BMJ Open Diabetes Research & Care

Applying machine learning approaches for predicting obesity risk using US health administrative claims database

Casey Choong , Alan Brnabic, Chanadda Chinthammit, Meena Ravuri, Kendra Terrell, Hong Kan

To cite: Choong C, Brnabic A, Chinthammit C, *et al*. Applying machine learning approaches for predicting obesity risk using US health administrative claims database. *BMJ Open Diab Res Care* 2024;**12**:e004193. doi:10.1136/bmjdrc-2024-004193

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjdrc-2024-004193).

Received 14 March 2024
Accepted 12 September 2024

Eli Lilly and Company, Indianapolis, Indiana, USA

Correspondence to
Dr Casey Choong;
choong_kar-chan@lilly.com

ABSTRACT

Introduction Body mass index (BMI) is inadequately recorded in US administrative claims databases. We aimed to validate the sensitivity and positive predictive value (PPV) of BMI-related diagnosis codes using an electronic medical records (EMR) claims-linked database. Additionally, we applied machine learning (ML) to identify features in US claims databases to predict obesity status.

Research design and methods This observational, retrospective analysis included 692 119 people ≥18 years of age, with ≥1 BMI reading in MarketScan Explorys Claims-EMR data (January 2013–December 2019). Claims-based obesity status was compared with EMR-based BMI (gold standard) to assess BMI-related diagnosis code sensitivity and PPV. Logistic regression (LR), penalized LR with L1 penalty (Least Absolute Shrinkage and Selection Operator), extreme gradient boosting (XGBoost) and random forest, with features drawn from insurance claims, were trained to predict obesity status (BMI≥30 kg/m^2) from EMR as the gold standard. Model performance was compared using several metrics, including the area under the receiver operating characteristic curve. The best-performing model was applied to assess feature importance. Obesity risk scores were computed from the best model generated from the claims database and compared against the BMI recorded in the EMR.

Results The PPV of diagnosis codes from claims alone remained high over the study period (85.4–89.2%); sensitivity was low (16.8–44.8%). XGBoost performed the best at predicting obesity with the highest area under the curve (AUC; 79.4%) and the lowest Brier score. The number of obesity diagnoses and obesity diagnoses from inpatient settings were the most important predictors of obesity. XGBoost showed an AUC of 74.1% when trained without an obesity diagnosis.

Conclusions Obesity prevalence is under-reported in claims databases. ML models, with or without explicit obesity, show promise in improving obesity prediction accuracy compared with obesity codes alone. Improved obesity status prediction may assist practitioners and payors to estimate the burden of obesity and investigate the potential unmet needs of current treatments.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Obesity is underestimated in administrative claims databases; a significant proportion of individuals with obesity do not have a diagnosis code for the condition recorded in their claims data. Machine learning methods can optimize predictability and may be used to predict disease status, diagnosis, and clinical variables using real-world evidence.

WHAT THIS STUDY ADDS

⇒ This real-world data analysis confirmed the upward trend in obesity prevalence from 2013 to 2019. While most obesity codes in administrative claims data accurately identify cases of obesity (ie, high positive predictive value), they consistently miss a significant proportion of individuals with this disease (ie, low sensitivity).

⇒ This study addressed the limitation of underutilization of obesity codes in claims data via a machine learning model, enabling more accurate identification of individuals with obesity in administrative claims data.

⇒ The most important predictors of obesity were the number of obesity diagnoses, obesity diagnoses from inpatient settings, diagnosis of obstructive sleep apnea, diagnosis of hypertension, and the use of antidiabetic or antihypertension agents.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Improved obesity status prediction could help healthcare professionals and payors estimate the burden of the disease and evaluate current obesity treatment strategies.

INTRODUCTION

Obesity remains a major health crisis in the USA. About half of the adult US population is likely to have obesity by 2030.[1] Obesity is associated with multiple chronic conditions such as type 2 diabetes (T2D), cardiovascular diseases (CVD), metabolic syndrome, chronic kidney disease, metabolic dysfunction-associated steatotic liver disease (MASLD), certain types of cancer, obstructive sleep apnea (OSA), osteoarthritis and various psychological conditions.[2–4] Consequently, obesity imposes a significant economic burden.[5]

The obesity status of an individual can be assessed and classified by measuring body mass index (BMI).[6] Based on their BMI, the

US Centers for Disease Control and Prevention classifies people as underweight (below 18.5 kg/m^2), normal (18.5 to <25 kg/m^2), overweight (25 to <30 kg/m^2), class 1 (30 to <35 kg/m^2), class 2 (35 to <40 kg/m^2) or class 3 (≥40 kg/m^2) obesity.[7] Despite its use as a key anthropometric measure of obesity status, BMI is not adequately recorded in US administrative claims databases.[8] When the International Classification of Disease (ICD) diagnosis codes for obesity are included in claims data, obesity status can be identified with a high positive predictive value (PPV), but in reality where these diagnosis codes are underused, it results in low sensitivity.[8–11] This limits the use of administrative claims data in studying obesity as an exposure, confounder or effect modifier of interest in weight-related health service research.[12]

The inherent low sensitivity of administrative claims data in defining obesity can be addressed by predicting the risk of obesity using real-world data (RWD), that is, data related to patient health status or the delivery of health care, including medical claims, and electronic medical records (EMR), or data from registries and digital health technology.[13] RWD is imperative to researchers, health technology assessment agencies, payors, and other stakeholders to assess and mitigate the risk of obesity-related comorbidities and to better guide clinical decisions.[14] Prediction modeling is often performed using classical statistical regression methods. However, these models may overlook complex associations when a higher number of variables are studied. In addition, choosing the right model is not straightforward when using these methods.[15] Machine learning (ML) methods address these limitations by maximizing predictability and effectively handling possible non-linear relationships and higher-order interactions.[16] ML methods can be used to develop algorithms to predict disease risk, diagnosis, and clinical variables.[17–21]

