history of present illness, physical exams, family history and problem list.

Concept extraction
Fit-for-purpose artificial intelligence techniques are used to extract concepts. To measure the accuracy of extraction of each concept that will be studied, a reference standard is developed by clinicians that consists of training, validation, and test annotations.

- The training and validation annotations are used to train deep learning models such as transformers to extract concepts from narrative text until extraction accuracy is at the level specified in the study protocol. The test annotations are used to test the performance of the model against the reference standard to ensure that the accuracy specified in the protocol can be achieved when the models are deployed on real-world data for concept extraction.
- For concepts that require inference, a combination of deep learning models including transformers and large language models (LLMs) are deployed to extract clinical concepts of interest. This extraction is also tested against a reference standard as described above.
- A named entity recognition (NER) model detects token spans within a text snippet that likely represent a medical entity such as a condition or a measurement. Different sequence classification models then determine the attributes of the extracted entity based on its surrounding context. These attributes, for example, include whether a condition is experienced by a patient or by someone else like a family member, and whether it is an ongoing or a past condition.
- For findings that contain measurements, a generative LLM is used to extract a measurement’s numerical values, units, and the date when a measurement is taken.

Concept normalization
The extracted phrases are matched against an internal database of clinical concepts to normalize identified information to known concepts. Approximate string retrieval with cosine similarity is implemented to map variants of a concept to its canonical form. For instance, trouble sleeping is a commonly used expression in place of insomnia. Statistical measures such as concept frequency and association table are employed to disambiguate terms that can have different meanings in different contexts. Final concepts are normalized to standard concepts in the OMOP vocabulary[2], which is an ontology of industry-standard terminologies such as SNOMED, ICD-10, LOINC, and RxNorm.

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eAppendix 3. Description of Reference Standard Creation and Accuracy Determination

Manual annotation was performed on a subset of patient encounters. A set of variables relevant to the therapeutic area was selected in advance of annotation.

Annotators were required to hold a clinical degree and have at least six months of annotation experience. Two annotators reviewed every selected encounter and annotated every selected variable. Annotations could include present = true or false, experienced = true or false, current = true or false, and other attributes appropriate to the variable. Annotators were not allowed to guess. For example, "HC" may not be assumed to be hypercholesterolemia. Annotators were allowed to infer. For example, documentation of ejection fraction 60% in one location and heart failure in another may be inferred to be heart failure with preserved ejection fraction. Inferences were required to be marked as inference with references to all data, whether structured or unstructured, that led to that inference.

Annotations were locked after the encounter was annotated. The two annotators were blinded to each other's annotations. Once both annotators locked their annotations, inter-rater reliability was measured using Cohen's kappa score. A kappa score of 0.7 was required for a reference standard to be considered sufficient. Conflicts were reviewed and resolved by committee, which included a clinical informaticist.

Traditional and advanced data for each of the selected encounters were compared against the reference standard with a non-match counted as an error. Given that acute conditions are specific to point in time, a match in the current flag was also required. For example, the text "history of asthma exacerbation", if interpreted as asthma exacerbation, current=true, would be marked as an error due to the incorrect current flag.

For each variable, data accuracy was quantified as recall, precision, and F1 score. Recall or sensitivity was estimated as the proportion of records with the variable identified through the traditional and advanced approaches, separately, divided by the number of records with the variable identified according to the reference standard. Precision or positive predictive value was estimated as the proportion of variables identified through the traditional and advanced approaches, separately, that were identified through the reference standard. Precision is preferred over specificity because of the high true negative rate for conditions occurring in any given patient, which can heavily influence specificity. Given the variety of controlled vocabularies in routine data, such as SNOMED, ICD-10, RxNorm, and LOINC, mapping was used where required to ensure more granular concepts were properly handled (e.g., severe asthma as a form of asthma). Select code sets used across variable types are listed in **eTable 3**.

The F1 score was used as a summary measure of precision and recall. The F1 score, calculated as 2 x ([precision x recall]/ [precision + recall]), is the weighted harmonic mean of precision and recall and considered to be more informative especially when there is a low prevalence of a variable, which is referred to as a class imbalance.[3]

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.

eReferences.

[1] Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory. Accessed November 20, 2024.

[2] OHDSI standard vocabularies, https://github.com/OHDSI/Vocabulary-v5.0/wiki. Updated October 1, 2024. Accessed December 6, 2024.

[3] Williams CKI. The Effect of Class Imbalance on Precision-Recall Curves. *Neural Comput*. 2021;33(4):853-857. doi: 10.1162/neco_a_01362

Copyright 2025 Riskin D et al. *JAMA Network Open*. This is an open access article distributed under the terms of the CC-BY-NC-ND License, which does not permit alteration or commercial use, including those for text and data mining, AI training, and similar technologies.