



Advanced models for improved prediction of opioid-related overdose and suicide events among Veterans using administrative healthcare data

Ralph Ward^{1,2} · Erin Weeda^{1,3} · David J. Taber^{1,4} · Robert Neal Axon^{1,5} ·
Mulugeta Gebregziabher^{1,2}

Received: 12 May 2021 / Revised: 21 August 2021 / Accepted: 19 October 2021 /
Published online: 2 November 2021

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021

Abstract

Veterans suffer disproportionate health impacts from the opioid epidemic, including overdose, suicide, and death. Prediction models based on electronic medical record data can be powerful tools for identifying patients at greatest risk of such outcomes. The Veterans Health Administration implemented the Stratification Tool for Opioid Risk Mitigation (STORM) in 2018. In this study we propose changes to the original STORM model and propose alternative models that improve risk prediction performance. The best of these proposed models uses a multivariate generalized linear mixed modeling (mGLMM) approach to produce separate predictions for overdose and suicide-related events (SRE) rather than a single prediction for combined outcomes. Further improvements include incorporation of additional data sources and new predictor variables in a longitudinal setting. Compared to a modified version of the STORM model with the same outcome, predictor and interaction terms, our proposed model has a significantly better prediction performance in terms of AUC (84% vs. 77%) and sensitivity (71% vs. 66%). The mGLMM performed particularly well in identifying patients at risk for SREs, where 72% of actual events were accurately predicted among patients with the 100,000 highest risk scores compared with 49.7% for the modified STORM model. The mGLMM's strong performance in identifying true cases (sensitivity) among this highest risk group was the most important improvement given the model's primary purpose for accurately identifying patients at most risk for adverse outcomes such that they are prioritized to receive risk mitigation interventions. Some predictors in the proposed model have markedly different associations with overdose and suicide risks, which will allow clinicians to better target interventions to the most relevant risks.

Keywords Opioid epidemic · Risk prediction model · Decision support · Opioid safety

✉ Ralph Ward
ralph.ward@va.gov

Extended author information available on the last page of the article

1 Introduction

The national epidemic of adverse events related to opioid abuse continues to be an urgent problem despite efforts in recent years to reduce the rate of opioid prescriptions and to implement treatment interventions (Centers for Disease Control and Prevention 2019). Compared to the general population, Veterans suffer a disproportionate impact from the combined effects of chronic pain, opioid use disorder (OUD), and mental health disorders due to their exposure to physical and mental trauma (Seal et al. 2012; Vowles et al. 2019; Nahin et al. 2017). The Veterans Health Administration (VHA) implemented the Opioid Safety Initiative in 2013; this required steps to achieve lower opioid prescription rates and lower mean doses for remaining prescriptions (Gellad et al. 2017). The VHA and Department of Defense (DoD) promoted joint guidelines in 2017 for managing chronic pain that required careful management of opioid dose and duration (U.S. VHA and DoD, 2017). Following these initiatives, several key opioid risk indicators for prescription opioids improved. Yet overdose deaths among Veterans continued to increase from 14.47 per 100,000 person-years in 2010 to 21.08 per 100,000 person-years in 2016, even while the percentage who received a prescription opioid within 90 days of overdose death declined from 54% in 2010 to 26% in 2016 (Lin et al., 2019). This trend suggested increased use of illicit opioids. More recently, the COVID-19 pandemic is likely to further exacerbate this negative trend in opioid overdose deaths (Holland et al. 2021), highlighting the continued critical importance of VHA programs providing treatment for Veterans with OUD, or who are at substantial risk for developing OUD.

The VHA Office of Mental Health and Suicide Prevention (OMHSP) developed a risk prediction model in 2018 designed to identify patients most at risk for overdose (OD) and suicide-related events (SREs) related to opioid use. This model became the basis for a clinical decision support tool called the Stratification Tool for Opioid Risk Mitigation (STORM) (Oliva et al. 2017), and relies solely on electronic medical record (EMR) data for risk prediction and stratification. STORM has been implemented throughout the VHA as a risk mitigation tool to help prioritize patients at greatest risk for an opioid-related adverse outcome. Patients with opioid use have their risk scores updated each night from daily electronic medical record extracts; this score helps to drive a dashboard that summarizes the patient's current risk factors, risk mitigation strategies, non-pharmacological pain treatments and the status of healthcare appointments. The model was developed from a single year of EMR data (2010), and estimated a patient's risk for the combined outcome (OD or SRE) in the following year. The combined OD-SRE outcome was used due to concerns that it may be difficult to clinically distinguish the two events. Predictor variable categories included demographics, prior outcomes or treatments related to opioid risk, prescriptions (opioid or concomitant medications with known risk), substance use disorders (SUD), mental health disorder diagnoses, and medical comorbidities. The STORM authors recognized several limitations, including the potential drawbacks for using a combined outcome for OD and SRE with predictors which may have differing effects for each outcome. The authors also suggested that improved model prediction accuracy might be achieved by using an expanded set of variable types and data sources (Oliva et al. 2017). Building on these observations, we sought to develop alternative models that improve prediction performance through the following steps: (1) Improved model design: we used an advanced joint model to separately account for each outcome with a covariance structure that allowed for capturing the interdependencies within each patient over multiple years. This permitted us to simultaneously assess each predictor's association with the two outcomes, such that one

predictor could be strongly associated with one outcome but have a weaker (or even protective) association with the other outcome. This was expected to help produce improved prediction performance compared to models with a single combined outcome, or compared to running separate models for each outcome. (2) Incorporated new predictor variables associated with adverse opioid outcomes based on a literature search, including adding chronic pain as a comorbid condition, expanding the types of opioid and concomitant medications, incorporating positive urine lab screening results and positive screening results for depression or suicide risk, and expanding the types of variables used to account for a patient's location based on zip code and census tract identifiers. These new predictor variables were all expected to improve prediction accuracy due to prior evidence that suggested their association with adverse opioid outcomes. (3) Used more recent (2014–2019) longitudinal data for training and validating models which reflects recent trends in improved opioid safety policies and the increased use of illicit opioids. This was expected to improve prediction performance since our prediction models would be trained on information that more closely matched current day trends. (4) Incorporated Centers for Medicare and Medicaid Services (CMS) prescription and healthcare utilization data: evidence suggests Veterans who receive opioids from both the VHA and Medicare systems are at greater risk of prescription OD death (Moyo, et al. 2019). By enriching our dataset with risk factors and previous outcomes from CMS records we expected to have more information about Veterans who used both systems. This in turn was expected to drive prediction improvements. Our a-priori hypotheses were that substantial gains in prediction accuracy and risk stratification could be achieved through these combined measures. Such improvements would provide more accurate and actionable information to clinicians for use in mitigating opioid-related risks.

2 Methods

As described in detail below, we developed new prediction models that account for the four improvements described above under the generalized linear mixed model (GLMM) and multivariate generalized linear mixed model (mGLMM) frameworks. We compared their performance to that of a modified STORM model, for which we replicated most aspects of the original STORM model. All models were trained and validated with the same longitudinal dataset that incorporated both VHA and CMS data. The modified STORM model's predictor variables were limited to those used in the original STORM model, while the GLMM and mGLMM predictors also included a large number of new variables in addition to the original STORM variables.

