Federal Aviation Administration

DOT/FAA/AM-22/10
Office of Aerospace Medicine
Washington, DC 20591

Pilot Medical Certification Period Health State Forecasts

Felix Bradbury
Greg Chesterton
Stanley C. Chin
Kunal K. Sarkhel
David Slater
Zory Slater
Ben Wellner

The MITRE Corporation, McLean, VA

November 2022

Technical Report

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications Web site:
(www.faa.gov/go/oamtechreports)

Technical Report Documentation Page

1. Report No.
DOT/FAA/AM-22/10

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Pilot Medical Certification Period Health State Forecasts

5. Report Date
September 30, 2022

6. Performing Organization Code

7. Author(s)
F. Bradbury; G. Chesterton; S. Chin; K. Sarkhel; D. Slater; Z. Slater;
B. Wellner. All authors from The MITRE Corporation

8. Performing Organization Report No.
MITRE PBWP Ref 4_80-2.C.1-1

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.
DTFAWA-10-C-00080

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:
Bradbury (0000-0001-8445-5480); Chesterton (0000-0003-2439-1299); Chin (0000-0002-0460-5942);
Sarkhel (0000-0003-4721-9416); D. Slater (0000-0002-3160-0389); Z. Slater (0000-0002-7618-215X);
B. Wellner (0000-0003-2689-990X)
Technical report DOI: https://doi.org/10.21949/1524436

16. Abstract

The Federal Aviation Administration (FAA) Office of Aerospace Medicine supports research to use available healthcare data to inform policies regarding pilots' medical certifications. The MITRE Corporation's Center for Advanced System Development (MITRE CAASD) was asked to examine methods in advanced data analytics and machine learning (ML) to inform such risk-based decision-making. As an initial step in assessing the potential predictive value of commercially available healthcare data, the FAA provided MITRE CAASD the IBM MarketScan dataset—a large set of commercial healthcare claims records. Using this dataset, we developed methods for identifying health status and changes in health status across conditions; for measuring changes in health status among enrollees with diabetes mellitus (DM); and for measuring the onset of new cases of DM, traumatic brain injury, sleep apnea, and chronic obstructive pulmonary disease. We developed a repeatable workflow and modeled these conditions using a wide range of ML methods. We conclude that ML-based predictive modeling of health conditions from IBM MarketScan data is feasible and informative. However, additional clinical information from commercially available electronic health records would likely improve accuracy and more closely align with future FAA needs.

17. Key Word
Aviation medicine, safety and health, medical claims data

18. Distribution Statement
Document is available to the public through the National Transportation Library: https://ntl.bts.gov/ntl

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
36

22. Price

Abstract

The Federal Aviation Administration (FAA) Office of Aerospace Medicine supports research to use available healthcare data to inform policies regarding pilots' medical certifications. The MITRE Corporation's Center for Advanced System Development (MITRE CAASD) was asked to examine methods in advanced data analytics and machine learning (ML) to inform such risk-based decision-making. As an initial step in assessing the potential predictive value of commercially available healthcare data, the FAA provided MITRE CAASD the IBM MarketScan dataset—a large set of commercial healthcare claims records. Using this dataset, we developed methods for identifying health status and changes in health status across conditions; for measuring changes in health status among enrollees with diabetes mellitus (DM); and for measuring the onset of new cases of DM, traumatic brain injury, sleep apnea, and chronic obstructive pulmonary disease. We developed a repeatable workflow and modeled these conditions using a wide range of ML methods. We conclude that ML-based predictive modeling of health conditions from IBM MarketScan data is feasible and informative. However, additional clinical information from commercially available electronic health records would likely improve accuracy and more closely align with future FAA needs.

Executive Summary

The MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) was tasked to examine methods in advanced data analytics and machine learning (ML) to inform a next-generation pilot medical certification safety management system. As an initial step in assessing the potential predictive value of commercially available healthcare data, the Federal Aviation Administration (FAA) provided MITRE with the IBM MarketScan dataset—an extensive set of commercial healthcare claims. Using this dataset, we developed methods for identifying health status and changes in health status across conditions; for measuring changes in health status among enrollees with diabetes mellitus (DM); and for measuring the onset of new cases of DM, traumatic brain injury (TBI), sleep apnea, and chronic obstructive pulmonary disease (COPD).

We developed a repeatable workflow and modeled these conditions using a wide range of ML methods, including individual and ensemble methods, and examined their accuracy using multiple criteria. Our primary finding is that these predictive models replicate or exceed existing models based on claims data; however, they generally have poor precision and recall statistics as measured by well-known metrics such as the F1 and Matthews Correlation Coefficient scores. Area Under the Receiver-Operating Characteristics Curve (AUC) scores, which measure the overall ability of the model to correctly classify cases at different threshold levels, varied from 0.67 (TBI) to 0.844 (COPD).

We also examined Airman Medical Certification forms, but because the integration of these data would necessitate a significant degree of natural language processing (NLP) and because NLP is beyond the scope of work, we did not include these in our preliminary analyses. Instead, MITRE CAASD focused on the immediate problem of representing and modeling healthcare outcomes from claims and eligibility data.

We conclude that ML-based predictive modeling of health conditions from MarketScan data is feasible and informative. However, additional information from electronic health records and the ability to more closely link clinical conditions to pilot medical certification would likely improve accuracy and more closely align with future FAA needs.

Table of Contents

List of Figures

List of Tables

1 Introduction

The Federal Aviation Administration's (FAA's) Office of Aerospace Medicine (AAM) is actively engaged in activities with numerous challenges that demand better use of medical data for timely, risk-based pilot medical certification decision-making in an environment of rapid change in both healthcare and aerospace operations. In particular, the AAM is aligning its practices with FAA's Safety Management System (SMS) policy.

To implement SMS and conduct risk-based decision-making, policymakers require risk assessments of identified hazards. AAM must manage the risk that a pilot determined to be fit for duty for the certification period will instead become medically unfit during that period. At a more granular level, a concern is that a pilot's chronic condition or set of conditions is such that there is an unacceptably high likelihood of a health episode that could impact pilot performance during the certification period.

Safety risk assessment requires estimates of likelihood. FAA Order 8040.4 includes proposed likelihood definitions for large commercial and small aircraft categories of operations. FAA encourages using quantitative methods for their objectivity, although qualitative data and expert judgment are acceptable. Qualitative judgment varies from person to person, which can introduce variance and uncertainty in decision-making. Quantitative analysis, on the other hand, minimizes subjective analytical variation and is suitable for cases where the outcome of interest can be modeled, and data support that model.

AAM seeks to research, develop, and validate tools, techniques, and procedures—particularly in big data and machine learning (ML)—that will form the technological foundations to implement a next-generation pilot medical certification safety management system. This research leverages large commercially available healthcare datasets and current big data analytics to enable precision-based aeromedical risk assessments that cannot be developed from existing agency medical certification data that are limited in quantity and quality.

As an initial step in assessing the potential value of commercially available healthcare data, MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) was tasked to develop and evaluate models to predict changes in pilot health status using medical claims data. MITRE CAASD developed methods for identifying changes in health status, modeled these outcomes using a wide range of ML methods, and generated prediction models at the individual level for overall chronic health status change, the onset of specific conditions, and the increasing severity of specific conditions.

In this preliminary analysis, we used commercial claims data from the IBM MarketScan dataset for our models. We demonstrated that claims data can be used in conjunction with ML methods to predict the onset and changes in severity in some conditions of interest, including diabetes mellitus (DM), traumatic brain injury (TBI), obstructive sleep apnea (OSA), and chronic obstructive pulmonary disease (COPD).

2 Literature Review

A substantial body of research now exists on the application of ML in predicting individual medical outcomes from administrative claims, electronic health records (EHRs), and other clinical data sources. These models show promise, with some outcomes, data types, and modeling methods proving more challenging than others.

In general, model performance depends on the medical outcome of interest. Models predicting unanticipated inpatient and emergency department visits show a poorer fit in all data and modeling environments, especially when compared to those predicting near-term mortality. Predictions of the onset of new conditions or increasing disease complexity or severity of existing conditions yield results between these two extremes. Modelers have pursued numerous feature selection and dimensionality reduction techniques, although none exhibit definitively superior performance. Although modeling techniques and methods vary, XGBoost showed the most promising results in terms of area under the receiver operating characteristic curve (AUC, or C-score), a commonly used metric of model goodness-of-fit and discriminatory power. Adding EHR or other clinical data generally improved model performance, although Desai et al. (2020) found slight improvement in some scores for heart failure modeling. A summary of select current models and methods is shown in Table 1.

