ORIGINAL RESEARCH

Body Mass Index Variable Interpolation to Expand the Utility of Real-world Administrative Healthcare Claims Database Analyses

Bingcao Wu • Wing Chow • Monish Sakthivel • Onkar Kakade • Kartikeya Gupta • Debra Israel • Yen-Wen Chen • Aarti Susan Kuruvilla

B. Wu (✉) • W. Chow • D. Israel • Y.-W. Chen

Janssen Scientific Affairs, LLC, Titusville, NJ, USA

Email: bwu34@its.jnj.com

M. Sakthivel • O. Kakade • K. Gupta • A. Susan Kuruvilla

Mu Sigma Business Solutions, LLC, Bengaluru, India

ABSTRACT

Introduction: Administrative claims data provide an important source for real-world evidence (RWE) generation, but incomplete reporting, such as for body mass index (BMI), limits the sample sizes that can be analyzed to address certain research questions. The objective of this study was to construct models by implementing machine learning (ML) algorithms to predict BMI classifications (≥30, ≥35, and ≥40 kg/m^2) in administrative healthcare claims databases and then internally and externally validate them.

Methods: Five advanced ML algorithms were implemented for each BMI classification on a random sampling of BMI readings from the Optum PanTher Electronic Health Record database (2%) and the Optum Clinformatics Date of Death (20%) database, while incorporating baseline demographic and clinical characteristics. Sensitivity analyses with oversampling ratios were conducted. Model performance was internally and externally validated.

Results: Models trained on the Super Learner ML algorithm (SLA) yielded the best BMI classification predictive performance. SLA model 1 utilized sociodemographic and clinical characteristics, including baseline BMI values; the area under the receiver operating characteristic curve (ROC AUC) was approximately 88% for prediction of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2 (internal validation), while accuracy ranged from 87.9% to 92.8% and specificity ranged from 91.8% to 94.7%. SLA model 2 utilized sociodemographic information and clinical characteristics, excluding baseline BMI values; ROC AUC was approximately 73% for prediction of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2 (internal validation), while accuracy ranged from 73.6% to 80.0% and specificity ranged from 71.6% to 85.9%. The external validation on the MarketScan Commercial Claims and Encounters database yielded relatively consistent results with slightly diminished performance.

Conclusion: This study demonstrated the feasibility and validity of using ML algorithms to predict BMI classifications in administrative healthcare claims data to expand the utility for RWE generation.

Keywords: Administrative healthcare claims databases; BMI classification; Body mass index; Machine learning; Predictive models; Real-world evidence generation

Key Summary Points

Why carry out this study?

- The sizeable underreporting of body mass index (BMI) data in administrative healthcare claims databases impedes the comprehensive study of the population with obesity and improved methodology is needed.
- To address this need for improved methodology, we have harnessed machine learning techniques to interpolate BMI variable data.

What was learned from the study?

- Based on this study, machine learning algorithms can be applied to administrative healthcare claims data to predict BMI classifications with high validity.
- This novel approach can be leveraged across multiple therapeutic areas to better understand variations in BMI related disease risk, treatment outcomes, healthcare resource use and costs in real-world settings.
- The strategic machine learning approach undertaken in this study may also be relatively easily applied to the development of similar predictive models for other underreported clinical variables in administrative healthcare claims databases.

DIGITAL FEATURES

This article is published with digital features, including a summary slide, to facilitate understanding of the article. To view digital features for this article go to https://doi.org/10.6084/m9.figshare.13359923.

INTRODUCTION

Real-world evidence (RWE) generated from administrative healthcare claims databases are valuable to understand patient characteristics, health outcomes, and health economics at the population-level [1, 2]. Such administrative healthcare claims database analyses are increasingly being utilized for clinical evidence generation and complement the evidence generated from randomized clinical trials and other clinical intervention studies [1, 2]. The findings of claims-based studies are informative to many healthcare system stakeholders, including providers and payers, federal and local government agencies, pharmaceutical/medical device companies, and patients [1, 2]. Although administrative healthcare claims database analyses provide several advantages (eg, large heterogenous populations, rare event capture, low cost, short timeframes for completion) [2], they also have limitations, including the incomplete reporting of certain clinical variables in the data sources. This creates obstacles to the comprehensive and accurate understanding of patient characteristics and outcomes.

One notable example of such a clinical variable is body mass index (BMI), a biometric measure that has been used in the risk assessment of many health conditions, with a BMI of 30 kg/m^2 or greater indicating the medical condition of obesity in adults and greater health risk [3, 4]. National organizations in the US, such as the Centers for Disease Control and Prevention, have stratified obesity into 3 severity classifications, BMIs 30 to <35 kg/m^2, BMIs 35 to <40 kg/m^2, and BMIs ≥40 kg/m^2, which are reflective of increasing health risks [3, 4]. BMI is predictive of greater risk for multiple disease conditions, including metabolic syndrome, type II diabetes, cardiovascular disease, some cancers, liver and kidney disease, arthritis, asthma, and depression, as well as greater risk for all-cause mortality [4-6]. Additionally, variations in BMI are predictive of healthcare resource utilization and costs [7-9]. The health risks associated with obesity and its high prevalence in the US [4] necessitates study of populations with obesity on several inter-related facets, such as population sociodemographic and clinical characteristics, current and emerging health outcomes and costs, value of therapeutic interventions, patient-drug/procedure interactions, etc. However, in an administrative healthcare claims database analysis, in which 746,763 health plan members in years 2013 through 2016 were included, it was reported that BMI value diagnoses were coded for only 14.6% [10]. The sizeable underreporting of BMI data in administrative healthcare claims databases impedes the comprehensive study of the population with obesity and improved methodology is needed.

