Validation of an Electronic Medical Record–Based Algorithm for Identifying Posttraumatic Stress Disorder in U.S. Veterans

Kelly M. Harrington[1,2], **Rachel Quaden**[1], **Murray B. Stein**[3,4], **Jacqueline P. Honerlaw**[1], **Shadha Cissell**[3], **Robert H. Pietrzak**[5,6], **Hongyu Zhao**[7,8], **Krishnan Radhakrishnan**[7,9], **Mihaela Aslan**[7,10], **John Michael Gaziano**[1,11], **John Concato**[7,10], **David R. Gagnon**[1,12], **Joel Gelernter**[5,6,13], **Kelly Cho**[1,11] **VA Million Veteran Program and Cooperative Studies Program**

[1]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, Massachusetts, USA

[2]Department of Psychiatry, Boston University School of Medicine, Boston, Massachusetts, USA

[3]Psychiatry Service, VA San Diego Healthcare System, San Diego, California, USA

[4]Departments of Psychiatry and Family Medicine & Public Health, University of California San Diego, La Jolla, California, USA

[5]Psychiatry Service, VA Connecticut Healthcare System, West Haven, Connecticut, USA

[6]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut, USA

[7]VA Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, Connecticut, USA

[8]Department of Biostatistics, Yale University School of Public Health, New Haven, Connecticut, USA

[9]Department of Internal Medicine, University of Kentucky College of Medicine, Lexington, Kentucky, USA

[10]Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA

[11]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

[12]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA

[13]Departments of Genetics and Neuroscience, Yale University School of Medicine, New Haven, Connecticut, USA

Abstract

We developed an algorithm for identifying US veterans with a history of posttraumatic stress disorder (PTSD) using the Department of Veterans Affairs (VA) electronic medical record (EMR).

Correspondence concerning this article should be addressed to Kelly M. Harrington, Massachusetts Veterans Epidemiology and Research Information Center (MAVERIC), VA Boston Healthcare System, 150 S. Huntington Ave. (151-MAV), Boston, MA 02130, USA. kelly.harrington@va.gov.

This work was motivated by the need to create a valid EMR-based phenotype to identify thousands of cases and controls for a genome-wide association study of PTSD in veterans. We used manual chart review ($n = 500$) as the gold standard. For both the algorithm and chart review, three classifications were possible: *likely PTSD*, *possible PTSD*, and *likely not PTSD*. We used Lasso regression with cross-validation to first select statistically significant predictors of PTSD from the EMR and then to generate a predicted probability score of being a PTSD case for every participant in the study population. Probability scores ranged from $0 – 1.00$. Comparing the performance of our probabilistic approach (Lasso algorithm) to a rule-based approach (ICD algorithm), the Lasso algorithm showed modestly higher overall percent agreement with chart review compared to the ICD algorithm (80% vs. 75%), higher sensitivity (.95 vs. .84), and higher overall accuracy (AUC = .95 vs. .90). We applied a 0.7 probability cut point to the Lasso results to determine final PTSD case and control status for the VA population. The final algorithm had a 0.99 sensitivity, 0.99 specificity, 0.95 positive predictive value, and 1.00 negative predictive value for PTSD classification (grouping *possible PTSD* and *likely not PTSD*) as determined by chart review. This algorithm may be useful for other research and quality improvement endeavors within the VA.

Widespread implementation of electronic medical record (EMR) systems provides opportunities for transforming population-based research by enabling efficient, cost effective collection of data on a large scale, and thus helps to address a rate-limiting step for genetic research: the need for large sample sizes (Charles, Gabriel, & Furukawa, 2014; Smoller, 2017). Specifically, the development of clinical phenotypes (i.e., observable traits such as height or blood type, the presence of a disease, or the response to a medication; Newton et al., 2013) derived from EMR data and the linkage of EMRs with biobanks creates a valuable data resource for genetic and other biomarker discovery (Olson et al., 2014; Smoller, 2017). However, one major challenge that interferes with capitalizing on these resources is the need to demonstrate the validity of phenotypes extracted from the EMR (Newton et al., 2013; Smoller, 2017; Wojczynski & Tiwari, 2008). Thus, there is a critical need for highly accurate EMR phenotyping algorithms to advance genomic and mechanistic studies in megabiobanks, such as the UK Biobank and the Million Veteran Program (MVP), whereby it is not feasible to carefully and prospectively assess every participant for the disease of interest. The main objective of the current project was to develop an EMR-based algorithm for identifying posttraumatic stress disorder (PTSD) cases and controls in US veterans for a genome-wide association study (GWAS) of PTSD which is being conducted within the MVP.

Investigators in the Department of Veterans Affairs (VA) are well-positioned to use EMR-based phenotypes for clinical research given the large size of the patient population, the national scope of the healthcare system, and the wealth of over 15 years of longitudinal data available. Large VA EMR databases have been used to conduct extensive clinical and epidemiological research on a variety of priority disease domains in the VA. Diagnostic data captured in billing codes (typically based on International Classification of Diseases [ICD-9 or ICD-10]) are readily available and commonly used in EMR-derived phenotype algorithms to classify patients with specific diseases but the validity of phenotypes solely based on billing codes is questionable (Smoller, 2017). PTSD is a high-priority, complex phenotype that is prevalent in the VA population. As summarized in Supplemental Table 1, a number of

studies have evaluated the validity of PTSD diagnoses found in the VA EMR (Abrams et al., 2016; Frayne et al., 2010; Gravely et al., 2011; Holowka et al., 2014; Magruder et al., 2005). These prior studies have shown that ICD codes from the VA EMR can be used to identify PTSD, albeit with varying degrees of accuracy. Inconsistent results of PTSD validation studies using VA data can be attributed, in large part, to divergent diagnostic criteria, assessment methods, sampling procedures, and reference standards for comparing EMR-derived diagnoses.

The primary aim of this study was to develop and validate an EMR algorithm for identifying lifetime (ever) PTSD in a sample of Veterans Health Administration (VHA) service users. We sought to improve upon prior algorithms that relied exclusively on ICD codes by developing a multivariable prediction model of PTSD that assigns a probabilistic score of PTSD caseness from 0 – 1.00 for every participant in the study population. The advantages of using predicted probabilities of PTSD caseness instead of a dichotomous classification include the ability to apply a probability score of PTSD to the entire population, greater flexibility in selecting the most appropriate cut point(s) for defining cases and controls, and potential reusability of the phenotype in future studies of PTSD. For example, investigators may prioritize having a more precise PTSD case definition (by setting a higher threshold) at the cost of decreasing the sample size. Another innovative aspect of our study was using a three-level outcome for both algorithm definitions as well as classifications by expert chart reviewers to capture an intermediate category with less diagnostic certainty, reflecting real-world clinical presentations of subthreshold PTSD or insufficient information available to make a definitive PTSD diagnosis. We validated the PTSD classification algorithm using a training dataset of 500 expert-reviewed medical records labeled as *likely PTSD* (case), *possible PTSD* (neither case nor control), and *likely not PTSD* (control). We compared the performance of our probabilistic PTSD algorithm against a rule-based approach (ICD codes only), hypothesizing that our probabilistic algorithm would outperform the ICD algorithm.

A secondary aim of this paper was to demonstrate how we applied the probabilistic PTSD algorithm to (a) the VHA population and (b) the MVP population in a genomic study of PTSD in combat-exposed veterans. We determined the optimal cut points for classifying PTSD cases and controls based on estimated misclassification rates, operating characteristics, and sample sizes. Last, we used MVP survey data related to PTSD symptoms to further refine and validate the final classification algorithm for the GWAS of PTSD in MVP.

