American Journal of Epidemiology

Vol. 192, No. 2
https://doi.org/10.1093/aje/kwac182
Advance Access publication:
November 4, 2022

Practice of Epidemiology

Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning

David S. Carrell*, Susan Gruber, James S. Floyd, Maralyssa A. Bann, Kara L. Cushing-Haugen, Ron L. Johnson, Vina Graham, David J. Cronkite, Brian L. Hazlehurst, Andrew H. Felcher, Cosmin A. Bejan, Adee Kennedy, Mayura U. Shinde, Sara Karami, Yong Ma, Danijela Stojanovic, Yueqin Zhao, Robert Ball, and Jennifer C. Nelson

* Correspondence to Dr. David Carrell, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: david.s.carrell@kp.org).

Initially submitted August 11, 2021; accepted for publication October 11, 2022.

We sought to determine whether machine learning and natural language processing (NLP) applied to electronic medical records could improve performance of automated health-care claims-based algorithms to identify anaphylaxis events using data on 516 patients with outpatient, emergency department, or inpatient anaphylaxis diagnosis codes during 2015–2019 in 2 integrated health-care institutions in the Northwest United States. We used one site's manually reviewed gold-standard outcomes data for model development and the other's for external validation based on cross-validated area under the receiver operating characteristic curve (AUC), positive predictive value (PPV), and sensitivity. In the development site 154 (64%) of 239 potential events met adjudication criteria for anaphylaxis compared with 180 (65%) of 277 in the validation site. Logistic regression models using only structured claims data achieved a cross-validated AUC of 0.58 (95% CI: 0.54, 0.63). Machine learning improved cross-validated AUC to 0.62 (0.58, 0.66); incorporating NLP-derived covariates further increased cross-validated AUCs to 0.70 (0.66, 0.75) in development and 0.67 (0.63, 0.71) in external validation data. A classification threshold with cross-validated PPV of 79% and cross-validated sensitivity of 66% in development data had cross-validated PPV of 78% and cross-validated sensitivity of 56% in external data. Machine learning and NLP-derived data improved identification of validated anaphylaxis events.

anaphylaxis; electronic health records; health outcome identification; machine learning, supervised; postmarketing product surveillance; predictive modeling

Abbreviations: ARIA, Active Risk Identification and Analysis System; AUC, area under the receiver operating characteristic curve; BART, Bayesian additive regression trees; CI, confidence intervals; EHR, electronic health record; FDA, Food and Drug Administration; ICD-10-CM, *International Classification of Diseases Tenth Revision, Clinical Modification*; ICD-9-CM, *International Classification of Diseases Ninth Revision, Clinical Modification*; KPNW, Kaiser Permanente Northwest; KPWA, Kaiser Permanente Washington; LASSO, least absolute shrinkage and selection operator; NLP, natural language processing; PPV, positive predictive value.

Anaphylaxis is a rare but severe allergic reaction with rapid onset, often caused by exposure to medications, food, or venom, and can be life-threatening (1). Lifetime US prevalence estimates range from 0.05% to 2% (2) and incidence is increasing (3–7). Anaphylaxis mortality rates are stable overall but are increasing for medication exposure (8), the most common cause of fatal anaphylaxis (1, 9, 10).

Models capable of accurately identifying anaphylaxis in real-world health data could contribute importantly to medical product safety surveillance. Since 2008, the Sentinel Initiative of the US Food and Drug Administration (FDA) has developed a powerful set of analytical tools and the world's largest multisite distributed database to monitor the safety of FDA-approved medical products using real-world data (11–13). A key component of Sentinel is the Active Risk Identi-

fication and Analysis (ARIA) System (12, 14, 15), designed to support semiautomated targeted surveillance of medical products without the need to manually review medical records. ARIA analyses have successfully informed many FDA regulatory decisions (16) but have been insufficient for monitoring health outcomes of interest with delayed or complex clinical presentation (17, 18) such as anaphylaxis (19).

To date, most attempts to accurately identify anaphylaxis events using automated algorithms applied to structured medical claims data have had limited success (20). This is in part due to the challenges of diagnosing a condition with diverse clinical presentation (19), reliance on structured medical claims data (21, 22), and the practice of "rule-out" coding (23). An algorithm published by Walsh et al. in 2013 (22) intended to improve identification of anaphylaxis in FDA safety surveillance studies using structured diagnosis, procedure, and encounter data. It achieved a positive predictive value (PPV) of 63.1% when evaluated on 122 potential anaphylaxis events across 8 health-care settings (22). This was an improvement over prior algorithms (24) but remained insufficient for use in FDA ARIA analyses, which generally require PPV $\geq$80%. In general, diagnosis codes appear to identify anaphylaxis events with high sensitivity but lack specificity (25). Misclassification of anaphylaxis events is a major barrier to disease surveillance efforts and can prevent clinicians from identifying actionable health risks.