To address this need for improved methodology, we have harnessed machine learning (ML) techniques to interpolate BMI variable data. ML is a rapidly advancing field and refers to algorithms and statistical methodologies that are used to build analytical models based on systems learning from data, identifying patterns, and yielding decisions [11]. Such statistical tools, including gradient boosted decision trees, least absolute shrinkage and selection operator (LASSO) regression, random forest, and artificial neural networks (NN), can be applied to raw data sets for the imputation of missing data, replacement of outliers, feature extraction, statistical classification, and optimization of predictive model accuracy [12]. Among other applications, ML techniques have been shown in multiple RWE studies to be useful for model development for the prediction of diagnoses, clinical variables, and disease risk [12-20]. The objective of this study was to construct models by implementing ML algorithms to predict BMI classifications (≥30, ≥35, and ≥40 kg/m^2) in administrative healthcare claims databases and then internally and externally validate them, and thereby expand the utility for RWE generation of administrative healthcare claims database analyses.

METHODS

Data Sources

Three real-world US administrative healthcare databases were utilized in this study, the Optum PanTher Electronic Health Record database (Optum EHR), the Optum Clinformatics Date of Death (Optum DOD) database, and the IBM MarketScan Commercial Claims and Encounters (IBM CCAE) database. Both the Optum EHR and DOD databases were used for model development and validation purposes. The IBM CCAE database was used as the external validation database in this study. All datasets were from databases of de-identified patient data and so ethics committee approval was not required.

The Optum EHR multi-dimensional database contains de-identified information on outpatient visits, diagnostic procedures, medications, laboratory results, hospitalizations, clinical notes, and patient outcomes primarily from Integrated Delivery Networks. The EHR data encompasses >80 million patients with ≥7 million from each US census region. The database contains a provider network of over 140,000 providers at >700 hospitals and 7,000 clinics with broad geographical representation.

The Optum DOD longitudinal administrative claims database is comprised of claims data from United Healthcare (UHC) fully insured patients, UHC administrative services only, Medicaid, and legacy Medicare Choice membership. The data includes integrated enrollment, inpatient, outpatient, and

outpatient pharmacy claims for >80 million unique de-identified members since 2000.

The IBM CCAE database is a longitudinal administrative claims database comprised of de-identified data from individuals enrolled in employer-sponsored insurance health plans. The data includes inpatient, outpatient, and outpatient pharmacy claims, as well as enrollment data, from large employers and health plans who provide private healthcare coverage to >140 million employees, their spouses, and dependents.

Study Methodology Flow

The Optum EHR and the Optum DOD databases were used to supply training datasets for the 5 advanced ML algorithms that were implemented to construct the predictive models of each BMI classification. The constructed predictive models were then internally validated on the Optum databases and externally validated on the IBM CCAE database. The methodology flow of this study is depicted in Fig. 1 and involved 6 steps, 1) data extraction, 2) feature aggregation, 3) exploratory data analysis, 4) feature engineering, 5) modelling and sensitivity analysis, and 6) model selection. The primary goal was to implement predictive models to interpolate BMI classifications within claims data representative of large populations.

Data Extraction

All datasets for the study populations were extracted from the Optum EHR, Optum DOD, and IBM CCAE databases during January 1, 2013 to December 31, 2019, based on the latest data available at the time of assessment. All datasets were from databases of de-identified patient data. A BMI reading was identified either from a BMI observation (numeric value) in the Optum EHR dataset or from an International Classification of Diseases (ICD)-9/10 diagnosis code indicating a BMI classification in the claims data sources. Each BMI reading (observation/diagnosis) during the study intake period from January 1, 2014 to December 31, 2019 was indexed on the event date as a reading and one person may have contributed multiple readings. Sociodemographic information, including age, gender, US region, and US regional division, and clinical characteristics, including all recorded medical diagnoses, medications, and procedures, were extracted at the index date and during the corresponding 12-month baseline periods, separately for each index BMI reading. Additionally, BMI readings for each quarter prior to the index reading were extracted. Data extraction was performed on disease agnostic populations (ie, not a subset population with a specific disease). The codes and descriptions of all

sociodemographic information and clinical characteristics extracted for the study populations are provided in the online supplement.

Feature Aggregation

The diagnosis codes (ICD-9/-10) and procedure codes (ICD-9/-10; Current Procedure Terminology [CPT-4] codes; Healthcare Common Procedure Coding System [HCPCS] codes) were grouped using Clinical Classification Software (CCS), while medication codes were grouped using the Generic Product Identifier (GPI). Such groupings were used to increase the ease of computation and clinical interpretation.