Table 1: Recent Literature on Machine Learning and Predictive Models for Healthcare Outcomes

Study	Outcome	Method	Result
Administrative Claims Data			
Croon et al. (2022)	Scoping review of 16 ML models on heart failure (HF) readmissions	Review of 16 studies on readmission, using various models and data	AUC 0.61–0.79, with the highest AUC using features from EHR and imaging data; the claims-based model achieved AUC of 0.64
Desai et al. (2020)	HF: mortality, high cost, hospitalization, home days loss	Compared logistic to ML, with gradient boost showing best performance; added EHR data for modest gains	Gradient boost AUC from claims: mortality 0.73, hospitalization for HF 0.75, high cost 0.73, home days loss 0.79
Lewis et al. (2021)	Preventable hospitalizations in HF patients	Sequential deep learning on tokenized monthly vectors	AUC 0.78
MacKay et al. (2021)	30-day mortality (CMS AI challenge)	Gradient boost ML on Medicare claims data using HCC, CCS, DRG	Mortality: 0.88; 17 adverse events 0.88–0.86; 30-day rehospitalization 0.73
Non-traditional Data Sources			
Chang and Chen (2021)	Increase in self-reported disease severity using app data from Flaredown (Kaggle)	XGBoost and ensemble models	F1 of 0.93 reported

Study	Outcome	Method	Result
Dinh et al. (2019)	Onset of diabetes, CVD from NHANES survey data including lab results	XGBoost ensemble models using survey and demographics only, and adding lab values	GBDT diabetes: 0.86 AUC, 0.95 with labs; Ensemble CVD: 0.83, 0.84 with labs
James et al. (2021)	Onset of dementia in memory clinic patients	NACC uniform data set including sociodemographics, functional status, symptoms, neuropsychological test battery, various models including XGBoost	XGBoost: AUC 0.92, accuracy 0.92
EHR Data			
Bose et al. (2021)	Pediatric onset of chronic asthma	EHR data, various ML models tested	XGBoost: AUC 0.81
Molani et al. (2022)	COVID-19 severity score dichotomized to mild/severe	Logistic Regression, Random Forest, GBDT, AdaBoost models, all had similar AUC	GBDT AUC 0.78, true positive 0.75 for younger population, AUC 0.81 and true positive 0.73 for older
Rajkomar et al. (2018)	In-hospital mortality, length of stay (LOS), discharge diagnosis	Vector of time-sequenced tokens using FHIR for each patient visit, using all EHR data, including free text. Used LSTM, TANN, a NN with boosted time-based decision "stumps"	In-hospital mortality: AUC 0.95; LOS: AUC 0.86; discharge dx AUC 0.87

In addition, we examined the current literature on ML analyses for the four specific conditions tested in the current study: DM, COPD, OSA, and TBI. Claims-based ML modeling of DM and COPD were directly used in designing and considering our model approach. ML uses in OSA and TBI tend to focus on clinical data, imaging, monitoring, and other sensor data; only those studies directly relating to claims or EHR data are included in Table 2.

Table 2: Recent Literature on Machine Learning and Predictive Models for Healthcare Outcomes in Diabetes, COPD, OSA, and TBI populations

Study	Outcome	Method	Result
Diabetes Onset and Severity			
Ravaut et al. (2021)	Onset of diabetes	Claims data XGBoost	AUC 0.80
Anderson et al. (2015)	Onset of diabetes	EHR data ensemble model	AUC 0.76
Tuppad and Patil (2022)	Evaluation of ML in diabetes risk assessment, diagnosis, prognosis/disease progression	Review of 86 articles, including claims, EHR, clinical data	Diagnosis: AUC 0.70 to 0.90 using different data and models Prognosis: 0.76–0.87 all, including clinical features

Study	Outcome	Method	Result
COPD Onset and Severity			
Muro et al. (2021)	Onset of COPD from employee annual checkup medical records	XGBoost on behavioral, clinical, lab features	AUC 0.96 with accuracy of 91.7%, sensitivity of 84.5%, specificity of 96%
Lu and Uddin (2021)	Onset of COPD, CVD from Australian claims data	Graph-based weighted patient network constructed from diagnosis codes clustered to Elixhauser categories; modeled using Graph Attention Network	CVD accuracy 0.93, CPD accuracy 0.89
Goto et al. (2019)	COPD readmissions, claims	Logistic regression compared with deep neural network on claims data	AUC of 0.61–0.645
Min et al. (2019)	COPD readmissions, claims	Extensive comparison of models and features comparing knowledge-driven clinically derived features, and data-mining features	Knowledge features: 0.61–0.64; data-driven features: 0.64-0.65
Sleep Apnea			
Ramachandran and Karuppiah (2021)	Sleep apnea diagnosis survey of ML	Biomedical markers, ECG/EEG, etc. outputs predominate, e.g., Artificial Neural Network to model diagnosis from ECG signals with 85% accuracy	Survey article
Mencar et al. (2020)	Sleep apnea increase in severity	Demographics, spirometry, gas exchange, symptoms with dichotomized outcome, using random forest	Accuracy of 44.7%
Traumatic Brain Injury			
Chan et al. (2020)	Total medical expense in TBI cohort using administrative data	Linear regression model finds importance of pre-injury health status on costs (severity)	R2 not reported
Matsuo et al. (2020)	In-hospital mortality for TBI patients	EHR data with many clinical features; random forest	0.90 AUC with 100% sensitivity and 72.3% specificity, 91.7% accuracy

A description of the relevant articles with brief summaries of each can be found in Appendix C.

3 Data sources

3.1 Healthcare Claims Data

The study used the IBM MarketScan Commercial Claims and Encounters Database of national medical claims data for privately insured individuals and their dependents, age 65 or younger,[1] for January 1, 2017, through December 31, 2019. The study population was further limited to those aged 18 years and older who were fully covered by their insurance in 2017, 2018, and 2019. The resulting dataset contains 281,655,673 diagnosis codes, 515,676,057 procedure codes, and 125,992,515 prescriptions for 8,119,171 individuals.

MarketScan is commonly used in research settings to examine questions such as drug utilization, clinical care practices, socio-demo-geographic variations in treatment and outcomes, and other retrospective analyses. A MedLine search identifies over 800 articles published in 2021 and 2022 using or referencing MarketScan.

Healthcare claims data contain billing information for inpatient and outpatient visits, including the type of provider, service site, visit date, diagnoses, procedures, and prescriptions. Data are encoded according to several standard schemas: diagnoses are indicated using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), inpatient procedures use the ICD-10 Procedure Coding System (ICD-10-PCS), outpatient procedures use Current Procedural Terminology (CPT-4), and prescriptions use the National Drug Code as well as therapeutic class and therapeutic group. The MarketScan data also include summary and detailed enrollment files that provide details on insurance enrollment coverage, date of birth, and patient gender.

All data are commercially provided and anonymized to comply with Health Insurance Portability and Accountability Act requirements. As a result, human subject research review and approval by an Institutional Review Board are not required.

3.2 Common Data Model

We tested the mapping of MarketScan data elements to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).[2] The CDM aligns disparate datasets into standard vocabularies so that research is comparable, generalizable, and yields reproducible results. CDM datasets can be easily expanded to include additional data elements from other sources, such as laboratories, eligibility files, pharmacy files, or EHRs. The contents of CDM data types can be mapped to known dictionaries to create clinically meaningful insights and analyses, for example, National Drug Codes to drug ingredients.

[1] See https://www.ibm.com/products/marketscan-research-databases

[2] See Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, Londhe A, Zhu V, Ryan PB. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553-64. doi: 10.1093/jamia/ocu023. Epub 2015 Feb 10. PMID: 25670757; PMCID: PMC4457111. The effort is further described at https://ohdsi.github.io/CommonDataModel/. Note that the Common Data Model should not be confused with the acute inpatient charge data master which also uses the abbreviation, “CDM;” the latter is the basis of facility pricing for services, the former is a data management framework.

Despite these potential benefits, translating MarketScan to the OMOP CDM requires third-party support; we determined that the cost of such support outweighs the value of our specific research objectives.[3]

3.3 Feature Selection

In this initial phase, we used ICD-10-CM diagnosis codes, CPT-4 procedure codes, and National Drug Codes directly as binary (on/off) indicators and used the publicly available Healthcare Cost and Utilization Project (HCUP) Clinical Classification Software (CCS; beta-version) tools to reduce the dimensionality of the extensive diagnosis and procedure code space.[4]

The CCS tool maps over 70,000 specific ICD-10-CM diagnosis codes and over 10,000 CPT-4 procedure codes into clinically meaningful aggregate categories. We used a Python implementation of these tools (the package `hcuppy`) to generate approximately 300 CCS treatment features from CPT-4 procedure codes and approximately 500 CCS clinical features from ICD-10-CM diagnosis codes.[5] We used therapeutic groups to aggregate individual prescription drugs, as provided in the MarketScan dataset.

The HCUP groupings of diagnosis codes greatly reduced the computational and time requirements for model training and tuning. We performed several specific tests comparing models trained directly with ICD-10 and CPT-4 procedure codes with those trained using HCUP groups and found similar model performance. As such, this report focuses on models trained with HCUP features.