2.1 Population

The study period was January 1, 2014 to December 31, 2019. Veterans were included who: (1) had one or more orally administered opioid prescriptions (cough medicines excluded) filled by VHA or Medicare part-D pharmacies from January 1, 2014 through December 31, 2018, or (2) had an OUD diagnosis, or (3) had an overdose involving any drug or prescription medication during the study timeframe. The patient's index date was the date of the earliest opioid prescription, OUD diagnosis or overdose during the study period. Each year of patient data following the index date was used to predict outcomes in the following year. Patients were followed until death or until December 31, 2019.

2.2 Data sources

National Veterans Health Administration (VHA) Corporate Data Warehouse (CDW) data were merged with corresponding CMS Medicare inpatient, outpatient and pharmacy data obtained from the VA Information Resource Center (VIREC). Joint VA/DoD Mortality Data Repository and VA Suicide Prevention Application Network (SPAN) data were used to identify suicide-related events and cause of death (Center of Excellence for Mortality Data Repository 2020).

2.3 Outcomes

The outcome variables were suicide-related events (suicide ideation, attempt, or completed suicide) and prescription medication overdose (OD). These outcomes were modeled as separate binary variables in the multivariate generalized linear mixed model (mGLMM) and as a single (combined) binary variable in the modified STORM and GLMM models to indicate that either a SRE or OD had occurred. Suicide-related events and prescription medication overdose events were identified by ICD-9-CM or ICD-10-CM code definitions established by the VA Stratification Tool for Opioid Risk Mitigation (STORM) model (Oliva et al. 2017). Note that prescription medication overdose events were not limited to those involving opioids; this is consistent with the STORM model. Every suicide related event was included regardless of whether it was suspected to be related to an overdose.

2.4 Predictor variables

Variables were defined for each one-year period following a patient's index date until death or censoring. The predictor variables from one year were used to predict outcomes in the following year. The selection of the additional variable types used in the GLMM and mGLMM models to supplement the original STORM variables was informed by a thorough literature search for predictors likely to be associated with OD or SRE (Centers for Disease Control and Prevention 2019; Bohnert et al. 2011, 2014; Lo-Ciganic et al., 2019, Cochran et al. 2017); variable selection also depended on their availability within the VHA or CMS datasets. Supplemental Table 1 provides comparison listings of the predictor variables used in the mGLMM and GLMM models versus the modified STORM model.

We considered other variable selection methods, including penalized regression methods and various machine learning algorithms. However, our earlier research in this area demonstrated these methods were not likely to provide substantially improved prediction performance over a thorough literature search (Ward et al. 2018).

2.4.1 Demographics

Demographic predictors included age, sex, race-ethnicity, marital status, and percent service-related disability. Location variables included the VA station offering primary care that was closest to the Veteran's residence, the Veteran Integrated Service Network (VISN), and other variables based on the Veteran's zip code or census tract. These were the rural–urban location based on Rural Urban Commuting Area (RUCA) code classifications (University of Washington 2019), the Area Deprivation Index (University of Wisconsin School of Medicine Public Health 2015), the Area Health Resources Files (U.S. Health Resources and Services Administration 2019), and County Health Rankings from

the University of Wisconsin Population Health Institute (2019). All location variables were time-varying, such that we used the patient's last recorded location in a given year to fix location for that year.

2.4.2 Opioid prescriptions

Opioid dose was calculated by determining morphine equivalent daily dose (MEDD) (Centers for Disease Control and Prevention 2018) and reported as mean MEDD and mean opioid days per year. Total opioid prescription fills per year for long-acting (LA) and short-acting (SA) opioids were determined. Cumulative days per year were also determined for any opioids and for long-acting (LA) opioids. Fills by opioid type were also determined for fentanyl (LA and SA), hydrocodone (LA and SA), hydromorphone (LA and SA), meperidine, methadone, morphine (LA and SA), oxycodone (LA and SA), oxymorphone (LA and SA), tapentadol (LA and SA), and tramadol (LA and SA). Annual opioid fills and mean MEDD per year from CMS part-D claims were also determined. We also accounted for whether a patient received any part-D opioids at any point during the study timeframe.

2.4.3 Other medications

Total annual prescription fills were identified for medications classes that may increase the risk for suicide or overdose events when prescribed with opioids. These classes included antidepressants, antipsychotics, benzodiazepines, duloxetine, gabapentin, muscle relaxants, stimulants, and zolpidem. Total annual cumulative days were determined for antidepressants, benzodiazepines with opioids, and gabapentin.

2.4.4 Medical comorbidities

Elixhauser comorbidities were determined based on previously validated ICD code definitions (Quan et al. 2005). Comorbidities were carried forward from the first year reported; for example, if a patient had a diabetes diagnosis at baseline, it was assumed to exist in each following year even if the diagnosis code did not recur in a following year. Since comorbid chronic pain is frequently seen in Veterans with opioid use, we included it as a predictor in our models. Veterans meeting criteria for chronic pain within any 365 day period since the index date were identified using a validated algorithm that included the following steps (Tian Zlateva and Anderson 2013): (1) a single occurrence of an ICD-9 or ICD-10 code shown to be highly likely to represent chronic pain, or (2) two or more occurrences of ICD codes shown to be likely to represent chronic pain, separated by at least 30 days, or (3) receipt of at least 90 days of opioid medication, or (4) one occurrence of an ICD code likely to represent chronic pain AND two or more numeric pain scores of 4 or higher more than 30 days apart. Patients were not considered to have chronic pain until 90 days had passed after any surgery. The ICD-9 and ICD-10 codes used in the chronic pain algorithm are provided in supplemental table 6. Finally, we reported sleep apnea, osteoporosis and a history of falls and accidents due to their previous association with increased risk of overdose (Oliva et al. 2017).

2.4.5 Mental health and substance use comorbidities

Mental health conditions included bipolar disorder, post-traumatic stress syndrome (PTSD), major depression, and other mental health disorders. Substance use disorders (SUD) included opioid, alcohol, tobacco, sedatives, stimulants, cannabis, hallucinogens, and other substances. Similar to medical comorbidities, conditions were carried forward from the first year reported, with one exception: we distinguished between an OUD diagnosis that occurred in a prior year with a newly diagnosed OUD in the current prediction year.

2.4.6 Laboratory data

Positive urine lab results were included for cocaine, antidepressants, amphetamines, alcohol, benzodiazepines, oxycodone, opiates, cannabis, morphine, methadone, fentanyl, barbiturates, buprenorphine, and phencyclidine.

2.4.7 Prior events

Prior outcomes or treatments in earlier years were used as predictors; these events include prior OD, prior SRE, prior SUD treatment, prior inpatient mental health treatment, and prior detoxification treatment.

2.4.8 Healthcare utilization

Annual total numbers of outpatient visits, inpatient stays and emergency department visits were determined for both the VHA and CMS systems.

2.4.9 Mental health screening results

Patient Health Questionnaire-9 (PHQ-9) (Kroenke, Spitzer and Williams 2001), the Brief Addiction Monitor (BAM) scores (Cacciola et al. 2013), and Columbia-Suicide Severity Rating Scale (C-SSRS) scores (Posner et al. 2011) were used as predictors. PHQ-9 is a measure of depression severity, where a higher score indicates greater severity. BAM is a measure to assess risk and protective factors related to substance use disorders; the difference between subscale scores for risk factors and protective factors was used in our analyses, such that a more positive value indicated risk factors outweighed protective factors. Positive suicide risk screening results were collected from C-SSRS scores (positive responses to questions 3, 4, 5 or 8 indicated positive risk) and from item 9 of the PHQ-9 (any score > 0 indicated positive risk).