To overcome these limitations models must better discriminate between actual and potential anaphylaxis events. Applying machine learning to rich, text-based electronic health record (EHR) data has been successful in other clinical domains (26–31), and Bi et al. (32) have argued in the *Journal* that such methods may improve accuracy in epidemiologic investigations. Text mining has been shown to improve ascertainment of other rare conditions such as rhabdomyolysis (33). Ball et al. (34) and Yu et al. (35) have found that information extracted from EHRs about exposures, comorbidities, presenting symptoms, treatments, and disease severity appearing in chart notes may be useful for identifying actual anaphylaxis. Recent studies have used natural language processing (NLP) and machine-learning methods to improve the identification of anaphylaxis events but did not include external populations to validate their findings (21, 36).

Our motivation and the motivation of the 2013 Walsh study (22) are the same: to improve automated algorithms for identifying anaphylaxis in postmarketing FDA safety studies. Our novel contributions are incorporating rich EHR data and using data-driven machine-learning methods for model development. We evaluate our models using V-fold cross-validated performance metrics, as well as assessing performance in external data. The latter helps assess generalizability to other settings within the Kaiser network. This work is part of a larger effort to enhance FDA Sentinel medical product safety surveillance capability (17).

METHODS

Setting and study sample

This predictive modeling activity used data from Kaiser Permanente Washington (KPWA); we used data from Kaiser Permanente Northwest (KPNW) for external validation. KPWA delivers care to about 700,000 people in Washington State and Idaho through integrated care ("HMO") and traditional insurance plans, with most hospital and emergency department care provided through contracts with non-Kaiser regional hospitals, thereby constituting a diverse source of inpatient data. KPNW delivers integrated care to about 600,000 people in northwest Oregon and southwest Washington State, including outpatient, emergency department, and inpatient care.

We planned to generate gold-standard outcome data for approximately 250 potential anaphylaxis events in each study site (500 total), the maximum number feasible within our study's resource constraints. In KPWA, we used data from health-care encounters to develop structured data covariates, NLP-derived covariates, and multiple models for identifying validated anaphylaxis events. To assess the reproducibility (external validation) of these models, we then applied them without modification to data from KPNW.

To identify potential anaphylaxis events during October 1, 2015, through March 31, 2019, for patients $\geq$1 year of age, we translated the *International Classification of Disease*, *Ninth Revision*, *Clinical Modification* (ICD-9-CM), code sets and logic published by Walsh et al. (22) (Web Appendix 1, available at https://doi.org/10.1093/aje/kwac182) to *International Classification of Disease*, *Tenth Revision*, *Clinical Modification* (ICD-10-CM), codes following the method described by Fung et al. (37), including a recommended clinician review and curation step (Web Appendix 2 and Web Tables 1 and 2). Potential anaphylaxis events included emergency department or inpatient encounters with anaphylaxis diagnoses ("path 1"), and outpatient encounters with anaphylaxis diagnoses plus, on the same day from any setting, diagnosis codes for bronchospasm, stridor, or hypotension, and/or procedure codes for cardiopulmonary resuscitation, epinephrine, or diphenhydramine injection (our "path 2"). Each patient contributed 1 potential anaphylaxis event: their earliest qualifying path-1 encounter or, if none existed, their earliest qualifying path-2 encounter. Our path-1 and path-2 criteria are thus comparable to Walsh's criteria A and B, respectively, facilitating a direct comparison of our results to those reported by Walsh in these 2 strata.

The KPWA sample (n = 239) included 161 path-1 encounters in 15 non-Kaiser, externally operated regional hospitals from whom we obtained medical records (59% of all qualifying path-1 encounters), and 78 path-2 encounters documented in the KPWA EHR (a 42% random sample). The KPNW sample included all 277 path-1 and -2 encounters documented in the KPNW EHR.