[3] Examples of MarketScan data mapped to CDM include Molinaro A, DeFalco F. Empirical assessment of alternative methods for identifying seasonality in observational healthcare data. BMC Med Res Methodol. 2022 Jul 2;22(1):182. doi: 10.1186/s12874-022-01652-3. PMID: 35780114; PMCID: PMC9250712; Khera R, Schuemie MJ, Lu Y, Ostropolets A, Chen R, Hripcsak G, Ryan PB, Krumholz HM, Suchard MA. Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies. BMJ Open. 2022 Jun 9;12(6):e057977. doi: 10.1136/bmjopen-2021-057977. PMID: 35680274; PMCID: PMC9185490; Williams RD, Reps JM, Kors JA, Ryan PB, Steyerberg E, Verhamme KM, Rijnbeek PR. Using Iterative Pairwise External Validation to Contextualize Prediction Model Performance: A Use Case Predicting 1-Year Heart Failure Risk in Patients with Diabetes Across Five Data Sources. Drug Saf. 2022 May;45(5):563-570. doi: 10.1007/s40264-022-01161-8. Epub 2022 May 17. PMID: 35579818; PMCID: PMC9114056. Note that all these studies explicitly combine and compare data from multiple databases and sources.

[4] The software tools can be found at the AHRQ HCUP site, https://www.hcup-us.ahrq.gov/tools_software.jsp

[5] The python hcuppy package documentation can be found at https://pypi.org/project/hcuppy/

4 Modeling Approaches

For this work, we focused on three categories of outcomes: the onset of specific chronic conditions, the relative change in severity for specific conditions, and the change in cumulative chronic conditions. We examined these outcomes in four specific conditions: DM, TBI, OSA, and COPD.

A variety of modeling approaches are available for predicting condition onset and severity. Several modeling techniques were evaluated for this project, including the following:

- Event history analysis: Also known as cross-sectional, time series, or panel data analysis, this approach has its roots in biomedical engineering and mechanical engineering and reflects the marriage of ordinary least squares (OLS) regression and time-series modeling.
- Auto-regressive integrated moving averages (ARIMA): This simple yet powerful approach creates a linear equation that describes and forecasts time series data.
- Linear classification algorithms such as Poisson regression and logistic regression: These techniques can be used to categorize the presence or absence of an outcome, such as a new disease state or multiple disease states.
- Product-limit analysis, also known as the Kaplan-Meier approach: This descriptive method of survival analysis generates a population survival curve and essential statistics such as the median survival time.

There are various reasons why the above models were not selected for use. In general, these models were not chosen for adoption because of one or more of the following:

- The model may not perform well with sparse or missing data.
- The relationship between the outcome of interest and the covariates cannot be assumed linear; in such cases, the OLS estimator is biased and does not meet maximum likelihood criteria.
- There may be a lack of appropriate "time-to-failure" outcomes in the data for survival analysis.

For this report, we focused on two problem formulations using claims data aggregated within each year for each patient: (continuous) regression and binary classification. To evaluate multiple modeling approaches, we utilized the MITRE high-performance computing environment and the AutoML package `autoGluon`. `AutoGluon` enables easy-to-use and easy-to-extend AutoML with a focus on automated stack ensembling, deep learning, and real-world applications spanning image, text, and tabular data.[6]

4.1 Modeling Overall Change in Severity

For modeling overall changes in individual health status, we focused on modeling the yearly change in the number and severity of chronic conditions. We created a weighted chronic condition score based on the HCUP groupings of diagnosis codes. For each HCUP group, we

[6] https://auto.gluon.ai/stable/index.html

referred to the clinical expertise of the team to select the groups that represent chronic conditions and to assign each group a severity score between one and three. For example, the presence of a condition in HCUP group Cancer of the Prostate was given a weight of 3, while the group Asthma was given a weight of 2. This approach is intended as a proof of concept for measuring cumulative health of individuals and was used to assess the overall disease burden faced by an individual and to measure changes in that burden over time.

To model severity over time, we evaluated two approaches: predicting an individual's severity yearly and predicting a new measure called delta-severity that we defined to capture the relative change in an individual's number and severity of chronic conditions each year. Formally, delta-severity is defined as

$$delta - severity(year) = \frac{NC(year) - NC(year-1)}{NC(year-1)} * \frac{SC(year) - SC(year-1)}{SC(year-1)},$$

where $NC(year)$ is the cumulative number of chronic conditions for an individual for all years, including $year$, and $SC(year)$ is the cumulative severity.

For example, if an individual had two chronic conditions with total severity 5 in 2017 and had a new chronic condition with severity 2 in 2018, their delta-severity score would be

$$delta - severity = \frac{3-2}{2} * \frac{7-5}{5} = 0.20.$$

As this measure captures the number of new conditions that have developed over time and the change in severity from one year to the next, we could track and model the progression of disease with a single value.

For severity modeling, we computed the cumulative number of chronic conditions and their severity for each of the three years of the dataset, then used those values to compute the *delta severity* measure described above for three models:

1. Predict *delta severity* for 2017–2018 based on 2017 data
2. Predict *delta severity* for 2018–2019 based on 2018 data
3. Predict *delta severity* for 2018–2019 based on 2017 and 2018 data

This task was treated as a regression problem, where models were tuned based on mean absolute error. Each dataset was split for 75% training and 25% testing. The `autoML` package `auto gluon` was used to train, tune, and evaluate multiple types of models.

Results for the *delta-severity* measure are shown in Table 3, and results for chronic condition severity are shown in Table 4. In both cases, models were compared with dummy classifiers. For delta-severity, we found that our models did not outperform a dummy classifier that predicts the median value for all individuals; in addition, the very low r2 values for all models, including the dummy median classifier indicates overall poor fit and possible non-linearity in explanatory variables.

Table 3: Chronic Condition Delta-Severity Best Model Performance

Dataset	Outcome	Approach	MSE	MAE	r2
2017	2018	AutoGluon	187.918	2.265	0.044
2018	2019	AutoGluon	76.235	1.328	0.021

2017, 2018	2019	AutoGluon	83.709	1.108	0.015
2017	2018	Dummy Median	201.500	2.370	0.030
2018	2019	Dummy Median	78.920	1.060	0.010
2017, 2018	2019	Dummy Median	86.180	1.130	0.010

We also developed models to predict an individual's severity level in each year, given their previous severity level. Models were compared against a dummy classifier that predicted each individual's severity level would not change. As shown in Table 4, our models outperformed the best dummy variable models, and all models showed a high r2 goodness-of-fit measure.

Table 4: Chronic Condition Severity Best Model Performance

Dataset	Outcome	Approach	MSE	MAE	r2
2017	2018	AutoGluon	36.685	4.389	0.770
2018	2019	AutoGluon	29.441	3.826	0.856
2017, 2018	2019	AutoGluon	27.983	3.905	0.865
2017	2018	Dummy 2017 severity	82.950	6.530	0.480
2018	2019	Dummy 2018 severity	59.980	5.370	0.710
2017, 2018	2019	Dummy 2018 severity	58.710	5.280	0.720

The results are not entirely unexpected. Given the nature of chronic disease, the conditions, treatment, and prescriptions observed in a previous year provide substantial predictive power in chronic conditions observed in the following year. However, these models provide relatively little insight into the change in severity level when all possible chronic conditions are considered simultaneously in the model. This can be seen by considering the mean absolute error, which indicates that the average error in severity is nearly four at best. In the next section, we discuss modeling onset and increases in severity for four specific conditions.

4.2 Modeling Individual Conditions

Four conditions were selected for analysis: DM, TBI, OSA, and acute myocardial infarction (later revised to COPD). Cohorts were selected based on ICD-10-CM diagnosis codes and CPT-4 procedure codes:

- Diabetes:[7] ICD-10-CM E08xx, E09xx, E10xx, E11xx, E13xx, O24.0, O24.1, O24.3, O24.8, Z4681, Z9641, Z86.31

- TBI:[8] ICD-10-CM S02.0, S02.1, S02.3, S02.7, S02.8, S02.9, S06, S06.0, S06.1, S06.2, S06.3, S06.4, S06.5, S06.6, S06.8, S06.9, S07, S07.0, S07.1, S07.8, S07.9, S099, T060

[7] Based on Glasheen (2017) and extended to include pre-existing conditions, encounters related to insulin pumps, and personal history codes

[8] McChesney-Corbeil J, Barlow K, Quan H, Chen G, Wiebe S, Jette N. Validation of a Case Definition for Pediatric Brain Injury Using Administrative Data. Can J Neurol Sci. 2017 Mar;44(2):161-169. doi: 10.1017/cjn.2016.419. Epub 2017 Jan 20. PMID: 28103959.