2.5 Data analysis

2.5.1 Summary statistics

Summary statistics (mean, proportion, etc.) for the patient demographic and clinical characteristics were computed by SRE and OD outcomes. We did not report p values because nearly every comparison was highly significant due to the large population size.

2.5.2 Statistical models

Modified STORM model First, we ran a modified version of the original STORM model (Oliva et al. 2017) with our longitudinal dataset using the same combined binary outcome (OD or SRE) and the same predictor and interaction terms. The original STORM model was run on a single year of predictor data (2010), and was based on a GLMM approach with a logit link and random effects for VHA station and VISN to account for clustering by facility. Our modified STORM model differed in several respects from the original: our dataset was longitudinal with up to 5 years of data per patient, included both VHA and CMS data sources, and included a random intercept for each patient. Finally, we expanded the STORM patient inclusion criteria to also include those with a prior OD or OUD diagnosis; the original model included those with any opioid prescription. Although the modified model did not perfectly replicate the original STORM model design, it served as a baseline model against which we could compare potential prediction performance improvements.

GLMM with new predictors The next model used the same GLMM approach and combined outcome as the modified STORM model, but included the wide range of new predictors described earlier, including healthcare utilization variables, chronic pain diagnosis, mental health screening results, positive urine lab results, additional opioid and other medication classes, and location variables. Like the STORM model, this was a generalized linear mixed model with a logit link, a single binary outcome (OD or SRE), but also included a random intercept for each patient.

Multivariate generalized linear mixed model (mGLMM) This proposed approach used a joint modeling design to model both outcomes (OD and SRE) jointly through two GLMMs that were linked through a shared random intercept. In addition to the joint effect of an exposure on both outcomes, this also provided marginal effects of an exposure on each of the outcomes while accounting for their interdependence. This mGLMM approach also provided a way to model unobserved heterogeneity in variance caused by subject and cluster effects, by correlated repeated measures and by multiple related outcomes. This approach may be particularly useful in our case, where the OD and SRE outcomes are highly correlated. The joint model has been shown to reduce bias and standard errors when compared to models which use a single combined (binary) outcome, or compared to using separate models for each outcome (Gebregziabher et al. 2012; Jaffa et al. 2016, Gebregziabher et al. 2018). Let $y_{ij}^k = (y_{i1}^k, \dots, y_{im}^k)$ be a vector of outcome responses for subject i ($i = 1, \dots, n$) at times j ($j = 1, \dots, m$) for outcome k ($k = 1, \dots, K$). The GLMM for each outcome Y_i^k is:

$$E(Y_i^k | X_i, Z_i) = g^{-1}(X_i \beta^k + Z_i b_i^k)$$

where $b_i^k \sim N(0, G^k)$ and X_i and Z_i are fixed and random effect covariates, respectively, and β^k and b_i^k are fixed and random effect parameters, respectively. G^k is a qxq covariance

matrix, where q is the dimension of the random effects vector. Here we use a logit link with the binary outcomes. For two outcomes (OD and SRE), we induced correlation between two responses in the joint model through shared random intercepts:

$$\log \text{it}\left(Y_{ij}^1\right) = \beta_0^1 + b_0 + \beta_1^1 t_j + \beta_2^1 x_i,$$

$$\log \text{it}\left(Y_{ij}^2\right) = \beta_0^2 + b_0 + \beta_1^2 t_j + \beta_2^2 x_i,$$

The shared random intercept model builds an assumption that the OD and SRE outcomes are related through some latent process. By modeling the two outcomes jointly, we are able to simultaneously assess each predictor's association with the two outcomes, such that one predictor could be strongly associated with one outcome but have a weaker (or even protective) association with the other outcome. This is expected to help produce improved prediction performance compared to models with a single composite outcome, or compared to running separate models for each outcome. For example, Liu et al. (2008) showed that running independent models that ignored possible correlation between related outcomes could result in biased estimates.

2.5.3 Variable selection

For the GLMM and mGLMM models, variable selection steps included checking for collinearity between closely-related covariates; for example, annual opioid days, mean MEDD and cumulative opioid days were all strongly correlated, and mean MEDD was used in the models to represent opioid dose. Potential interactions between key variables were tested for significance in exploratory analyses, but were not retained in the final model because they did not improve overall prediction performance. In developing prediction models, we were less concerned with identifying and eliminating unimportant predictors (model parsimony). As Steyerberg (2010) discussed, parsimony in prediction models is not an important concern when we have large patient cohorts; instead, we are primarily concerned with prediction accuracy in patient data that was not used to train the model. Hastie et al. (2009) also provided similar conclusions. Parsimony can become important in explanatory models when we want the model to be interpretable; parsimony is also needed when the number of patients is small relative to the number of candidate predictors: neither condition applies in our case.

2.5.4 Prediction performance

We used tenfold cross validation to validate prediction model performance. Model discrimination performance was assessed using validation data following model development using training data. Performance statistics included: (1) comparison of area under the receiver operator curve (AUC) statistics using the DeLong test (DeLong et al. 1988); (2) sensitivity, specificity, positive predictive value, negative predictive value, and number needed to evaluate (Romero-Brufau et al. 2015) were reported at the optimized prediction threshold as determined by the maximum Youden Index (Youden et al. 1950). The Youden Index is based on a balance of sensitivity and specificity and the maximum value occurs where the receiver operator curve reaches a maximum height above the diagonal line that

represents 50% prediction performance (random chance). While AUC provided a measure of overall prediction performance, improvements in sensitivity were also emphasized since this is a measure for how well a model can identify true OD and SRE cases while minimizing the number of missed cases. Measures that assessed how well the model avoided false positive predictions (specificity and positive predictive value) were considered less important in this setting because correctly identifying cases enables the delivery of potentially life-saving interventions, while false positive results added little risk. Model calibration performance was assessed by stratifying risk predictions into low risk (below the maximum Youden Index), medium risk (between the maximum Youden Index and the highest 10th percentile), and high risk (top 10th percentile). We then compared the models' actual rates for each outcome across the three strata. We also assessed performance when the risk stratification cutpoint was set such that the patients with the highest 100,000 predicted risk scores were screened in order to compare how many OD or SRE events would be correctly identified at this cutpoint (sensitivity) versus how many would be false negative or false positive predictions. Finally, we assessed prediction performance for subgroups by race-ethnicity, sex, and age groups to evaluate whether predictions for these groups were less effective, and to what degree. All analyses were performed using SAS, version 9.4 (SAS Institute, Inc.).

2.5.5 Net benefit analysis

We conducted net benefit analysis (also called decision curve analysis) to compare models in terms of their screening benefit. The net benefit (NB) for treatment is defined as:

$$NB(p_t) = \frac{TP}{n} - \frac{FP}{n} \left(\frac{p_t}{1 - p_t} \right)$$

where n is the total number of patients, TP and FP are the number of true or false positives, and p_t is a risk threshold (Vickers 2006). NB could be viewed as the difference between the benefit for screening the correct group (true-positives) and the weighted cost for screening the wrong group. Vickers et al. (2006) wrote that this approach is intended to determine if a new model provides any clinical benefit, and should not be considered a strict measure of predictive performance.