Method

Participants

This study was undertaken as a part of a GWAS of PTSD which is being conducted within the Million Veteran Program (MVP). MVP, as a mega-biobank established within the national VA healthcare system, collects blood specimens and questionnaires from consented veteran volunteers and links them with consent for EMR research use, as has been described in detail elsewhere (Gaziano et al., 2016). The source population consisted of veterans who have utilized the VHA for medical care since EMR implementation ($N = 16,770,849$). As shown in Figure 1, after excluding 608,666 veterans with ICD-9/10 codes for bipolar

disorder or schizophrenia because of concerns about reliability of PTSD symptom reporting, we censused all veterans with at least one ICD-9/10 code for PTSD (*n* = 1,400,839). To create a *likely not PTSD* (control) group, we also censused veterans with no ICD-9/10 codes for PTSD (*n* = 14,768,648). To ensure that a patient was not classified as a control due to limited use of VHA services, we required at least 3 months in the VHA system and at least 10 unique visit days. The MVP study population (*N* = 74,091) consisted of MVP enrollees who have been genotyped, completed a survey indicating that they have experienced combat or served in a warzone, and did not have ICD codes for bipolar disorder or schizophrenia.

Chart review sample.—A total of 500 veterans (including 300 who also enrolled in MVP) were randomly selected for chart review using a stratified sampling strategy based on our initial working algorithm (ICD algorithm) to ensure adequate representation of MVP participants and VHA enrollees, and our three outcome categories, for training the PTSD model (see "ICD algorithm definition" below for details). We generated chart review assignments for 400 veterans using two selection criteria: (1) the proportion of MVP participants to VHA enrollees was set at 60:40 and (2) the prevalence of ICD algorithm outcome categories were fixed to 45% *likely PTSD*, 30% *possible PTSD*, and 25% *likely not PTSD.* In addition, 50% of MVP participants (*n* = 150) were selected because they were positive for combat exposure. Finally, we selected a random sample of 100 charts to estimate the prevalence of PTSD in the study population (50 MVP participants, 50 VHA enrollees). After excluding 15 veterans with insufficient information for classification, the final chart review sample was comprised of 485 veterans (293 MVP participants, 192 VHA enrollees), 458 men, 27 women, 342 Whites, 88 Blacks, and 55 other race/missing (see Table 1).

Procedure

This study was approved and reviewed annually by the Institutional Review Boards (IRBs) at three U.S. Department of Veterans Affairs Healthcare System facilities [*locations removed for blind review*]. The IRBs granted waivers of informed consent and HIPAA authorization for access to protected health information required to conduct this VA database study. The MVP protocol was initially approved by the VA Central IRB in 2010 and reviewed annually. MVP enrollees provided written consent, gave a blood sample, and completed self-report questionnaires.

Data sources.—The algorithm was defined using variables available in the VA EMR. We obtained patient sociodemographic information, ICD-9/10 codes for mental health diagnoses, PTSD screening measures, patient flags (combat, military sexual trauma), and medication prescriptions from the VA Corporate Data Warehouse (CDW) in November 2016. For consistency, we only used CDW data that was available prior to the chart review date for a given participant. We also used MVP questionnaire data (i.e., self-report of current PTSD symptoms) to further refine and validate our classification of PTSD cases and controls among MVP participants included in a genomic study of PTSD.

Algorithm development process.—First, we created a rule-based working algorithm using ICD codes based on literature review (see Supplemental Table 1) and consultation with experts in PTSD. The "ICD algorithm" was defined using only ICD-9/10 codes for

PTSD. The ICD algorithm was used to create a base population (depicted in the upper righthand corner of Figure 1) of patients who were grouped according to their likelihood of having PTSD [i.e., *likely PTSD* (case), *possible PTSD*, and *likely not PTSD* (control)]. Second, the ICD algorithm was used to select a stratified random sample of 500 veterans from the base population for manual chart review. Third, we developed a multivariable prediction model of PTSD ("Lasso algorithm") using VA EMR data and generated a probabilistic score of PTSD caseness from 0 – 1.00 for every participant in the study population. The Lasso model was trained using chart validated cases as the "gold standard."

Chart review validation.—Blinded, independent chart reviews and abstraction were performed by five subject matter experts with a minimum of six years of experience in the diagnosis and treatment of PTSD (3 PhDs in clinical psychology, 1 psychiatrist, 1 licensed clinical social worker). All chart reviewers followed the same detailed protocol and manual chart abstraction form (see supplementary materials) and underwent intensive training in its use including six gold standard cases during the initial calibration phase. All EMR chart reviews were conducted between January 2015 and May 2016. The abstractors selected one of the following classifications of patient's PTSD status: *likely PTSD* (significant evidence of a lifetime PTSD diagnosis), *possible PTSD* (weak or indeterminate evidence of a lifetime PTSD diagnosis or presents with subclinical level of PTSD symptomatology), *likely not PTSD* (no evidence of a lifetime PTSD diagnosis), or *insufficient data available*. A random sample of 25% of the 500 charts ($n = 125$) was independently reviewed by two raters to assess inter-rater agreement for PTSD diagnosis. Discrepancies between raters on classifications of PTSD diagnosis were reviewed via regular teleconferences to reach final consensus. A weighted kappa for lifetime PTSD diagnosis was calculated for each of the possible two rater pairs (Cohen, 1968). We considered a kappa of 0.80 to be excellent. Kappa statistics ranged from 0.75 to 0.87 (*substantial agreement* to *almost perfect agreement*) (Landis & Koch, 1977) and percent agreement ranged from 75.7% to 85.7% (mean = 79.4%). Importantly, none of the rater discrepancies on chart classifications were between *likely PTSD* and *likely not PTSD*.

Measures

ICD algorithm definition.—We defined *likely PTSD* (case) by either 1 inpatient ICD-9/10 code for PTSD (ICD-9 code 309.81 and ICD-10 codes F43.10, F43.11, F43.12 listed as the primary or secondary diagnosis), or at least 2 outpatient ICD-9/10 codes for PTSD within any one-year window by a mental health professional (VA clinic stop codes 501–599). We defined *likely not PTSD* (control) as the absence of ICD-9/10 diagnostic codes for PTSD by any VHA clinic or specialty, during any VHA visit (inpatient or outpatient) or on the Problem List. We defined *possible PTSD* by having only 1 outpatient ICD-9/10 code for PTSD within a year by a mental health professional or by having only outpatient ICD-9/10 code(s) for a PTSD diagnosis made by a non-mental health professional.

Lasso algorithm definition (candidate predictors).—We tested various candidate predictors available from the VA CDW in the Lasso model, including PTSD screening instruments, psychiatric comorbidities, medications, and demographic characteristics.

***DSM-IV PTSD Checklist (PCL)* (Weathers, Litz, Herman, Huska, & Keane, 1993).—**The PTSD Checklist (PCL) (Weathers et al., 1993) is a 17-item self-report measure of PTSD symptoms based on DSM-IV criteria with solid psychometric properties including good temporal stability, internal consistency, test-retest reliability, and convergent validity (Wilkins, Lang, & Norman, 2011). Veterans were classified as having or not having a PCL score in their chart (no = 0, yes = 1) because scores are not missing at random in the EMR (i.e. lack of a PCL may indicate insufficient reason to screen a patient for PTSD).

