Journal of the American Medical Informatics Association, 27(1), 2020, 136–146
doi: 10.1093/jamia/ocz173
Advance Access Publication Date: 24 October 2019
Research and Applications

Research and Applications

Imputation and characterization of uncoded self-harm in major mental illness using machine learning

Praveen Kumar [1,2] Anastasiya Nestsiarovich [1] Stuart J. Nelson [3] Berit Kerner [4] Douglas J. Perkins [1] and Christophe G. Lambert [1,2,5]

[1]Center for Global Health, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA, [2]Department of Computer Science, University of New Mexico, Albuquerque, New Mexico, USA, [3]Biomedical Informatics Center, Department of Clinical Research & Leadership, George Washington University, Washington, DC, USA, [4]Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA, and [5]Translational Informatics Division, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA

Corresponding Author: Christophe G. Lambert, PhD, University of New Mexico Health Sciences Center, Department of Internal Medicine, Center for Global Health and Division of Translational Informatics, MSC10-5550, 915 Camino de Salud NE, Albuquerque, NM 87131, USA; cglambert@unm.edu

Received 13 June 2019; Revised 16 August 2019; Editorial Decision 4 September 2019; Accepted 9 September 2019

ABSTRACT

Objective: We aimed to impute uncoded self-harm in administrative claims data of individuals with major mental illness (MMI), characterize self-harm incidence, and identify factors associated with coding bias.
Materials and Methods: The IBM MarketScan database (2003-2016) was used to analyze visit-level self-harm in 10 120 030 patients with $\geq$2 MMI codes. Five machine learning (ML) classifiers were tested on a balanced data subset, with XGBoost selected for the full dataset. Classification performance was validated via random data mislabeling and comparison with a clinician-derived "gold standard." The incidence of coded and imputed self-harm was characterized by year, patient age, sex, U.S. state, and MMI diagnosis.
Results: Imputation identified 1 592 703 self-harm events vs 83 113 coded events, with areas under the curve $>$0.99 for the balanced and full datasets, and 83.5% agreement with the gold standard. The overall coded and imputed self-harm incidence were 0.28% and 5.34%, respectively, varied considerably by age and sex, and was highest in individuals with multiple MMI diagnoses. Self-harm undercoding was higher in male than in female individuals and increased with age. Substance abuse, injuries, poisoning, asphyxiation, brain disorders, harmful thoughts, and psychotherapy were the main features used by ML to classify visits.
Discussion: Only 1 of 19 self-harm events was coded for individuals with MMI. ML demonstrated excellent performance in recovering self-harm visits. Male individuals and seniors with MMI are particularly vulnerable to self-harm undercoding and may be at risk of not getting appropriate psychiatric care.
Conclusions: ML can effectively recover unrecorded self-harm in claims data and inform psychiatric epidemiological and observational studies.

Key words: self-harm, suicide, machine learning, coding, electronic health records

INTRODUCTION

Suicide is 1 of the 10 leading causes of death in the United States and continues to have increasing incidence.[1,2] There are approximately 25 suicide attempts for every suicide death[3] with self-harming behavior being a major risk factor for subsequent suicide.[4,5] A total of 82.7% of those who attempt suicide have a concurrent mental disorder.[6] Thus, timely identification of self-harming behavior in mentally ill individuals is an essential leverage point to reduce mortality. However, only 41%-52% of adults receive outpatient care within 30 days after an emergency department visit for self-harm.[7,8]

Inadequate coding of suicidality and self-harm in medical records has been consistently reported.[9–12] Different injury reporting standards across healthcare organizations and geographical regions create additional challenges.[13–17] The lack of robust recording inhibits its appropriate screening diagnostics, referrals, and treatment.[11,18] Undercoding of self-harm also impedes the ability to estimate event prevalence and reduces the statistical power to perform time-to-event comparative effectiveness pharmacotherapy studies.

Machine learning (ML) methods have been used both to predict self-harm and to impute its presence as a missing phenotype. For the former purpose, some have used natural language processing (NLP) on clinical notes, while others have applied regression methods, random forests, or Bayesian models on electronic health record data with International Classification of Diseases-Ninth Revision (ICD-9) and -Tenth Revision (ICD-10) codes.[19–26] Artificial neural network, NLP-based models,[27–29] a hybrid ML and rule-based approach,[30] and manually developed statistical algorithms[31] have been applied to infer the presence of suicidality and self-harm. Limitations of these previous studies include small sample sizes, restrictive subpopulations (military, youths, pregnant subjects), and a relatively limited number of covariates. To our knowledge, this is the first investigation describing coded vs imputed incidence of self-harm.

The objective of this study was to apply ML algorithms at the visit level to impute self-harm events that were uncoded in claims data of individuals with major mental illness (MMI) (schizophrenia, schizoaffective disorder, major depressive disorder, and bipolar disorder). Model information was used to identify factors associated with coding discrepancies, and to characterize coded vs imputed self-harm incidence in various demographic groups. The term *self-harm* hereafter includes both suicide attempts and self-harming behavior.

MATERIALS AND METHODS

Because our study was limited to a single dataset for approach development and validation, and self-harm incidence estimates, we took a multiprong strategy to both validate the ML models and to verify we were not overfitting (Figure 1). In the following subsections, we first describe the target dataset for imputation as well as data subsets and transformations used to compare ML methods, assess prediction with different classes of covariates, and confirm recovery of deliberately mislabeled meta-visits. A per-person model was created to ensure within-individual information leakage was not occurring, and a 70%/30% train-test validation model was built to confirm that prediction performance was not explained by overfitting. To contrast ML-based and human-derived visit classification, a "gold standard" was established using the expertise of 3 clinicians. Investigation was done into important covariates for self-harm prediction, and a coding-bias-model was created to see which variables were associated with highly certain imputed self-harm cases not being coded. Finally, detailed comparisons of coded vs imputed self-harm incidence were made by patient age, sex, MMI category, and U.S. state of residence. To support replication, we have made our source code available (https://gitlab.com/PCORIUNMPUBLIC/self_harm_imputation).

Data source

The IBM Health Analytics MarketScan commercial claims and encounters database (2003-2016), containing information on 136 978 978 commercially insured U.S. individuals up to 65 years of age was transformed to the OMOP CDMv5 (Observational Medical Outcomes Partnership Common Data Model)[32] format using ETL-CDMBuilder.[33] To capture complete information related to 1 clinical event, we combined consecutive inpatient, emergency room, and outpatient visits with no gap of >1 day into meta-visits.

Data staging, phenotyping, and covariate selection

A total of 10 120 030 patients (32.9% male and 67.1% female) with ≥2 diagnostic codes for MMI during the observation period were selected, corresponding to 519 590 773 unique meta-visits. After excluding meta-visits consisting of purely outpatient visit(s), which had a negligible number of self-harm events, a total of 20 783 244 (4.0%) meta-visits were analyzed as a full dataset, corresponding to 6 037 479 unique patients (31.9% male and 68.1% female) (Figure 1).

A self-harm phenotype was defined by the presence of ICD-10-Clinical Modification (ICD-10-CM) codes X7{1-9}*, X8{0-3}*, ICD-9-Clinical Modification (ICD-9-CM) codes E95{0-9}*, SNOMED (Systematized Nomenclature of Medicine) codes 59274003, 276853009, 418420002, and their descendants. If 1 of these codes was present during a meta-visit, the meta-visit was labeled as class 1; otherwise, it was labeled as class 0.