Recently, Wu *et al*[8] applied four ML algorithms: Catboost, random forest, penalized logistic regression with L1 penalty (Least Absolute Shrinkage and Selection Operator (Lasso)), and artificial neural networks to predict BMI classifications in a claims database. The study identified a Super-Learner algorithm (SLA) that leveraged predictions from four ML algorithms through logistic regression, with an area under the receiver operating characteristic curve (AUC ROC) of approximately 88% for the prediction of BMI classifications of 30–40 kg/m^2. However, the study was conducted using EMR and claims databases independently with the claims data source as the target, which may not accurately reflect actual BMI values due to shortcomings pertaining to the claims database.

In the current study, we assessed the validity of obesity diagnosis codes in claims data and their concordance with the EMR data using a large US claims-EMR-linked database with residents across regions. Furthermore, we applied ML algorithms to identify features in the US claims database to predict obesity status using BMI recorded in the EMR as the gold standard.

RESEARCH DESIGN AND METHODS

Data source

This observational, retrospective study used data from the Merative MarketScan Explorys Claims-EMR data set. The data set was built by combining MarketScan administrative claims and Explorys EMR data and consists of 6.5 million individuals with claims-EMR linked records from more than 400 000 providers and physicians from all four US census regions. This data set comprises comprehensive longitudinal patient records including healthcare cost information, outpatient prescription fills including specialty pharmacy, coverage eligibility, vital signs and other biometrics, medical and surgical history, laboratory results, implantable devices, patient-reported outcomes, inpatient drugs, ambulatory prescriptions, clinical events, and procedures.

Study participants

The study included individuals aged ≥18 years with at least one valid BMI value (LOINC (Logical Observation Identifiers Names and Codes): 39156-5 and 89270-3) recorded in the claims-EMR data set from January 2013 to December 2019. The date of the last valid BMI record was set as the index date. Individuals were required to be commercially insured or covered under a Medicare Supplemental plan with continuous enrolment during the baseline period (ie, 12 months prior to, or on the index date).

Validity of the obesity diagnosis codes

Obesity prevalence over the study period was determined through EMR (based on the BMI value (>30 kg/m^2)) and claims data (based on the presence of obesity diagnoses codes recorded for any healthcare encounter, ie, outpatient, emergency room (ER) and inpatient visits at baseline), divided by the number of subjects that met the inclusion criteria, respectively. The validity of the obesity diagnosis codes, ICD-Clinical Modification (CM)–9 and ICD-CM-10, was evaluated in terms of their sensitivity and PPV.

Candidate variables and feature aggregation

Model features included in the prediction model were demographic characteristics such as age, sex, BMI, payor, and index year measured at the index date; clinical characteristics such as diagnoses codes, procedure codes, medication codes, and Charlson Comorbidity Index measured at baseline; and healthcare resource utilization such as the number of inpatient, ER, and outpatient visits at baseline. (A full list of features used in the prediction models is available in online supplemental table 4.) The main binary outcome of interest was obesity, defined as BMI≥30 kg/m^2 recorded in the EMR.

The features obtained from the data set were grouped into relevant code hierarchies or clinical concepts to increase the ease of computation and clinical interpretation. The diagnosis codes (ICD-CM-9, ICD-CM-10), and procedure codes (Current Procedure Terminology

and Healthcare Common Procedure Coding System (HCPCS)) were grouped using Clinical Classification Software, a web-based analytics software.[22] Medications in the National Drug Code and HCPCS were grouped using the Cerner Multum drug database.[23] Analytical data sets were created using the Instant Health Data platform (Panalgo, Boston, Massachusetts, USA) and preprocessed with the cohorts, target variables, and features using the data science module within the platform. All features were carried forward without selection during preprocessing.

Model evaluation and testing

The ML algorithms used for feature selection were Lasso, extreme gradient boosting (XGBoost), and random forest. The data set was randomly split into two partitions; the training/validation (60%/20%) set and the test (holdout) (20%) set. Hyperparameters were assessed using five-fold cross-validation in the training/validation set. The models were trained and evaluated multiple times with different data splits to obtain a reliable assessment of their performance for each hyperparameter combination. Models were initially evaluated using their default hyperparameter settings (online supplemental table 1) but were then tuned using a random grid search to optimize the predictive ability of the models. The best-performing model was then selected for further tuning.[24]

Model performance was evaluated using the test set. The primary metric of evaluation of the model performance was the AUC ROC. Other metrics investigated were PPV (accuracy), sensitivity (recall), precision, Youden Index, F1 score, negative predictive value (NPV), and specificity.[8] Validation of the best-performing model was conducted using the test set which was not used during the model's training and validation. This approach provided a good proxy for how the model would perform with new data.

Feature importance ranking and risk score computation

A total of 479 features were entered into the ML models. The best-performing model was used to select and rank a feature's importance based on said feature's relative importance to risk prediction using the training data set. To assess the probability of an individual meeting the criteria for obesity, the final model calculated a risk score for each individual in the cohort. This risk score was aligned with the known BMI from the EMR to then classify the risk score distribution relative to the true BMI.