***DSM-IV Primary Care PTSD Screen (PC-PTSD)* (Prins et al., 2003).—**The PC-PTSD is a four-item self-report measure corresponding to the four factors associated with the DSM-IV PTSD construct. Items are scored dichotomously as either 0 or 1 (0 = no, 1 = yes). The PC-PTSD has demonstrated good test-retest reliability and good diagnostic efficiency in primary care settings (Prins et al., 2003). We used a 3-level variable of the PC-PTSD (never administered; administered with a score of 0, 1 or 2; administered with a score of 3 or 4) based on evidence of an optimal cut point of 3 (Tiet, Schutte, & Leyva, 2013).

PTSD-related variables.—We tested several other predictors related to PTSD diagnosis including VA service-connected disability rating for PTSD (no = 0, yes = 1), absence/presence of 2 or more outpatient ICD-9/10 codes for PTSD within a year by a mental health professional, absence/presence of 1 or more inpatient ICD-9/10 codes for PTSD, count of ICD-9/10 codes for PTSD associated with mental health clinic stop codes (501–599), count of ICD-9/10 codes for PTSD associated with PTSD clinic stop codes (516, 519, 540–542, 561, 562, 580, 581), and count of ICD-9/10 codes for PTSD associated with primary care physician clinic stop codes (323, 348, 350).

Psychiatric comorbidities and medications.—We tested comorbid depression and anxiety disorder (absence = 0, presence = 1) in the Lasso model, defined by having either one inpatient or two outpatient ICD-9/10 codes (list of ICD codes is available upon request). Prescription counts of psychiatric medications were also evaluated, grouped by medication class (antidepressants, antipsychotics, sedatives/hypnotics, mood stabilizers, and prazosin).

Demographic characteristics.—We tested age (continuous), sex (male = 0, female = 1), ethnicity (NA = 0, Hispanic = 1, not Hispanic = 2), and race (NA = 0, White = 1, Black = 2, Other = 3).

Trauma exposure.—Trauma exposure was not included in the original Lasso model due to concerns about reliability and missingness of these variables in CDW. However, we tested the combat flag (no = 0, yes = 1) and military sexual trauma (MST) flag (no = 0, yes = 1) as predictors in an alternative model to evaluate the impact on accuracy of the algorithm.

Probability of caseness.—Every participant in the study population received a probability of PTSD caseness ranging from 0 – 1.00 based on the Lasso probabilistic model.

Data Analysis

Chart review sample characteristics.—We compared the gold standard chart review *likely PTSD* and *likely not PTSD* groups on demographic and clinical characteristics. Specifically, we tested for statistically significant differences between the two groups using the chi-square test and t-test for frequencies and means, respectively. Using the same statistical approach, we also tested for significant differences between MVP participants and VHA enrollees by three PTSD categories: *likely PTSD*, *possible PTSD*, and *likely not PTSD*.

Lasso algorithm model development and evaluation.—First, we evaluated possible collinearity among candidate predictors for the Lasso algorithm using Spearman rank-correlation coefficients. Next, we used Least Absolute Shrinkage and Selection Operator (Lasso) penalized regression with 10-fold cross-validation to fit a prediction model for PTSD. Lasso penalized regression imposes an L1 penalty on the size of coefficients and sets some coefficients to zero, thus retaining only the most important predictors in the model without causing overfitting and multiple testing issues of other selection methods (e.g., stepwise models) (Tibshirani, 1996). Lasso analyses were completed using the R package GLMNET, which has built-in cross-validation (cv.glmnet) (Friedman, Hastie, & Tibshirani, 2010). The coefficients for each covariate and outcome were used to determine the probability for membership in each outcome category (i.e. *likely PTSD*, *possible PTSD*, *likely not PTSD*). Since Lasso can handle multinomial distributions, we entered the three-level PTSD outcome in the regression model. Lasso generated a probability for each of the three outcome categories for every participant; the total probability sums to 1.0. Participants were then classified into one of the three categories based on the maximum probability generated.

Participants with missing age or sex were removed from the analysis ($n = 1,443$). For all other variables tested in the model, missingness was incorporated into the coding of the variable (e.g., PCL score was coded as absent/present [0/1]). To evaluate the rank order of predictors, we standardized the coefficients for *likely PTSD* by multiplying the standard deviation times the beta value. All programs for participant selection and analysis were checked by an independent data analyst for quality control.

Performance characteristics.—We calculated classification metrics (percent agreement, Cohen's weighted kappa, sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], and area under the curve [AUC]) to evaluate the performance of the ICD and Lasso algorithms against the chart review gold standard. We also calculated the AUC (or c-statistic) to directly compare the performance of an alternative Lasso model (including trauma variables) and to evaluate whether the accuracy of the algorithms differed for the MVP and VHA subsets of the chart review sample. Given these metrics require classifying subjects into dichotomous categories (e.g., case and control), we took three approaches to calculate the performance characteristics of each algorithm. First, we dropped all *possible PTSD*. Second, we grouped *possible PTSD* with *likely PTSD*. Third, we grouped *possible PTSD* with *likely not PTSD*. We applied the same rule to algorithm and chart review classifications for all three approaches.

Weighted population estimates.—We over-sampled *likely PTSD* and *possible PTSD* groups to ensure we had a sufficient number of veterans with PTSD symptoms for training the model. To account for this over-sampling of *likely PTSD* and *possible PTSD*, we weighted each individual based on their selection for chart review (i.e. ICD algorithm case status) (Scholer et al., 2007). The weighted estimates better represent the algorithm performance in the entire VHA population. We used bootstrapping with these weights to obtain estimates and confidence intervals for sensitivity, specificity, PPV, and NPV in the VHA population.

Developing the cut points.—Using the probabilities of case status generated by Lasso, we next determined the optimal cut points for the final classification of PTSD cases and controls for the VHA and MVP study populations. We derived the optimal thresholds for cases and controls based on estimated misclassification rates, operating characteristics, and sample sizes. Specifically, we conducted a sensitivity analysis to determine the most appropriate cut points for the predictive probabilities (i.e., 0.6, 0.7, or 0.8). To estimate the misclassification rates associated with applying various cut points to the study population, we used weights derived from the ICD algorithm status of the chart review participants (as previously described). We calculated the operating characteristics for each cut-off value in the overall VHA study population as well as the MVP study population. Given our use of three outcome categories, we also calculated the operating characteristics under two conditions (i.e., *possible PTSD* grouped with *likely PTSD* vs. *likely not PTSD*).

Refining algorithm for the MVP study.—We used PCL scores from MVP survey data (available for everyone in the MVP cohort) to further refine and validate the *likely not PTSD* (control) group classified by the Lasso algorithm. We considered various cut points for the PCL (30–50), and assessed tradeoffs between improved accuracy versus loss in sample size for the GWAS of PTSD analysis.

Results

Characteristics of the chart review sample

Table 1 shows the demographic and clinical characteristics for the final chart review sample ($n = 485$), after excluding 15 veterans with insufficient data available in the electronic medical record. The sample is comprised of 40.8% veterans classified by chart review as *likely PTSD*, 17.3% as *possible PTSD*, and 41.9% as *likely not PTSD*. The sample of 100 randomly selected chart reviews yielded a lifetime PTSD prevalence estimate of 18.0%. The rate of PTSD was higher among women veterans (44.4%) than men (40.6%). Approximately 73% ($n = 352/485$) of the chart reviewed sample had evidence of combat exposure based on expert chart review, presence of the combat flag in the EMR, and/or self-report on the MVP survey. Compared to *likely not PTSD* (controls), *likely PTSD* (cases) were significantly younger; more ethnically diverse; more likely to have a history of combat service, military sexual trauma, and service connection for PTSD; and more likely to have depressive and anxiety disorders. When the chart review sample was stratified by MVP status, we found that the MVP group was generally comparable to the VHA group with three exceptions. Among MVP participants, *likely PTSD* (cases) were significantly older, *likely not PTSD*

(controls) were significantly less racially and ethnically diverse, and *possible PTSD* cases were more likely to have depression relative to the VHA group (see Supplemental Table 2).