Exploratory Data Analyses

To understand the distribution of data, extensive exploratory data analyses were performed to identify any data anomalies and reduce data dimensions. Table 1 shows the results of the different models across the 3 databases.

Feature Engineering

Given that both the dependent variables (BMI classifications) and all potential features, except age, were dichotomous, random forest methods and Chi-square tests were used to identify the features that were significantly associated with each BMI classification. Firstly, the random forest algorithm was used to rank the features by feature importance score. Then, the top ranked features were cross validated using the Chi-square test. Feature selection was performed in the Optum EHR and DOD databases, separately. Due to the constraint of computation power and the large available sample size, only 2% of random samples from the Optum EHR database and 20% from the Optum DOD database at a time were used in the feature selection analyses. To reduce selection bias, the random samples were bootstrapped 5 times performing the same analyses in each iteration. Out of the entire 1,266 available features from the Optum EHR and DOD databases, 379 features that were consistently identified across the two databases and 5 iterations were finally selected for the predictive models (Fig. 2). On the 379 features, again a feature selection process was carried out and the top 100 features were selected (Table 2). Since the models performed better when they were trained on the set of 100 features, all 5 of the ML algorithms were trained using the top 100 selected features.

Modelling and Sensitivity Analysis

Binary classification models were developed for the following BMI classifications: BMI=30 kg/m^2 (model output=1, if BMI ≥30; =0 if BMI <30); BMI=35 kg/m^2 (model output=1, if BMI ≥35; =0 if BMI <35); BMI=40 kg/m^2 (model output=1, if BMI ≥40; =0 if BMI <40). Considering some patients may have historical BMI data available in the baseline, which could be a strong predictor, while others do not, 2 models were developed for each of the BMI classifications to account for these 2 scenarios. The first model (model 1) included the baseline BMI feature in addition to the other 100 selected features and was only trained among patient cohorts with baseline BMI data available. The second model (model 2) was built on the 100 selected features and was trained on patient cohorts without baseline BMI data. Four mathematically different algorithms were implemented on the models, Catboost, random forest, LASSO, and NN. Catboost and random forest provide nonlinearity due to their tree-based approach, while LASSO is a linear model. NN provide a computation intensive approach based on various activations. With these 4 mathematically different algorithms, we ensured use of varied ML techniques to address our research objective. Additionally, both models 1 and 2 were trained with a novel automated (self-assigned/calculated) weighted prediction approach (Super Learner algorithm; SLA), which leveraged the prediction from the 4 different ML algorithms through a logistic regression with 5 bootstrapped random samples from the Optum EHR and DOD databases.

In addition to use of varied ML techniques, several sensitivity analyses were performed to pursue optimal model performance. First, as previously mentioned, the models were examined using the full 379 features versus only using the top 100 features; the latter yielded better performance due to less overfitting. Second, model performance was compared when measuring clinical characteristic features on a quarterly basis versus on a yearly basis during the 12-month baseline period. The results indicated better performance using the yearly measured features. In addition, due to the rarity of BMI ≥35 and ≥40 kg/m^2 classifications in the populations, an oversampling technique was applied to improve the model sensitivity; 3 oversampling ratios, 50/50, 60/40, and 70/30, were evaluated while creating the training datasets. Furthermore, hyperparameter tuning was performed for all the algorithms to maximize the performance of the models. Lastly, the models were trained separately on male and female cohorts; however, no significant improvement was observed compared to the models training on the gender-combined cohort.

Predictive Model Performance

The performance of predictive models 1 and 2 was internally examined in the Optum databases and externally tested in the IBM CCAE database (Fig. 3). The performance was assessed by area under the receiver operating characteristic curve (ROC AUC), F1 score, accuracy, negative predictive value (NPV), specificity, precision, and recall.

RESULTS

The best algorithms of the models and the oversampling ratios across all the iterations of BMI classifications are shown in Table 3. The SLA on the top 100 features was the best ML algorithm for both models with a 50/50 oversampling ratio for the BMI ≥30 kg/m^2 classification and a 60/40 oversampling ratio for the BMI ≥35 and ≥40 kg/m^2 classifications.

Internal Validation

Implementing the SLA on model 1 and internally validating on the Optum DOD database, yielded ROC AUC values of approximately 88% for prediction of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2, while accuracy ranged from 87.9% to 92.8%, F1 score ranged from 77.3% to 87.7%, and specificity ranged from 91.8% to 94.7% (Fig. 4). Implementing the SLA on model 2 and internally validating on the Optum DOD database, yielded ROC AUC values of approximately 73% for prediction of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2, while accuracy ranged from 73.6% to 80.0%, F1 score ranged from 48.1% to 74.6%, and specificity ranged from 71.6% to 85.9% (Fig. 5). Detailed predictive performance results of models 1 and 2 trained on the Optum DOD database and internally validated on the Optum DOD database are shown in Supplementary Table 1.