- OSA:[9] ICD-10-CM G47.30, G47.31, G47.32, G47.33, G47.34, G47.35, G47.36, G47.38, G47.39; CPT-4 G0398, G0399, G0400, A7027, A7028, A7029, A7030, A7031, A7032, A7033, A7034, A7035, A7036, A7037, A7038, A7039, A7044, A7046, A4604, E0601, E0562, E1356, E1357, E1358, E1390, E1399
- Acute myocardial infarction: ICD-10-CM I21xx, I22xx
- COPD: ICD-10-CM J41xx, J42xx, J43xx, J44xx

We used the cohort definitions for each condition to compute which enrollees met the cohort criteria in each of the three calendar years. For onset prediction, we constructed three datasets for each condition:

1. Predict onset in 2018 based on 2017 data
2. Predict onset in 2019 based on 2018 data
3. Predict onset in 2019 based on both 2017 and 2018 data

For each dataset, we constructed a negative cohort consisting of a random sample of enrolled individuals who did not have the condition in either the outcome or predictor years. The dataset was then split into 75% training and 25% testing. Models were trained on HCUP-grouped ICD-10-CM diagnosis codes and CPT-4 procedure codes as well as age, sex, and the therapeutic group for all prescribed medications. The `autoML` package `autogluon` was utilized for training and tuning models automatically, allowing us to evaluate multiple different modeling approaches quickly. To measure the performance of classifiers, we focused on the Area under the Receiver Operating Characteristics Curve (ROCC, AUC, or C-statistic). We also report the F1 and Matthews Classification Coefficient (MCC) scores for a prediction threshold of 0.5.[10] This threshold was chosen arbitrarily to report performance. Future work could include measuring the best decision threshold to meet FAA goals concerning false positives and negatives.

4.2.1 Diabetes

Diabetes was modeled for both onset of new cases among a cohort without a diabetes diagnosis in the previous year and increasing severity in a cohort of patients diagnosed with diabetes.

4.2.1.1 Diabetes Onset

For diabetes onset, binary classification models were trained based on the three different datasets utilizing the HCUP grouped binary predictor variables. Model performance was tuned and measured based on AUC. We found that the AUCs on the holdout test sets were statistically equivalent for all three datasets (see Table 5).

Table 5: Diabetes Onset Best Model Performance

[9] Based on https://support.apriadirect.com/hc/en-us/articles/360022526374-What-are-the-HCPCS-CPT-or-billing-codes-related-to-CPAP-Sleep-Apnea-and-Oxygen-

[10] Given a confusion matrix with True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the F1 score is calculated as 2TP/(2TP + FN + FP). The MCC is calculated as ((TN*TP) – (FP*FN)/((sqrt((TN + FN) * (FP + TP) * (TN + FP) * (FN + TP)). The latter is preferred when classifying negative values correctly is important. See for example Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Jan 2;21(1):6. doi: 10.1186/s12864-019-6413-7. PMID: 31898477; PMCID: PMC6941312.

Dataset	Outcome	AUC	F1	MCC
2017	2018	0.792	0.327	0.348
2018	2019	0.792	0.323	0.347
2017, 2018	2019	0.783	0.238	0.275

Furthermore, we can see in Table 6 that model performance is equivalent across model types, and ensembling does not improve performance for this task.

Table 6: Diabetes Onset Predict 2018 Model Performance

Model	AUC
WeightedEnsemble_L3	0.791974
XGBoost_BAG_L2	0.791927
CatBoost_BAG_L2	0.791909
LightGBMXT_BAG_L2	0.791901
LightGBMLarge_BAG_L2	0.791897
LightGBM_BAG_L2	0.791885
WeightedEnsemble_L2	0.791776
NeuralNetTorch_BAG_L2	0.791685
RandomForestEntr_BAG_L2	0.791338
CatBoost_BAG_L1	0.791097
RandomForestGini_BAG_L2	0.791043

4.2.1.2 Diabetes Severity

To measure increases in diabetes severity, we calculated the Diabetes Complications Severity Index (DCSI) as implemented in ICD-10 codes by Glasheen et al. (2017). The DCSI is a validated and commonly used method for assessing the severity of complications in diabetes and has been used to predict or risk-adjust models of costs, hospitalization, and mortality.

For each enrolled individual that met the diabetes inclusion criteria, we calculated the cumulative DCSI score yearly and the change in score between subsequent years (see Figure 1). We then constructed models that used the current diabetes severity score, age, sex, the HCUP grouped procedure, diagnosis codes, and a therapeutic group for medications as features to predict the diabetes severity score in the subsequent year. More specifically, we considered two prediction tasks:

1. Binary Classification: Predict if an individual's severity will increase next year
2. Regression: Predict an individual's severity score next year

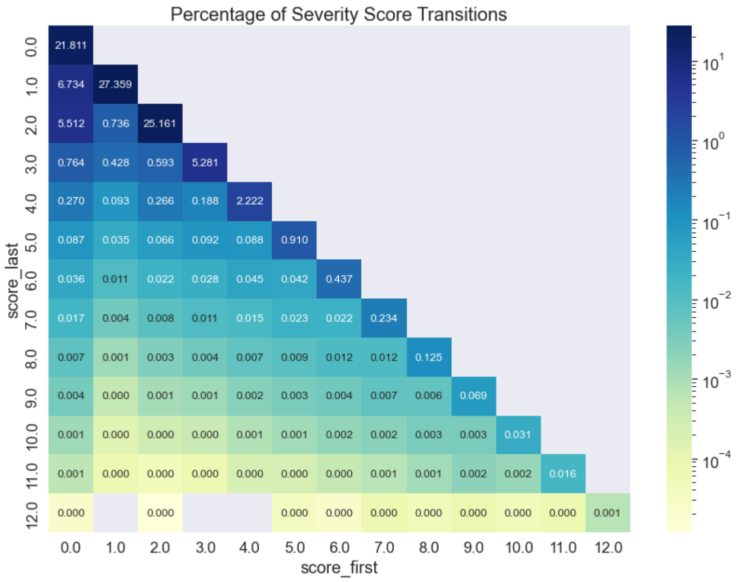

Figure 1: Diabetes Severity Score Yearly Changes

For diabetes severity binary classification models (that is, to predict whether diabetes severity increased), we trained models based on two different datasets using the HCUP grouped binary predictor variables. Model performance was tuned and measured based on AUC. We found that the AUCs on the test sets were statistically equivalent for both datasets (see Table 7). As with the onset modeling, we found that model performance was equivalent across model types and with ensembling.

Table 7: Diabetes Severity Binary Increase Model Performance

Dataset	Outcome	AUC	F1	MCC
2017	2018	0.733	0.930	0.051
2018	2019	0.727	0.931	0.052

To predict the cumulative diabetes severity using regression models, we trained based on the two different datasets using the HCUP grouped binary predictor variables. Model performance was tuned and measured based on mean absolute error (MAE). Model performance was compared against three dummy classifiers: predict mean diabetes score, median diabetes score, and no change in diabetes score. We found that the MAEs for both datasets were statistically equivalent to the dummy no-change classifier (see Table 8), suggesting that our ability to predict diabetes severity score is no better than predicting that the score does not change.

Table 8: Diabetes Severity Regression Model Performance

Dataset	Outcome	MAE	r2	MSE	Model
2017	2018	0.22	0.63	0.38	autogluon

Dataset	Outcome	MAE	r2	MSE	Model
2017	2018	0.81	0.00	1.58	dummy median
2017	2018	0.82	0.00	1.58	dummy mean
2017	2018	0.21	0.73	0.43	dummy no change
2018	2019	0.28	0.76	0.38	autogluon
2018	2019	0.71	0.04	1.05	dummy median
2018	2019	0.73	0.00	1.02	dummy mean
2018	2019	0.20	0.61	0.39	dummy no change

4.2.2 Traumatic Brain Injury

In the absence of clinical data allowing for the construction of TBI severity indices such as the Glasgow Outcomes Scale, we modeled only the onset of new TBI cases.

Due to the nature of TBI and its relation to physical trauma caused by accidents or other unpredictable events, we expected this condition to have the least predictability in the initial onset. Indeed, we found the lowest AUC scores for these models and poor F1 and MCC scores, as shown in Table 9. Future work should focus on severity and specific health outcomes, such as episodes of loss of consciousness or changes in behavior and other personality disorders.

Table 9: TBI Onset Best Model Performance

Dataset	Outcome	AUC	F1	MCC
2017	2018	0.677	0.006	0.042
2018	2019	0.670	0.012	0.054
2017, 2018	2019	0.677	0.003	0.029

4.2.3 Sleep Apnea

As with TBI, in the absence of clinical data (sleep studies, sleep diaries, documentation of disturbed sleep patterns and breathing, etc.) used to assess and document the severity of OSA cases, we modeled only the initial onset of OSA. The results of our models are shown in Table 10.