Lasso algorithm

The results of the Lasso procedure are shown in Table 2; unstandardized and standardized coefficients are reported for all variables retained in the model. All classes of psychiatric medications, the count of ICD-9/10 codes for PTSD associated with PTSD clinic stop codes, race, and comorbid anxiety disorder were dropped from the model. Results of the alternative Lasso model that added the MST screen and combat flag are provided in Supplemental Table 3. We found that combat and MST were retained by the Lasso regression and the main change was that sex dropped out of the model. Only three male veterans were reclassified among the 485 charts reviewed according to the alternative model: two *possible PTSD* and one *likely not PTSD* were reclassified as *likely PTSD* (each probability of being a PTSD case was just over 0.5). The AUC value for the alternative model was identical to the original Lasso model (AUC = .95) which suggests that the addition of these two variables did not improve the overall accuracy of the algorithm.

Agreement between algorithms and chart review

Table 3 displays 3×3 contingency tables comparing patient classifications according to the ICD algorithm, the expanded Lasso algorithm, and chart review. The percent agreement between the ICD algorithm and the gold standard PTSD diagnosis was 75.3% ($n = 365/485$) and kappa (κ) = 0.72 (95% CI: 0.68 – 0.77) which indicates *substantial agreement*. The percent agreement of Lasso algorithm-defined PTSD with chart review classifications was 80.2% (389/485) and κ = 0.77 (95% CI: 0.73 – 0.81) which suggests *substantial agreement.* The percent agreement between the ICD algorithm and the Lasso algorithm for lifetime PTSD diagnosis was 75.7% and κ = 0.74 (95% CI: 0.70 – 0.78), suggesting *substantial agreement.* Applying the algorithms to the entire VHA cohort, the prevalence of PTSD was slightly higher according to the Lasso algorithm (14.9%) than the ICD algorithm (12.9%). Figure 2 shows the predicted probabilities of PTSD caseness compared with chart review classifications. The median probabilities of PTSD caseness were .92, .14, and .03 for chart reviewed *likely PTSD*, *possible PTSD*, and *likely not PTSD*, respectively. The *likely not PTSD* group had the highest level of agreement between chart classifications and algorithm predictive probabilities, whereas there was considerably more variation for *possible PTSD* and *likely PTSD* groups. However, there were 7 outliers in the *likely not PTSD* group which had moderate probabilities of being a *likely PTSD* case (.53 – .60). Each outlier had one or more of the strongest predictors of PTSD present in the EMR, whereas chart reviewers did not find any significant evidence of PTSD in the charts. Notably, all outliers would be excluded from analysis when applying the optimal cut point of 0.7 for PTSD cases and controls.

Performance characteristics

The performance characteristics for identifying lifetime PTSD diagnosis using the ICD algorithm and the Lasso algorithm are presented in Table 4, separately for the chart review sample and the VHA study population. Both algorithms performed optimally when participants with diagnostic uncertainty (i.e., *possible PTSD*) were excluded from

calculations and there was a trend for both algorithms to perform better when grouping *possible PTSD* with *likely not PTSD* (versus *likely PTSD*). For direct comparison of the two algorithms, we will focus on the performance characteristics based on grouping *possible PTSD* with *likely not PTSD* (controls) because it represents an intermediate scenario between the most optimistic and pessimistic results. The Lasso and ICD algorithms had comparable specificity (.84 vs. 87), PPV (.57 vs. .59), and NPV (.99 vs. .96), whereas the Lasso algorithm had higher sensitivity (.95 vs. .84) and slightly higher accuracy (AUC = .95 vs. .90). After applying population weights to account for over-sampling the *likely PTSD* and *possible PTSD* groups, the specificity (.95 vs. .97) and PPV (.72 vs. .82) improved for both the Lasso- and ICD-algorithms, respectively.

We also evaluated whether the algorithms performed differently in the chart review sample of MVP participants versus VHA enrollees. Similar accuracy was observed in the MVP and VHA subgroups for both the ICD algorithm (AUC = .91 vs. .89) and the Lasso algorithm (AUC = .95 vs. .94). Table 5 shows estimated misclassification rates and corresponding operating characteristics using a range of cutoffs for probability of PTSD caseness (0.6 to 0.8) for the VHA and MVP study populations (see Supplemental Table 4 for summary of cut points considered for controls). A cutoff of 0.7 appears to be optimal for both VHA and MVP, given that the estimated misclassification rates leveled off without improvement in accuracy to balance the additional loss of sample size that would accompany a higher probability cutoff. After applying a probability cut point of 0.7 for both PTSD cases and controls, the operating characteristics overall looked similarly strong for both the VHA and MVP populations with all metrics   .95 except for the PPV in the MVP population (.87).

After deriving the optimal cut points for the final MVP cohort, we used PCL scores from MVP survey data to help further validate MVP Lasso algorithm-defined controls. We selected a threshold score of 30 on the PCL from the MVP survey to minimize false negative classifications. As a result, we excluded veterans from the final MVP control sample who screened positive for PTSD according to the MVP-administered PCL despite having little or no indication of PTSD in the VA EMR. Although this step reduced the final MVP control sample by 28.3% (from *n* = 46,884 to 33,609), it increased our confidence that we identified a group of "true controls" for inclusion in the MVP GWAS of PTSD analysis.

Discussion

The current study developed and evaluated an EMR algorithm for identifying lifetime PTSD in a sample of VHA service users that assigns a likelihood of PTSD caseness ranging from 0 – 1.00. We validated our probabilistic PTSD algorithm ("Lasso algorithm") using a training dataset of expert-adjudicated medical record reviews on 500 veterans and found *substantial agreement* with chart reviewers' classifications (Cohen's κ = .77). As previously described, three classifications were possible for algorithm definitions and chart reviews: *likely PTSD* (case), *possible PTSD*, and *likely not PTSD* (controls). For simplicity, we have focused our discussion on the set of results based on grouping *possible PTSD* with *likely not PTSD*. The Lasso algorithm showed high sensitivity (.95) and high accuracy (AUC = .95) compared to time and resource-intensive chart review. The predictors which appeared to contribute the most to probability of PTSD caseness in the Lasso model were (1) presence of service

connection for PTSD, (2) presence of at least two outpatient ICD codes for PTSD by a mental health professional within a year, and (3) the PC-PTSD screen. We acknowledge that variable retention depends largely on what other variables are entered in the predictive model, as demonstrated by sex dropping out of the model when the combat flag and MST screen were added to the model. Similarly, elimination of a variable from the model does not indicate a lack of association with PTSD, rather that it does not add to our predictive ability.

This paper also describes the method we used for determining the optimal cut points for classifying PTSD cases and controls in the overall VHA population and in the MVP population within the context of a genomic study of PTSD which motivated this work. We found an optimal probability cut point of 0.7 for *likely PTSD* (cases) and *likely not PTSD* (controls) in the VHA study population to minimize both false positive and false negative classifications, which yielded high sensitivity, specificity, PPV, and NPV (.95). The optimal probability threshold was the same (0.7) in the MVP population of combat-exposed veterans and showed similarly strong operating characteristics with the exception of a lower PPV than was observed in the VHA population (.87 vs. .95).