A total of 185 234 unique covariates were selected for the analysis, including patient age, sex, meta-visit start year, and 9 feature classes: *Manually Curated, Procedure, Condition, Drug, Billing Code Position, Device, Observation, Measurement, and Ancestor terms*. The following vocabularies were used to analyze the dataset: (1) ICD-9-CM, ICD-10-CM, and SNOMED for diagnoses/conditions/observations; (2) ICD-9-CM Volume 3 (ICD-9-CM V3), ICD-10-Procedure Coding System (ICD-10-PCS), and Current Procedural Technology, Fourth Edition (CPT-4) for procedures; and (3) RxNorm ids for drugs. The procedure codes from ICD-9-CM V3 and CPT-4 vocabularies were mapped to ICD-10-PCS concepts. All ICD-9-CM and ICD-10-CM diagnosis codes were mapped to SNOMED equivalents. The billing code position covariates captured visit complexity (higher numbers implies more activity). A set of covariates was manually created by clinical experts via aggregating codes of different types and vocabularies based on similar clinical significance. None of the codes used for the self-harm phenotyping were included in the list of covariates. On average, each meta-visit had 115 features. A sparse data matrix comprising 20 783 244 meta-visit rows and 185 234 covariate columns was generated.

For extensive experimentation purposes, we created a smaller balanced dataset comprising all 83 113 class 1 meta-visits and a randomly selected set of 83 113 class 0 meta-visits.

Machine learning classification algorithms

The following 5 ML classifiers were trained and tested on the balanced dataset: tree-based XGboost[34] (Balanced-data-model), logistic regression, random forest, decision tree, and linearSVC.

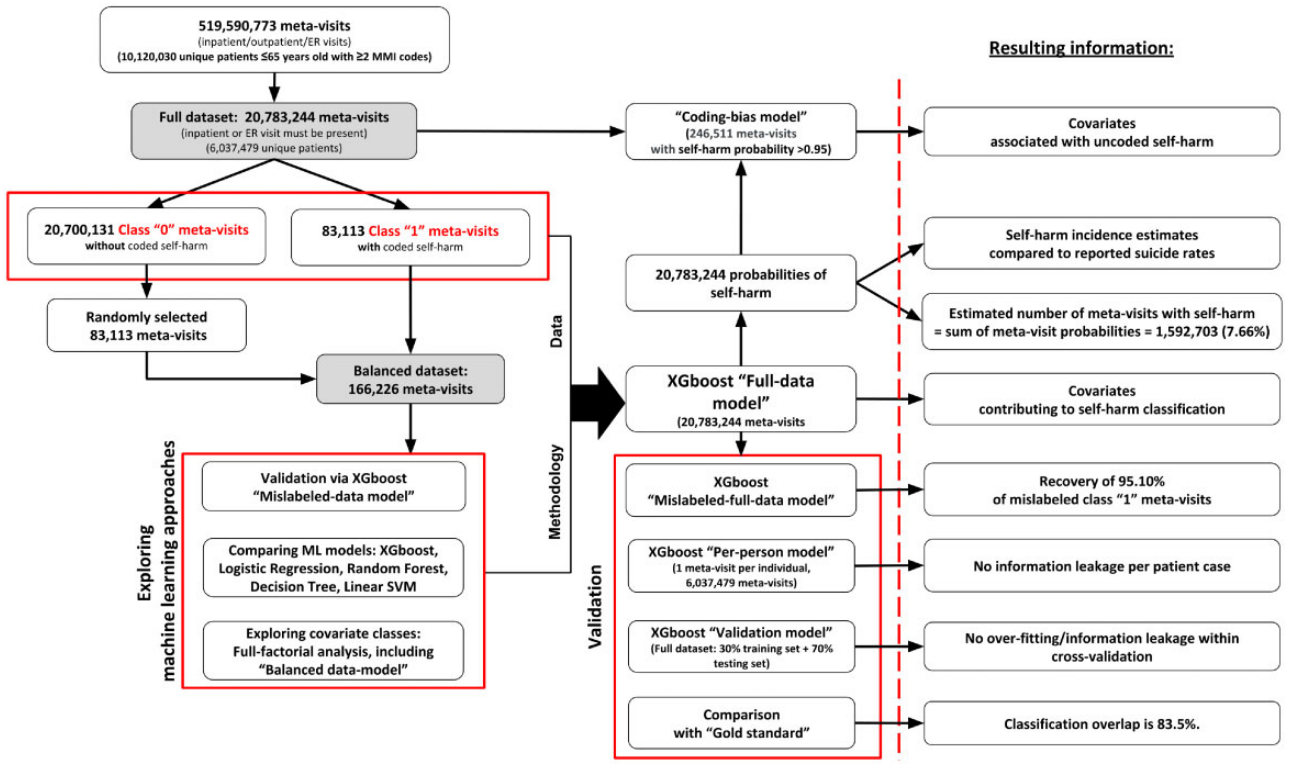

Figure 1. Study schema. The target dataset for phenotype imputation and self-harm incidence estimation was the full dataset. Machine learning (ML) approaches were first explored using the balanced dataset, with which we assessed the ability to recover deliberately mislabeled meta-visits, compared the performance of 5 ML algorithms, and explored the importance of covariate classes. The XGboost ML method was chosen for all subsequent modeling. The Full-data-model was used to characterize self-harm incidence, but several additional models were used to validate our approach and derive additional insight. The Mislabeled-full-data-model assessed whether deliberately misclassified meta-visits could be recovered with class imbalance. The Per-person model ensured prediction performance was not skewed by within-individual information leakage. The validation model verified that prediction performance was not explained by overfitting. The gold standard comparison contrasted ML classification with that of clinical experts. The XGboost model was also used to identify the most important meta-visit classification covariates, as well as features associated with uncoded self-harm via building the Coding-bias-model. ER: emergency room; MMI: major mental illness; SVM: support vector machine.

Five-fold cross-validation was repeated 100 times to report the area under the receiver-operating characteristic curve (AUC-ROC), accuracy (ACC), and Matthews correlation coefficient (MCC)[35] with 90% confidence intervals. The scikit-learn ML library was used for all classifiers.[36]

Owing to computing limitations, only the XGboost model was trained and tested on the full dataset (Full-data-model) and 5-fold cross-validation was repeated 10 times to report the AUC-ROC, accuracy, and MCC with 80% confidence intervals. The other ML algorithms were either too slow or failed to execute on the full dataset.

We used default values of parameters for the ML methods comparison, with final optimization of XGboost parameters done through the grid search cross-validation method on the balanced dataset.

Computational validation of the machine learning classification algorithms

To validate recovery of uncoded self-harm, we randomly mislabeled half of the class 1 meta-visits in the balanced dataset as class 0 and built the XGboost Mislabeled-data-model with 5-fold cross-validation, reporting performance using the original labels. We repeated the converse of this experiment by changing label 0 to 1 in a randomly selected half of the class 0 meta-visits. We also randomly mislabeled half of the class 1 meta-visits in the full dataset as class 0, building the XGboost Mislabeled-full-data-model with 5-fold cross-validation, and reported performance using the original labels.

On average, there were 3.44 meta-visits per person in the full dataset. To verify that our classification models did not overfit due to within-individual information leakage, we built the XGboost Per-person-model with 5-fold cross-validation on the dataset comprising 1 randomly selected meta-visit per person (6 037 479 meta-visits).

To ensure there was no overfitting within the Full-data-model, the XGboost Validation-model was trained on a random 70% subset of meta-visits of the full dataset and validated on the remaining 30%.