Sensitivity analysis

A series of sensitivity analyses were performed on participants with BMI≥35 kg/m^2 and ≥40 kg/m^2 as binary target variables. A shorter baseline period of data (ie, 3 months and 6 months) was used to assess the model performance in the absence of 12 months of data. The binary classification model was trained and validated on a model that excluded BMI/weight-related diagnoses (eg, ICD-CM-10: overweight or obesity: E68*; BMI: Z68*) from the input feature and the model performance was reported.

Ethical review and regulatory considerations

This research was conducted in strict compliance with all state, local, and federal regulatory requirements. The study execution was consistent with Good Clinical Practices, Good Epidemiological Practices, the International Convention on Harmonization, Health Insurance Portability and Accountability Act regulations, the US Department of Health and Human Services, the Office of Human Research Protection and any applicable Internal Review Board guidelines. When applicable, the research was conducted in accordance with the regulations of the US Food and Drug Administration as described in 21 Code of Federal Regulations 50 and 56 and all applicable laws. Participants' data were de-identified to protect their privacy.

RESULTS

Demographics and clinical characteristics

A total of 3 532 946 participants had a BMI value recorded in the database; 2 781 248 were aged ≥18 years, of whom 692 119 had continuous enrolment in the baseline period. The cohort attrition is depicted in online supplemental figure 1. Of those with continuous enrolment, 276 646 (40.0%) people had BMI>30 kg/m^2 in the EMR database and 96 427 (13.9%) had an obesity diagnosis code indicating BMI>30 kg/m^2 or an obesity-related diagnosis code in the administrative claims data. The demographics and characteristics are given in table 1. The mean (SD) age of the participants was 50.8 (17.8) years, and the mean (SD) BMI was 29.5 (7.1) kg/m^2. There was a higher proportion of women compared with men (55.5% vs 44.5%) across the subsets. The characteristics of people with obesity in the EMR and the claims subsets differed from each other. Relative to those classified to have obesity based on BMI in the EMR, those with claims diagnosis codes tended to be older, have a higher BMI, higher disease burden, and have Medicare Supplemental insurance.

Prevalence, sensitivity, and PPV

Obesity prevalence, sensitivity, and PPV, as derived from the ICD-CM from claims from 2013 to 2019, are shown in figure 1. The PPV of the diagnosis codes remained relatively high over the study period with an average of 88% (ranging from 85.4% to 89.2%), while sensitivity remained low, with an average of 31% (ranging from 16.8% to 44.8%). The stratified table in online supplemental table 2 provides information on the coding of obesity-related diagnoses in the claims data by the measured BMI, as recorded in EMR.

Model prediction

The performance of the predictive models in the training/validation data set, as examined and assessed by AUC, average precision, PPV (accuracy), sensitivity (recall), precision, F1 score, NPV, and specificity, is presented in table 2. XGBoost had a higher AUC of 79.4% and a greater accuracy of 73.5% compared with

Table 1 Demographics and cohort characteristics

	Overall	EMR		Claims	
		People with BMI<30 kg/m^2	People with BMI>30 kg/m^2	People without obesity diagnosis code	People with obesity diagnosis code
N, %	692 119 (100)	415 473 (60)	276 646 (40)	595 692 (86)	96 427 (14)
Age, mean (SD)	50.8 (17.8)	50.7 (19.3)	50.9 (15.4)	50.4 (18.2)	53.2 (15.1)
BMI, mean (SD)	29.5 (7.1)	25.0 (3.2)	36.2 (5.8)	28.2 (6.2)	37.0 (7.3)
Female	384 024 (55.5)	232 589 (56.0)	151 435 (54.7)	325 795 (54.7)	58 229 (60.4)
BMI group*					
Underweight	11 750 (1.7)	11 750 (2.8)	NA	115 486 (1.9)	264 (0.3)
Normal	181 415 (26.2)	181 415 (43.7)	NA	179 804 (30.2)	1611 (1.7)
Overweight	222 308 (32.1)	222 308 (53.5)	NA	212 723 (35.7)	9585 (9.9)
Class 1 Obesity	147 962 (21.4)	NA	147 962 (53.5)	115 956 (19.5)	32 006 (33.2)
Class 2 Obesity	73 636 (10.6)	NA	73 636 (26.6)	48 151 (8.1)	25 485 (26.4)
Class 3 Obesity	55 048 (8.0)	NA	55 048 (19.9)	27 572 (4.6)	27 476 (28.5)
CCI, mean (SD)	0.7 (1.6)	0.7 (1.6)	0.7 (1.5)	0.7 (1.5)	1.2 (1.9)
CCI categories					
0	494 388 (71.4)	303 543 (73.1)	190 845 (69)	441 026 (74.0)	53 362 (55.3)
1	78 366 (11.3)	41 384 (10.0)	36 982 (13.4)	62 194 (10.4)	16 172 (16.8)
2	54 941 (7.9)	30 895 (7.4)	24 046 (8.7)	43 340 (7.3)	11 601 (12.0)
3+	64 424 (9.3)	39 651 (9.5)	24 773 (9.0)	49 132 (8.3)	15 292 (15.9)
Payor					
Commercial	555 977 (80.3)	324 933 (78.2)	231 044 (83.5)	479 584 (80.5)	76 393 (79.2)
Medicare	136 142 (19.7)	90 540 (21.8)	45 602 (16.5)	116 108 (19.5)	20 034 (20.8)
Index year					
2013	53 434 (7.7)	34 178 (8.2)	19 256 (7.0)	49 713 (8.4)	3721 (3.9)
2014	69 434 (10)	44 028 (10.6)	25 406 (9.2)	63 777 (10.7)	5657 (5.9)
2015	100 863 (14.6)	62 778 (15.1)	38 085 (13.8)	90 622 (15.2)	10 241 (10.6)
2016	170 261 (24.6)	102 323 (24.6)	67 938 (24.6)	147 846 (24.8)	22 415 (23.3)
2017	140 843 (20.4)	81 889 (19.7)	58 954 (21.3)	118 160 (19.8)	22 683 (23.5)
2018	50 683 (7.3)	30 607 (7.4)	20 076 (7.3)	42 665 (7.2)	8018 (8.3)
2019	106 601 (15.4)	59 670 (14.4)	46 931 (17.0)	82 909 (13.9)	23 692 (24.6)