Our final classification algorithm (i.e., after applying cut points of 0.7 to PTSD cases and controls, and thus excluding participants with less decisive probabilities) compares favorably with all of the previous studies we reviewed that used ICD codes from the VA EMR to identify PTSD (Abrams et al., 2016; Frayne et al., 2010; Gravely et al., 2011; Holowka et al., 2014; Magruder et al., 2005). Our algorithm performed most similarly to Abrams et al.'s (2016) PTSD algorithm that was defined by 3 outpatient ICD-9 PTSD codes and used chart review as the reference standard. Compared to our final algorithm, Abrams and colleagues' algorithm carried a slightly lower sensitivity (.98 vs. .99), lower specificity (.97 vs. .99), and higher PPV (.97 vs. .95).

We also compared the performance of our probabilistic approach (Lasso algorithm) to a rule-based approach (ICD algorithm) against the gold standard chart review classifications. Relative to the ICD algorithm, the Lasso algorithm classified many more chart review participants as *likely PTSD* (cases) or *likely not PTSD* (controls), and markedly fewer participants as *possible PTSD* (11 vs. 118). The Lasso algorithm showed modestly higher overall percent agreement with chart review compared to the ICD algorithm (80% vs. 75%) and the operating characteristics were quite similar for the two algorithms. The ICD algorithm showed slightly higher specificity, PPV, and NPV, whereas the Lasso algorithm had higher sensitivity (.95 vs. .84) and overall accuracy (AUC = .95 vs. 90). The Lasso algorithm yielded more *likely PTSD* (cases) and *likely not PTSD* (controls) than the ICD algorithm (see Figure 1), and thus can afford greater statistical power with a larger sample size. However, when it is desirable to increase the precision and specificity of case/control definitions by excluding individuals with diagnostic uncertainty, this can result in a significant drop in sample size. For example, in the MVP genomic case-control study, Lasso identified 22,651 *likely PTSD* cases but the sample size decreased to 16,490 after we applied the $p > 0.7$ cut point (shown in Table 5).

Taken together, these findings suggest that ICD codes may be sufficient for identifying PTSD cases and controls depending on the goals of a study (Bauer et al., 2015). That is, the

simpler rule-based ICD algorithm may be preferable for studies that can afford to eliminate individuals with intermediate diagnostic certainty, do not require a likelihood of having PTSD for every patient in the population, or have limited resources for data extraction and analysis. The probabilistic approach, which defines PTSD caseness continuously (0 – 1.00), may be most desirable for studies designed to analyze PTSD as a quantitative trait. The predictive PTSD probability scores can be used in a variety of ways in line with the specific study aims. For example, directly modeling the algorithm-derived probability of being a case has been shown to improve test power to detect phenotype-genotype association and effect estimation compared with algorithms that yield dichotomous predictions (Sinnott et al., 2014). For the proposed genomic case-control study, the MVP population has a sufficient number of participants to allow those with less decisive probabilities to be excluded from analysis. If it were necessary to classify PTSD status for all participants, such as in a longitudinal cohort study, a fixed cutoff can be chosen which maximizes sensitivity and specificity. Defining PTSD caseness as a predicted probability also offers potential reusability of the phenotype in future studies.

Strengths and Limitations

To our knowledge, this study represents the first to develop an EMR-based algorithm for identifying PTSD in veterans using VHA services that uses a probabilistic (statistical modeling) approach rather than a rule-based approach that relies on ICD codes. Our stratified sample provides a sufficient range of subjects to train the algorithm, while providing sampling weights to account for deliberate oversampling and to allow accurate projection of statistics to the general VHA population. Another innovative aspect of our study involved using a three-level outcome for both algorithm definitions as well as classifications by expert chart reviewers. Although the *possible PTSD* category reflects the reality that there is sometimes insufficient information available in the EMR to make a definitive diagnosis, we acknowledge that our use of a third intermediate category had the unintended consequence of complicating the interpretation of results, given that current methods for evaluating performance characteristics are based on dichotomous outcomes.

Several limitations of this study must be acknowledged. First, these findings were based on data collected from and about VHA service users, including a subset of veterans voluntarily enrolled in the MVP, and may not generalize to the entire VHA population or to veterans who predominantly utilize other healthcare settings. In addition, the over-representation of White male combat veterans in the validation sample may limit the generalizability of the algorithm to other subgroups of VHA users (e.g., women and minorities). This concern was partially alleviated by our finding that both algorithms had nearly identical levels of accuracy in the MVP and VHA subgroups, despite a higher rate of combat exposure among MVP participants in the chart review sample. Further, a recent study demonstrated that the demographic profile of participants in MVP is representative of the broader VHA population including mean age and gender distributions (Nguyen et al., 2018). Second, EMR-based algorithms are limited by the accuracy and completeness of the EMR data used to define the disease or condition of interest, and thus are subject to misclassification (Chubak, Pocobelli, & Weiss, 2012; Kim et al., 2012). Third, this study relied on EMR review to derive a "gold standard" for validation which depends on clinicians to properly recognize and diagnose

PTSD cases. To address this limitation, we are conducting external validation of the algorithm using the Clinician-Administered PTSD Scale for DSM-5 (CAPS-5) (Weathers et al., 2018), which is considered a more general "gold standard" PTSD assessment tool. The CAPS-5 validation project is well underway and, if successful, will lend further support for the accuracy and validity of our algorithm.

Conclusions

In conclusion, the current study describes an EMR-based algorithm that accurately identifies lifetime PTSD in veterans. Our algorithm generated probabilistic scores for PTSD for every participant in the VHA study population which affords flexibility in applying cut points to achieve the desired level of accuracy for classifying PTSD cases and controls and potential reusability of the phenotype in future studies. An important future direction for research is to refine our algorithm by applying high-throughput phenotyping and natural language processing (NLP) techniques to improve the efficiency and accuracy of defining phenotypes based on EMR data (Yu et al., 2015). Although designed for use in a research setting, a potential future clinical application of this algorithm is development of a comprehensive screening tool (e.g., a "PTSD calculator" that generates a veteran's predicted probability score of having PTSD) to aid VA clincians in assessment and treatment planning.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Abrams TE, Vaughan-Sarrazin M, Keane TM, & Richardson K (2016). Validating administrative records in post-traumatic stress disorder. International Journal of Methods in Psychiatric Research, 25, 22–32. doi: 10.1002/mpr.1470 [PubMed: 26082028]

Bauer MS, Lee A, Miller CJ, Bajor L, Li M, & Penfold RB (2015). Effects of diagnostic inclusion criteria on prevalence and population characteristics in database research. Psychiatric Services, 66, 141–148. doi: 10.1176/appi.ps.201400115 [PubMed: 25320943]

Charles D, Gabriel M, & Furukawa MF (2014). Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008–2013. ONC Data Brief, no. 16. https://www.healthit.gov/sites/default/files/oncdatabrief16.pdf

Chubak J, Pocobelli G, & Weiss NS (2012). Tradeoffs between accuracy measures for electronic healthcare data algorithms. Journal of Clinical Epidemiology, 65, 343–349 e342. doi: 10.1016/j.jclinepi.2011.09.002 [PubMed: 22197520]

Cohen J (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213–220. [PubMed: 19673146]

Frayne SM, Miller DR, Sharkansky EJ, Jackson VW, Wang F, Halanych JH, … Keane TM (2010). Using administrative data to identify mental illness: what approach is best? American Journal of Medical Quality, 25, 42–50. doi: 10.1177/1062860609346347 [PubMed: 19855046]

Friedman J, Hastie T, & Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33, 1–22. [PubMed: 20808728]

Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, … O'Leary TJ (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. Journal of Clinical Epidemiology, 70, 214–223. doi: 10.1016/j.jclinepi.2015.09.016 [PubMed: 26441289]