Gold standard creation

At both study sites, we conducted medical records reviews to determine whether potential events met National Institute of Allergy and Infectious Disease (NIAID) clinical criteria for anaphylaxis (gold standard) (19). At KPWA these determinations were made by 2 physician adjudicators (J.S.F., M.A.B.) using methods described previously (25). To improve efficiency, at KPNW 2 trained abstractors rendered

determinations following a protocol developed by KPWA and KPNW clinicians and implemented in RedCAP (available at https://github.com/kpwhri/Sentinel-Anaphylaxis). A KPNW clinician (A.H.F.) adjudicated any charts that abstractors considered ambiguous. KPNW abstractor determinations were concordant in 88% of a random subset of 16 charts independently reviewed. The same notes used for chart review were used to generate NLP-derived data.

Structured data and covariates

Guided by clinical domain expertise (J.S.F., M.A.B.), informatics knowledge (D.S.C., R.L.J.), and observed distributions in KPWA data—but without knowledge of gold-standard determinations—we manually engineered structured covariates we judged useful for distinguishing actual from potential anaphylaxis events in automated models. Constrained by our sample size, we operationalized a modest-sized set of 47 structured covariates. These included demographics, setting for the qualifying healthcare encounter, potential cause of anaphylaxis (food, medication, blood product or vaccine, or unspecified), recent immunotherapy or imaging with intravenous contrast, clinical interventions (e.g., epinephrine, vasopressors, cardiopulmonary resuscitation, intubation, tryptase labs, immunology follow-up), history of allergic reactions, and history of competing diagnoses, including angioedema, chronic respiratory disease, and serious infection (Web Appendix 3 and Web Table 3). We operationalized these covariates identically in KPWA and KPNW Sentinel Common Data Model data, standardized representations of encounters, diagnoses, procedures, and medications extracted from medical claims records and EHR equivalents (13, 17, 38).

NLP-derived data and covariates

Based on review and discussion of 50 KPWA charts, our clinical (J.S.F., M.A.B.) and informatics (D.S.C., B.L.H., D.J.C.) experts manually curated an initial custom dictionary of concepts they considered useful for identifying validated anaphylaxis events. Botsis and Ball (39) showed that automated approaches can similarly be used to identify relevant anaphylaxis concepts. We therefore augmented this initial dictionary with Unified Medical Language System (UMLS) (40) concepts appearing in $\geq$3 of 6 published anaphylaxis knowledge base articles (41–46) following the "concept collection" method described by Yu et al. (47). Our final dictionary (Web Appendix 4 and Web Table 4) was enriched with synonyms from the UMLS Metathesaurus (40, 48) synonyms (e.g., dyspnea synonyms "breathing difficulty" and "shortness of breath"), and synonyms discovered in KPWA charts by NLP-assisted manual review (49) (e.g., the misspelled dyspnea synonym "couldn't breath").

We then identified all instances of these dictionary terms in the KPWA corpus using a locally developed NLP system implementing an approach similar to that of Apache cTAKES (50, 51) and a KPWA-tailored version of the ConText algorithm (52) to distinguish affirmative mentions (i.e., those not negated, historical, hypothetical or about someone other than the patient). The KPWA corpus for path-1 encounters originated as paper print-outs of EHR notes from 15 regional non-Kaiser medical facilities, which we converted to electronic text using Tesseract optical character recognition (53) and a context-sensitive spelling correction model (54, 55) trained on KPWA chart notes. The KPWA path-2 corpus (as well as the entire KPNW corpus) was extracted from the KPWA (and KPNW) Epic Clarity EHR databases in electronic format.

Next, in consultation with clinical experts (J.S.F., M.A.B.), informaticists (D.S.C., B.L.H., D.J.C.) used KPWA's NLP-generated data to manually engineer a large set of candidate NLP-derived covariates. Illustrative examples of these covariates included:

- Rules indicating signs and symptoms from multiple organ systems (e.g., evidence of skin/mucosal tissue involvement and either respiratory compromise or reduced blood pressure)
- Mentions of individual symptom categories (e.g., skin/mucosal tissue involvement)
- Counts of affirmative anaphylaxis mentions normalized by chart length
- Mentions of cardiopulmonary resuscitation
- Mentions of multiple epinephrine injections

We implemented multiple versions of many covariates; some requiring only one mention of a concept while others required at least 2 mentions, a commonly used "counting rule" (56). We engineered 468 candidate NLP-derived covariates.