External Validation

The external validation on the IBM CCAE database yielded relatively consistent results with slightly diminished performance as expected. Implementing the SLA on model 1 and externally validating on the IBM CCAE database, yielded ROC AUC values ranging from 78.7% to 83.6% for prediction of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2, while accuracy ranged from 84.0% to 90.0%, F1 score ranged from 66.9% to 81.8%, and specificity ranged from 90.5% to 95.5% (Supplementary Table 2). Implementing the SLA on model 2 and externally validating on the IBM CCAE database, yielded ROC AUC values ranging from 67.1% to 71.4% for prediction of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2, while accuracy ranged from 69.5% to 74.4%, F1 score ranged from 40.6% to 69.7%, and specificity

ranged from 70.6% to 83.7% (Supplementary Table 2). Detailed predictive performance results of models 1 and 2 trained on the Optum DOD database and Optum EHR databases and internally and externally validated are shown in Supplementary Tables 2-5.

DISCUSSION

In this study, we implemented multiple ML algorithms to construct and optimize predictive models and then applied the models to administrative healthcare claims databases to predict BMI ranges and thereby expand the coverage of BMI data in such data sources. The 2 SLA-based models exhibited the best predictive capabilities of BMI classifications of ≥30, ≥35, and ≥40 kg/m^2. Model 1 (ROC AUC values of approximately 88% across the 3 predicted BMI classifications [internal validation]; 79%-84% [external validation]), which included baseline BMI data, performed better than model 2, which did not include baseline BMI data. However, in the absence of baseline BMI data, model 2 yielded satisfactory performance, with ROC AUC values of approximately 73% across the 3 predicted BMI classifications (internal validation); 67%-71% (external validation). Applying these predictive models to administrative healthcare claims data sources in real-world database studies will potentially produce a better understanding for researchers, healthcare providers, payers, patients, and other healthcare system stakeholders of variations in sociodemographic data, health outcomes, healthcare costs, responses to therapeutic interventions, patient-drug/procedure interactions, etc. among persons with different BMI classifications.

Prior research studies using administrative healthcare data sources have repeatedly shown that BMI is substantially underreported [10, 14, 21, 22]. Martin et al conducted a study (2002-2008 patient cohorts) in which an administrative database was referenced to a clinical registry database, and reported low sensitivity (7.75%) but high specificity (98.98%) for detecting obesity based on diagnosis (ICD-10 diagnosis codes E65-E68) in the administrative data source [22]. The obesity prevalence in the administrative database was only 2.4% compared to the 20.3% prevalence observed among a patient cohort in the clinical registry database [22]. In a more recent study (2013-2016) of administrative EHR and claims data (Optum Integrated Claims database), Ammann et al reported that among 746,763 plan members, 14.6% had BMI-related diagnoses coded [10]. In this study the ICD-9/-10 codes had a satisfactory predictive value (>70%) across different BMI classifications, meaning that the claims-based diagnoses were fairly accurate [10]. However, their sensitivity was relatively low at <30% [10]. This low sensitivity may in part be attributed to a skewed BMI distribution in the claims data towards the morbid

and obese population due to those individuals with a BMI in the normal range or being mildly overweight not having a recorded ICD-9-10 diagnosis and thus are underrepresented. Other reasons for limitations of BMI data in administrative healthcare claims databases include that obesity remains underrecognized as an actual disease and there is a coding focus during data extraction on more obvious clinical disease categories; in physician notes as well, the term obesity is frequently not mentioned or more loosely termed [22]. In light of the findings of Martin et al and Ammann et al, the predictive models of BMI classifications constructed in our study with ML using 100 key patient features significantly improve the prediction of obesity of patient cohorts represented in administrative healthcare claims databases, especially the sensitivity (ie, the low sensitivity and skewed BMI distribution only will impact the imbalance of the BMI classifications in the training data, which was partially addressed herein through oversampling techniques). The high specificity of the claims-based BMI data ensures a high degree of internal validity during the model development and validation.

Based on cross-sectional National Health and Nutrition Examination Survey data, among adults (>20 years of age) in the US in 2015-2016, the prevalence of obesity was 39.6% and among youths (2-19 years of age) it was 18.5% [23]. Although the increase in obesity may appear to be stabilizing, at least compared to 2013-2014 survey data in the US, the obese population represents a significant proportion of the overall US population. Large administrative healthcare claims data sources with ML algorithms implemented can supplement such nationwide survey data to provide more complete datasets of subpopulations, in this instance also stratified by BMI classes. Moreover, the large amount of sociodemographic and clinical characteristic data contained in such sources can be helpful to understand the prevalence of obesity, as well as the extent of underdiagnosis, across many different subpopulations (ie, US geographic regions, age groups, health insurance types, disease categories, etc). Together with the multitude of comorbidities (diabetes, cardiovascular disease, cancer, etc) associated with obesity and increased healthcare costs corresponding with higher BMI classifications [4-9], it is important to use big data sources to understand at the population-level variations in health outcomes of those with obesity; the stratification by BMI classifications may provide a more in-depth understanding of the impact of obesity severity on health outcomes as well. The utility of this RWE generation, particularly when combined with other big data technologies (eg, genomics, metabolomics, information collected by personal monitoring devices-GPS, Fitbit), can be explored under the infrastructures of health system disease management and public health surveillance and interventions [24, 25].