Table 10: Sleep Apnea Onset Best Model Performance

Dataset	Outcome	AUC	F1	MCC
2017	2018	0.758	0.238	0.220
2018	2019	0.741	0.158	0.170
2017, 2018	2019	0.751	0.142	0.163

4.2.4 Chronic Obstructive Pulmonary Disease

As with TBI and OSA, in the absence of clinical data (spirometry measurement, measures of difficulty with exercise or other activities of daily living, etc.) used to measure the severity of COPD, we modeled only the initial onset of COPD. Model results are shown in Table 11. Although the goodness of fit is reasonably good, the F1 score and MCC are not.

Tailoring the model to improve its performance depends on the user's tolerance for false positive results and on the process by which an automated flag might be used operationally to identify patients at risk for developing COPD. Specifically, various other cut points could be used, or a measure such as "precision at k" could be used to identify those most at risk. These may prove promising avenues for future research.

Table 11: COPD Onset Best Model Performance

Dataset	Outcome	AUC	F1	MCC
2017	2018	0.844	0.183	0.242
2018	2019	0.844	0.131	0.194
2017, 2018	2019	0.844	0.132	0.194

4.3 Measuring Model Performance

For classification models, AUC scores were the primary methods of comparing performance across differing modeling strategies within a condition and for comparing the overall effectiveness of predictive modeling across conditions. In keeping with existing literature, we found the best-performing models used ensemble or boosting approaches, as seen in diabetes onset prediction. We also found that diabetes and COPD had the highest AUC scores for onset prediction, compared with OSA and TBI.

As noted under the COPD results, we will explore other ways of assessing model performance that align with how the prediction would be used, as a form of screening tool among many, for example, or as a high precision tool that would only identify those most at risk for developing the condition or having increased in severity.

We focused on MAE and the coefficient of determination R squared (R2) for regression models. MAE was chosen over mean squared error (MSE) because it is less sensitive to outliers.

4.3.1 Model Explainability

In these initial exploratory analyses, we did not focus on interpretability testing. For all models trained, we did calculate the permutation importance for each feature. Permutation importance is defined as the decrease in the model score when a particular feature is randomly shuffled. We utilized permutation importance as a model validity check, where we looked at the most important features to ensure we were not including codes (or groupings of codes) that should have been part of the cohort definition.

Although permutation importance measures how important a feature is to the performance of a particular model, it does not necessarily reflect the intrinsic predictive value of the feature itself.

Future work could include integrating feature importance and model explainability into the analytic pipeline.

5 Lessons Learned

Medical claims and patient eligibility data, also known as health services utilization data, are derived from reimbursement information and are required to determine payment by Medicaid, Medicare, or commercial payers. Claims data were not designed to understand the way patients and providers make decisions, nor were claims data designed to facilitate the prediction of medical outcomes.

In Appendix A, we provide additional detail on the potential value of incorporating commercially available EHR data in the analysis. Briefly, compared with claims data, EHR data would provide the following for predictive analytics purposes:

1. More complete and robust identification of condition onset and condition exacerbation
2. More timely insights, which would enable early diagnosis of at-risk patients much earlier than using claims that have 30-day, or longer, lag times in filing and processing
3. Richer data about a condition as opposed to just a diagnosis code, including laboratory results, results of sleep studies, imaging results and notes, clinical notes, and patient-based instruments, including activities of daily living, symptoms, behavioral health, pain scales, etc.
4. Richer data about the behavior and degree of compliance of the patient
5. Clinically relevant information that provides actionable results for patients and providers

Examples from peer-reviewed research indicate that adding clinical data to claims data related to heart failure hospitalizations significantly improved mortality prediction and helped many hospitals improve their performance ratings.[11]

Lastly, we describe lessons learned and potential next steps in developing predictive health outcomes models for FAA.

5.1.1 Results from Modeling

While claims data were never designed to be used for predicting clinical outcomes or for tracking the changes in health status over time, we did find them to add some degree of insight, including:

- According to standard model performance criteria, claims data are usable in predictive ML models.
- Claims data provide, at best, a proxy for the outcome of interest—the health status of a pilot and the impact of health status on flight capabilities.
- The ICD-10 coding system includes approximately 140,000 codes for procedures and diagnoses. Although dimension reduction can be accomplished with various methods, no single dominant technique exists.

[11] For example, see Hammill BG, Curtis LH, Fonarow GC, Heidenreich PA, Yancy CW, Peterson ED, Hernandez AF. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. Circ Cardiovasc Qual Outcomes. 2011 Jan 1;4(1):60-7. doi: 10.1161/CIRCOUTCOMES.110.954693. Epub 2010 Dec 7. PMID: 21139093.

- As estimated here using claims data alone, predictive models had reasonable AUC scores but poor F1 and MCC scores using a default threshold of 0.5. MCC scores, which balance false negatives and positives, were particularly poor, even with relatively low F1 scores.
- Existing FAA aeromedical examination data provide an essential end outcome but are not currently linkable to any other source of information for predictive modeling purposes.

5.1.2 Limitations of Existing Data

An overall discussion on the limits of administrative claims data is presented in Appendix A. The claims dataset used in this analysis—the MarketScan dataset—is quite large and robust in terms of linkage to inpatient, outpatient, and pharmaceutical claims, along with basic demographic information. As noted earlier, over 800 studies in PubMed in the past two years cite or use MarketScan data. Nevertheless, detailed aspects of patient demographics—including race, income, and place of work—are missing, as are clinical outcomes, vital signs, laboratory results, and other potentially valuable indicators.

Only three years of MarketSscan data were provided in the extract used for this study. This limits the ability to track long-term chronic illness or establish a long baseline before the first diagnosis of a condition. Additional years are available and would improve on these aspects of analysis.

5.1.3 Next Steps: Data Augmentation

Opportunities for augmenting existing data include linking claims data to other data sources, including the IBM Explorys EHR data, examining ways of using non-claims data to validate or extend findings, and expanding on modeling methods to incorporate other potentially informative aspects of existing or linked data. The following section describes some potential methods and approaches; it is not intended to be exclusive or prescriptive.

5.1.3.1 Develop Existing FAA Aeromedical Examination Data

The primary outcome of interest—pilot capability given health conditions—was not assessed directly here. Instead, various proxies were considered, including changes in overall health status given a specific diagnosed condition, the initial onset of a condition, and overall health status as described by chronic disease conditions. These measures are only indirectly related to pilot capabilities on the flight deck.

The FAA has aeromedical examination information for pilot certification using the FAA-specific diagnostic codes. While these data cannot be directly linked to claims information, FAA codes could be cross-referenced the ICD-10 coding system. The resulting data set could be used to validate the linkages between conditions and pilot certification outcomes.

5.1.3.2 Improve Feature Set and Model Disease Severity by Incorporating Electronic Health Record Data

As noted in Appendix A, EHRs have a wealth of structured and unstructured clinical data extensively used in predictive modeling and research. EHR data are particularly useful for predicting severity changes and incorporating lifestyle and other data elements not captured in claims. A summary of potential ways EHR data could be used for modeling is shown in Table 12.

Table 12: Potential Methods for Incorporating EHR Data into Predictive Models

Condition	EHR Data Elements	Purpose in Modeling
Any Condition	Blood pressure; body mass index; history of smoking, alcohol use, exercise; clinical notes, discharge summaries, other free text	Features, risk adjustment, severity
OSA	Results of sleep studies; home sleep tests including blood oxygen, airflow, breathing patterns; sleep diaries or other self-reported data	Severity
Diabetes	HbA1c score, comprehensive metabolic panel results, microalbumin urine test results, presence of neuropathy and vision problems	Severity
TBI	Glasgow Coma Scale score, Glasgow Outcome Score, Glasgow Outcome Score Extended, Activities of Daily Living, other instruments, e.g., Minnesota Multiphasic Personality Inventory, Patient Health Questionnaire-9, etc.	Severity
COPD	Global Initiative for Chronic Obstructive Lung Disease (GOLD) score by spirometric readings; Age Dyspnea, and Airflow Obstruction (ADO) index; Modified Medical Research Council (mMRC) dyspnea scale and FEV1 via spirometry.	Severity

5.1.3.3 Link Claims Data to Other Sources

A variety of external data sources have been incorporated into claims-based modeling. A commonly used method is to attach geographically based social determinants of health indicators to individuals residing in particular Zip Codes, using US Census American Community Survey and other data sources.[12] Other data linkages, for example, to Centers for Medicare and Medicaid Services (CMS) hospital star ratings, are also possible.[13] However, these general geographic methods would fail to provide individual-level data for pilot health state predictive modeling.