3 Results

3.1 Patient characteristics

Table 1 provides a summary of patient demographic variables and a history of prior events stratified by prescription OD or SRE outcomes. There were 1,744,667 Veterans in the cohort, of which 165,680 (9.5%) had at least one prescription medication overdose during the study period, and 97,688 (5.6%) had at least one SRE (Table 1). Most variables were strongly associated with both outcomes but in some cases with opposing effects. For example, in the oldest age group (> 65) a larger percentage had at least one OD compared to those under age 30 (12.3% vs. 4.5%), but a smaller percentage had at least one SRE (2.8% vs. 10.1%). Similarly, men were more likely to experience at least one OD but less likely to experience a SRE compared to women. Nearly 85% of the full cohort met the chronic pain definition, and these patients were more likely to experience an OD or SRE: 7.8% and 17.5% with 'likely' or 'highly

Table 1 Population characteristics*

		≥ 1 overdose N(%)	≥ 1 suicide-related event N(%)	Overall N(%)
Group size		165,680 (9.5)	97,688 (5.6)	1,744,667
Race-ethnicity	non-Hispanic white	122,429 (10.2)	64,073 (5.3)	1,203,231 (69)
	non-Hispanic Black	31,688 (8.4)	23,681 (6.3)	375,726 (21.5)
	Hispanic	6738 (6.9)	6207 (6.4)	97,052 (5.6)
	Other	4525 (6.6)	3727 (5.4)	68,658 (3.9)
Sex	Female	12,866 (8)	11,772 (7.3)	160,905 (9.2)
	Male	152,514 (9.6)	85,916 (5.4)	1,583,762 (90.8)
Age	Under 30	3468 (4.5)	7806 (10.1)	76,982 (4.4)
	30–50	18,185 (5.3)	27,567 (8.1)	341,299 (19.6)
	51–65	61,206 (9.4)	43,441 (6.6)	653,531 (37.5)
	Over 65	82,521 (12.3)	18,874 (2.8)	672,855 (38.6)
Service-related disability	< 50%	87,999 (9.6)	42,347 (4.6)	913,268 (52.3)
	≥ 50%	77,381 (9.3)	55,341 (6.7)	831,399 (47.7)
Marital status	Unmarried	87,299 (10)	64,118 (7.3)	877,311 (50.3)
	Married	78,081 (9)	33,570 (3.9)	867,356 (49.7)
Urban rural location	Rural or highly rural	60,120 (9.7)	30,717 (5)	619,142 (35.5)
	Urban	105,260 (9.4)	66,971 (6)	1,125,525 (64.5)
Prior events				
Prior suicide-related event		22,977 (29.7)	39,087 (50.5)	77,401 (4.4)
Prior overdose		52,033 (40.5)	22,093 (17.2)	128,479 (7.4)
Prior MH treatment		12,518 (32.7)	24,816 (64.9)	38,264 (2.2)
Prior SUD treatment		14,391 (23.6%)	24,037 (39.4%)	60,952 (3.5%)
Opioids from CMS partD		31,090 (14.8)	14,852 (7.1)	209,716 (12)
Chronic pain	Not diagnosed	11,768 (4.3)	5096 (1.9)	272,190 (15.6)
	Likely	84,331 (7.8)	46,802 (4.3)	1,077,561 (61.8)
	Highly likely	69,281 (17.5)	45,790 (11.6)	394,916 (22.6)

*Percentages in OD and SRE columns represent proportion of that subgroup having the outcome; percentages in total column represent that subgroup's proportion of the full population

likely' chronic pain experienced at least one OD, compared with 4.3% among those without chronic pain. For SREs, these values were 4.3%, 11.6%, and 1.9%, respectively. Patients with a prior OD or SRE were significantly more likely to have future events. For example, 29.7% of those with a prior SRE had at least one overdose event, and 50.5% had at least one additional SRE. Among patients with a previous inpatient mental health treatment, 32.7% had at least one post-treatment OD event and 64.9% experienced at least one SRE. Veterans with opioids from Medicare sources were somewhat more likely to experience an OD (14.8%) or a SRE (7.1%) compared to the overall population. Supplementary tables 2–5 provide a full listing of the distributions of predictor variables used in the GLMM and mGLMM models stratified by prescription OD or SRE outcomes; similar to Table 1, there are often distinct differences in these distributions when outcome groups are compared.

Table 2 Prediction performance measures

	mGLMM	GLMM	Modified STORM
Area under the ROC curve (AUC) and 95% CI*	0.837 (0.836, 0.838)	0.801 (0.80, 0.802)	0.774 (0.772, 0.776)
Sensitivity	0.71	0.69	0.66
Specificity	0.81	0.76	0.76
Precision (PPV)	0.09	0.135	0.13
Negative predictive value (NPV)	0.99	0.98	0.98
Number needed to evaluate	10.6	7.4	7.8
Maximum Youden score (optimized threshold probability)	0.52	.45	.42

Sensitivity, specificity, precision, negative predictive value, and number needed to evaluate were determined at the maximum Youden Index, an optimized threshold that occurs where the receiver operator curve reaches a maximum height above the diagonal line that represents 50% prediction performance (random chance). Sensitivity is the ratio of correctly screened (true positive) cases to the total actual cases; specificity is the ratio of correctly screened non-cases (true negatives) to the total non-cases; precision is the number of true positive cases to the total number screened positive; negative predictive value is the total number of true negative cases to the total number screened negative. Number needed to evaluate is the inverse of precision, and is the number of screened cases for each true positive. While AUC provides a measure of overall prediction performance, improvements in sensitivity are emphasized over other measures since it measures how well a model can identify true OD and SRE cases while minimizing the number of missed cases. Measures that assessed how well the model avoided false positive predictions (specificity and positive predictive value) were considered less important in this setting because correctly identifying cases enables the delivery of potentially life-saving interventions, while false positive results added little risk

*Delong test for difference in AUC: mGLMM vs. GLMM: $p < 0.0001$; GLMM vs. STORM: $p < 0.0001$; mGLMM vs. STORM: $p < 0.0001$

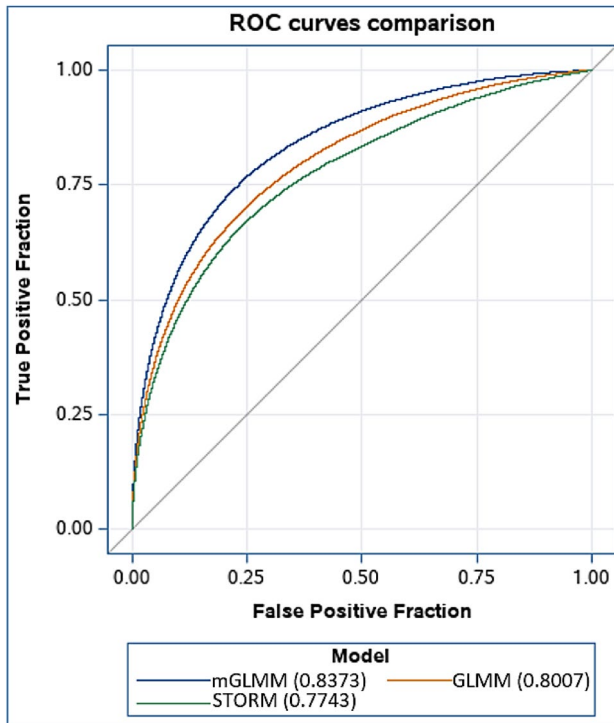


Fig. 1 Area under the ROC curve comparison for mGLMM, GLMM and STORM replication models, with AUC values in parenthesis

3.2 Prediction performance

Table 2 summarizes prediction performance statistics from the modified STORM, GLMM and mGLMM models. The 95% confidence intervals for area under the ROC curve (AUC) do not overlap, and indicate the mGLMM model has significantly better prediction performance (AUC 0.837, 95% CI: 0.836, 0.838) compared to the GLMM (AUC 0.801, 95% CI: 0.800, 0.802) and modified STORM (AUC 0.774, 95% CI: 0.772, 0.776) models. Results from Delong tests for differences in AUC between each model pairing were highly significant ($p < 0.0001$ for all). A comparison of ROC curves (Fig. 1) provides a similar conclusion. The mGLMM model's sensitivity and specificity are superior as determined at the optimized Youden score cutpoint, but precision, negative predictive value results are slightly lower but roughly similar to the comparison models.