The results of this study and application of the constructed predictive models of BMI classifications in secondary database analyses have certain limitations. Firstly, the clinical utility of BMI in the assessment of risks for obesity-related comorbidities is well recognized at the population level; however, it can sometimes be less useful in assessing the risks of obesity-related comorbidities among individual patients due to the heterogeneity in fat distribution across people in general, males and females, age groups, race/ethnic groups, etc., [26]. Despite having some shortcomings as a clinical biomarker, BMI is a widely accepted useful tool in clinical practice, especially when applied with other clinical measures of cardiometabolic risk factors [26]. Secondly, the databases we utilized are comprised of administrative healthcare data mostly from a single channel of insured members (all age groups) across the US, and thus, the predictive models may not be generalizable to healthcare systems outside of the US, and the performance may vary among subpopulations in specific states or regions, or specific age groups (eg, pediatric population). However, the Optum DOD database does contain a portion of patients with Medicaid and legacy Medicare Choice membership. Also, the administrative healthcare data sources we used to build and train the predictive models are subject to potential coding errors, inconsistencies, and incompleteness. Additionally, the presence of a diagnosis code on a medical claim does not guarantee positive presence of a disease, as the diagnosis code may be incorrectly coded or included as a rule out criteria.

While other methodologies, such as simpler regression analyses, may be useful for prediction of BMI classifications in certain instances, our primary goal was to implement and optimize predictive models to interpolate BMI classifications within claims data representative of large populations. The constructed predictive models provided a robust solution to achieve our research objective based on several strengths. First, we used both EHR and claims data sources, giving both provider and payer perspectives, to select more comprehensively features that were significantly associated with BMI classifications, which helped to reduce the potential intrinsic information bias of using one data source type caused by the data collection mechanism. Second, we constructed 2 models for the scenarios of with and without BMI history to best leverage BMI history to improve model performance. In addition, we implemented 4 mathematically different ML algorithms, of which we discovered during this process that some performed better than others. Because of this variability in performance, which by itself provides an assessment of the different ML algorithms, we then combined them into the SLA. The more sophisticated SLA demonstrated superiority over the other singularly used ML algorithms; the SLA has also been shown in other analyses to be the optimal tool for constructing such predictive models [27,

28]. Such an approach may additionally be more efficient when the number of covariates is large for assessing the multiple covariate interactions and correlation terms than simpler statistical approaches. Furthermore, the various sensitivity analyses we conducted strengthened the model selection decisions. Lastly, we externally validated the models in another large nationally representative claims database, which demonstrated the predictive models' performance stability and external validity. When a study budget and technical infrastructure are not constraining, this approach may be utilized over other simpler techniques to provide the optimal prediction model.

CONCLUSIONS

This study demonstrated the feasibility and validity of using ML algorithms to predict BMI classifications in administrative healthcare claims data to expand the utility for RWE generation. Furthermore, it was a relatively straight-forward approach to access BMI information in claims-based data sources. This novel approach to predict BMI classifications in administrative healthcare claims data can be leveraged across multiple therapeutic areas to better understand variations in BMI related disease risk, treatment outcomes, healthcare resource use and costs in real-world settings, as well as be leveraged for other clinical variables that may be underreported in administrative healthcare claims data sources.

ACKNOWLEDGEMENTS

The authors would like to thank Helen Hardy of Janssen Scientific Affairs and Aakash Bhargava, Simran Modi, and Thapa Anshul Chandrasingh of Mu Sigma, LLC for their contribution in the research concept development and data exploration phase of this study.

Funding: This research, preparation of the manuscript, and the journal's Rapid Service and Open Access fees, were funded by Janssen, LLC.

Medical Writing Assistance: The authors would like to thank Nancy Connolly of Janssen Scientific Affairs for her medical writing support in preparation of this manuscript and also Jay Lin and Melissa Lingohr-Smith of Novosys Health for their writing contributions to this manuscript and acknowledge that this contribution was supported by Janssen, LLC.

Authorship: All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Author Contributions: All authors contributed to study conception and design; Bingcao Wu, Monish Sakthivel, Onkar Kakade, Kartikeya Gupta, and Aarti Susan Kuruvilla contributed to data acquisition, analysis, and interpretation; all authors contributed to drafting and revising of the manuscript and have given their approval for this manuscript version to be published**.**

Disclosures: Bingcao Wu, Debra Israel, and Yen-wen Chen are employees of Janssen Scientific Affairs, LLC and are stockholders in Johnson & Johnson. Wing Chow was an employee of Janssen Scientific Affairs, LLC at the time of this study. Monish Sakthivel, Onkar Kakade, and Kartikeya Gupta are employees of Mu Sigma, Business Solutions, LLC. Aarti Susan Kuruvilla was an employee of Mu Sigma, Business Solutions, LLC at the time of this study.

Compliance with Ethics Guidelines: All datasets were from databases of de-identified patient data and so ethics committee approval was not required.

Data Availability: The data that support the study are available within the article and online supplementary material. The codes used in this study are provided in the online supplement.