Clinician validation: creating a gold standard for meta-visit classification

To contrast the ML-based and human-derived classification of patient encounters, a total of 200 meta-visits were selected based on the Full-data-model probabilities (threshold of 0.5). Random selection from the full dataset could result in a very low number of self-harm codes captured (either coded or imputed), as it was a rare event. Thus, 50 meta-visits were randomly selected from each of 4 possible categories: with self-harm coded and imputed, coded but not imputed, imputed but not coded, and neither coded nor imputed. Three clinicians manually classified these meta-visits as either class 1 or class 0 by independently reviewing their features. Clinicians were aware of neither the self-harm code presence nor the ML

classification, but knew that all patients were diagnosed with MMI. Self-harm evidence was assessed based on the presence of poisoning or open wound of wrist, forearm, neck, or body torso. Current suicidal ideation or "accidents" were not sufficient conditions to label a meta-visit as having "self-harm," if coded alone. Some labeling discrepancies among clinicians were related to recreational drug overdose or poisoning, which, after discussion, were considered as not being indicative of self-harm when used alone. Also discussed were codes for cause of body part injury, such as evidence of iatrogenic effect, body site inflammation, allergy testing etc. Classification agreement was assessed between the XGboost model, each clinician, and the gold standard consensus reached among experts after joint discussion. The gold standard was not considered as absolute "truth" relative to ML classification, but rather as an alternative judgement on the same data source.

Important covariates used by the XGBoost classifier

To identify covariates with the greatest contribution to Full-data-model classification, we computed relative gain, relative weight, and relative cover for each covariate.

We also performed a full-factorial analysis (Full-factorial-models) on the balanced dataset using 510 combinations of 9 covariate classes to determine which contributed the most to model performance. The only 2 combinations that were not tested were (1) no class of covariate was selected and (2) only ancestors were selected (at least 1 class of covariates is needed to have class ancestors).

Features associated with uncoded self-harm

To identify the features influencing the likelihood of a meta-visit being coded as self-harm, we selected all meta-visits with class 1 probability assigned $\geq$0.95 by the Full-data-model, and built an XGboost model (Coding-bias-model) to classify whether these meta-visits were coded or not. To assess the relative importance and directionality (as a main effect) of a particular variable to coding self-harm, a log-ratio score was calculated for the 100 variables with highest gain scores. The log-ratio score indicates the degree of over- or underrepresentation of a variable in class 1 vs class 0 meta-visits, computed as: log((A*C/B), 2), where A is number of uncoded meta-visits with the covariate divided by number of coded meta-visits with the covariate, B is number of uncoded meta-visits, and C is number of coded meta-visits.

Characterization of self-harm incidence and patterns related to patient features

The incidence of coded and imputed self-harm was computed as a function of patient age, sex, meta-visit start year, state of residence, and MMI type. The incidence of coded self-harm was computed as the number of meta-visits with coded self-harm divided by the number of years of patient observation. The incidence of imputed self-harm was the sum of class 1 probabilities of meta-visits predicted by the Full-data-model divided by the number of years of patient observation.

To explore the observed sex-related differences in self-harm coding, we additionally compared the fraction of meta-visits with coded psychotherapy and suicidal or harmful thoughts by sex. The respective codes were selected from the list of covariates which contributed the most to the Full-data-model classification, and were analyzed in all meta-visits, and in meta-visits with coded self-harm.

RESULTS

Patient characteristics

The average patient age was 39.8 years for male individuals and 39.9 years for female individuals in the full dataset. The fraction of patients with different types of MMI was: 85.94% major depressive disorder, 5.37% bipolar disorder, 8.26% more than 1 MMI, 0.29% schizophrenia, and 0.14% schizoaffective disorder.

Performance of machine learning classification models

Out of 20 783 244 meta-visits recorded over 29 799 203 years of patient observation, the XGboost Full-data-model probabilities of self-harm (class 1) summed to 1 592 703 (7.66%), corresponding to an overall imputed annual incidence of 5.34%. The annual coded incidence was 0.28%. Of all ~20 million meta-visits, 842 263 (4.05%) had class 1 probability >0.5 and 246 511 (1.19%) had class 1 probability $\geq$0.95. Also, of 83 113 meta-visits coded for self-harm, 79 882 (96.11%) had class 1 probability >0.5 and 62 929 (75.71%) had class 1 probability $\geq$0.95. Performance of the XGboost-based ML models trained and tested on different datasets are shown in Table 1. The performance of the 5 different ML classification algorithms applied to the balanced dataset is shown in Table 2.

Comparison of machine learning meta-visits classification with the gold standard

The pairwise agreement between the Full-data-model, individual clinical experts, and the consensus gold standard is shown in Table 3.

Out of 200 meta-visits, 79 were categorized as class 1 by the gold standard. Among 100 meta-visits with documented self-harm, clinical experts labeled 52 as such, whereas in 100 meta-visits without documented self-harm, clinicians classified only 27 as having self-harm. The overall agreement between ML and gold standard was 84%.

Important covariates used by the XGBoost classifier

Out of 185 234 covariates, only 2205 (1.19%) contributed to the Full-data-model. The 15 covariates with the highest gain scores are shown in Table 4.

Classification results of the Full-factorial-models are shown in Supplementary Table S1. The model exclusively built with condition covariates had only slightly worse performance compared with a model built with all covariate classes (AUC 0.988 vs 0.991). Adding higher-order ancestor concepts had a negligible effect on the model performance, except for procedures, in which the AUC score increased from 0.800 to 0.828 after adding the ICD-10-PCS ancestor terms. When the uncommon and poorly predictive Device covariates were used alone, the AUC and accuracy were 0.51, and the MCC was 0.1.

Features associated with uncoded self-harm

The 100 covariates which contributed most to the Coding-bias-model are in Supplementary Table S2. Among the factors associated with higher likelihood of self-harm coding were intoxication and poisoning, accidents, asphyxiation, chest and head surgical repair, wrist wound, self-harming thoughts, depression, and psychotherapy. Features associated with lower likelihood of coded self-harm included substance dependence or abuse, heroin poisoning, neurological disorder, brain visualization, vehicle accidents, and falls.

Table 1. Classification performance of different XGboost-based classification models on different sets of meta-visits in patients with major mental illness

XGboost model	Validation method	Dataset	Accuracy	MCC	AUC-ROC
Full-data-model	5-fold cross-validation repeated 10 times	Full dataset with ∼20 million meta-visits	$0.960 \pm 4 \times 10^{-3}$	$0.297 \pm 2 \times 10^{-4}$	$0.990 \pm 4 \times 10^{-4}$
Per-person-model	5-fold cross-validation	Full dataset subset of ∼6 million meta-visits with 1 random meta-visit per person	0.966	0.334	0.991
Validation-model	5-fold cross-validation on the training set	70% random meta-visits from the full dataset	0.964	0.298	0.991
	Testing on the validation set	Remaining 30% of meta-visits from the full dataset	0.963	0.296	0.990
Balanced-data-model	5-fold cross-validation repeated 100 times	Balanced dataset with 166 000 meta-visits	$0.964 \pm 2 \times 10^{-4}$	$0.928 \pm 4 \times 10^{-4}$	$0.991 \pm 4 \times 10^{-4}$
Mislabeled-data-model	5-fold cross-validation. Original labels of meta-visits were used for assessing performance	Half of the class 1 meta-visits mislabeled in the balanced dataset	0.962	0.924	0.989
		Half of the class 0 meta-visits mislabeled in the balanced dataset	0.963	0.926	0.991
Mislabeled-full-data-model		Half of the class 1 meta-visits mislabeled in the full dataset	0.974	0.347	0.991
Coding-bias-model	5-fold cross-validation	All meta-visits from the full dataset with class 1. Probability threshold ≥0.95	0.679	0.306	0.738
Full-factorial-models		Balanced dataset; only condition covariates	0.957	0.914	0.988
		Balanced dataset; only hand-curated covariates	0.927	0.853	0.977
		Balanced dataset; only billing code position covariates.	0.788	0.577	0.875
		Balanced dataset; only observation covariates	0.775	0.562	0.813
		Balanced dataset; only procedure covariates	0.708	0.440	0.800
		Balanced dataset; only measurement covariates	0.589	0.245	0.594
		Balanced dataset; only drug covariates	0.550	0.192	0.586
		Balanced dataset; only device covariates	0.516	0.099	0.514