All data are presented as n, (%) unless specified otherwise. Obesity as indicated by BMI≥30 kg/m^2 in the EMR, and ICD codes in the claims data. There could be overlap in some people who were classified to have obesity in the EMR and/or the claims.
*Underweight (<18.5 kg/m^2); normal (18.5 to <25 kg/m^2); overweight (25 to <30 kg/m^2); class 1 obesity (30 to <35 kg/m^2); class 2 obesity (35 to <40 kg/m^2); class 3 obesity (≥40 kg/m^2) (CDC, 2022).
BMI, body mass index; CCI, Charlson Comorbidity Index; CDC, Centers for Disease Control and Prevention; EMR, electronic medical record; ICD, Classification of Disease; N, total population; NA, not applicable.

the other models tested. The Brier score among the tested models ranged from 0.17 to 0.20, with XGBoost scoring the lowest. The performance of XGBoost was validated in the test/holdout data set with the highest AUC of 79.4% (online supplemental table 3).

Feature importance

Figure 2 shows the feature importance plot generated by XGBoost in the training/validation data set. The features identified by XGBoost to positively impact obesity risk prediction in the training/validation data set were: the number of obesity diagnoses, obesity diagnoses from inpatient settings, diagnosis of OSA, hypertension, T2D, hyperglycemia, pre-diabetes, neurocognitive disorders such as delirium, dementia, amnestic and other cognitive disorders, MASLD and metabolic dysfunction-associated steatohepatitis (MASH), and use of antidiabetic agents, antihypertensive combinations, and diuretics such as bumetanide, mannitol, furosemide, torsemide, and bone resorption inhibitors, presence of obesity ER visit, and older age. The presence of overweight, underweight,

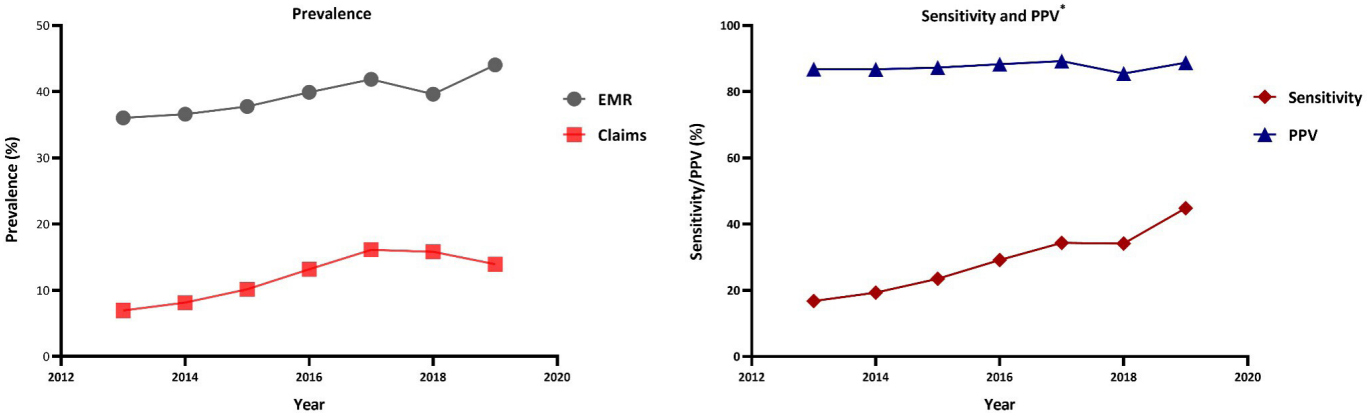

Figure 1 Obesity prevalence, sensitivity and PPV from ICD-CM from 2013 to 2019. *Sensitivity=TP/(TP+FN); PPV=TP/(TP+FP). BMI was used to determine obesity status in the EMR; obesity diagnosis codes were used to determine obesity status in the claims. BMI, body mass index; EMR, electronic medical record; FN, false negative; FP, false positive; ICD-CM, International Classification of Diseases-Clinical Modification; PPV, positive predictive value; TP, true positive.

diagnosis of osteoporosis, acne, melanocytic nevi, and examinations encounter were found to negatively impact the risk prediction.

Risk prediction

The probability of an individual meeting the criteria for obesity (BMI≥30 kg/m^2) by the XGBoost model as compared with the actual BMI classification, is shown in figure 3. The predicted probability (SD) of 0.72 (0.26) was observed in people in the class 3 obesity cohort, while a predicted probability (SD) of 0.20 (0.16) was observed in people in the underweight cohort. The model indicated that individuals with higher predictive values have a greater likelihood of meeting the BMI criteria for obesity.