5.1.3.4 Leverage Temporal Nature of Diagnoses and Conditions

Multiple approaches have been applied to leverage information from the temporal order of events documented in the EHR. Two potential methods are a Reverse Time Attention (RETAIN) model using a time-decay to weight more recent visits in a sequence of visits as recorded in the EHR (over 900 citations) and the more recent BEHRT (Bidirectional Encoder Representations from Transformers for EHRs) with over 100 citations.[14] In each case, EHR events (diagnoses,

[12] See for example Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. J Am Med Inform Assoc. 2020 Nov 1;27(11):1764-1773. doi: 10.1093/jamia/ocaa143. PMID: 33202021; PMCID: PMC7671639. Blewett LA, Call KT, Turner J, Hest R. Data Resources for Conducting Health Services and Policy Research. Annu Rev Public Health. 2018 Apr 1;39:437-452. doi: 10.1146/annurev-publhealth-040617-013544. Epub 2017 Dec 22. PMID: 29272166; PMCID: PMC5880724.

[13] See for example Kurian N, Maid J, Mitra S, Rhyne L, Korvink M, Gunn LH. Predicting Hospital Overall Quality Star Ratings in the USA. Healthcare (Basel). 2021 Apr 20;9(4):486. doi: 10.3390/healthcare9040486. PMID: 33924198; PMCID: PMC8074583.

[14] The RETAIN model was first described in Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attenuation Mechanism. Adv Neaural Inf Process Syst. 2016 Jan. 30th Annual Conf on Neural Inf Process Syst. https://arxiv.org/abs/1608.05745 and https://github.com/mp2893/retain. The BEHRT model is described in Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R,

treatments, etc.) are featurized, encoded, and arrayed in time-order, so existing techniques for modeling event data in time or sequence can be modified and applied.

Other approaches, including graph and embedding techniques, have also been used and could be explored further in the EHR and claims data.

Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):7155. doi: 10.1038/s41598-020-62922-y. PMID: 32346050; PMCID: PMC7189231 with code in the repository https://github.com/deepmedicine/BEHRT

6 Conclusion

Overall, we demonstrated that we can intake and use a large commercial claims database for ML, and claims data can support usable predictive models to inform risk analysis during aeromedical certification decision making. We were able to construct prediction models for overall chronic disease burden and the onset of diabetes, TBI, OSA, and COPD with varying degrees of success. Diabetes and COPD were particularly amendable to predictive modeling of onset. We also found success in predicting increases in diabetes comorbidities and severity. We believe this initial exploratory work provides a solid foundation for future work in predicting disease onset and impacts on health outcomes that could pose aeromedical hazards, especially as we can link additional data for feature development and measurement of severity at the individual level.

Appendix A Value of Incorporating EHR Data

Administrative claims data represent only the tip of the iceberg, analytically speaking. The additional incremental predictive value would be provided by EHR, sociodemographic, and consumer transactional data.

Table A-1 below gives a side-by-side comparison of medical admin claims and EHR data.

Table A-1: EHR and Administrative Claims Data Comparisons

	EHR Data	Admin Claims Data
Data	Limited: Captures only the portion of care provided by doctors using the EHR	Broad: Captures information from all doctors/providers caring for a patient
Patients	All patients (including uninsured)	Captures insured patients only
Pharmacy	Contains on that a physician prescribed a drug but not whether or not it was filled/refilled	An accurate record of all prescriptions that were filled including dates of refills
Non-Prescription Drugs	Present	Not present
Data Richness	Rich: Lab results, vital signs, patient surveys, habits (smoking, etc.), problem list, etc.	Limited: Diagnosis, procedures

A.1 The Value of Administrative Claims Data

Administrative claims data provide value in health services research across a range of attributes:

Clinical Validity. Claims data contains information about covered services used by members in a program or line of business. Examples include admission and discharge dates, diagnoses, and procedures.

Source of Care. Demographic data, such as age, date of birth, race, place of residence, and date of death, are also included in these administrative datasets and are mainly considered reliable and valid. Files containing this information about all enrolled beneficiaries are known as "denominator" files.

Cost Effectiveness. Conducting research using claims data is a cost-effective way to analyze a large population segment, especially when considering the alternative of requesting individual patients' medical charts. The data also allow access to claims information across multiple providers for a given beneficiary while providing a consistent reporting format.

Ability to Link to External Data Sources. Below is a list of some external datasets that can be linked to the utilization and enrollment data. Because of privacy concerns, the typical linkage mechanism occurs at the U.S. Zip Code level:

- U.S. Census
- Cancer registries (e.g., Surveillance, Epidemiology, and End Results Program)
- Other providers (e.g., VA, Medicaid)

- National death index/State vital statistics
- Surveys (e.g., Health and Retirement Study)
- Provider information

Depending on the availability of identifying/common variables, other external data sources may be linked to the data. Linking can occur either at the group level (based on geography, place of service, etc.) or at the person level (through Social Security Number or other identification).

Data Availability. Medicare and commercial data files are complete and available relatively quickly after the close of a given calendar or fiscal year. For example, enrollment information for each calendar year contained in the Master Beneficiary Summary File is generally available the following fall. Similarly, calendar year utilization files are more than 98% complete by the summer of the following year and available for release soon after.

A.2 The Limitations of Administrative Claims Data

Record of Care Received. Conditions must be diagnosed to appear in the utilization files; however, some diseases, such as hypertension, depression, and diabetes, are often underdiagnosed. In addition, while the files provide a reliable record of the care received by the beneficiary, they do not provide information on the care needed. It is difficult to study disease recurrence in detail since all the data may reveal is the start of a new treatment. Another critical point is that services that providers know in advance will be denied may be inconsistently submitted as bills and, therefore, inconsistently recorded in the files.

Diagnosis Information. In some cases, diagnosis information may not be comprehensive enough to allow detailed analysis. For example, a cancer diagnosis can be found as an ICD-9 diagnosis code in the data (e.g., lung cancer is 162.xx), but no information on stage or histology is included in the claims data. While the data contain information on chronic diseases, knowing that someone has a chronic disease does not reveal how long they have had the condition (incidence vs. prevalence) or the severity of their condition.

Lack of Diagnosis Codes in Prescription Drug Event Files. Another limitation related to diagnosis information is that prescription drug event (PDE) files contain no diagnosis codes. Because many drugs and procedures have multiple indications, it can be challenging to interpret the reason for a given prescription.

Procedures by Care Setting - Inconsistencies in Use of Coding. Different care settings use different coding systems for procedures treated in inpatient and outpatient settings. For example, inpatient care is coded using ICD-10-PCS procedure codes, while physician/supplier and durable medical equipment data are coded using CPT and Healthcare Common Procedure Coding System (HCPCS) codes. Furthermore, hospital outpatient care is coded as a mix of CPT and revenue center (hospital billing center) codes. Currently, there exists a less-than-perfect crosswalk between ICD-10-PCS codes and CPT codes.

Limited Clinical Information. Physiological measurements such as blood pressure, pulse, and cardiac ejection fraction are absent from the utilization files. In addition, results of common tests such as prostate screens, angiography, and pathological tests are not included. Exact timing of events can be difficult to discern. Specifically, the time from admission to a given event or timestamps for dates of service cannot be found in the data.

Exclusions in Utilization Data. Outlined below are several types of services and care that are not contained in the claims data:

- Covered services for which claims are not submitted and, therefore, not included in the data (e.g., immunizations provided through a grocery-store chain or an employer-wellness clinic).
- Some services are not covered and would, therefore, not be included.
- Prior to the release of Medicare Advantage, encounter data contained little information (and of largely unknown quality).
- Encounter data reflect capitated arrangements and do not include information on payments to providers.

A.3 Incremental Value of EHR Data

So then, exactly what incremental value does the EHR data add beyond claims data?

- More complete and robust condition identification.
- More timely insights enable early diagnosis of at-risk patients much earlier than using claims.
- Sentinel monitoring, which entails post-approval safety of medications.
- Richer data about a condition as opposed to just a diagnosis.
- Richer data about the behavior and degree of compliance of the patient.
- A higher degree of actionable information from which to initiate and monitor treatment.

Healthcare organizations traditionally rely on EHR and medical/pharmacy claims data to generate insights into patient populations and treatment utilization in the real world. These real-world data are then used to guide the development, launch, and commercialization of new therapeutics and medical devices. However, while claims data capture care utilization across the healthcare system, it does not include information regarding treatment outcomes, details of diagnostic evaluations, or the patient experience.