To mitigate risks of model overfitting, an informaticist (D.S.C.) used frequency distributions in the KPWA sample (without knowledge of gold-standard determinations) and expert judgment to select a reduced set of 100 NLP-derived covariates (Web Appendix 5 and Web Tables 5 and 6). Additionally, so that we could explore during model development whether clinical expertise alone was an effective approach to covariate selection, two clinicians (M.A.B., J.S.F.) selected an alternative set of 25 NLP-derived covariates (16 of which were not among the 100 informaticist-selected covariates; Web Appendix 5 and Web Table 6). Although some gold-standard data is commonly used during NLP covariate engineering (57–59), we conducted all NLP development without knowledge of gold-standard determinations, thereby reserving all gold-standard data for model development and validation.

Analytical data set

We used SAS, version 9.4 (SAS Institute, Inc., Cary, North Carolina), and the KPWA NLP system to generate the 43 structured and 116 NLP-derived covariates using the KPWA Sentinel Common Data Model data and corpus. We packaged and transported to KPNW via GitHub (60) this same SAS code and NLP system and used it to generate identical versions of the structured and NLP-derived covariates using KPNW Sentinel Common Data Model data and corpora. To avoid missing data issues, we defined all covariates—structured and NLP-derived—as counts (including zero) or binary indicators of whether evidence specified in the covariate's operational definition existed.

Machine-learning modeling methods and evaluation

We used R (version 3.6.3; The R Foundation, Indianapolis, Indiana) (61) to develop several models for each of 3 partially overlapping sets of covariates: 1) structured covariates only, 2) structured plus all 116 NLP-derived covariates, and 3) structured plus 25 clinician-selected NLP covariates. For each set of covariates, we evaluated 25 machine-learning algorithms. The first 24 were pairings between 3 covariate selection strategies and eight algorithms. The 3 covariate selection strategies were:

1. Least absolute shrinkage and selection operator (LASSO) (62), retaining covariates with nonzero coefficients in the model that minimized the cross-validated loss
2. Partitioning around medoids (63), using the silhouette width to identify the optimal number of clusters between 5 and 20, then retaining each cluster medoid
3. Retaining all variables (Retain-All)

The 8 modeling algorithms were:

1. Logistic regression
2. Elastic net, area under the receiver operating characteristic curve (AUC) loss (Elastic Net) (62)
3. Gradient-boosting machine with a maximum tree depth of 4, AUC loss (gradient-boosting machine 1) (64)
4. Gradient-boosting machine with a maximum tree depth of 2, AUC loss (gradient-boosting machine 2) (64)
5. Bayesian additive regression trees with regularization parameter $k = 1$ (BART1) (65, 66)
6. Bayesian additive regression trees with regularization parameter $k = 2$ (BART2) (65, 66)
7. Neural network with 1 node in one hidden layer (neural network 1) (67, 68)
8. Neural network with 2 nodes in one hidden layer (neural network 2) (67, 68)

Our twenty-fifth algorithm was Super Learner (SL), an ensemble method that calculated an optimal weighted combination of predictions from the other 24 models (69, 70).

Prior to modeling, we excluded from consideration 27 (4 of 47 structured and 23 of 116 NLP) covariates that had low variation, defined as $\geq$99% of observations having the same value and correlation with the outcome $<$0.05 (Web Appendix 6).

We evaluated cross-validated performance of the KPWA-developed models in KPWA data (15-fold cross-validated) and, separately, in KPNW data. Our global performance metric was cross-validated area under the receiver operating characteristic curve (cross-validated AUC) weighted to account for path-specific sampling probabilities. For the best performing models, we also calculated cross-validated sensitivity, specificity, positive predictive value (PPV), negative predictive value, F1 score (the harmonic mean of PPV and sensitivity), and F0.5 score (the harmonic mean of PPV and 0.5 $\times$ sensitivity) at cutpoints between the 10th and 95th quantiles of predicted risk from each model. To identify the important predictors in the model we ranked them by variable importance defined as the marginal mean difference in predicted probability associated with a 1-unit change in a binary covariate or a 1-standard-deviation change in a nonbinary covariate.

Our primary analysis compared the best performing model using only structured data covariates to the best performing model using structured data and all NLP-derived covariates. In secondary analyses, we evaluated the performance of: 1) the best performing model using all structured data covariates and the clinician-selected set of 25 NLP-derived covariates, and 2) a main-terms logistic regression model using only structured data covariates, representing a conventional model approach.

This Sentinel activity is a public health surveillance activity conducted under the authority of the FDA and, accordingly, was not subject to institutional review board oversight (71–73).