Sensitivity analysis

Trained and validated binary classification models for BMI≥35 kg/m^2 and ≥40 kg/m^2 had a higher AUC of 81.0% and 83.6%, respectively, using the XGBoost model. The models showed the topmost common predictive features were similar to those presented in the main analysis of BMI classification ≥30 kg/m^2. For BMI>35 kg/m^2, the number of obesity diagnoses and sleep apnea were the most important predictors of obesity, followed by overweight, use of antidiabetic agents, and use of antihypertensive medications. For BMI>40 kg/m^2, the number of obesity diagnoses and sleep apnea were the most important predictors of obesity, followed by the use of antidiabetic agents, overweight, and use of diuretic medications.

The XGBoost model performed better with 12 months of baseline data with an AUC of 79.4%, compared with 3-month and 6-month baseline data. On retraining the model without a baseline BMI/weight-related diagnosis, the XGBoost model yielded a satisfactory performance with an AUC value of 74.1%. The top five features identified were sleep apnea (positive impact), antidiabetic agents (positive impact), hypertension (positive impact), use of antihypertensive combination medications

Table 2 Performance of predictive models fitted to the training set (60%) and evaluated on the validation set (20%)

Model type	No. of features	AUC (%)	Average precision (%)	Brier score	PPV (%)	Sensitivity (%)	Precision (%)	Youden Index (%)	F1 score	NPV (%)	Specificity (%)
XGBoost model	479	79.4	75.6	0.174	73.5	61.7	68.8	43.0	0.651	76.1	81.4
Reg logistic model	458	76.5	71.5	0.187	71.5	60.2	65.7	39.2	0.628	74.8	79.0
Logistic model	481	76.4	71.5	0.187	71.4	60.2	65.5	39.0	0.627	74.8	78.8
Random forest model	478	76.1	72.3	0.198	71.9	57.5	67.4	39.0	0.621	74.2	81.4
Lasso model	58	75.3	70.9	0.194	72.0	54.0	69.3	38.0	0.607	73.3	84.1

Reg logistic model, logistic regression model; higher AUC, better distinction between patients with and without obesity; lower Brier score, greater accuracy; higher recall, better maximization of the number of true positives; higher precision, better minimization of false positives; Youden Index of >50%, higher F1 score, better performance of model; higher specificity, better identification of negative results.
AUC, area under the curve; Lasso, Least Absolute Shrinkage and Selection Operator; NPV, negative predictive value; PPV, positive predictive value; XGBoost, extreme gradient boosting.