3.2.1 Prediction performance in subgroups

Supplemental Fig. 1 a provides a comparison of AUC performance by race-ethnicity, sex, and age groups for the modified STORM model (results shown in gray) and mGLMM (results in color). STORM AUC values range from 0.738 to 0.822 (difference of 8.4%), and mGLMM values range from 0.818 to 0.859 (difference of 4.1%). Performance differences

by race-ethnicity were reasonably small for all groups: performance for non-Hispanic Blacks and other race-ethnicity groups was slightly better than for the whole group; for Hispanics performance was slightly improved in STORM but slightly lower in mGLMM. Similarly, for females, performance was slightly better in STORM but slightly lower in mGLMM. The largest subgroup differences occur by age group: performance was better for those younger than 65, but was 1.9% worse in mGLMM and 4.9% worse in STORM.

3.3 Net benefit analysis

Supplemental Fig. 1b provides net benefit results. This approach is another way to balance the gains from correctly identifying cases (TP) against the costs for incorrectly predicting patients are high risk (FP). When measured at the actual incidence of SRE or OD, the net benefit (NB) from screening is the vertical distance along the dashed line between the green and red curves. The NB values are strongly positive for both models, but are largest for mGLMM.

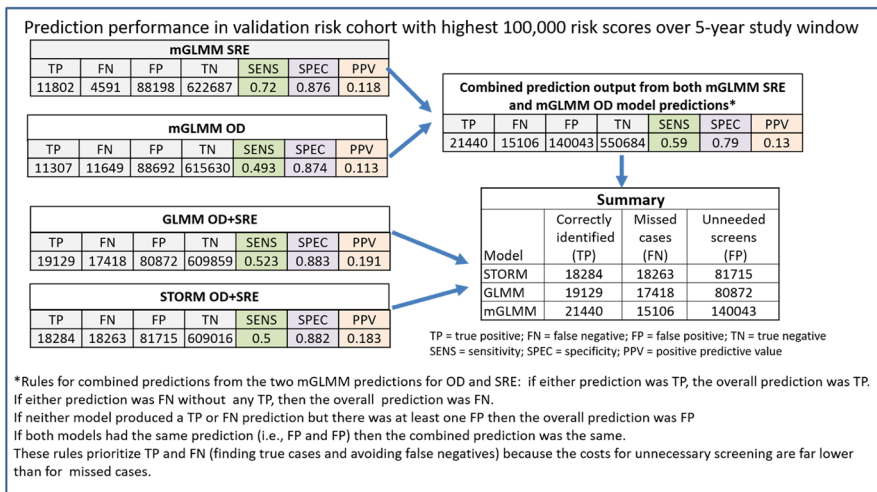


Fig. 2 We assessed prediction performance when the risk stratification cutpoint was set such that the patients with the highest 100,000 predicted risk scores were screened in order to compare how many overdose or suicide events would be correctly identified at this cutpoint (sensitivity) versus how many would be false negative or false positive prediction. For each model’s predictions in a validation cohort, we show numbers of true and false positives (TP, FP), true and false negatives (TN, FN), with corresponding sensitivity (SENS), specificity (SPEC), and positive predictive values (PPV). In the mGLMM model predicting SRE, SENS=72%, meaning nearly 3 of 4 actual SREs were correctly identified in the validation data, compared with 50% in the modified STORM model. For OD, the mGLMM and STORM models produced very similar results (49.3% and 50%, respectively). Since the mGLMM model produced two predictions per patient (OD and SRE), we combined them using rules that prioritized true positives and false negatives (see footnote). The combined mGLMM OD-SRE predictions included 21,440 TP results, or 17.3% more than from modified STORM (18,240 TP) and 12.1% more than GLMM (19,129 TP). False positive (FP) results were 140,043, 80,872, and 81,715 for mGLMM, GLMM, and STORM, respectively, indicating more patients would be unnecessarily screened using the mGLMM approach in order to produce the gains in TP cases

3.4 Risk stratification performance

Figure 2 provides a comparison of prediction performance when the risk stratification cut-point was set such that the patients with the highest 100,000 predicted risk scores were all screened; this allowed us to compare how many overdose or suicide events would be correctly identified at this cutpoint (true positives) versus how many would be false negative or false positive predictions when we focused on this highest risk group. In the mGLMM model predicting SRE, 72% of the actual SREs were correctly predicted in the validation data, compared with 50% in the modified STORM model. For OD, the mGLMM and STORM models produced very similar results (49.3% and 50%, respectively). Since the mGLMM model produced two predictions per patient (OD and SRE), we combined them using rules that prioritized true positives and false negatives; for example, patients with a true positive (TP) prediction for either OD or SRE were counted as TP overall, while patients with any false negative (FN) without any TP were counted as FN overall. Other rules assigned results when there was not a TP or FN result (see Fig. 2 footnotes). As shown in the summary table of Fig. 2, the combined mGLMM OD-SRE predictions included 21,440 TP results, or 17.3% more than from modified STORM (18,240 TP) and 12.1% more than GLMM (19,129 TP). False positive (FP) results were 140,043, 80,872, and 81,715 for mGLMM, GLMM, and STORM, respectively, indicating substantially more patients would be unnecessarily screened using the mGLMM approach in order to produce the important gains in TP cases. Supplemental Fig. 1.c shows the distribution of patients among the top 100k risk groups from the STORM and mGLMM models. There were a total of 97,776 patients from the 3 high risk groups (STORM, mGLMM-OD and mGLMM-SRE), with 21,013 (21.5%) common to all groups. Note that one patient could be screened more than once due to the longitudinal design. Most of the STORM patients (93%) were accounted for by the other models (4208 were unique to STORM), while 41% of the full group were only accounted for by the mGLMM model. This indicated that the mGLMM predicted a substantial number of other patients were at high risk while also accounting for most STORM high risk patients. Supplemental Fig. 1.d provides a comparison of calibration performance, which was satisfactory for all models.

3.5 Case examples comparing model predictions

Figure 3 provides two case examples that demonstrate differences in STORM and mGLMM model responses as each patient's risk profile increases from one year to the next. For each patient, results are also shown for 4 different age ranges since changes in OD and SRE risks are different as a patient ages. The left side of each arrow shows the first year's predicted risk percentile, and the right side shows the model's estimate from the current year. At each point, a patient's risk score was calculated using each model's parameter estimates and the associated predictor variable values; the inverse logit function was applied to this linear predictor to produce the risk score estimate. Risk scores were expressed as percentiles to indicate relative risk among all patients in the opioid risk population.