Gravely AA, Cutting A, Nugent S, Grill J, Carlson K, & Spoont M (2011). Validity of PTSD diagnoses in VA administrative data: comparison of VA administrative PTSD diagnoses to self-reported PTSD Checklist scores. Journal of Rehabilitation Research and Development, 48, 21–30. [PubMed: 21328160]

Holowka DW, Marx BP, Gates MA, Litman HJ, Ranganathan G, Rosen RC, & Keane TM (2014). PTSD diagnostic validity in Veterans Affairs electronic records of Iraq and Afghanistan veterans. Journal of Consulting and Clinical Psychology, 82, 569–579. doi: 10.1037/a0036347 [PubMed: 24731235]

Kim HM, Smith EG, Stano CM, Ganoczy D, Zivin K, Walters H, & Valenstein M (2012). Validation of key behaviourally based mental health diagnoses in administrative data: suicide attempt, alcohol abuse, illicit drug abuse and tobacco use. BMC Health Services Research, 12, 18. doi: 10.1186/1472-6963-12-18 [PubMed: 22270080]

Landis JR, & Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174. [PubMed: 843571]

Magruder KM, Frueh BC, Knapp RG, Davis L, Hamner MB, Martin RH, … Arana (2005). Prevalence of posttraumatic stress disorder in Veterans Affairs primary care clinics. General Hospital Psychiatry, 27, 169–179. doi: 10.1016/j.genhosppsych.2004.11.001 [PubMed: 15882763]

Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, … Denny (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of American Medical Informatics Association, 20, e147–154. doi: 10.1136/amiajnl-2012-000896

Nguyen XT, Quaden RM, Song RJ, Ho YL, Honerlaw J, Whitbourne S, … Cho K (2018). Baseline characterization and annual trends of body mass index for a mega-biobank cohort of US Veterans 2011–2017. Journal of Health Research and Reviews. Advance online publication. doi: 10.4103/jhrr.jhrr_10_18

Olson JE, Bielinski SJ, Ryu E, Winkler EM, Takahashi PY, Pathak J, & Cerhan JR (2014). Biobanks and personalized medicine. Clinical Genetics, 86, 50–55. doi: 10.1111/cge.12370 [PubMed: 24588254]

Prins A, Ouimette P, Kimerling R, Cameron R, Hugelshofer D, Shaw-Hegwer J, … Sheikh J (2003). The primary care PTSD screen (PC-PTSD): development and operating characteristics. Primary Care Psychiatry, 9, 9–14.

Scholer MJ, Ghneim GS, Wu S, Westlake M, Travers DA, Waller AE, … Wetterhall SF (2007). Defining and applying a method for improving the sensitivity and specificity of an emergency department early event detection system. AMIA Annual Symposium Proceedings Archive, 651–655.

Sinnott JA, Dai W, Liao KP, Shaw SY, Ananthakrishnan AN, Gainer VS, … Cai T (2014). Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. Human Genetics, 133, 1369–1382. doi: 10.1007/s00439-014-1466-9 [PubMed: 25062868]

Smoller JW (2017). The use of electronic health records for psychiatric phenotyping and genomics. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. doi: 10.1002/ajmg.b.32548

Tibshirani R (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58, 267–288.

Tiet QQ, Schutte KK, & Leyva YE (2013). Diagnostic accuracy of brief PTSD screening instruments in military veterans. Journal of Substance Abuse Treatment, 45, 134–142. doi: 10.1016/j.jsat. 2013.01.010 [PubMed: 23506779]

Weathers FW, Bovin MJ, Lee DJ, Sloan DM, Schnurr PP, Kaloupek DG,…Marx BP (2018). The Clinician-Adminstered PTSD Scale for DSM-5 (CAPS-5): Development and initial psychometric evaluation in military veterans. Psychological Assessment, 30, 383–395. doi:10.1037/pas0000486 [PubMed: 28493729]

Weathers FW, Litz BT, Herman D, Huska JA, & Keane TM (October 1993). The PTSD Checklist (PCL): Reliability, Validity, and Diagnostic Utility. Paper presented at the Annual Convention of the International Society for Traumatic Stress Studies, San Antonio, TX.

Wilkins KC, Lange AJ, & Norman SB (2011). Synthesis of the psychometric properties of the PTSD checklist (PCL) military, civilian, and specific versions. Depression and Anxiety, 28, 596–606. doi: 10.1002/da.20837. [PubMed: 21681864]

Wojczynski MK, & Tiwari HK (2008). Definition of phenotype. Advances in Genetics, 60, 75–105. doi: 10.1016/S0065-2660(07)00404-X [PubMed: 18358317]

Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, … Cai T (2015). Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. Journal of the American Medical Informatics Association, 22, 993–1000. doi: 10.1093/jamia/ocv034 [PubMed: 25929596]
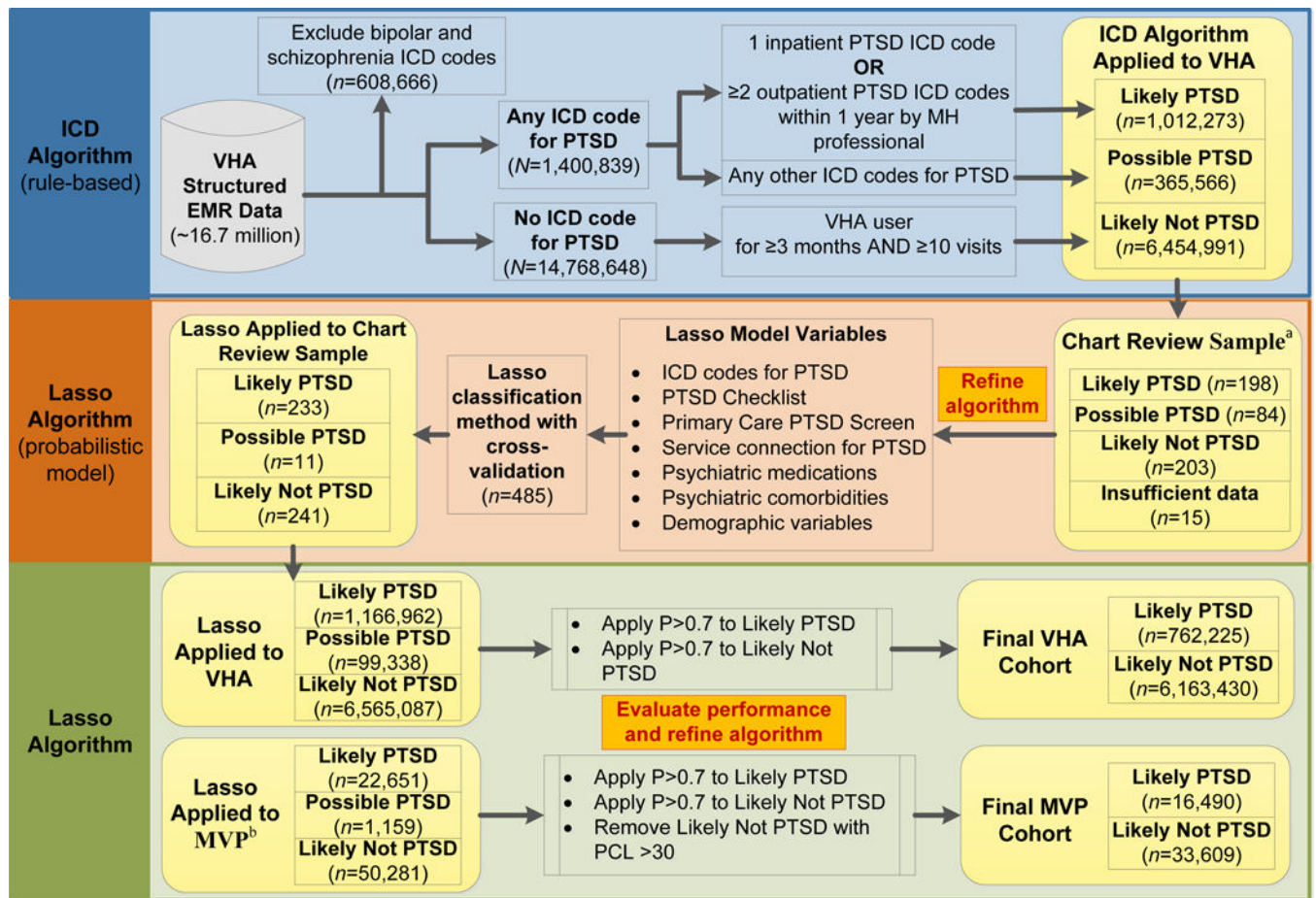
Figure 1.