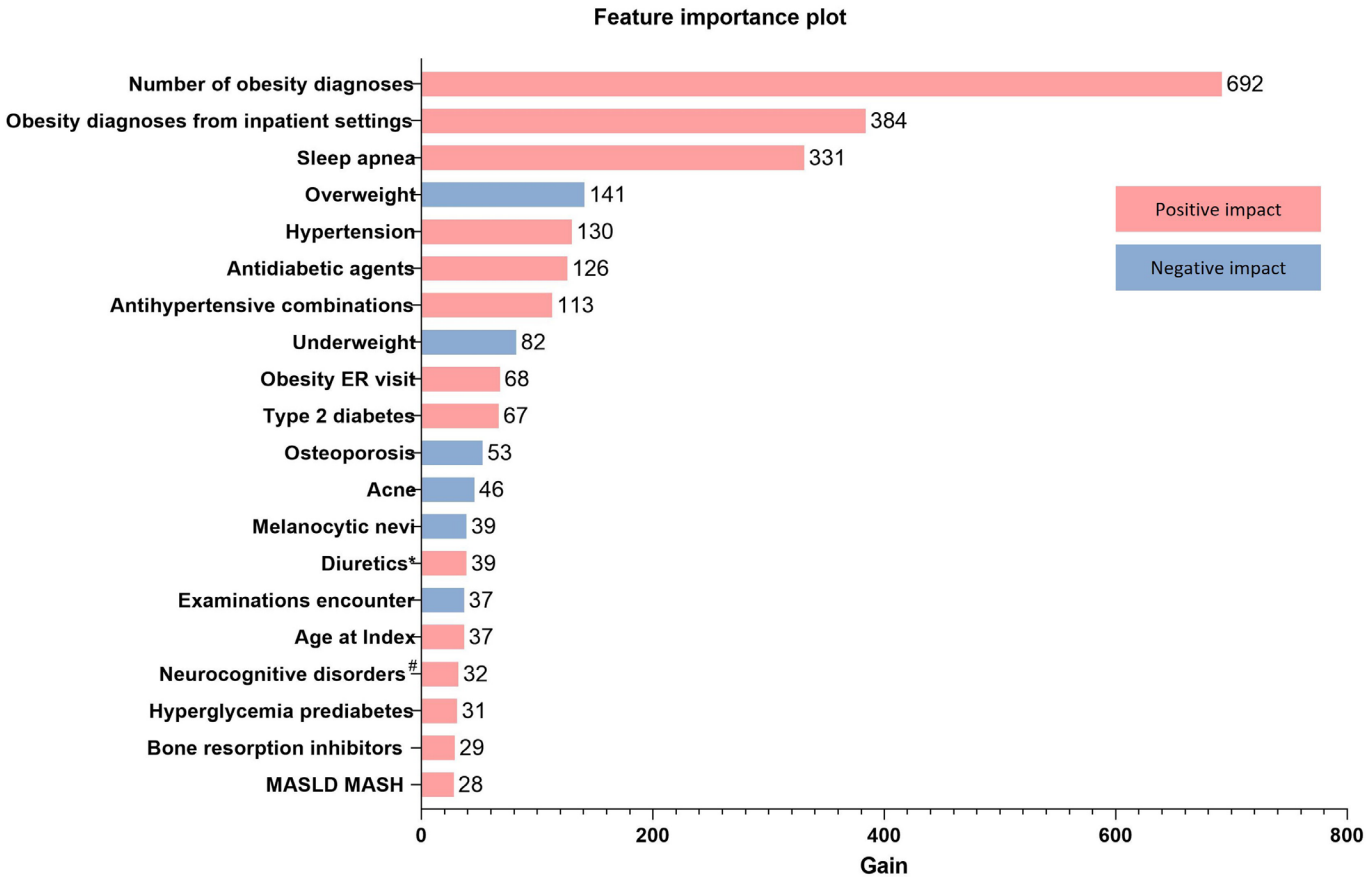

Figure 2 Top 20 features identified from XGBoost in the train/validation data set. Gain is the relative contribution of the corresponding feature to the model. All features were captured at baseline, except when indicated as index. *Diuretics include bumetanide, mannitol, furosemide and torsemide. #Neurocognitive disorders include delirium, dementia, amnestic and other cognitive disorders. ER, emergency room; MASH, metabolic dysfunction-associated steatohepatitis; MASLD, metabolic dysfunction-associated steatotic liver disease; XGBoost, extreme gradient boosting.

(positive impact), and osteoporosis (negative impact; online supplemental figure 2).

DISCUSSION

The growing prevalence of obesity necessitates the exploration of risk prediction and prevention strategies for obesity. Administrative claims databases can be comprehensive and inexpensive sources of RWD for epidemiological studies as they establish the prevalence and incidence of various chronic diseases across large and demographically diverse populations.[25] However, previous studies have shown that the use of these data sources may result in an incorrect estimate of obesity prevalence due to underutilization of the diagnosis codes.[8 11 26 27]

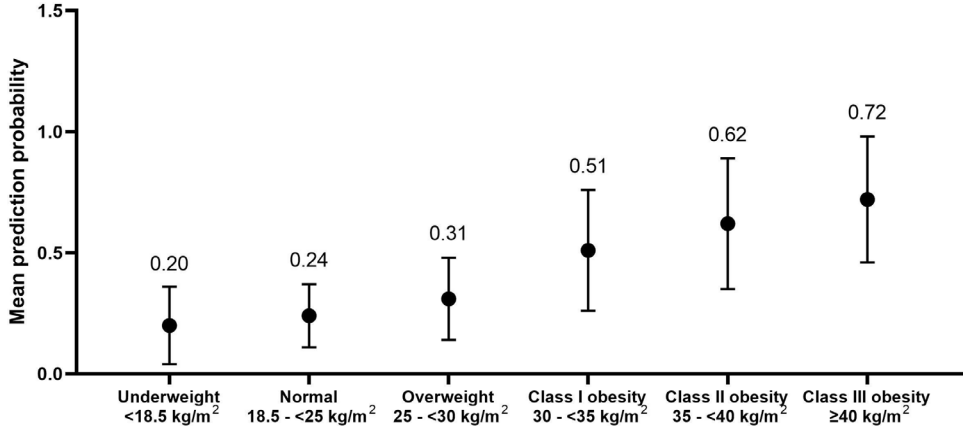

Figure 3 Predicted probability from XGBoost model versus actual BMI classification. BMI, body mass index; XGBoost, extreme gradient boosting.

The current study assessed the validity of obesity diagnosis codes in an administrative claims database and reported high PPV and low sensitivity, in concordance with previous studies.[11 12] Ammann *et al* reported low specificity and high PPV of the ICD-CM-9 and ICD-CM-10 BMI-related diagnosis codes for identification of patients with overweight or obesity.[11] Ammann *et al* also demonstrated higher sensitivity of ICD-CM-10 coding compared with ICD-CM-9 coding[11] which was confirmed by Suissa *et al*,[28] a finding potentially attributable to improved coding practices and reimbursement requirements. Moreover, the accuracy or PPV of obesity diagnosis codes was higher among patients with obesity-related complications such as diabetes or hypertension,[27] and the probability of having an obesity-related diagnosis code in claims data increased with comorbidity burden and hospitalization.[11]

In the current study, older individuals, those with severe obesity, or those with a higher disease burden were more likely to have an obesity diagnosis code recorded in the claims data. Possible reasons could be because these individuals may be more likely to seek medical care, or healthcare providers may code obesity for those with greater severity and burden, as they may consider obesity to be a driving diagnosis for the high disease burden. Indeed, people with an obesity diagnosis code were more likely to have increased healthcare utilization, including hospitalizations, emergency room visits, outpatient visits, and increased usage of medications compared with those without obesity diagnosis codes.[28] However, despite the increased usage of obesity diagnosis codes in the claims database, the true obesity prevalence was still underestimated in the claims database compared with EMR data in the current study. Furthermore, the sensitivity was low as only approximately 31% of people with BMI indicative of obesity in the EMR had a corresponding diagnosis in claims. These results further emphasize the magnitude of obesity code underutilization and its impact on assessing obesity in claims data.