Patient #1. Last year, patient #1 had a baseline history of PTSD with diagnoses that were consistent with 'likely chronic pain', and had used short-acting hydrocodone (MME 25 mg/day) for more than 90 days. In the current year, numerous additional risks were added: chronic pain is now 'highly likely', severe depression concurrent with an elevated PHQ-9 score and antidepressant prescriptions were added, and short-acting hydromorphone was prescribed. While the STORM (shown in blue) and MGLMM-OD

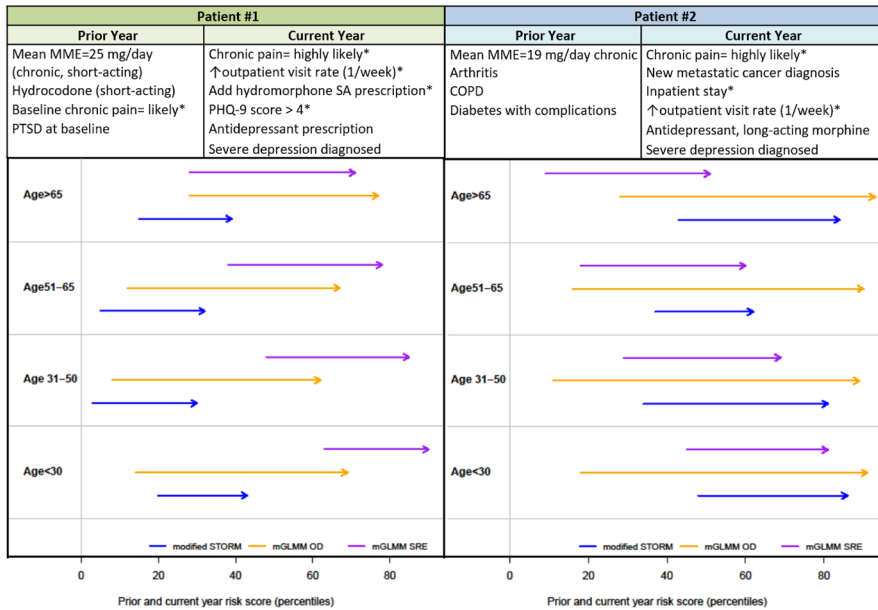


Fig. 3 Case examples. These examples demonstrate differences in model responses as each patient’s risk profile increases from one year to the next. For each patient, results are also shown for 4 different age ranges. The left side of each arrow shows the first year’s predicted risk percentile, and the right side shows the model’s estimate from the current year after new risk factors were added. At each point, a patient’s risk score was calculated using each model’s parameter estimates and the associated predictor variable values; the inverse logit function was applied to this linear predictor to produce the risk score estimate. Risk scores were expressed as percentiles to indicate relative risk among all patients in the opioid risk population. STORM model results are shown in blue; mGLMM OD results in orange, and mGLMM SRE results in purple

risk estimates (orange) were generally similar in year one (left side of arrows), the mGLMM-SRE risk (purple) was substantially higher if the patient were younger than 65. In the current year, the mGLMM model predicts risks for both SRE and OD have increased substantially, but suicide risks are again highest except for those 65 and older. These differences in OD and SRE risks given a patient’s age could inform clinical intervention decisions. The mGLMM risk estimates are consistently much higher than STORM estimates because chronic pain, PHQ-9 scores, hydromorphone, and outpatient visits are not accounted for by the STORM model, but have a strong combined impact in the mGLMM model.

Patient #2. This patient had a baseline history of arthritis, chronic obstructive pulmonary disease, and diabetes with complex complications, and had taken a short-acting opioid (MME 19 mg/day) for at least 90 days. This year, the patient has a new diagnosis of metastatic cancer, is highly likely to have chronic pain, had an inpatient stay and increased outpatient utilization, has a long-acting morphine prescription, and a severe depression diagnosis with an antidepressant prescription. As expected, both models predict substantial risk increases, but the mGLMM model provides additional insights which could inform clinical decisions. OD risk (orange) is now higher than suicide risk (purple) in all age groups; this difference is relatively small in the youngest group (91% vs. 81%) but is much larger in the oldest group (93% vs. 51%).

3.6 mGLMM model results

Supplementary Figs. 2–7 show the estimated effects for mGLMM model predictors for each outcome as odds ratios and 95% confidence intervals; overdose and suicide results are shown in blue and gold, respectively. The relative effect sizes give an indication for the magnitude of change expected in a patient's risk score should that predictor appear in the medical record. The predictors most strongly associated with suicide risk were new engagement in substance use disorder treatment, a prior suicide event, a new OUD diagnosis, a diagnosis for chronic pain, new inpatient MH treatment, mental health conditions (depression, bipolar disorder and other MH conditions), and positive screening for suicide risk via PHQ-9 item 9 or C-SSRS. For overdose, the strongest predictors were newly diagnosed OUD, prior OD event, a new engagement in substance use disorder treatment, 'highly likely' chronic pain, metastatic cancer or lymphoma, and amphetamine use disorder. Conversely, a prior engagement in substance use disorder treatment was strongly protective for both outcomes; treatment with MOUD or a positive urine lab result for methadone were similarly protective.

There were distinct risk differences by outcome for some predictors; for example, inpatient mental health treatment in the year prior to the outcome was significantly associated with higher suicide risk in the following year, but posed a somewhat reduced overdose risk (Suppl. Figure 2). There were also distinct differences in outcome risks by age group and race-ethnicity (Suppl. Figure 7). Medical comorbidities generally were most strongly associated with overdose risk while suicide risk was often lower (Suppl. Figure 3). Conversely, suicide risk was often stronger for mental health conditions and substance use disorders (Suppl. Figure 4). Among concomitant medication classes, antidepressants, antipsychotics, benzodiazepines, and duloxetine were all strongly associated with both outcomes, but were strongest for SRE (Suppl. Figure 6). Among opioid classes, short acting hydrocodone was associated with suicide risk but the overdose risk was smaller. Long- and short acting morphine were significant overdose risks, where the risk was highest in the long-acting version. Zolpidem and long-acting fentanyl were also strongly associated with OD. Short-acting tramadol was associated with suicide risk but the overdose risk was substantially lower.

In some cases, risks were substantially different when a predictor was newly observed compared to an earlier occurrence. For example, an OUD diagnosis in a prior year was protective for both outcomes, possibly associated with treatment, while a new (current-year) OUD diagnosis was among the strongest predictors for either outcome (Suppl. Figure 2). As mentioned earlier, a similar trend was seen for prior versus new SUD treatments.

4 Discussion

We sought to develop alternative risk prediction models that improve on the original STORM model's performance among Veterans with current opioid use, a history of OUD, or a prior prescription medication OD. Based on comparisons using tenfold cross validation, our proposed mGLMM model exhibited significantly better AUC results based on highly significant DeLong tests, with corresponding gains in sensitivity when compared against a modified STORM model and a GLMM that incorporated a multitude of new predictors. There was a 3.5% increase in AUC between STORM and GLMM, and a 4.5% increase between GLMM and mGLMM, or 8.1% overall. This suggests the performance

improvements in mGLMM came from both adding new predictors (STORM to GLMM) and adding an advanced joint model (GLMM to mGLMM). The mGLMM performed particularly well in identifying patients at risk for SREs, where 72% of actual events were accurately predicted among patients with the 100,000 highest risk scores compared with 49.7% for the modified STORM model. The combined mGLMM OD-SRE predictions for this highest risk cohort correctly identified 17.3% more true cases than the modified STORM model. The mGLMM's strong performance in identifying true cases (sensitivity) among this highest risk group was the most important improvement given the model's primary purpose for accurately identifying patients at most risk for adverse outcomes such that they are prioritized to receive risk mitigation interventions. On the other hand, the mGLMM model had a lower positive predictive performance, meaning more patients were falsely identified as high risk such that some intervention resources would be used on patients who may not need them. However, this drawback is less important since false positive screenings add little, if any patient risk.