REFERENCES

1. Xia AD, Schaefer CP, Szende A, et al. RWE framework: an interactive visual tool to support a real-world evidence study design. Drugs Real World Outcomes. 2019;6:193–203.
2. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. J Multidiscip Health. 2018;11:295–304.
3. Centers for Disease Control and Prevention. Defining adult overweight and obesity. https://www.cdc.gov/obesity/adult/defining.html. Accessed 11 Aug 2020.
4. Office of the Surgeon General (US). The Surgeon General's call to action to prevent and decrease overweight and obesity. Office of Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, National Institutes of Health, U.S. Department of Health and Human Services. Rockville, MD: Office of the Surgeon General; 2001. https://www.ncbi.nlm.nih.gov/books/NBK44206/. Accessed 11 Aug 2020.
5. Pi-Sunyer X. The medical risks of obesity. Postgrad Med. 2009;121:21–33.
6. Stommel M, Schoenborn CA. Variations in BMI and prevalence of health risks in diverse racial and ethnic populations. Obesity. 2010;18:1821–6.
7. Kamble PS, Hayden J, Collins J, et al. Association of obesity with healthcare resource utilization and costs in a commercial population. Curr Med Res Opin. 2018;34:1335–43.
8. Elrashidi MY, Jacobson DJ, St Sauver J, et al. Body mass index trajectories and healthcare utilization in young and middle-aged adults. Medicine (Baltimore). 2016;95:e2467.
9. Andreyeva T, Sturm R, Ringel JS. Moderate and severe obesity have large differences in health care costs. Obes Res. 2004;12:1936–43.
10. Ammann, EM, Kalsekar I, Yoo A, et al. Validation of body mass index (BMI)-related ICD-9-CM and ICD-10-CM administrative diagnosis codes recorded in US claims data. Pharmacoepidemiol Drug Saf 2018;27:1092–100.
11. SAS Institute Inc. Analytics Insight. Evolution of machine learning. https://www.sas.com/en_us/insights/analytics/machine-learning.html. Accessed 11 Aug 2020.
12. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst. 2018;42:92.
13. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38:1805–14.

14. Jauk S, Kramer D, Leodolter W. Cleansing and imputation of body mass index data and its impact on a machine learning based prediction model. Stud Health Technol Inform. 2018;248:116–23.
15. Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018;9:515.
16. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.
17. Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–7.
18. Mueller L, Berhanu P, Bouchard J, et al. Application of machine learning models to evaluate hypoglycemia risk in type 2 diabetes. Diabetes Ther. 2020;11:681–99.
19. Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed. 2011;104:443–51.
20. Dipnall JF, Paco JA, Berk M, et al. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. PLoS One. 2016;11:e0148195.
21. Quan H, Li B, Saunders D, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008;14:1424–41.
22. Martin BJ, Chen G, Graham M, et al. Coding of obesity in administrative hospital discharge abstract data: accuracy and impact for future research studies. BMC Health Serv Res. 2014;14:70.
23. Hales CM, Carroll MD, Fryar CD, et al. Prevalence of obesity among adults and youth: United States, 2015-2016. NCHS Data Brief. 2017(288):1–8.
24. Thesmar D, Sraer D, Phinheiro, et al. Combining the power of artificial intelligence with the richness of healthcare claims data: opportunities and challenges. PharmacoEconomics. 2019;37:745–52.
25. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health. 2018;39:95–112.
26. Cornier M-A, Després J-P, Davis N, et al. Assessing adiposity: a scientific statement from the American Heart Association. Circulation. 2011;124:1996–2019.
27. Naimi AI, Balzer LB. Stacked generalization: an introduction to Super Learning. Eur J Epidemiol. 2018;33:459–64.
28. Rose S. Mortality risk score prediction in an elderly population using machine learning. Am J Epidemiol. 2013;177:443–52.