Previous studies showed that people with severe obesity were more likely to have a BMI-related diagnosis code in administrative data relative to those of normal weight.[9 11 27] People with obesity diagnosis codes in the claims database are more likely to have class 3 obesity than those without obesity diagnosis codes, suggesting that diagnosis codes may not be recorded for people with class 1 obesity.[27] Underutilization of obesity diagnosis codes could also result from other factors, such as physicians not considering obesity to be a disease, or the obesity diagnosis not being based on an objective measurement of BMI, thus only capturing cases of severe obesity.[25] Taken together with the findings of the current study, these factors emphasize the importance of careful consideration of the diagnosis codes used for inclusion criteria in observational studies using claims data, and how diagnosis codes may impact outcomes such as disease prevalence or incidence. The use of diagnosis codes to help provide a "snapshot" for the better capture of obesity in the identification of target populations is critical to improving public health surveillance and research studies that use these databases, given the established association of obesity with several chronic diseases.[25]

Lately, Wu *et al*[8] developed two models applying an SLA: model 1 with recorded baseline BMI values and model 2 with demographics and clinical characteristics data, excluding baseline BMI values, to predict obesity in people of all age groups. Model 1 reported better performance than model 2 with a higher AUC ROC (88% vs 73%), accuracy (ranging from 87.9% to 92.8% versus 73.6% to 80.0%) and specificity (ranging from 91.8% to 94.7% versus 71.6% to 85.9%). However, a notable limitation of Wu *et al*[8] is that the study interpolated BMI from claims data that under-reports BMI. In the current study, we tested the predictive performance of five ML models to differentiate people with and without obesity. Of all the models tested, the XGBoost model demonstrated moderate-to-strong performance in predicting obesity risk. A higher AUC ROC and lower Brier score of the XGBoost model translated to better distinction of people with and without obesity with greater accuracy; however, a lower Brier score (0.17 in the current study) does not necessarily imply higher calibration.[29] To prevent poor calibration, we followed some of the strategies outlined in Van Calster *et al*.[30] For example, we included a sufficient sample size for the number of predictors, we used Lasso regression, a penalized regression technique, and we employed a simpler model that did not include too many interaction terms.

Furthermore, candidate variables that predict obesity risk were identified in the US administrative claims database using BMI recorded in the EMR. The XGBoost model ranked corresponding features by their relative contribution to the model in terms of "gain", calculated by averaging each feature's contribution for each tree in the model. Features were listed in descending order of their predictive power, as variables at the top contribute more to the model than the ones at the bottom. The highest-ranked predictors of obesity were relatively consistent across the ML algorithms used. Nevertheless, predictor importance did differ slightly between sensitivity analyses such as different lengths of baseline periods (3 or 6 months) and BMI targets of >35 kg/m^2 and >40 kg/m^2. Interestingly, results from ML models that excluded baseline BMI/weight-related diagnosis were able to predict obesity status based on other risk factors of obesity such as sleep apnea and use of antidiabetic medications. The findings potentially fill the gap of missing obesity status in administrative claims databases.

Besides the finding that the number of obesity diagnoses and diagnoses from inpatient settings were found to be the most important predictors of obesity, diagnosis of OSA, hypertension, and use of antidiabetic or antihypertension medications suggests that these factors are strongly associated with obesity and can be used to identify people at high risk for the condition. Interestingly, this study highlighted the presence of dermatological conditions such as acne and melanocytic nevi

to negatively impact obesity risk. No epidemiological evidence is available suggesting a relationship between obesity and sebum production, and thus the pathogenesis of acne. In one of the largest risk factor studies conducted on the prevalence of melanocytic nevi among children and adolescents in Baltic countries, the condition was found to be associated with higher BMI.[31] Aside from this, little to no evidence exists assessing the impact of these conditions on obesity risk.

Limitations of this study include the availability of all potential predictors in the databases investigated, such as race, region, and physical activity of the participants as well as the possible inaccuracies and misclassification of key variables like obesity diagnosis codes. Second, the database consists of administrative healthcare data of primarily large employer sponsor-insured individuals from a convenience sample of the population across the USA. While BMI is not the best indicator of obesity, it is a helpful tool for obesity screening and health assessment in clinical practice. It is a standardized metric used worldwide, a simple way to identify individuals who may be at risk for weight-related health problems, prompting further evaluation. Other anthropometric measures of adiposity, such as waist circumference or fat-to-muscle ratio, may provide a more complete picture of the obesity status of a patient. However, these measures are usually not captured in EMRs, and it would be difficult, if not impossible, to exclude specific populations, such as athletes with high BMI but without obesity,[32] from the analyses. Therefore, the results may not be generalizable to all populations. Lastly, further research is warranted to carry out external validation of the models using different databases.

Identifying potential candidate variables that predict obesity using RWD may help decision-makers understand the impact of obesity at the population level, helping them to identify appropriate levers to implement policy measures to mitigate risks. Furthermore, ML methods can help improve obesity status prediction and thus assist practitioners and payors to estimate the burden of the condition, investigate the potential unmet need for current treatment, and determine the economic value of new treatments, at both individual and population levels. Moreover, as obesity is a major risk factor for several chronic conditions, predicting its occurrence using RWD will ensure greater accuracy in risk estimates for morbidity and mortality associated with its comorbidities. Future research could focus on validating the ML model in other populations and evaluating predictive scores for obesity-related complications, such as CVD and mortality in administrative claims data. For example, Njei *et al* recently used an explainable machine learning model for high-risk MASH prediction and compared its performance with well-established biomarkers such as MASLD fibrosis scores.[33] The XGBoost model in the study had high sensitivity, specificity, AUC, and accuracy for identifying high-risk MASH. Furthermore, BMI was one of the top five predictors of high-risk MASH. Future studies are needed on how the predictive risk of obesity may change over time with obesity intervention or treatment.

CONCLUSIONS

Obesity is under-reported in administrative claims databases. Applying ML approaches to RWD may help predict obesity status and thus estimate the burden of the condition more accurately. The current study demonstrated moderate-to-strong predictive performance of XGBoost model in identifying people at high risk of obesity using claims data. The computed predictive value helped differentiate people in the claims data based on their obesity status, thus expanding the utility of BMI as a variable in the data source. Improved prediction of obesity status could assist practitioners and payors to estimate the burden of the condition and investigate the potential unmet need of current treatment at individual and population levels, which may lead to better prevention and treatment strategies for obesity.