These prediction gains were achieved by modeling the OD and SRE outcomes separately in the joint model such that the association between each predictor and both outcomes was assessed simultaneously. The STORM authors had used a combined OD-SRE outcome due to concerns that it may be difficult to distinguish the two events clinically; our results suggest that the two outcomes were in fact distinguishable in the administrative data such that a patient's predicted risk for one outcome was often substantially different than for the other.

Although the mGLMM is more complex and less parsimonious than the STORM model, the performance gains justify the added costs of obtaining more predictor variables and requiring greater computational power, both of which are readily available within the VHA systems. Since the model's primary purpose is to predict patient risks rather than to explain associations, parsimony and interpretability are less important. The largest computational cost is incurred only once, when parameter estimates are determined, such that frequently updated predictions for new patients can be determined very rapidly from those estimates.

Our work replicated the original STORM model conclusions concerning risk factors for OD and SRE, but also provided new important insights. By examining OD and SREs separately, we demonstrated that some risk factors could have markedly different associations with each outcome and thus a patient could be at substantially greater risk for one outcome over the other. Our model may enable clinicians to target interventions based on the most pertinent outcome, which is notable because risk mitigation options for SREs and OD can differ. For instance, examples of evidence-based interventions for SREs may include patient outreach, warm hand-offs or enhanced recruitment efforts into mental health treatment programs, and means reduction efforts; while for OD, examples include naloxone kits, SUD treatment programs, medication for opioid use disorder (MOUD), opioid dose reduction, and alternative approaches to pain control (Oliva et al. 2017).

We demonstrated that prediction performance lagged in certain subgroups, most notably for those older than age 65 (4.9% worse in STORM and 1.9% worse in mGLMM); this group comprised nearly 39% of the risk population. The STORM authors noted that multiple tailored models for subgroups were not used (Oliva et al. 2017). Our results suggest that prediction models tailored to subgroups would improve equity in risk screening. Further research is needed to better explain these differences.

Our work has several limitations. First, our modified STORM model had the same predictor variables and interaction terms as the original model but we used a newer 5-year longitudinal Veteran cohort, and there is no guarantee that the predicted risk scores are

identical to actual STORM scores based on the 2010 dataset. However, comparisons of effect sizes, risk stratification and calibration results from the original model (Oliva et al. 2017) provide evidence that our modified version is a reasonably close approximation. Next, there is a multi-year lag in cause of death data that could contribute to missed overdose or suicide deaths. There is also a time-lag for the availability of CMS data, and we used a single variable to capture any prior CMS opioid provision throughout the study period since these data would not be available on the daily basis needed for STORM implementation. Our study cohort is generally comprised of older men (76.1% were older than 50 and 90.8% were men), and our results may not be readily generalizable to other populations. Finally, we did not use electronic medical record text notes in these analyses, and substantial prediction gains are likely if new information involving OD and SRE risks were captured from this source.

There are numerous areas for future work. As discussed above, further research is needed to improve prediction performance among subgroups, particularly those 65 and older. Related to this, our work predicted Veterans were at increased risk for overdose if they had opioids from a Medicare source at any point during the study period; this result should be studied in further depth in a sub-group analysis limited to Veterans who are 65 or older. Similarly, given the presence of non-VA opioid prescriptions from Medicare was a strong predictor, future studies should explore incorporating other non-VA opioid data sources into the STORM model (e.g., prescription data from commercial insurers or data from state prescription drug monitoring programs). Next, work by Lo-Ciganic et al. (2019) suggests that machine learning methods may provide some additional prediction gains, and we are interested in developing an ensemble of top-performing models, such that each member ‘votes’ on a patient’s risk prediction. It may also be useful to examine patients’ risk trajectories over time to determine if key factors can be identified that predict a rapid escalation in risk.

In conclusion, our findings show that the proposed mGLMM prediction model is a better alternative to the current STORM model for identifying patients at high risk for adverse health outcomes related to opioid use due to refinements added with the mGLMM model design and the inclusion of new predictors. These refinements, paired with targeted timely clinical interventions, could further reduce mortality and morbidity in a Veteran population that is especially vulnerable to the expanding opioid epidemic.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10742-021-00263-7>.

Acknowledgements The authors gratefully acknowledge generous assistance provided during study design or manuscript preparation from Drs. Esther Meerwijk, Thomas Bowe, Susana Martins and Jodie Trafton (VA Program Evaluation and Resource Center) and from Dr. Elizabeth Santa Ana (Health Equity and Rural Outreach Center of Innovation).

Authors’ contributions RW, MG and RNA conceptualized and designed the study, interpreted analyses, and wrote, reviewed, and revised the manuscript. EW and DJT interpreted analyses, and reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Funding This research was supported by the Veterans Health Administration VISN 7 Research Operations Council in a 2019 Research Development Award.

Data availability The data used in these analyses are protected health information, and were provided by the VHA Corporate Data Warehouse and the VA Information Resource Center (VIREC). These data are not publicly available.

Code availability Code is available from the authors upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This research was approved by the Medical University of South Carolina IRB.