RESULTS

Cross-validated results in KPWA data

Table 1 summarizes patients included in the study. Out of 239 potential anaphylaxis events at KPWA, 154 (64%) met clinical criteria for validated anaphylaxis; 180 of 277 (65%) potential events at KPNW met validation criteria. Sampling fractions and corresponding weights for rescaling results to all study-eligible encounters are in Web Appendix 7 and Web Table 7. After excluding low-variation covariates, 43 structured and 93 NLP-derived covariates remained. Depending on the covariate selection method (LASSO, partitioning around medoids, or Retain-All) and data set (structured or structured and all NLP), 7–43 structured covariates and 13–93 NLP-derived covariates were included in models (Web Appendix 8 and Web Tables 8 and 9).

There were no statistically significant standardized mean differences between covariates from the development (KPWA) and validation (KPNW) sites (Web Appendix 9 and Web Table 10). Associations between each candidate predictor and the outcome are reported in Web Appendix 10 and Web Table 11.

Weighted cross-validated AUCs and 95% confidence intervals (CIs) for KPWA models considering only structured data, or structured data and all NLP data, are summarized in Table 2. Overall model performance varied considerably by algorithm, variable selection method, and data set, but the addition of NLP-derived covariates clearly improved performance. The best performing structured data model, a neural net with 1 node in one hidden layer using LASSO variable selection, yielded a weighted cross-validated AUC of 0.62 (95% CI: 0.58, 0.66; Table 2, A). When structured data were augmented with all NLP-derived data, the BART2 model with LASSO variable selection performed best, yielding a cross-validated AUC of 0.71 (95% CI: 0.66, 0.76; Table 2). The same BART2 model with Retain-All variable selection performed similarly (cross-validated AUC = 0.70, 95% CI: 0.66, 0.75). The Super Learner did not improve performance further. Variable importance rankings for selected models are in Web Appendix

Table 1. Characteristics of Patients and Potential Anaphylaxis Events[a] in the Study Samples for the Algorithm Development Site (Washington) and the External Validation Site (Pacific Northwest), Kaiser Permanente, 2015–2019

Characteristic	KPWA (*n* = 239[b])		KPNW (*n* = 277)	
	No.	%	No.	%
Sex				
Female	139	58.2	171	61.7
Male	100	41.8	106	38.3
Age group, years				
1–19	64	26.8	52	18.8
20–65	136	56.9	184	66.4
≥66	39	16.3	41	14.8
Race				
Asian	35	14.6	24	8.7
Black or African American	19	8.0	12	4.3
Other/multiple races	9	3.8	4	1.4
White	155	64.9	212	76.5
Unknown	21	8.8	25	9.0
Ethnicity				
Hispanic	13	5.4	24	8.7
Eligibility path				
Path 1: ED/inpatient anaphylaxis diagnosis	161	67.4	163	58.8
Path 2: outpatient anaphylaxis diagnosis	78	32.6	114	41.2
Validated anaphylaxis				
Yes	154	64.4[c]	180	65.0[d]
No	85	35.6	97	35.0

Abbreviations: ED, emergency department; KPNW, Kaiser Permanente Northwest; KPWA, Kaiser Permanente Washington.

[a] Each patient contributed 1 potential anaphylaxis event to the study sample.

[b] Structured data were available for 239 KPWA patients, and natural language processing data were available for 236 patients; natural language processing data was not generated for 3 KPWA patients because their clinical notes were not available.

[c] Percentage "Yes" = positive predictive value; the 95% confidence interval is 58%–70%.

[d] Percentage "Yes" = positive predictive value; the 95% confidence interval is 59%–71%.

11 and Web Table 12. For the model considering structured and all NLP covariates, the 5 most important structured covariates (all negatively associated with actual anaphylaxis) were: 1) number of prior years with allergic reaction diagnoses, 2) allergic reaction diagnosis in the prior year, 3) same-day exposure to any imaging procedure, and discharge prescriptions for 4) antihistamines or 5) corticosteroids (Web Appendix 11, Web Table 12, and Web Appendix 3). The 5 most important NLP-derived covariates (all positively associated with actual anaphylaxis) were: 1) ≥2 affirmative mentions of hypotension; 2) any description of respiratory compromise and reduced blood pressure in close proximity to mentions of either sudden onset, epinephrine administration, anaphylaxis diagnosis, or admission for observation; 3) ≥2 affirmative mentions of skin/mucosal involvement and either respiratory compromise or reduced blood pressure in close proximity to mentions of anaphylaxis diagnosis; 4) ≥2 affirmative mentions of wheezing; and 5) descriptions of skin/mucosal involvement and reduced blood pressure in close proximity to mentions of either sudden onset, epinephrine administration, anaphylaxis diagnosis, or admission for observation (Web Appendix 11, Web Table 12, and Web Appendix 5).