The results for the Full-data-model and the Balanced-data-model are shown with 80% and 90% confidence intervals, respectively. AUC-ROC: area under the receiver-operating characteristic curve; MCC: Matthews correlation coefficient.

Table 2. Classification performance of 5 different machine learning algorithms on the balanced dataset of patients with major mental illness, using 5-fold-cross-validation with 100 repetitions and reported with 90% confidence intervals

Machine learning model/performance	XGboost balanced-data-model with optimized parameters	XGboost balanced-data-model with default parameters	Logistic regression	Random forest	Decision tree	LinearSVC
Accuracy	$0.964 \pm 2 \times 10^{-4}$	$0.961 \pm 2 \times 10^{-4}$	$0.963 \pm 3 \times 10^{-4}$	$0.946 \pm 1 \times 10^{-3}$	$0.947 \pm 7 \times 10^{-4}$	$0.959 \pm 3 \times 10^{-4}$
MCC	$0.928 \pm 4 \times 10^{-4}$	$0.922 \pm 4 \times 10^{-4}$	$0.926 \pm 6 \times 10^{-4}$	$0.892 \pm 3 \times 10^{-3}$	$0.896 \pm 1 \times 10^{-3}$	$0.919 \pm 7 \times 10^{-4}$
AUC-ROC	$0.991 \pm 4 \times 10^{-4}$	$0.990 \pm 2 \times 10^{-4}$	$0.990 \pm 1 \times 10^{-4}$	$0.982 \pm 6 \times 10^{-4}$	$0.948 \pm 7 \times 10^{-4}$	$0.988 \pm 1 \times 10^{-4}$

Optimized parameters for XGboost model: max_depth = 6, base_score = 0.5, gamma = 0, max_delta_step = 0, min_child_weight = 2, objective = 'binary: logistic', booster = 'gbtree', subsample = 0.6, scale_pos_weight = total negative class/total positive class, colsample_bytree = 1, colsample_bylevel = 0.8, learning_rate = 0.04

Characterization of self-harm incidence and patterns related to patient features

Figure 2 shows the coded and imputed self-harm incidence by sex from 2003-2016. Both coded and imputed self-harm rose from 2006 onward. The incidence of coded self-harm ranged from 0.09% to 0.54% for male individuals and from 0.11% to 0.49% for female individuals throughout the observation period. The incidence of imputed self-harm ranged from 4.09% to 8.75% for male individuals

Table 3. The pairwise agreement between the XGboost Full-data model consensus gold standard and 3 clinicians regarding the presence of self-harm (with >0.5 probability) in 200 selected meta-visits of patients with major mental illness

Classifier	Full-data-model	Clinician 1	Clinician 2	Clinician 3	Gold standard
200 randomly selected meta-visits					
Full-data-model	1.00	0.81	0.80	0.78	0.84
Clinician 1	0.81	1.00	0.77	0.88	0.88
Clinician 2	0.80	0.77	1.00	0.76	0.86
Clinician 3	0.79	0.88	0.76	1.00	0.87
Gold standard	0.84	0.88	0.86	0.87	1.00
50 meta-visits where self-harm was neither coded nor imputed					
Full-data-model	1.00	0.98	1.00	0.96	1.00
Clinician 1	0.98	1.00	0.98	0.98	0.98
Clinician 2	1.00	0.98	1.00	0.96	1.00
Clinician 3	0.96	0.98	0.96	1.00	0.96
Gold standard	1.00	0.98	1.00	0.96	1.00
50 meta-visits where self-harm was not coded but imputed					
Full-data-model	1.00	0.54	0.68	0.60	0.54
Clinician 1	0.54	1.00	0.70	0.78	0.76
Clinician 2	0.68	0.70	1.00	0.76	0.82
Clinician 3	0.60	0.78	0.76	1.00	0.90
Gold standard	0.54	0.76	0.82	0.90	1.00
50 meta-visits where self-harm was coded but not imputed					
Full-data-model	1.00	0.74	0.64	0.66	0.88
Clinician 1	0.74	1.00	0.54	0.84	0.82
Clinician 2	0.64	0.54	1.00	0.50	0.68
Clinician 3	0.66	0.84	0.50	1.00	0.74
Gold Standard	0.88	0.82	0.68	0.74	1.00
50 meta-visits where self-harm was both coded and imputed					
Full-data-model	1.00	0.98	0.88	0.88	0.92
Clinician 1	0.98	1.00	0.86	0.90	0.94
Clinician 2	0.88	0.86	1.00	0.80	0.92
Clinician 3	0.88	0.90	0.80	1.00	0.88
Gold standard	0.92	0.94	0.92	0.88	1.00

Table 4. Covariates from the Full-data-model contributing most to XGboost meta-visit classification for self-harm presence

OMOP concept ID	SNOMED concept ID	Covariate	Relative gain	Relative cover	Relative weight
442562	75478009	Poisoning	0.3200	0.0273	0.0103
444100	46206005	Mood disorder	0.0359	0.0109	0.0142
440921	417746004	Traumatic injury	0.0226	0.0060	0.0026
—	—	External injury	0.0158	0.0062	0.0314
432586	74732009	Mental disorder	0.0139	0.0089	0.0163
73553	399269003	Arthropathy	0.0135	0.0041	0.0014
4168335	416462003	Wound	0.0099	0.0093	0.0047
438028	7895008	Poisoning by drug and/or medicinal substance	0.0082	0.0045	0.0116
4108646	283057008	Abrasion of upper limb	0.0079	0.0154	0.0005
444187	125643001	Open wound	0.0067	0.0051	0.0075
4130851	127278005	Injury of upper extremity	0.0063	0.0084	0.0059
4219871	399963005	Abrasion	0.0055	0.0014	0.0003
—	—	Psychiatric diagnosis	0.0048	0.0044	0.0012
4306645	83507006	Finding of thought content	0.0046	0.0014	0.0047
4111213	285261008	Dangerous and harmful thoughts	0.0042	0.0045	0.0075

The covariates are sorted by relative gain, which reflects the magnitude of covariate contribution to predicting the class of the meta-visit relative to other features (relative gain = gain of the covariate / ∑gain of all covariates). The weight indicates how many times the covariate was used to split the data across all trees in the model (relative weight = weight of the covariate / ∑weight of all covariates). The cover indicates the average number of observations in which the covariate was used to split the data across all trees in the model (relative cover = cover of the covariate ÷ ∑cover of all covariates).