References

- Bohnert, A.S.B., Ilgen, M.A., Galea, S., et al.: Accidental poisoning mortality among patients in the Department of Veterans Affairs Health System. *Med Care* **49**(4), 393–396 (2011)
- Bohnert, A.S.B., Ilgen, M.A., Trafton, J.A., Kerns, R.D., Eisenberg, A., Ganoczy, D., Blow, F.C.: Trends and regional variation in opioid overdose mortality among Veterans Health Administration patients, fiscal year 2001 to 2009. *Clin J Pain* **30**(7), 605–612 (2014)
- Cacciola, J.S., Alterman, A.I., DePhillippis, D., et al.: Development and initial evaluation of the Brief Addiction Monitor (BAM). *J Subst Abuse Treat.* **44**(3), 256–263 (2013). <https://doi.org/10.1016/j.jsat.2012.07.013>
- Center of Excellence for Mortality Data Repository. Joint Department of Veterans Affairs (VA) and Department of Defense (DoD) Suicide Data Repository – National Death Index (NDI). <http://www.dspo.mil/Portals/113/Documents/SDR%20Fact%20Sheet.pdf>; extract date 3/20/2020.
- Centers for disease control Factsheet on drug poisoning. Downloaded from https://www.cdc.gov/nchs/factsheets/factsheet_drug_poisoning.pdf, accessed 13 March 2019.
- Centers for Disease Control and Prevention. Analyzing Prescription Data and Morphine Milligram Equivalents (MME). Available at: <https://www.cdc.gov/drugoverdose/resources/data.html>; accessed June 10, 2018.
- Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report: Drug and opioid-involved overdose deaths – United States, 2013–2017. *67*(51–52), 1419–1427.
- Cochran, G., Gordon, A.J., Lo-Ciganic, W.H., Gellad, W.F., Frazier, W., Lobo, C., Donohue, J.M.: An examination of claims-based predictors of overdose from a large Medicaid program. *Med Care* **55**(3), 291–298 (2017)
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3), 837–845 (1988). <https://doi.org/10.2307/2531595>
- Gebregziabher, M., Zhao, Y., Dismuke, C., Axon, N., Hunt, K., Egede, L.E.: Joint Modeling of longitudinal multiple source cost data using multivariate generalized linear mixed models. *Health Serv. Outcomes Res. Method.* (2012). <https://doi.org/10.1007/s10742-012-0103-0>
- Gebregziabher, M., Eckert, M., Matthews, L., et al.: Joint modeling of multivariate hearing thresholds measured longitudinally at multiple frequencies. *Commun Stat Theory Methods* **47**(22), 5418–5434 (2018)
- Gellad, W.F., Good, C.B., Shulkin, D.J.: Addressing the opioid epidemic in the United States: lessons from the Department of Veterans Affairs. *JAMA Internal Med* **177**(5), 611–612 (2017)
- Hastie, T., Tibshirani, R., Friedman, J.: Elements of statistical learning: data mining, inference, and prediction. Springer, New York (2009)
- Holland, K.M., Jones, C., Vivolo-Kantor, A.M., et al.: Trends in US emergency department visits for mental health, overdose, and violence outcomes before and during the COVID-19 pandemic. *JAMA Psychiat.* (2021). <https://doi.org/10.1001/jamapsychiatry.2020.4402>
- Jaffa MA, Gebregziabher M, Jaffa AA, Luttrell D, Luttrell LM, and the DCCT/EDIC Study Group. Multivariate Generalized Linear Mixed Models With Random Intercepts to Analyze Cardiovascular Risk Markers in Type-1 Diabetic Patients. *J Appl Stat.*, 43(8):1447–1464 (2016)
- Kroenke, K., Spitzer, R.L., Williams, J.B.: (2001) The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* **16**(9), 606–613 (2001). <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Lin, L., Pelzman, T., McCarthy, J.F., Oliva, E.M., Trafton, J.A., Bohnert, A.S.B.: Changing trends in opioid overdose deaths and prescription opioid receipt among veterans. *Am J Prev Med* **57**(1), 106–110 (2019)

- Liu, L., Conaway, M.R., Knaus, W.A., Bergin, J.D.: A random effects four-part model, with application to correlated medical costs. *Comput. Stat. Data Anal.* **52**(9), 4458–4473 (2008)
- Lo-Ciganic, W.H., Huang, J.L., Zhang, H.H., Weiss, J.C., Wu, Y., Kwoh, C.K., Donohue, J.M., Cochran, G., Gordon, A.J., Malone, D.C., Kuza, C.C.: Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA Netw. Open* (2019). <https://doi.org/10.1001/jamanetworkopen.2019.0968>
- Moyo, P., Zhao, X., Thorpe, C.T., Thorpe, J.M., Sileanu, F.E., Cashy, J.P., Hale, J.A., et al.: Dual receipt of prescription opioids from the department of veterans affairs and medicare part d and prescription opioid overdose death among veterans: a nested case-control study. *Ann. Intern. Med.* **170**(7), 433–442 (2019)
- Nahin, R.L.: Severe pain in veterans: the effect of age and sex, and comparisons with the general population. *J. Pain.* **18**(3), 247–254 (2017)
- Oliva, E.M., Bowe, T.K., Martins, S., et al.: Development and applications of the Veterans Health Administration's Stratification Tool for Opioid Risk Mitigation (STORM) to improve opioid safety and prevent overdose and suicide. *Psych Services* **14**(1), 34–49 (2017)
- Posner, K., Brown, G.K., Stanley, B., et al.: The columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am. J. Psychiatry* **168**(12), 1266–1277 (2011)
- Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Ghali, W.A.: Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **43**(11), 1130–1139 (2005). <https://doi.org/10.1097/01.mlr.0000182534.19832.83>
- Romero-Brufau, S., Huddleston, J.M., Escobar, G.J., Liebow, M.: Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit. Care* **19**(1), 285 (2015). <https://doi.org/10.1186/s13054-015-0999-1>
- Seal, K.H., Shi, Y., Cohen, G., Cohen, B.E., Maguen, S., Krebs, E.E., Nyelan, T.C.: Association of mental health disorders with prescription opioids and high-risk opioid use in US Veterans of Iraq and Afghanistan. *JAMA, J. Am. Med. Assoc.* **307**(9), 940–947 (2012)
- Steyerberg, E. W. (2010). *Clinical Prediction Models A Practical Approach to Development Validation and Updating* 1st ed 2009. Springer, New York
- Tian, T.Y., Zlateva, I., Anderson, D.R.: Using electronic health records data to identify patients with chronic pain in a primary care setting. *J Am Med Inform Assoc* **20**, e275–e280 (2013). <https://doi.org/10.1136/amiajnl-2013-001856>
- U.S. Department of Veterans Affairs and U.S. Department of Defense (2017). VA/DoD Clinical Practice Guideline for Opioid Therapy for Chronic Pain. Version 3.0. Downloaded on October 21, 2019 from https://www.va.gov/HOMELESS/nchav/resources/docs/mental-health/substance-abuse/VA_DoD-CLINICAL-PRACTICE-GUIDELINE-FOR-OPIOID-THERAPY-FOR-CHRONIC-PAIN-508.pdf.
- U.S. Health Resources and Services Administration. Area Health Resources Files. <https://data.hrsa.gov/topics/health-workforce/ahrf>. Accessed 10 April 2019
- University of Washington. Rural Health Research Center. Rural-Urban Commuting Area Codes (RUCAs). Available at: <http://depts.washington.edu/uwruca/>. Accessed August 2, 2019.
- University of Wisconsin Population Health Institute (2019) County Health Rankings and Roadmaps. <http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation>. Accessed 10 April 2019.
- University of Wisconsin School of Medicine Public Health. 2015 Area Deprivation Index v2.0. Downloaded from <https://www.neighborhoodatlas.medicine.wisc.edu/> 10 April 2019.
- Vickers, A., Elkin, E.: Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Making*. **26**, 565–574 (2006)
- Vowles, K.E., Cusack K.J., Cardon K.E., McEntee M.L.: Integrated behavioral treatment for Veterans with co-morbid chronic pain and hazardous opioid use: A randomized controlled pilot trial. *J Pain*, published online November 2019
- Ward, R., Egede, L., Ramakrishnan, V., Frey, L., Axon, R., Gebregziabher, M.: An improved comorbidity summary score for measuring disease burden and predicting mortality with applications to two national cohorts. *Communications in Statistics: Theory and Methods* **48**(18), 4642–4655 (2018)
- Youden, W.J.: Index for rating diagnostic tests. *Cancer* **3**(1), 32–35 (1950)

Authors and Affiliations

Ralph Ward^{1,2}  · **Erin Weeda**^{1,3}  · **David J. Taber**^{1,4} · **Robert Neal Axon**^{1,5} · **Mulugeta Gebregziabher**^{1,2} 

¹ Health Equity and Rural Outreach Innovation Center, Ralph H. Johnson Veterans Affairs Medical Center, Charleston, SC, USA

² Department of Public Health Science, Medical University of South Carolina, Charleston, SC, USA

³ College of Pharmacy, Medical University of South Carolina, Charleston, SC, USA

⁴ Division of Transplant Surgery, College of Medicine, Medical University of South Carolina, Charleston, SC, USA

⁵ College of Medicine, Medical University of South Carolina, Charleston, SC, USA