Appendix B Abbreviations and Acronyms

Term	Definition
AAM	Aerospace Medical Research Division
ADO	Age Dyspnea, and Airflow Obstruction
AI	Artificial Intelligence
ARIMA	Auto-Regressive Integrated Moving Averages
AUC	Area Under the Curve
AUPRC	Area Under Precision-Recall Curve
BEHRT	Bidirectional Encoder Representations from Transformers
CAASD	Center for Advanced System Development
CCS	Clinical Classification Software
CDM	Common Data Model
CMS	Centers for Medicare & Medicaid Services
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease 2019
CPT	Current Procedural Terminology
CVD	Cardiovascular Disease
DCSI	Diabetes Complications Severity Index
DM	Diabetes Mellitus
DRG	Diagnosis Related Group
ECG	Electrocardiogram
EEG	Electroencephalogram
EHR	Electronic Health Record
FAA	Federal Aviation Administration
FHIR	Fast Healthcare Interoperability Resources
GBDT	Gradient Boosting Decision Tree
GOLD	Global Initiative for Chronic Obstructive Lung Disease
HbA1c	Glycated Hemoglobin
HCC	Hierarchical Condition Category
HCUP	Healthcare Cost and Utilization Project
HF	Heart Failure
ICD	International Classification of Diseases

LOS	Length of Stay
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MCC	Matthews Classification Coefficient
ML	Machine Learning
MSE	Mean Squared Error
mMRC	Modified Medical Research Council
NACC	National Alzheimer's Coordinating Center
NHANES	National Health and Nutrition Examination Survey
NLP	Natural Language Processing
NN	Neural Network
OLS	Ordinary Least Squares
OMOP	Observational Medical Outcomes Partnership
OSA	Obstructive Sleep Apnea
PDE	Prescription Drug Event
RETAIN	Reverse Time Attention
ROCC	Receiver Operating Characteristics Curve
SMS	Safety Management System
TANN	Time Aware Neural Network
TBI	Traumatic Brain Injury
WHO	World Health Organization

Appendix C Annotated Bibliography

Anderson, J. P., Parikh, J. R., Shenfeld, D. K., Ivanov, V., Marks, C., Church, B. W., Laramie, J. M., Mardekian, J., Piper, B. A., Willke, R. J., & Rublee, D. A. (2015). Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. *Journal of Diabetes Science and Technology*, *10*(1), 6–18. https://doi.org/10.1177/1932296815620200. PMID: 26685993; PMCID: PMC4738229. Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes - PMC (nih.gov)

- EHR data from Humedica for 2007–2012; 24,331 patients in integrated delivery networks. Modeled onset of DM from normal blood glucose, onset of prediabetes, onset of DM from prediabetes (all based on HbA1c scores). Ensemble model from normal blood glucose to DM: AUC 0.78.

Bose, S., Kenyon, C. C., & Masino, A. J. (2021). Personalized prediction of early childhood asthma persistence: A machine learning approach. *PloS One*, *16*(3), e0247784. https://doi.org/10.1371/journal.pone.0247784 PMID: 33647071; PMCID: PMC7920380. Personalized prediction of early childhood asthma persistence: A machine learning approach | PLOS ONE

- Included 9,934 patients with asthma diagnosed before age five. Pediatric EHR data follow-up for onset of chronic asthma. AUC 0.86 for Random Forest and XGBoost. Diagnoses grouped by adjusted clinical group. Only top 30 procedures used, medications grouped by therapeutic class, lab measures vetted by clinicians.

Chan, V., Hurst, M., Petersen, T., Liu, J., Mollayeva, T., Colantonio, A., Sutton, M., & Escobar, M. D. (2020). A population-based sex-stratified study to understand how health status preceding traumatic brain injury affects direct medical cost. *PloS One*, *15*(10), e0240208. https://doi.org/10.1371/journal.pone.0240208. A population-based sex-stratified study to understand how health status preceding traumatic brain injury affects direct medical cost | PLOS ONE

- Ontario Canada, administrative claims data. Predicted direct medical costs for TBI patients, with predictors demographics, if emergency, pre-injury health status, injury severity, mechanism of severity, etc. Linear regression models (R2 not reported).

Chang, Y., & Chen, X. (2021). Estimation of Chronic Illness Severity Based on Machine Learning Methods. *Wireless Communications and Mobile Computing*, Sep:1-13. https://doi.org/10.1155/2021/1999284 https://www.hindawi.com/journals/wcmc/2021/1999284/

- Using app data from Flaredown (published as part of Kaggle competition), self-reported severity predicted using demographics, trigger, other data from app. Did not report AUC but did report high F1 scores using XGBoost and ensemble methods, F1 of 0.93. https://www.kaggle.com/datasets/flaredown/flaredown-autoimmune-symptom-tracker

Croon, P. M., Selder, J. L., Allaart, C. P., Bleijendaal, H., Chamuleau, S. A. J., Hofstra, L., Isgum, I., Ziesemer, K. A., & Winter, M. M. (2022). Current state of artificial intelligence-based algorithms for hospital admission prediction in patients with heart failure: a scoping review. *European Heart Journal-Digital Health*, *3*(3), 415-425.; https://doi.org/10.1093/ehjdh/ztac035 https://academic.oup.com/ehjdh/advance-article/doi/10.1093/ehjdh/ztac035/6617146?login=true

- Scoping review on use of ML for heart failure predictive models. Sixteen studies on readmission: AUC 0.61-0.79, with highest using features from EHR and image data; claims-based model achieved AUC of 0.64. Readmission is difficult to predict; mortality is much easier to predict, as is highest cost outliers.

Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Network Open*, *3*(1), e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962. PMID: 31922560; PMCID: PMC6991258. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes - PubMed (nih.gov)

- Prognostic study of 9,502 Medicare-enrolled subjects with heart failure. All-cause mortality, hospitalization, top cost, home days loss greater than 25%, modeled using logistic and a variety of ML models. Only marginally better than logistic; Gradient Boost model performed best: mortality 0.73, hospitalization for HF 0.75, high cost 0.73, home days loss 0.79. Improved AUPRC when EHR data are added.

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, *19*(1), 211. https://doi.org/10.1186/s12911-019-0918-5. PMID: 31694707; PMCID: PMC6836338. https://pubmed.ncbi.nlm.nih.gov/31694707/

- Used NHANES interview and other data including blood pressure, labs, to predict incidence of diabetes and CVD. Ensemble model for CVD had AUC 0.83 without labs and 0.84 with labs; gradient boost for diabetes AUC 0.86 without labs and 0.95 with labs.

Glasheen, W. P., Renda, A., & Dong, Y. (2017). Diabetes Complications Severity Index (DCSI)-Update and ICD-10 translation. *Journal of Diabetes and Its Complications*, *31*(6), 1007–1013. https://doi.org/10.1016/j.jdiacomp.2017.02.018. PMID: 28416120. https://pubmed.ncbi.nlm.nih.gov/28416120/

- ICD-9 codes for secondary and primary diabetes plus all five ICD-10 diabetes categories were incorporated into an updated tool. Additional modifications were made to improve the accuracy of severity assignments. Tools were tested in a Medicare Advantage population. In type 2 subpopulation, prevalence steadily declined with increasing score according to the updated DCSI tool; the original tool resulted in an aberrant local prevalence peak at DCSI = 2. In type 1 subpopulation, score prevalence was greater in type 1 vs. type 2 subpopulations (3 vs. 0). Both instruments predicted current-year inpatient admissions risk and near-future mortality, using either purely ICD-9 data or a mix of ICD-9 and ICD-10 data.

Goto, T., Jo, T., Matsui, H., Fushimi, K., Hayashi, H., & Yasunaga, H. (2019). Machine Learning-Based Prediction Models for 30-Day Readmission after Hospitalization for Chronic Obstructive Pulmonary Disease. *COPD*, *16*(5-6), 338–343.

https://doi.org/10.1080/15412555.2019.1688278. PMID: 31709851. https://www.tandfonline.com/doi/full/10.1080/15412555.2019.1688278

- Predicting COPD readmissions within 30 days of discharge, in administrative claims data against logistic regression reference model. C-statistics only modestly successful – AUC of 0.61 to 0.645, positive predictive values <0.10. Followed up ML model with a lasso regression to test significance of specific variables.

James, C., Ranson, J. M., Everson, R., & Llewellyn, D. J. (2021). Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients. *JAMA Network Open*, *4*(12), e2136553. https://doi.org/10.1001/jamanetworkopen.2021.36553. PMID: 34913981; PMCID: PMC8678688. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2787228

- Using memory clinic patient-level data from NACC uniform data set including socio-demographics, family history, symptoms, functional score from neuropsychological test, predict onset of dementia. XGBoost showed accuracy of 0.92 and AUC 0.92. Outperforming regression models with relatively few variables in model.

Lewis, M., Elad, G., Beladev, M., Maor, G., Radinsky, K., Hermann, D., Litani, Y., Geller, T., Pines, J. M., Shapiro, N. L., & Figueroa, J. F. (2021). Comparison of deep learning with traditional models to predict preventable acute care use and spending among heart failure patients. *Scientific Reports*, *11*(1), 1164. https://doi.org/10.1038/s41598-020-80856-3. https://www.nature.com/articles/s41598-020-80856-3.pdf

- Twelve years of commercial claims for heart failure patients: preventable hospitalization in six-month period; emergency admission in six-month period; preventable costs (inpatient, emergency) in one-year period. Compared knowledge-driven and data-driven features. Knowledge: demographics, readmission, comorbidity indicators, etc. Generated word2vec embeddings on all medical codes for feature, both over entire patient history (FNN, GBM models) and using sequential monthly vectors (LSTM, CNN). Sequential deep learning model scored highest with AUC Preventable hospitalization: 0.78, Preventable ED: 0.68, Preventable costs: 0.73
- All models were 11 to 15 points higher than conventional regression model with knowledge-driven features, but range is still less than desirable. Also evaluate a "precision at k" score —for the top k% scored as being "at risk" how many predicted to have preventable admissions had an admission (PPV)—achieved scores above 40% at 1% top scores for inpatient admissions and emergency visits.