OMOP: Observational Medical Outcomes Partnership; SNOMED: Systematized Nomenclature of Medicine.

and from 3.49% to 7.09% for female individuals. For each year, the incidence for coded self-harm was comparable in both sexes, but the incidence of imputed self-harm was consistently higher in male than in female patients.

The patterns of coded and imputed self-harm in patients of different age and sex are shown in Figure 3. In younger age groups (12-21 years), the incidence of coded self-harm was higher in female individuals than in male individuals. From 38 to 65 years of age, the incidence

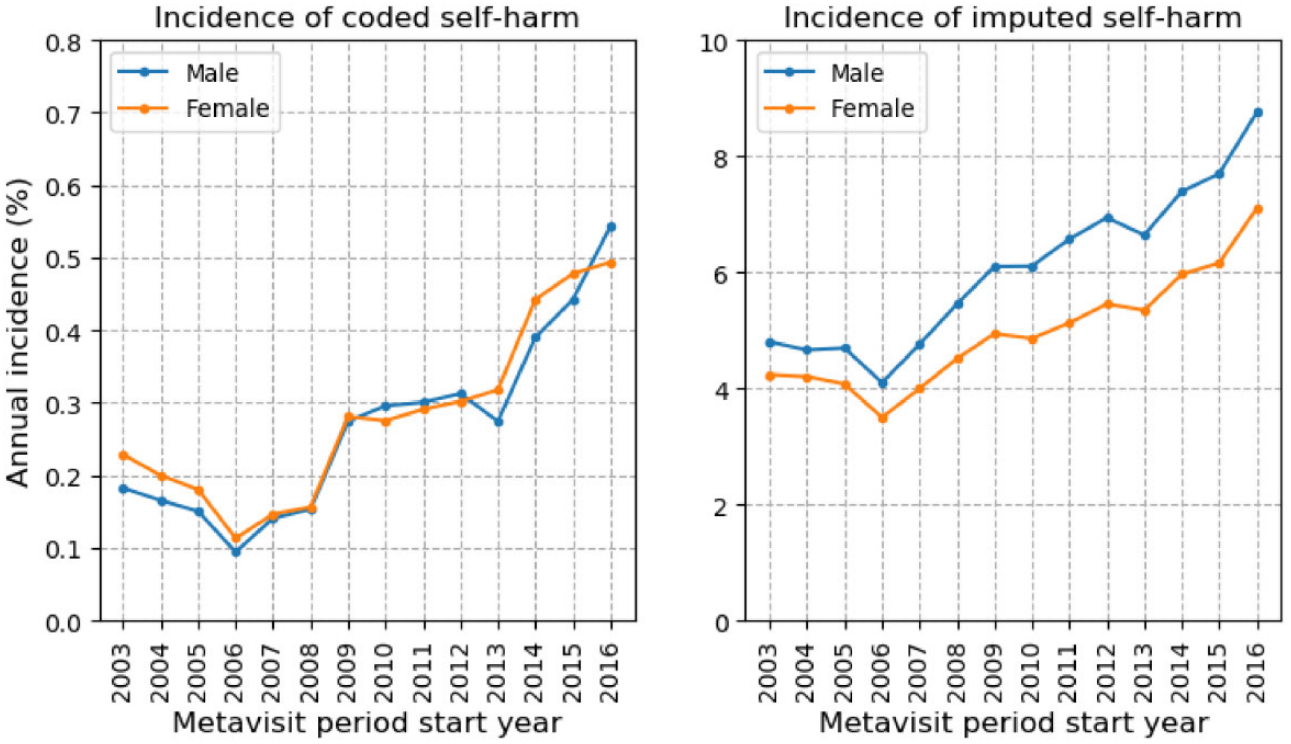

Figure 2. Self-harm meta-visits in patients with major mental illness of different sex per year. The left graph shows the annual percentage incidence of coded self-harm for male individuals (blue line) and female individuals (orange line); the right graph shows the annual percentage incidence of imputed self-harm for male individuals (blue line) and female individuals (orange line).

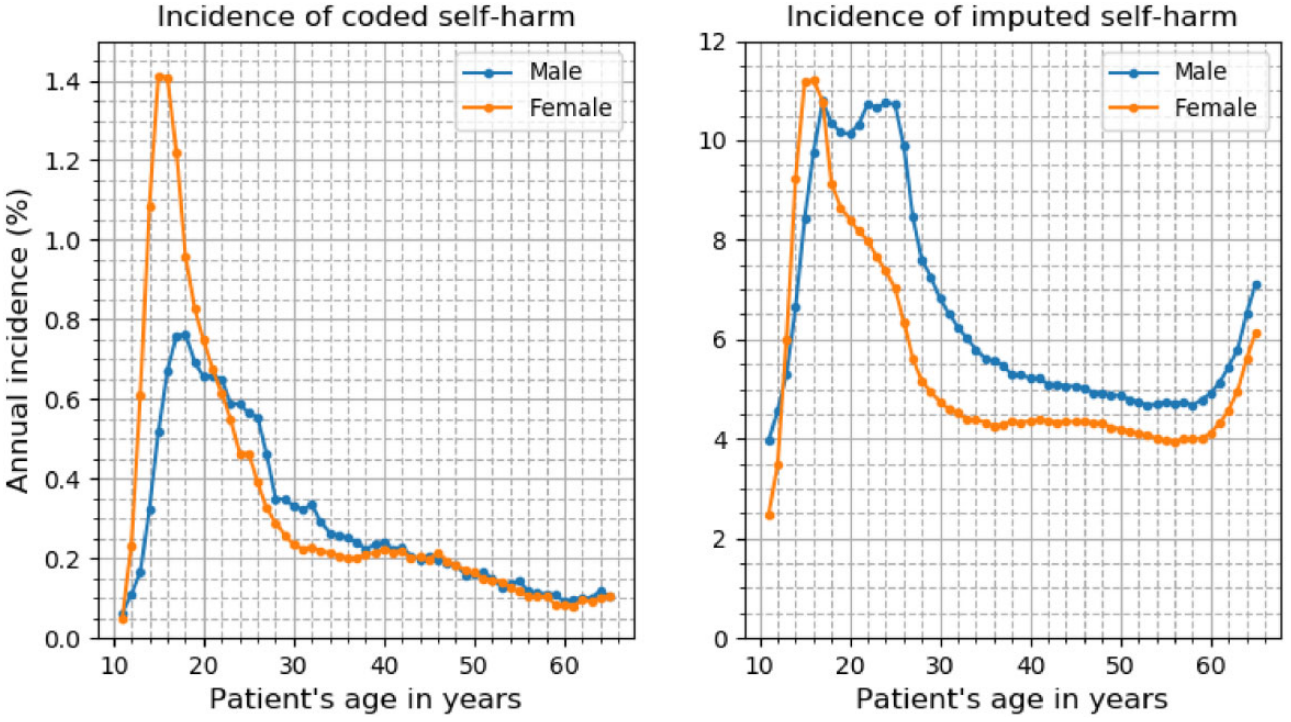

Figure 3. Self-harm meta-visits in patients with major mental illness of different age and sex. The left graph shows the annual percentage incidence of coded self-harm in male individuals (blue line) and female individuals (orange line). The right graph shows the annual percentage incidence of machine learning-imputed self-harm by sex.