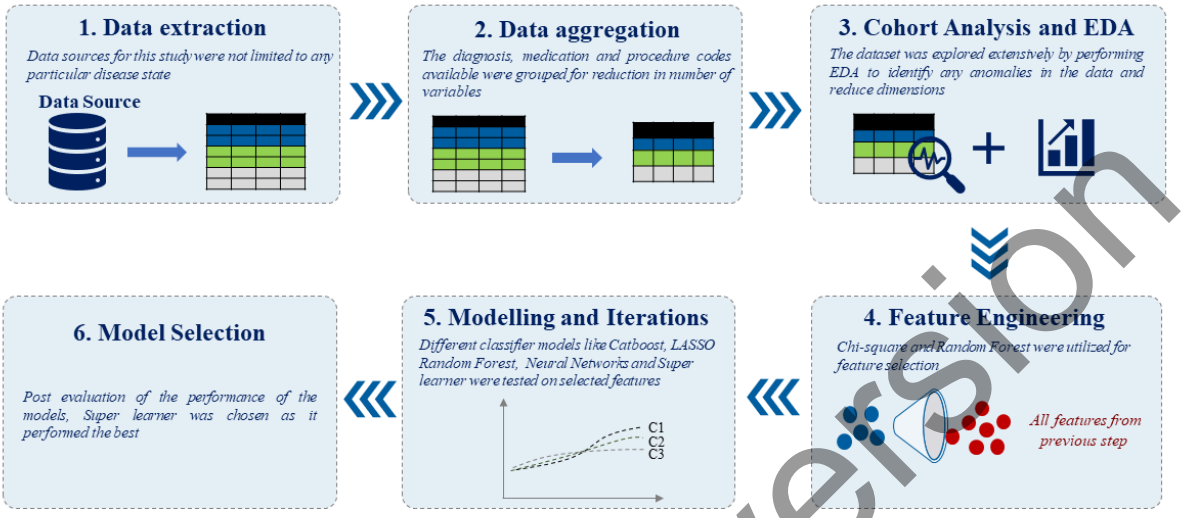

Fig. 1 Methodology flow

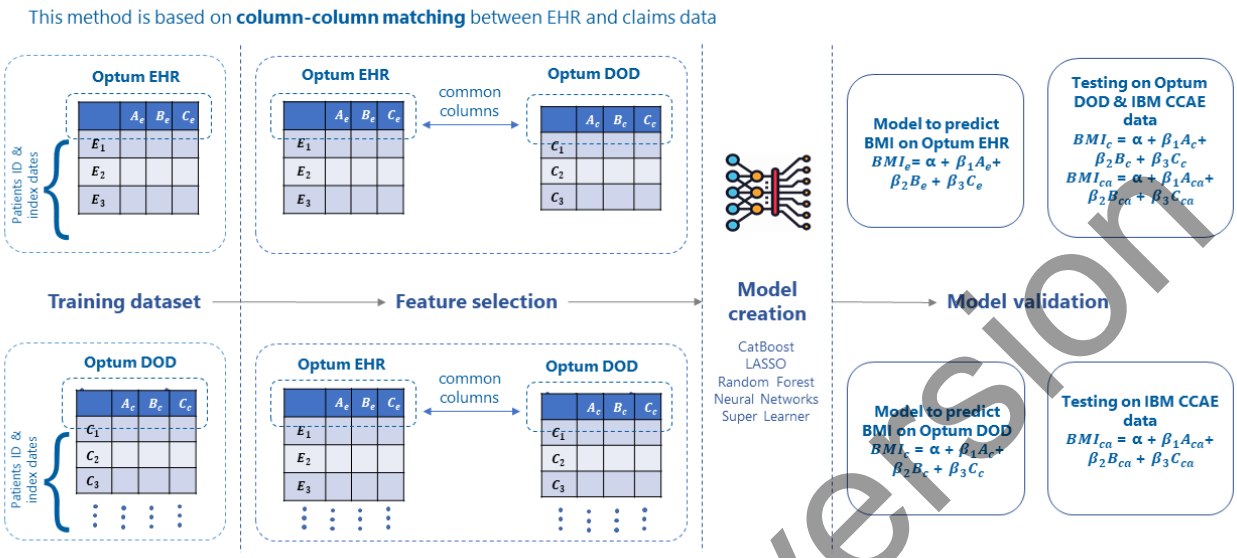

Fig. 2 Process flow of machine learning algorithm implementation for feature engineering