Lu, H., & Uddin, S. (2021). A weighted patient network-based framework for predicting chronic diseases using graph neural networks. *Scientific Reports*, *11*(1), 22607. https://doi.org/10.1038/s41598-021-01964-2. PMID: 34799627; PMCID: PMC8604920. https://www.nature.com/articles/s41598-021-01964-2

- Study restructured 1.2 million Australian claims records as graph neural network of patients sharing similar diagnoses (grouped by Elixhauser comorbidity categories), achieving accuracy of 0.93 for CVD and 0.89 for COPD onset prediction.

MacKay, E. J., Stubna, M. D., Chivers, C., Draugelis, M. E., Hanson, W. J., Desai, N. D., & Groeneveld, P. W. (2021). Application of machine learning approaches to administrative claims data to predict clinical outcomes in medical and surgical patient populations. *PloS*

One, *16*(6), e0252585. https://doi.org/10.1371/journal.pone.0252585. PMID: 34081720; PMCID: PMC8174683. https://pubmed.ncbi.nlm.nih.gov/34081720/

- CMS AI challenge entrant. Thirty-day mortality in Medicare beneficiaries with an inpatient hospitalization 2009–2011. Extreme gradient boosted tree performed best: AUC 0.88 for mortality, 0.80–0.86 for 17 adverse events. Rehospitalization AUC 0.73. (i.e, harder problem, more random). Demographics, state/county, ESRD/disability indicator, ICD-9-CM diagnoses grouped in HCC categories, procedures grouped in CCS categories, DRG code, linked to US Census data for patient Zip Code.

Matsuo, K., Aihara, H., Nakai, T., Morishita, A., Tohma, Y., & Kohmura, E. (2020). Machine Learning to Predict In-Hospital Morbidity and Mortality after Traumatic Brain Injury. *Journal of Neurotrauma*, *37*(1), 202–210. https://doi.org/10.1089/neu.2018.6276. Epub 2019 Sep 18. PMID: 31359814.

- Predicting in-hospital mortality using random forest on clinical features: Glasgow Coma Scale, blood pressure, pupillary response, computed tomography findings, lab values. Included 232 patients. RF for mortality: 0.90 AUC with 100% sensitivity and 72.3% specificity, 91.7% accuracy.

Mencar, C., Gallo, C., Mantero, M., Tarsia, P., Carpagnano, G. E., Foschino Barbaro, M. P., & Lacedonia, D. (2020). Application of machine learning to predict obstructive sleep apnea syndrome severity. *Health Informatics Journal*, *26*(1), 298–317. https://doi.org/10.1177/1460458218824725. Epub 2019 Jan 30. PMID: 30696334. https://journals.sagepub.com/doi/10.1177/1460458218824725?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

- Study of 313 patients with sleep apnea diagnosis, demographics, spirometry, gas exchange, symptoms as features to predict severity. Random Forest worked better for dichotomized outcome, regression model better for index value. Not very strong: accuracy of 44.7%

Min, X., Yu, B., & Wang, F. (2019). Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. Scientific reports, 9(1), 2362. https://doi.org/10.1038/s41598-019-39071-y. PMID: 30787351; PMCID: PMC6382784. https://pubmed.ncbi.nlm.nih.gov/30787351/

- Tested difference between "knowledge driven features"—HOSPITAL score, LACE index, other clinically relevant features—and "data driven features"—diagnosis codes (raw, first three digits of ICD-9-CM, CCS, HCC); procedures (CCS, BETOS, revenue codes); pharmacy (NDC mapped to Generic Therapeutic Codes); these features were either used as counts per period, present/not present in period, or a TF-IDF reduction. Used claims data from year before index admission to predict readmission within 30 days of hospital discharge.
- Tested temporal structure of data, including embedding based on time window per person, weighting event pairs based on time distance, and med2vec representation of time. In neural net models, sequence representation, matrix representation using binned time intervals, matrix representation with sparse matrix representation of continuous time.
- Model fit was generally poor. AUC scores by approach:

  - Knowledge features: 0.61–0.64
  - Data-driven features: 0.64–0.65
  - Gradient-boosting decision tree had highest score: 0.65
  - CNN and other neural networks: 0.65

Molani, S., Hernandez, P. V., Roper, R. T., Duvvuri, V. R., Baumgartner, A. M., Goldman, J. D., Ertekin-Taner, N., Funk, C. C., Price, N. D., Rappaport, N., & Hadlock, J. J. (2022). Risk factors for severe COVID-19 differ by age for hospitalized adults. *Scientific Reports*, *12*(1), 6568. https://doi.org/10.1038/s41598-022-10344-3. PMID: 35484176; PMCID: PMC9050669. https://www.nature.com/articles/s41598-022-10344-3

- EHR data from 6,906 hospitalized patients examined for COVID-19 outcomes. Using data present in first seven hours to predict mild/severe outcomes (dichotomized WHO severity scale). GBDT had highest AUC and accuracy scores, but all performed well (AUC 0.78– 0.83 for GBDT all ages, by age group).

Muro, S., Ishida, M., Horie, Y., Takeuchi, W., Nakagawa, S., Ban, H., Nakagawa, T., & Kitamura, T. (2021). Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study. *JMIR Medical Informatics*, *9*(7), e24796. https://doi.org/10.2196/24796. PMID: 34255684; PMCID: PMC8293159. Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study - PMC (nih.gov)

- Annual medical checkup records for employees, 1998–2019. Included 24,815 subjects. Predict onset of COPD diagnosis. XGBoost AUC 0.96 with accuracy of 91.7%, sensitivity of 84.5%, specificity of 96%. Lung function test results, smoking status were most important factors.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., Mossin, A., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*, 18. https://doi.org/10.1038/s41746-018-0029-1. PMID: 31304302; PMCID: PMC6550175. https://www.nature.com/articles/s41746-018-0029-1

- Collaboration between Google Health and UCSF Medical Center, modeled hospital EHR data for 216,221 hospitalizations. Outcomes modeled were in-hospital mortality, unplanned readmission within 30 days, long length of stay, discharge diagnosis. Created vector of time-sequenced tokens using FHIR for each patient visit, using all EHR data including free text. Used LSTM, TANN, a NN with boosted time-based decision "stumps."
- In-hospital mortality: AUC 0.95 (vs. 0.85 with logistic regression model); LOS: AUC 0.86 (vs. 0.76 for regression model); Discharge diagnosis: AUC 0.87
- Note that this is an easier problem than others—death in hospital from hospital data.

Ramachandran, A., & Karuppiah, A. (2021). A Survey on Recent Advances in Machine Learning Based Sleep Apnea Detection Systems. *Healthcare (Basel, Switzerland)*, *9*(7),

914. https://doi.org/10.3390/healthcare9070914. PMID: 34356293; PMCID: PMC8306425. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8306425/

- Qualitative assessment of multiple ML models and data sources, primarily focusing on clinical and sensor data (e.g., ECG, EOG, EEG; BP, HR; other sensor data) to diagnose OSA or provide risk scoring for severity. A summary of sensor data ML analyses is provided (Appendix A).

Ravaut, M., Harish, V., Sadeghi, H., Leung, K. K., Volkovs, M., Kornas, K., Watson, T., Poutanen, T., & Rosella, L. C. (2021). Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. *JAMA Network Open*, *4*(5), e2111315. https://doi.org/10.1001/jamanetworkopen.2021.11315. PMID: 34032855; PMCID: PMC8150694. Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes - PMC (nih.gov)

- Ontario administrative data set of 1,657,395 patients over 10 years. XGBoost decision tree model. Two-year block of patient history and yes/no diabetes five years later. Simulates a screening process to predict DM onset. Demographics, geographic linked data, usage, lab results, hospitalizations, prescription history, etc.

Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*, *2*(2), 22. https://doi.org/10.1007/s43674-022-00034-y. PMID: 35434723; PMCID: PMC9006199. Machine learning for diabetes clinical decision support: a review - PMC (nih.gov)

- Review of 86 studies on diabetes risk scoring, diagnosis, prognosis. All using clinical features derived from EHR, labs, medical imaging, other medical sensor data. AUCs vary from 0.70–0.90 for diagnosis, 0.76-0.87 for prognosis/disease progression.