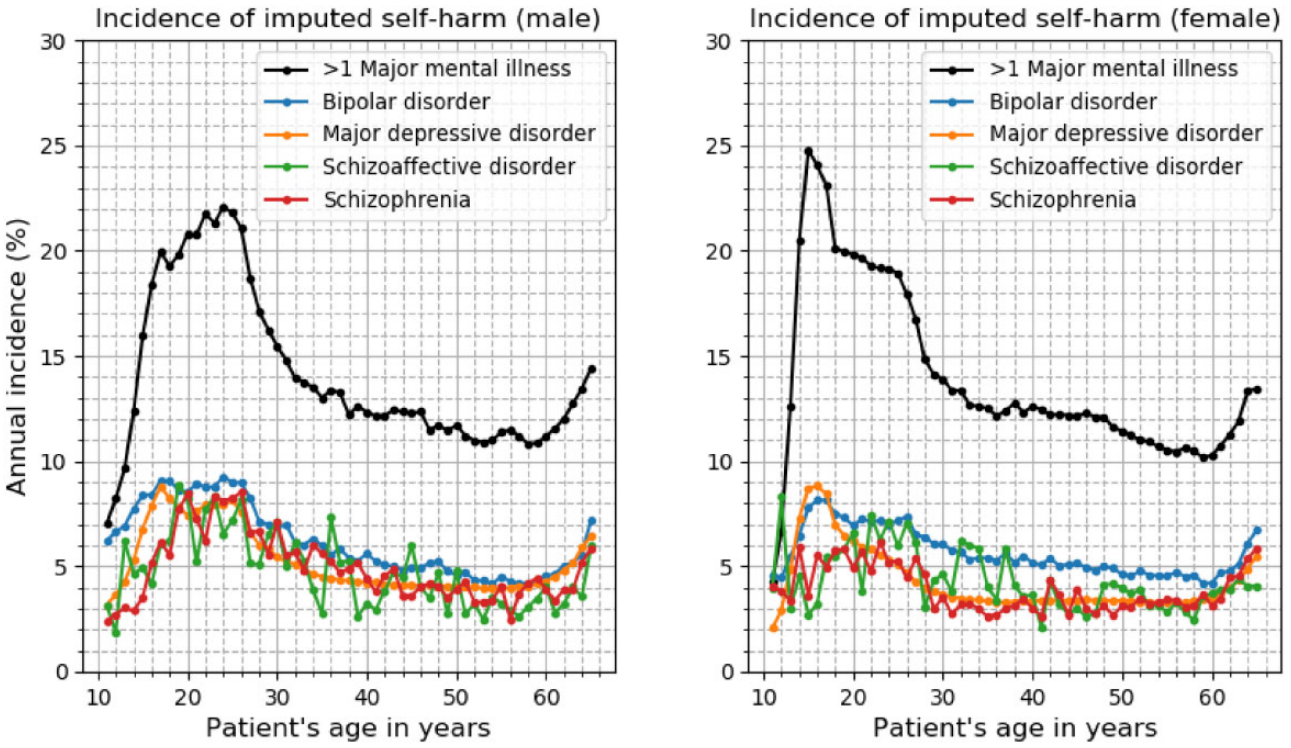

Figure 4. The annual incidence of meta-visits with machine learning-imputed self-harm by category of major mental illness. The left graph shows the data for male individuals and the right graph for female individuals.

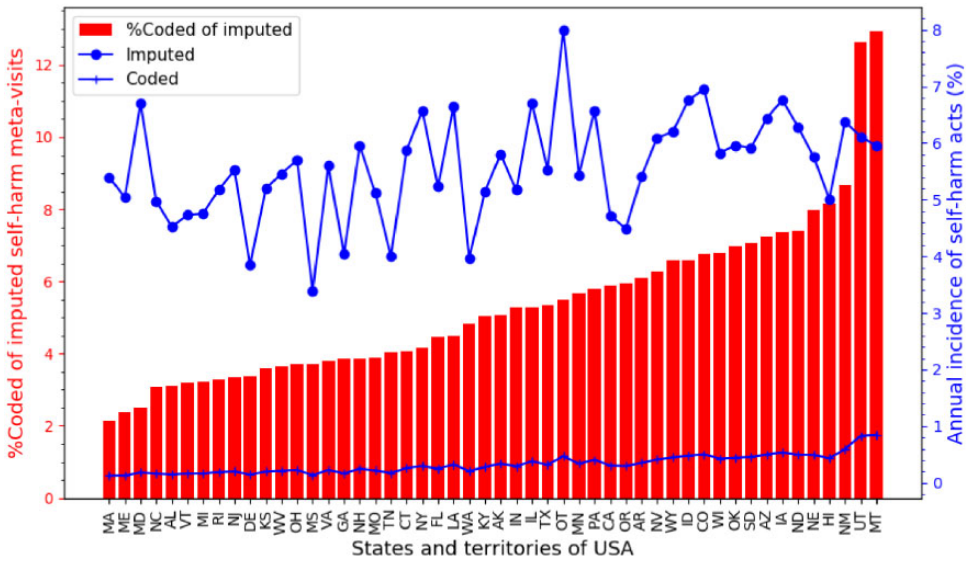

Figure 5. Meta-visits with self-harm in patients with major mental illness residing in different states and territories of the United States. The blue plots show the annual percentage incidence of meta-visits with imputed self-harm (blue dots) and coded self-harm (blue hatches). The red bars show the fraction of coded self-harm events among the imputed ones. Due to MarketScan license restrictions, data for South Carolina were excluded from the figure. OT: others (includes DC, Puerto Rico and other U.S. territories).

of coded self-harm was comparable in both sexes; for the interval of 22 to 37 years of age, it was markedly higher for male than female individuals. The incidence of imputed self-harm was higher in young female vs male individuals 13-16 years of age, and higher for male than female patients 17 years of age and older. The fraction of coded self-harm was higher in young individuals vs older ones, and in female vs male patients (Supplementary Figure S1). The incidence of coded and imputed self-harm based on patient age, not differentiated by sex, is shown in Supplementary Figure S2.

The fraction of meta-visits with coded "suicidal thoughts" was higher for male than for female individuals (3.1% vs 1.9%). Male patients also had more meta-visits coded with "dangerous and harmful thoughts" (3.2% vs 1.9%). Both psychotherapy and suicidal/harmful thoughts were coded at least 1.26 times more often in

male than female individuals when all meta-visits were analyzed. However, among meta-visits with coded self-harm, these differences were not observed (Supplementary Table S3).

The incidence of imputed self-harm by MMI category and patient sex is shown in Figure 4. The average percentage incidences of self-harm by MMI category over all ages were as follows (male vs female patients): >1 MMI: 14.11% vs 13.90%; bipolar disorder: 6.23% vs 5.68%; major depressive disorder: 5.25% vs 4.27%; schizoaffective disorder: 4.75% vs 4.31%; schizophrenia: 4.95% vs 3.90%. More details on each disease category can be found in Supplementary Figure S3.

The patterns of coded and imputed self-harm in different U.S. states are shown in Figure 5. The annual incidence of coded self-harm ranged from 0.13% to 0.84%, whereas imputed self-harm incidence ranged from 3.40% to 7.99% among different states. All but 2 U.S. states (Montana and Utah) coded <10% of the imputed self-harm. The lowest incidence of coded self-harm (0.13%) was observed in Massachusetts, where injury code reporting is mandated,[13] whereas the highest incidence of coded self-harm (0.84%) was observed in Montana, despite having no such mandates.

DISCUSSION

Self-harm events are underreported, but can be imputed using ML algorithms. Underreporting varies by sex and age, evincing potential coding bias. Self-harm incidence varies considerably across ages, by sex, and within MMI categories. The performance similarity of the Full-data-model with both the Validation-model and the Per-person-model gives assurance that overfitting was not occurring. Additionally, the near random performance of a model with sparse, low information content, device covariates, gives confidence that our cross-validation processes were not compromised.