Sampling and flow of participants for defining and validating algorithm to identify lifetime PTSD in veterans. VHA = Veterans Health Administration; EMR = electronic medical record; ICD = International Classification of Diseases; MH = mental health; MVP = Million Veteran Program; Lasso = Least Absolute Shrinkage and Selection Operator (regression procedure); P = probability.

[a]Sample includes 300 MVP participants. [b]MVP cohort requires combat exposure defined by MVP survey and available genotype.
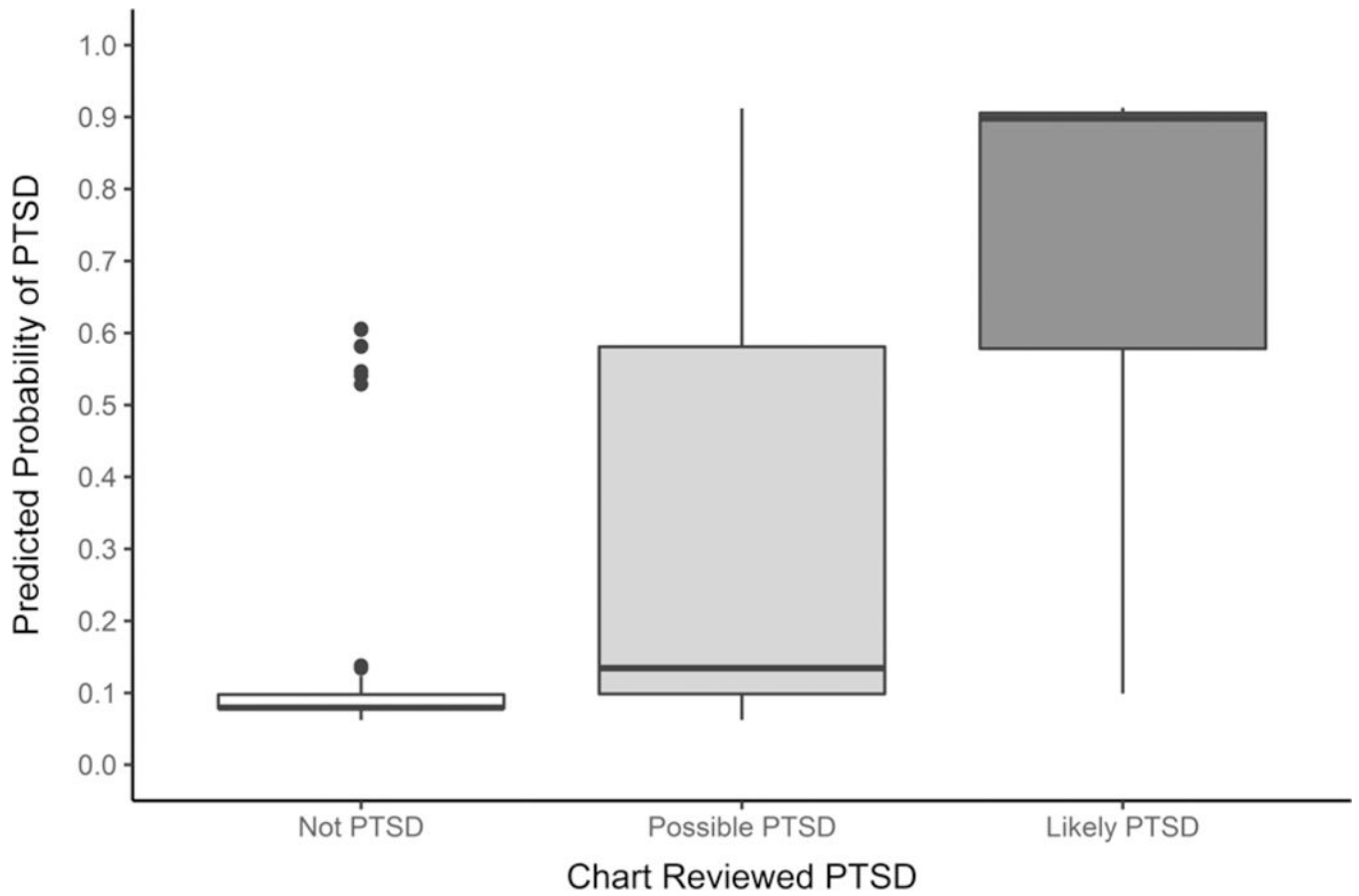
Figure 2.
Boxplot of chart reviewed PTSD case status by Lasso algorithm predicted probability of being a *likely PTSD* (case).

Table 1

Characteristics of the Chart Review Sample by PTSD Classification

Characteristic	Likely PTSD (*n* = 198)	Possible PTSD (*n* = 84)	Likely not PTSD (*n* = 203)	Overall (*N* = 485)
	n (%)	*n (%)*	*n (%)*	*n (%)*
MVP participant	120 (60.6)	45 (53.6)	128 (63.1)	293 (60.4)
Age, *M* (*SD*)	60.45 (15.00)	62.00 (16.02)	68.35 (14.39)*	64.03 (15.35)
Female sex	12 (6.1)	5 (6.0)	10 (4.9)	27 (5.6)
Race				
White	142 (71.7)	60 (71.4)	140 (70.0)	342 (70.5)
Black	39 (19.7)	16 (19.1)	33 (16.3)	88 (18.1)
Other/Missing	17 (8.6)	8 (9.5)	30 (14.8)	55 (11.3)
Hispanic/Latino ethnicity	23 (11.6)	3 (3.6)	7 (3.5)*	33 (6.8)
Missing ethnicity	10 (5.1)	6 (7.1)	25 (12.3)	41 (8.5)
Comorbid depression	116 (58.6)	42 (50.0)	56 (27.6)*	214 (44.1)
Comorbid anxiety	76 (38.4)	26 (31.0)	35 (17.2)*	137 (28.2)
Service connected (PTSD)	142 (71.7)	19 (22.6)	2 (1.0)*	163 (33.6)
PCL available (CDW)	62 (31.3)	21 (25.0)	6 (3.0)*	89 (18.4)
PC-PTSD screen				
Score of 0, 1, 2	68 (34.3)	39 (46.4)	159 (78.3)*	266 (54.9)
Score of 3, 4	89 (45.0)	32 (38.1)	11 (5.4)*	132 (27.2)
Missing	41 (20.7)	13 (15.5)	33 (16.3)	87 (17.9)
2 + OP PTSD ICD codes[a]	164 (82.8)	30 (35.7)	5 (2.5)*	199 (41.0)
1 + IP PTSD ICD codes	44 (22.2)	2 (2.4)	4 (2.0)*	50 (10.3)
No. PTSD dx by PCP, *M* (*SD*)	2.41 (4.16)	1.38 (2.42)	0.19 (1.02)*	1.30 (3.08)
No. PTSD dx by MHP, *M* (*SD*)	24.29 (60.46)	4.87 (12.60)	0.25 (1.35)*	10.86 (40.53)
Combat service (CDW)	48 (24.2)	13 (15.5)	19 (9.4)*	81 (16.2)
MST (positive screen)	12 (6.1)	5 (6.0)	4 (2.0)*	21 (4.2)

Note. PTSD = posttraumatic stress disorder; PCL = DSM-IV PTSD Checklist; CDW = Corporate Data Warehouse; PC-PTSD = Primary Care PTSD Screen; OP = outpatient; ICD = International Classification of Diseases; IP = inpatient; no. = number; dx = diagnosis; PCP = Primary Care Physician; MHP = mental health professional; MST = military sexual trauma.