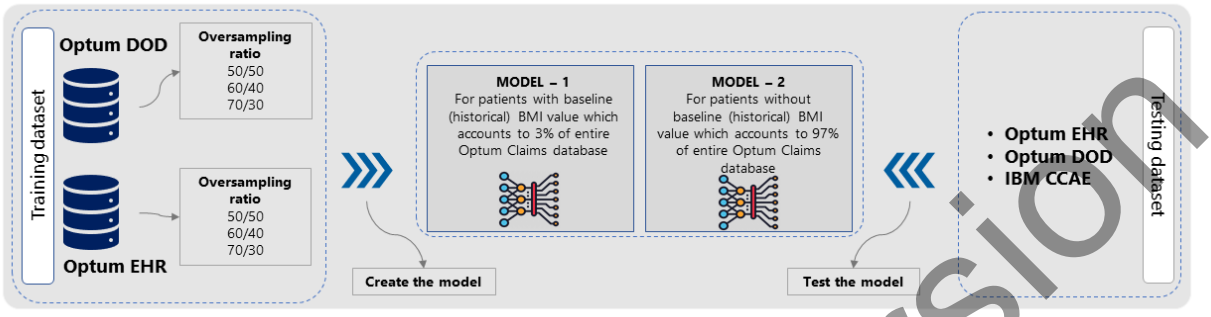

Fig. 3 Training and testing datasets

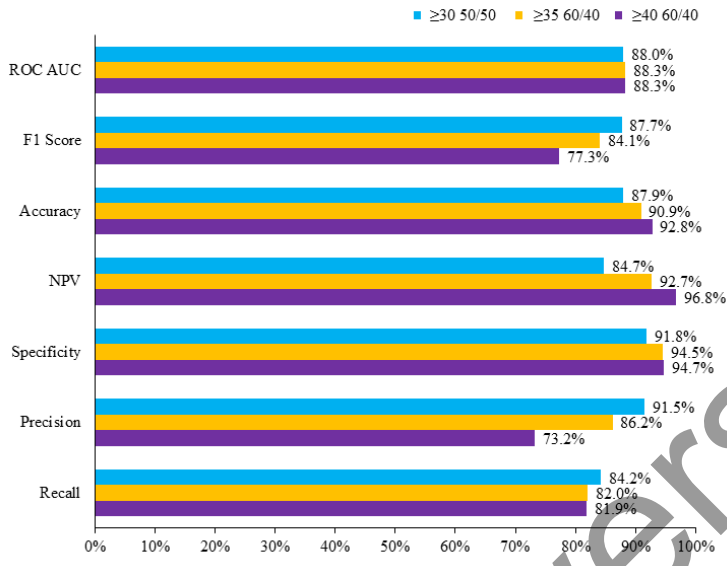

Fig. 4 Predictive performance results of model 1 trained on the Super Learner algorithm and internally validated on the Optum DOD database

ROC AUC: Area under the receiver operating characteristic curve; NPV: Negative predictive value

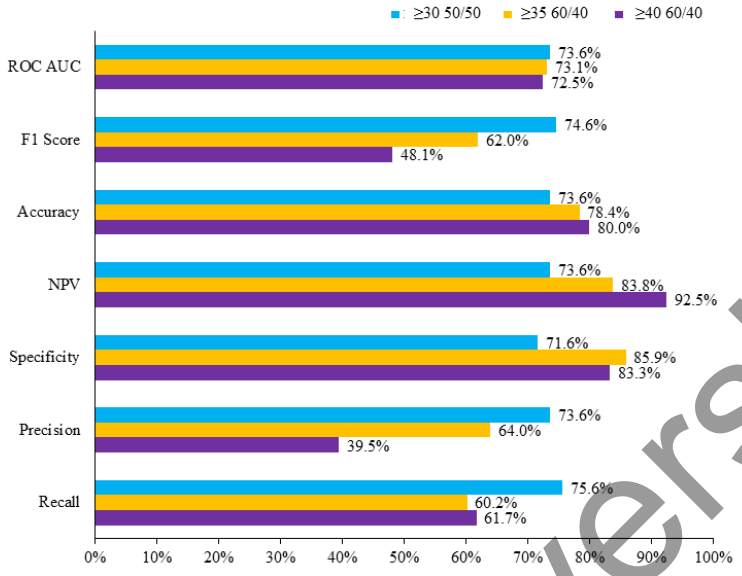

Fig. 5 Predictive performance results of model 2 trained on the Super Learner algorithm and internally validated on the Optum DOD database

ROC AUC: Area under the receiver operating characteristic curve; NPV: Negative predictive value

Table 1. Results of the different models across the 3 databases

Database	Optum EHR		Optum DOD		IBM CCAE	
No. of patients	37,011,188		5,280,836		6,332,087	
No. index BMI readings	343,711,980		16,316,746		15,147,663	
No. rows/columns in training and testing datasets	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
	6,800,000/ 123	6,800,000/ 111	3,300,000/ 123	3,300,000/ 111	3,400,000/ 123	3,400,000/ 111
Considered patient cases out of the patient cohort	2%		20%		22%	
Oversampling ratio in training data	50/50, 60/40, 70/30					
Age group						
<21 years	19%		7%		27%	
21-30 years	13%		4%		11%	
31-45 years	20%		12%		22%	
46-60 years	23%		22%		30%	
>60 years	25%		55%		10%	
BMI classification						
≥30 kg/m^2	51%		40%		45%	
≥35 kg/m^2	29%		20%		27%	
≥40 kg/m^2	16%		10%		16%	
US region						
South	24%		50%		51%	
Midwest	50%		21%		22%	
West	9%		20%		10%	
Northeast	13%		9%		16%	
Others	4%		0%		1%	

Table 2. Number of features selected for each BMI classification prediction

BMI classification	Features	No. of features before selection	Of the 379 selected features	Of the top 100 selected features
≥30 kg/m^2	Diagnoses	244	100	49
	Medications	739	100	34
	Procedures	283	179	17
≥35 kg/m^2	Diagnoses	244	109	60
	Medications	739	108	40
	Procedures	283	144	-
≥40 kg/m^2	Diagnoses	244	112	65
	Medications	739	101	35
	Procedures	283	139	-

Table 3. Best algorithm of the models and oversampling ratios across all the iterations of BMI classifications

BMI classification	Model	Algorithm model trained on	Model output	Oversampling ratio
≥30 kg/m^2	Model 1	Super Learner	1 if BMI ≥30 0 if BMI <30	50/50
≥30 kg/m^2	Model 2	Super Learner	1 if BMI ≥30 0 if BMI <30	50/50
≥35 kg/m^2	Model 1	Super Learner	1 if BMI ≥35 0 if BMI <35	60/40
≥35 kg/m^2	Model 2	Super Learner	1 if BMI ≥35 0 if BMI <35	60/40
≥40 kg/m^2	Model 1	Super Learner	1 if BMI ≥40 0 if BMI <40	60/40
≥40 kg/m^2	Model 2	Super Learner	1 if BMI ≥40 0 if BMI <40	60/40