Plots of cross-validated AUC for selected KPWA models are in Web Appendix 12. The best performing model incorporating the 25 clinician-selected NLP covariates (Web Appendix 5) achieved a cross-validated AUC of 0.67 (95% CI: 0.63, 0.71; additional details are in Web Appendix 12 and Web Table 13), roughly half-way between the best structured data model (cross-validated AUC = 0.62, 95% CI: 0.58, 0.66) and the best model incorporating all NLP-derived covariates (cross-validated AUC = 0.71, 95% CI: 0.66, 0.76; additional details are in Web Appendix 12 and Web Table 13).

Table 2. Cross-Validated Weighted[a] Area Under the Receiver Operating Characteristic Curve for Models Predicting Anaphylaxis Case Status Based on Structured Data and Structured and Natural Language Processing Data, Kaiser Permanente Washington, 2015–2019

Data Set and Algorithm	Covariate Selection Strategy					
	LASSO		PAM		Retain-All	
	AUC	95% CI	AUC	95% CI	AUC	95% CI
Structured data						
Logistic regression	0.58[b]	0.541, 0.627	0.58[b]	0.541, 0.627	0.56	0.522, 0.606
Elastic Net	0.59	0.544, 0.630	0.57	0.531, 0.615	0.61	0.562, 0.650
GBM 1	0.58	0.533, 0.623	0.57	0.528, 0.618	0.58	0.538, 0.624
GBM 2	0.56	0.513, 0.601	0.57	0.526, 0.614	0.60	0.557, 0.645
BART 1	0.59	0.544, 0.628	0.56	0.518, 0.602	0.59	0.552, 0.636
BART 2	0.59	0.552, 0.636	0.57	0.532, 0.616	0.59	0.551, 0.635
NNET 1	0.62[b]	0.578, 0.660	0.58	0.537, 0.627	0.58	0.533, 0.617
NNET 2	0.56	0.516, 0.602	0.53	0.489, 0.573	0.57	0.525, 0.609
Super Learner[c]			0.58[b] (95% CI: 0.537, 0.625)			
Structured and NLP data						
Logistic regression	0.64	0.598, 0.690	0.66[b]	0.615, 0.705	0.49	0.436, 0.536
Elastic Net	0.66	0.618, 0.710	0.65	0.605, 0.695	0.65	0.604, 0.694
GBM 1	0.60	0.559, 0.649	0.61	0.566, 0.654	0.68	0.630, 0.724
GBM 2	0.60	0.559, 0.649	0.62	0.577, 0.665	0.67	0.626, 0.718
BART 1	0.70	0.653, 0.747	0.66	0.611, 0.699	0.69	0.641, 0.731
BART 2	0.71[b]	0.663, 0.757	0.65	0.607, 0.697	0.70[c]	0.658, 0.750
NNET 1	0.57	0.527, 0.617	0.62	0.572, 0.662	0.58	0.537, 0.621
NNET 2	0.63	0.589, 0.677	0.65	0.608, 0.698	0.66	0.609, 0.701
Super Learner[c]			0.69[b] (95% CI: 0.642, 0.734)			

Abbreviations: AUC, area under the receiver operating characteristic curve; BART, Bayesian additive regression trees; CI, confidence interval; GBM, gradient-boosting machine; LASSO, least absolute shrinkage and selection operator; NLP, natural language processing; NNET, neural network; PAM, partitioning around medoids.

[a] Based on weighting of observations to account for biased sampling in Paths 1 and 2.

[b] These results are featured in the Results and Discussion section of this paper.

[c] Weighted combination of predictions from the other 24 models using this feature set.

Figure 1 and Table 2 summarize incremental improvement in performance attributable to machine-learning methods, augmenting structured data with NLP-derived EHR data, or both. The main-terms logistic regression approach applied to all 43 structured data covariates (Table 2, Retain-All), which we defined as a traditional "baseline" modeling approach, achieved a cross-validated AUC of 0.56 (95% CI: 0.52, 0.61). Logistic regression with both of the other covariate selection strategies offered improved cross-validated AUC (0.58, 95% CI: 0.54, 0.63). Applying more sophisticated machine-learning methods to the same structured data improved cross-validated AUC to 0.62 (95% CI: 0.58, 0.66; Table 2, algorithm neural network 1). Applying traditional regression methods to structured and all NLP-derived data improved cross-validated AUC to 0.66 (95% CI: 0.62, 0.71; Table 2). Incorporating both machine learning and NLP-derived data in algorithm development improved cross-validated AUC to 0.71 (95% CI: 0.66, 0.76; Table 2).