Acknowledgements The authors thank Elizabeth Eby (Eli Lilly and Company, IN, USA) for the critical review of the manuscript. Richa Kapoor, PhD, an employee of Eli Lilly Services India Pvt. Ltd., provided medical writing support which was funded by Eli Lilly and Company.

Contributors 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. Conception of the work: CCho, AB, HK. Design of the work: CCho. Acquisition of data for the work: CCho, KT. Analysis of data for the work: CCho. Interpretation of data for the work: CCho, AB, CChi, HK, MR. Critical revision/drafting of the work: All authors. Guarantor: CCho.

Funding This study was sponsored by Eli Lilly and Company, Indianapolis, USA.

Competing interests CCho, AB, CChi, KT and HK are employees and stockholders of Eli Lilly and Company, Indiana, USA. MR was an employee of Eli Lilly and Company at the time of the study.

Patient consent for publication Not applicable.

Ethics approval As this study uses only de-identified patient records, it was exempted from Institutional Review Board approval. A formal 'Consent to Release Information' form was not required as this was an observational study that used previously collected data and did not impose any form of intervention.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Casey Choong http://orcid.org/0000-0001-6476-784X

REFERENCES

1 Ward ZJ, Bleich SN, Cradock AL, *et al*. Projected U.S. State-Level Prevalence of Adult Obesity and Severe Obesity. *N Engl J Med* 2019;381:2440–50.
2 Ansari S, Haboubi H, Haboubi N. Adult obesity complications: challenges and clinical impact. *Ther Adv Endocrinol Metab* 2020;11:2042018820934955.
3 Kinlen D, Cody D, O'Shea D. Complications of obesity. *QJM* 2018;111:437–43.
4 Lin X, Li H. Obesity: Epidemiology, Pathophysiology, and Therapeutics. *Front Endocrinol (Lausanne)* 2021;12:706978.
5 Boye KS, Lage MJ, Terrell K. Healthcare outcomes for patients with type 2 diabetes with and without comorbid obesity. *J Diabetes Complications* 2020;34:107730.
6 Wharton S, Lau DCW, Vallis M, *et al*. Obesity in adults: a clinical practice guideline. *CMAJ* 2020;192:E875–91.
7 CDC. Defining adult overweight & obesity. 2023. Available: https://www.cdc.gov/obesity/basics/adult-defining.html [Accessed 31 Jan 2023].
8 Wu B, Chow W, Sakthivel M, *et al*. Body Mass Index Variable Interpolation to Expand the Utility of Real-world Administrative Healthcare Claims Database Analyses. *Adv Ther* 2021;38:1314–27.
9 Lloyd JT, Blackwell SA, Wei II, *et al*. Validity of a Claims-Based Diagnosis of Obesity Among Medicare Beneficiaries. *Eval Health Prof* 2015;38:508–17.
10 Peng M, Southern DA, Williamson T, *et al*. Under-coding of secondary conditions in coded hospital health data: Impact of co-existing conditions, death status and number of codes in a record. *Health Informatics J* 2017;23:260–7.
11 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.
12 Ammann EM, Kalsekar I, Yoo A, *et al*. Assessment of obesity prevalence and validity of obesity diagnoses coded in claims data for selected surgical populations: A retrospective, observational study. *Medicine (Baltimore)* 2019;98:e16438.
13 US Food and Drug Administration. Real-world evidence. Available: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [Accessed 27 May 2024].
14 Safaei M, Sundararajan EA, Driss M, *et al*. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Comput Biol Med* 2021;136:104754.
15 Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC Med Inform Decis Mak* 2021;21:54.
16 Ryo M, Rillig MC. Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere* 2017;8.
17 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.
18 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.
19 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.
20 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.
21 Zou Q, Qu K, Luo Y, *et al*. Predicting Diabetes Mellitus With Machine Learning Techniques. *Front Genet* 2018;9:515.
22 Healthcare Cost and Utilization Project AfHRaQU. Clinical classifications software refined CCSR for ICD-10-CM diagnoses. Rockville (MD), 2021. Available: https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp
23 Oracle. Drug database. United States: Oracle Cerner; 2023.
24 Mondal PK, Foysal KH, Norman BA, *et al*. Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers. *Sensors (Basel)* 2023;23:759.
25 Samadoulougou S, Idzerda L, Dault R, *et al*. Validated methods for identifying individuals with obesity in health care administrative databases: A systematic review. *Obes Sci Pract* 2020;6:677–93.
26 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.
27 Martin B-J, 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.
28 Suissa K, Schneeweiss S, Lin KJ, *et al*. Validation of obesity-related diagnosis codes in claims data. *Diabetes Obes Metab* 2021;23:2623–31.
29 Huang Y, Li W, Macheret F, *et al*. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020;27:621–33.
30 Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.
31 Kontautiene S, Stang A, Gollnick H, *et al*. The role of phenotype, body mass index, parental and sun exposure factors in the prevalence of melanocytic nevi among schoolchildren in Lithuania. *J Eur Acad Dermatol Venereol* 2015;29:1506–16.
32 Jonnalagadda SS, Skinner R, Moore L. Overweight athlete: fact or fiction? *Curr Sports Med Rep* 2004;3:198–205.
33 Njei B, Osta E, Njei N, *et al*. An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis. *Sci Rep* 2024;14:8589.