Recovering uncoded self-harm using the ML model

The Full-data-model probabilistically estimated the presence of 1 592 703 meta-visits with self-harm, whereas only 83 113 (5.22% of the estimate) were coded. Multiple ML experimental approaches demonstrated excellent performance. Even when half of the class 1 meta-visits in the full dataset were deliberately mislabeled, the ML model recovered 95.10% of them. This shows the effectiveness of ML models in recovering uncoded self-harm and supports earlier findings that suicidality is vastly underreported in administrative claims data.[16,31]

The overall agreement between the XGboost ML classification and the gold standard was 83.5% for the selected 200 meta-visits. Despite the fact the experts agreed with ML on only 54% of the uncoded but imputed self-harm meta-visits, many of these cases were not necessarily misclassified by the model given 10%-24% of interexpert variability.

Important covariates used by the XGBoost classifier

Poisoning, mood disorder, and traumatic injury were the 3 covariates with the highest contribution to the Full-data-model. Ingestion of dangerous substances is known as one of the most common means of suicide[37,38]; thus, it was expected to play an important role in providers assigning self-harming motives to patients. The evidence of intent underlying external injury remains a debatable question and is influenced by story representation of patients or witnesses. It is also a major source of discrepancies when computerized discharge data are compared with expert opinion.[9,10] Harmful thoughts were the most direct indicator of self-harming motives.

Features associated with uncoded self-harm

Meta-visits with drug abuse are less likely to have self-harm coded, presumably reflecting challenges distinguishing between accidental substance overdose driven by patient desire to experience psychotropic effect, and deliberate self-harming behavior. On the contrary, poisoning with non-narcotic substances, which are not expected to produce euphoric effects, were associated with higher probability of coding self-harm. Accidents, asphyxia, and damage to body areas commonly traumatized by self-harm (chest, head, wrist) can also be perceived as alarming signs of self-inflicted injury and can foster coding. Interestingly, vehicle accidents and falls were less likely to be coded with self-harm, probably being less likely to arouse provider suspicion. Higher medical attention to mental health (recognition of harmful thoughts via detailed interviews, psychotherapy) is associated with a higher likelihood of coding self-harm. This supports the common understanding that lack of psychological screening can lead to missed opportunities to discover patients in crisis and provide them with care.

Characterization of self-harm incidence and patterns related to patient features

The estimated incidence of coded self-harm in mentally ill individuals in our study was higher than reported for the general U.S. population (0.28% vs 0.14%).[39] However, imputed self-harm incidence was many times higher than both of these estimates, suggesting significant underreporting. Statistics on the incidence and prevalence of self-harm can be unreliable due to social taboo from disclosure[40] and due to general electronic health record use caveats.[41]

The annual incidences for coded and imputed self-harm followed a similar temporal trend, declining in 2006, then steadily rising each year, with a small decrease observed in 2013. This finding is consistent with reported suicide rate increases following the year 2006 in the general population,[39] and with rates of nonfatal self-harm increasing from 2001 to 2017.[39]

Our findings support the evidence on higher incidence of self-harm in adolescents and young adults,[42]. However, the fraction of uncoded self-harm increased with age (Supplementary Figure S1), and imputed self-harm incidence also increased in older individuals (>58 years of age). Thus, although young people have the highest risk of self-harming behavior, seniors are more vulnerable to self-harm being uncoded in billing data.

There is evidence on self-harm coding discrepancies by sex: male patients had a higher incidence of imputed self-harm than female (except for 13-16 years of age), a longer plateau of self-harm incidence maximum values, and a lower fraction of coded self-harm in meta-visits with imputed self-harm. Thus, the problem of underreporting of self-harming behavior is especially relevant for the male population with MMI, although this sex-related difference is hidden from direct observation in the billing data. For instance, the CDC reports lower age-adjusted rates of self-harm in male vs female individuals (118.39 vs 167.56 per 100 000).[39] This sex-related coding disparity could result from either patient or provider underreporting of self-harming intentions. Contrary to other research,[43,44] suicidal thoughts were coded more often in male than female patients in our study, suggesting potential provider assessment bias related to patient sex. Psychotherapy was also more commonly coded in male patients; thus, our findings challenge the common generalization

that male patients are less likely to disclose self-harm and seek psychiatric and psychological help.

Psychiatric Axis I disorders are identified in >80% of individuals presenting with self-harm.[45] Our data provide additional evidence that multiple comorbid MMI diagnoses are associated with several times higher annual incidence of coded and imputed self-harm. The poor prognosis of such patients may be a function of case complexity, or early misdiagnosis leading to inappropriate treatment, followed by subsequent rediagnosis.

Limitations

The findings from this study cannot be generalized to individuals with MMI attending purely outpatient visits, or those with completed suicides—the latter are rarely recorded in claims data. Although robust, data presented here are not representative of the entire U.S. population. For example, patients over 65 years of age were not included in the analyses. In addition, because Medicaid data were not available, a substantial population of severe and disabled MMI cases were not present in the analyses. The estimates of self-harm incidence reported here may, therefore, be lower than the actual population of individuals seeking care for mental health. An additional limitation of the study is the absence of clinical notes, which could limit the imputation power of ML models and verification of self-harm labeling. Last, yearly estimates of self-harm incidence could be impacted by changes in patient demographics since the mix of insurer data incorporated into the MarketScan data changed (mainly increased) over time.

CONCLUSION

- Only a small fraction (~1 of 19) of self-harm events comorbid with MMI were reported in U.S. administrative claims data.
- Machine learning classification models can effectively recover uncoded self-harm meta-visits, demonstrating excellent performance on multiple experiments with claims data including random meta-visit mislabeling.
- The incidence of imputed self-harm had 2 periods of elevation: at adolescence or young adulthood and >58 years of age.
- Self-harm undercoding steadily increases with age.
- The incidence of self-harm peaks sooner and drops earlier in young female vs male patients with MMI. However, male patients are more likely to have self-harm undercoded at all ages, indicating possible provider labeling bias related to sex stereotypes.
- Psychiatric indicators of suicidality (depression, harmful thoughts) and somatic events (poisoning, asphyxiation, chest, head, wrist traumas) are associated with higher self-harm coding in patients with MMI.
- Mentally ill individuals with comorbid substance abuse, including opioids, and neurological findings are less likely to have self-harm coded in their billing data.
- For all age groups, multiple comorbid MMI diagnoses were associated with 2-fold higher self-harm incidence.

FUNDING

This work was supported by the Patient-Centered Outcomes Research Institute award CER-1507-3160 (PI Lambert, 2016), and was part of the research project "Longitudinal Comparative Effectiveness of Bipolar Disorder Therapies" (NCT02893371). The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors or Methodology Committee.

AUTHOR CONTRIBUTIONS

All authors have fulfilled the criteria for authorship established by the International Committee of Medical Journal Editors and approved submission of the manuscript. PK, AN, and CGL were the primary authors of the article and SJN, BK, and DJP contributed to manuscript authorship and revision. PK and CGL performed all data analyses. AN, BK, and SJN contributed significantly to the "gold standard" portion of the article. CGL made substantial contributions to the conception and design of the study and secured study-specific funding. All coauthors participated in revising the manuscript critically, made important intellectual contributions, and approved the final version to be published.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Journal of the American Medical Informatics Association online.