External validation in KPNW data

We evaluated cross-validated AUC for each of the 75 KPWA-developed models in external data from KPNW (Web Appendix 13 and Web Table 14). Results confirmed that models using structured plus all NLP data were best at identifying cases. However, the BART2-LASSO model (for KPWA, cross-validated AUC = 0.71, 95% CI: 0.66, 0.76; for KPNW, cross-validated AUC = 0.63, 95% CI: 0.59, 0.66) generalized less well to KPNW than the BART2-Retain-All model (for KPWA, cross-validated AUC = 0.70, 95% CI: 0.66, 0.77; for KPNW, cross-validated AUC = 0.67, 95% CI: 0.63, 0.71). As shown in Figure 2, modest degradation in performance in KPNW data was evident toward the middle of the curve; model performance was comparable at both sites toward the tails of the curve.

Figure 3 plots PPV and sensitivity for the KPWA BART2 Retain-All model at cutpoints between the 10th and 95th percentiles of predicted risk in the KPWA and KPNW

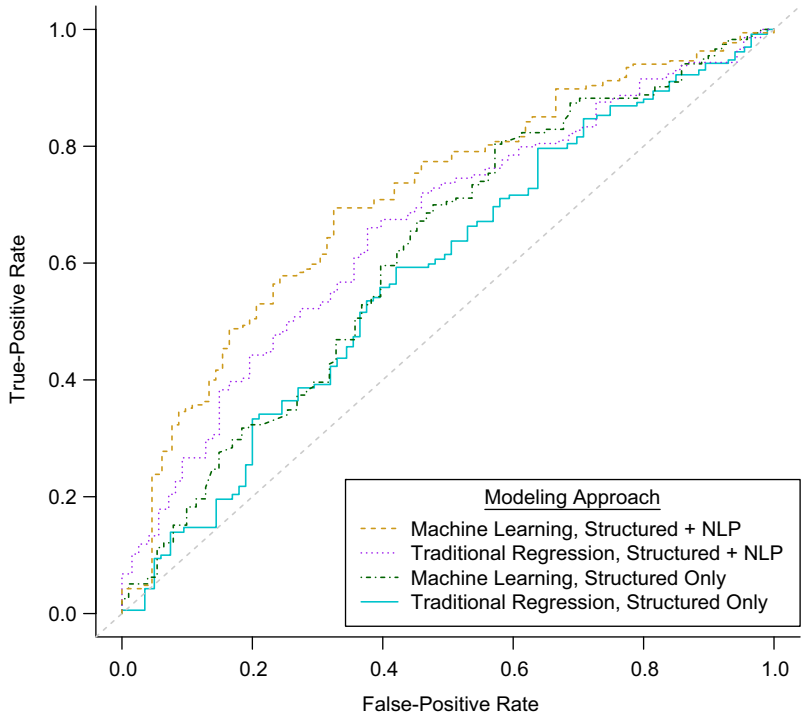

Figure 1. Weighted cross-validated area under the receiver operating characteristic curve for Kaiser Permanente Washington algorithms identifying actual anaphylaxis events in Kaiser Permanente Washington data (2015–2019) using the best machine-learning approach applied to structured and all natural language processing (NLP) data, traditional logistic regression approach applied to structured and all NLP data, machine-learning approach applied to structured data only, and traditional logistic regression approach applied to structured data only.

data sets. Table 3 presents the same information along with specificity, negative predictive value, F1 score, and F0.5 score. PPVs in both data sets generally increased gradually and in tandem from lows of 67% to highs of 86% (KPWA) or 98% (KPNW). Between the 35th and 50th percentiles of estimated risk, PPVs in both sites are nearly identical and increase gradually from 75% to near 80%. Sensitivity, in contrast, decreases rapidly in both sites from highs over 90% to lows below 10%, with KPNW's sensitivity 8%–10% lower than KPWA's between the 35th and 50th percentiles of estimated risk (Figure 3 and Table 3). The ultimate use of the classifications should guide the selection of a cutpoint to achieve desired tradeoffs in sensitivity, specificity, PPV, and negative predictive value. Web Appendix 14 and Web Table 15 present detailed performance metrics for selected models.