[a] 2 or more outpatient ICD codes for PTSD within 1 year by mental health professional.

* *p* < .05 (denotes Likely PTSD and Likely not PTSD groups are significantly different).

Table 2

Lasso Algorithm Model Coefficients for Likely PTSD, Possible PTSD, and Likely Not PTSD

Variable	Likely PTSD B	Likely PTSD β	Possible PTSD B	Likely not PTSD B
Intercept	−0.96		−0.39	1.35
Service connected for PTSD (no/yes)	2.20	1.00	0.19	−2.38
2 + outpt. PTSD ICD codes (no/yes)[a]	1.94	.95	0.24	−2.18
PC-PTSD (NA/score 0–2/score 3–4)	0.33	.22	0.10	−0.43
PTSD dx by PCP (ICD code count)	0.07	.22	0.06	−0.13
PTSD dx by MHP (ICD code count)	0.01	.21	−0.00	−0.00
1 + inpt. PTSD ICD codes (no/yes)	0.55	.17	−0.75	0.19
Comorbid depression (no/yes)	0.33	.16	0.17	−0.51
Ethnicity (NA/Hispanic/not Hispanic)	−0.26	.15	0.10	0.16
Age (continuous)	−0.01	.11	−0.00	0.01
PCL score available in CDW (no/yes)	0.03	.01	0.60	−0.64
Sex (male/female)	−0.00	.00	−0.00	0.01

Note. PTSD = posttraumatic stress disorder; outpt. = outpatient; ICD = International Classification of Diseases; PC-PTSD = Primary Care PTSD Screen; NA = not available/missing; dx = diagnosis; PCP = Primary Care Physician; MHP = mental health professional; inpt. = inpatient; PCL = DSM-IV PTSD Checklist.

[a] 2 or more outpatient ICD codes for PTSD within 1 year by mental health professional.

Table 3

Contingency Tables of Patient Classifications by Chart Review and by the Two Algorithms

	Chart Review Classification			
ICD Algorithm	**Likely PTSD**	**Possible PTSD**	**Likely not PTSD**	**Total**
Likely PTSD	166	30	7	203
Possible PTSD	31	45	42	118
Likely not PTSD	1	9	154	164
Total	198	84	203	485
Lasso Algorithm				
Likely PTSD	188	38	7	233
Possible PTSD	2	7	2	11
Likely not PTSD	8	39	194	241
Total	198	84	203	485

	Lasso Algorithm			
ICD Algorithm	**Likely PTSD**	**Possible PTSD**	**Likely not PTSD**	**Total**
Likely PTSD	197	2	4	203
Possible PTSD	33	9	76	118
Likely not PTSD	3	0	161	164
Total	233	11	241	485

Note. Of the 500 charts reviewed by content experts, 15 participants could not be classified due to insufficient information, leaving a final chart review sample of $n = 485$.

Table 4

Algorithm Performance Characteristics in the Chart Reviewed Sample and in the VHA Population

		Chart reviewed sample (N = 485)					VHA population (weighted bootstrapped results)			
		Se	Sp	PPV	NPV	AUC	Se [95% CI]	Sp [95% CI]	PPV[a] [95% CI]	NPV[a] [95% CI]
ICD Alg.	Drop Poss.[b]	.99	.96	.83	1.00	.98	.96 [.89, 1.0]	1.0 [.99, 1.0]	.96 [.90, 1.0]	1.0 [.99, 1.0]
	Poss.+ Case	.97	.76	.47	.99	.93	.75 [.66, .94]	.98 [.96, .99]	.88 [.80, .95]	.94 [.92, .96]
	Poss.+ Cont.	.84	.87	.59	.96	.90	.86 [.75, .95]	.97 [.96, .99]	.82 [.73, .91]	.98 [.97, .99]
Lasso Alg.	Drop Poss.[c]	.96	.97	.86	.99	.99	.96 [.90, 1.0]	1.0 [.99, 1.0]	.98 [.92, 1.0]	.99 [.98, 1.0]
	Poss.+ Case	.83	.96	.81	.96	.95	.65 [.56, .74]	1.0 [.99, 1.0]	.97 [.92, 1.0]	.91 [.88, .94]
	Poss.+ Cont.	.95	.84	.57	.99	.95	.95 [.89, 1.0]	.95 [.93, .97]	.72 [.61, .83]	.99 [.99, 1.0]

Note. The 3-level outcome was dichotomized to calculate performance metrics using three approaches: drop all possible PTSD, group possible PTSD with likely PTSD, and group possible PTSD with likely not PTSD. VHA = Veterans Health Administration; Se = sensitivity; Sp = specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the curve; CI = confidence interval; Alg. = Algorithm; Poss. = possible PTSD; Case = likely PTSD (case); Cont. = likely not PTSD (control).

[a]PPV and NPV calculated using prevalence estimate of 18% based on random sample of chart reviews (n = 100).

[b]
n = 328.

[c]
n = 397.

Table 5

Cut Points for Predictive Probabilities of PTSD Case Status and Estimated Misclassification Rates in the VHA and MVP Cohorts

Cohort	Prob. (case) cut-off	Cases (n)	Est. mis.[a] (%)	Est. mis.[b] (%)	Se[a]	Se[b]	Sp[a]	Sp[b]	PPV[a]	PPV[b]	NPV[a]	NPV[b]
VHA[c]	Lasso	1,166,962	23.9	6.4	.99	.78	.96	.99	.83	.93	1.0	.95
	> 0.6	933,249	17.3	4.2	.99	.74	.97	.99	.89	.96	1.0	.95
	> 0.7	762,225	8.7	0.6	.99	.71	.99	1.0	.95	.99	1.0	.94
	> 0.8	690,599	8.5	0.7	.99	.69	.99	1.0	.96	.99	1.0	.94
MVP[d]	Lasso	22,651	18.2	5.7	.98	.98	.92	.97	.73	.89	1.0	1.0
	> 0.6	19,058	12.9	1.4	.98	.98	.95	.99	.81	.97	1.0	.99
	> 0.7	16,490	9.2	1.5	.98	.97	.97	1.0	.87	.98	.99	.99
	> 0.8	15,416	9.6	1.6	.98	.97	.97	1.0	.87	.98	.99	.99

Note. VHA = Veterans Health Administration; MVP = Million Veteran Program; Prob. = probability; Est. mis. = estimated misclassification; Se = sensitivity; Sp = specificity; PPV = positive predictive value; NPV = negative predictive value.

[a] Group possible PTSD + controls.

[b] Group possible PTSD + cases.

[c] Based on n = 6,163,430 controls.

[d] Based on n = 46,884 controls.