COMPETING INTERESTS STATEMENT

None declared.

REFERENCES

1. WISQARS Leading Causes of Death Reports. https://webappa.cdc.gov/sasweb/ncipc/leadcaus10_us.html. Accessed March 26, 2019.
2. WISQARS Fatal Injury Reports. https://webappa.cdc.gov/sasweb/ncipc/mortrate10_us.html. Accessed March 26, 2019.
3. Crosby AE, Han B, Ortega LAG, *et al*. Suicidal thoughts and behaviors among adults aged ≥18 years–United States, 2008-2009. *MMWR Surveill Summ* 2011; 60: 1–22.
4. Crandall C, Fullerton-Gleason L, Aguero R, *et al*. Subsequent suicide mortality among emergency department patients seen for suicidal behavior. *Acad Emerg Med* 2006; 13 (4): 435–42.
5. Andover MS, Gibb BE. Non-suicidal self-injury, attempted suicide, and suicidal intent among psychiatric inpatients. *Psychiatry Res* 2010; 178 (1): 101–5.
6. Canner JK, Giuliano K, Selvarajah S, *et al*. Emergency department visits for attempted suicide and self harm in the USA: 2006-2013. *Epidemiol Psychiatr Sci* 2018; 27 (1): 94–102.
7. Olfson M, Marcus SC, Bridge JA. Emergency treatment of deliberate self-harm. *Arch Gen Psychiatry* 2012; 69 (1): 80–8.
8. Marcus SC, Bridge JA, Olfson M. Payment source and emergency management of deliberate self-harm. *Am J Public Health* 2012; 102 (6): 1145–53.
9. Bethell J, Rhodes AE. Identifying deliberate self-harm in emergency department data. *Health Rep* 2009; 20 (2): 35–42.
10. LeMier M, Cummings P, West TA. Accuracy of external cause of injury codes reported in Washington State hospital discharge records. *Inj Prev* 2001; 7 (4): 334–8.
11. Kemball RS, Gasgarth R, Johnson B, *et al*. Unrecognized suicidal ideation in ED patients: are we missing an opportunity? *Am J Emerg Med* 2008; 26 (6): 701–5.
12. Anderson HD, Pace WD, Brandt E, *et al*. Monitoring suicidal patients in primary care using electronic health records. *J Am Board Fam Med* 2015; 28 (1): 65–71.
13. Abellera J, Annest JL, Conn JM, *et al*. How states are collecting and using cause of injury data: 2004 update to the 1997 report. 2005. http://www.cste2.org/webpdfs/ECodeFinal3705.pdf Accessed August 30, 2017.

14. Annest JL, Fingerhut LA, Gallagher SS, *et al*. Strategies to improve external cause-of-injury coding in state-based hospital discharge and emergency department data systems: recommendations of the CDC Workgroup for Improvement of External Cause-of-Injury Coding. *MMWR Recomm Rep* 2008; 57 (RR-1): 1–15.
15. Carlson KF, Nugent SM, Grill J, *et al*. Accuracy of external cause-of-injury coding in VA polytrauma patient discharge records. *J Rehabil Res Dev* 2010; 47 (8): 689–97.
16. Lu CY, Stewart C, Ahmed AT, *et al*. How complete are E-codes in commercial plan claims databases? *Pharmacoepidemiol Drug Saf* 2014; 23 (2): 218–20.
17. Hoffman GJ, Hays RD, Shapiro MF, *et al*. Claims-based identification methods and the cost of fall-related injuries among US older adults. *Med Care* 2016; 54 (7): 664–71.
18. Ting SA, Sullivan AF, Miller I, *et al*. Multicenter study of predictors of suicide screening in emergency departments. *Acad Emerg Med* 2012; 19 (2): 239–43.
19. Poulin C, Shiner B, Thompson P, *et al*. Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS One* 2014; 9 (1): e85733.
20. Kessler RC, Warner CH, Ivany C, *et al*. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study to Assess Risk and rEsilience in Servicemembers (Army STARRS). *JAMA Psychiatry* 2015; 72 (1): 49–57.
21. Barak-Corren Y, Castro VM, Javitt S, *et al*. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry* 2017; 174 (2): 154–62.
22. Bhat HS, Goldman-Mellor SJ. Predicting adolescent suicide attempts with neural networks. *arXiv* 2017 Dec 1 [E-pub ahead of print].
23. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017; 5 (3): 457–69.
24. Simon GE, Johnson E, Lawrence JM, *et al*. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry* 2018; 175 (10): 951–60.
25. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J Child Psychol Psychiatry* 2018; 59 (12): 1261–70.
26. DelPozo-Banos M, John A, Petkov N, *et al*. Using neural networks with routine health records to identify suicide risk: feasibility study. *JMIR Ment Health* 2018; 5 (2): e10144.
27. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. *AMIA Annu Symp Proc* 2012; 2012: 1244–53.
28. Oh J, Yun K, Hwang J-H, *et al*. Classification of suicide attempts through a machine learning algorithm based on multiple systemic psychiatric scales. *Front Psychiatry* 2017; 8: 192.
29. Zhong Q-Y, Mittal LP, Nathan MD, *et al*. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. *Eur J Epidemiol* 2019; 34 (2): 153–62.
30. Fernandes AC, Dutta R, Velupillai S, *et al*. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci Rep* 2018; 8 (1): 7426.
31. Patrick AR, Miller M, Barber CW, *et al*. Identification of hospitalizations for intentional self-harm when E-codes are incompletely recorded. *Pharmacoepidem Drug Safe* 2010; 19 (12): 1263–75.
32. Voss EA, Makadia R, Matcho A, *et al*. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc* 2015; 22 (3): 553–64.
33. ETL-CDMBuilder. Github. https://github.com/OHDSI/ETL-CDMBuilder Accessed March 26, 2019.
34. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM; 2016: 785–94.
35. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975; 405 (2): 442–51.
36. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
37. Baker SP, Hu G, Wilcox HC, *et al*. Increase in suicide by hanging/suffocation in the U.S., 2000-2010. *Am J Prev Med* 2013; 44 (2): 146–9.
38. Hawton K, Bergen H, Cooper J, *et al*. Suicide following self-harm: findings from the multicentre study of self-harm in England, 2000-2012. *J Affect Disord* 2015; 175: 147–51.
39. WISQARS (Web-based Injury Statistics Query and Reporting System. 2019. https://www.cdc.gov/injury/wisqars/index.html. Accessed June 5, 2019.
40. McAllister M. Multiple meanings of self harm: a critical review. *Int J Ment Health Nurs* 2003; 12 (3): 177–85.
41. Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013; 51: S30–7.
42. Fliege H, Lee J-R, Grimm A, *et al*. Risk factors and correlates of deliberate self-harm behavior: a systematic review. *J Psychosom Res* 2009; 66 (6): 477–93.
43. Angst J, Hengartner MP, Rogers J, *et al*. Suicidality in the prospective Zurich study: prevalence, risk factors and gender. *Eur Arch Psychiatry Clin Neurosci* 2014; 264 (7): 557–65.
44. Guo L, Luo M, Wang W, *et al*. Association between nonmedical use of opioids or sedatives and suicidal behavior among Chinese adolescents: an analysis of sex differences. *Aust N Z J Psychiatry* 2019; 53 (6): 559–69.
45. Hawton K, Saunders K, Topiwala A, *et al*. Psychiatric disorders in patients presenting to hospital following self-harm: a systematic review. *J Affect Disord* 2013; 151 (3): 821–30.