DISCUSSION

In this study, which included an external validation population, we found that algorithms to identify anaphylaxis from EHR data can benefit substantially from NLP-derived data and machine-learning methods. Web Appendix 15 contains code and instructions for obtaining predictions the BART-2 Retain-All model. Incorporating NLP data improved both machine-learned algorithms (cross-validated AUC of 0.71 vs. 0.62) and traditional regression algorithms (cross-validated AUV of 0.66 vs. 0.58; Figure 1). Similarly, machine-learned algorithms outperformed the expert-curated Walsh algorithm (22), as well as traditional regression algorithms, when applied to the same data (with cross-validated AUCs of 0.71 vs. 0.66 in structured and NLP-derived data and 0.62 vs. 0.58 in structured data only; Figure 1).

Comparing our results to those of Walsh et al., the proportion of potential anaphylaxis events validated by manual review (i.e., PPV; see Table 1) at KPWA (64.4%, 95% CI: 58, 70) and KPNW (65.0%, 95% CI: 59, 71) are comparable to the proportion validated in corresponding portion of the Walsh sample (67.3%, 95% CI: 57.4, 76.2; criteria A and B only) (22, p. 1209). Walsh speculated that removing from their algorithm codes with observed PPVs $\leq$70% might improve its PPV, and doing so yielded an empirical PPV of 75.0% with empirical sensitivity of 66%; however, without external/cross-validation these empirical performance metrics may be overly optimistic (74). In contrast, our cross-validated model yields cross-validated PPV of 79% when cross-validated sensitivity is set at 66%, and cross-validated sensitivity of 74% when cross-validated PPV is set at 75%.

This work also underscores the challenges of accurately identifying anaphylaxis through automated algorithms. In KPWA data, a cutpoint corresponding to approximately 80% (78.7%) PPV had sensitivity of 65.8%; the same cutpoint in external (KPNW) data yielded 78.1% PPV but sensitivity was attenuated (55.6%; Table 3). Sensitivity associated with a desired level of PPV is an important consideration

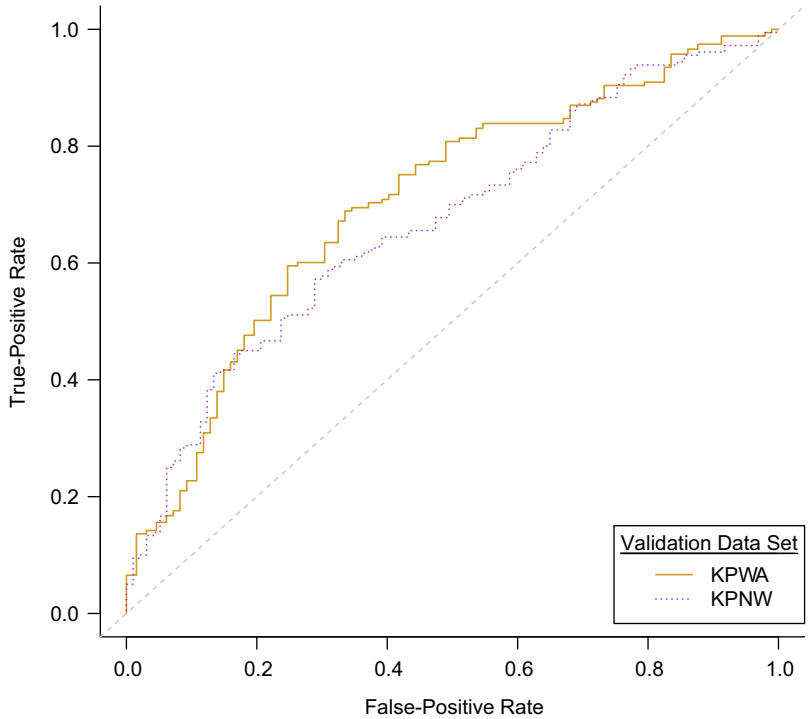

Figure 2. Cross-validated area under the receiver operating characteristic curve for the most generalizable high-performing machine-learning model developed at Kaiser Permanente Washington (KPWA, 2015–2019) using structured and natural language processing–derived data (Bayesian additive regression trees (BART) model 2 with Retain-All variable selection) evaluated in KPWA data and, separately, in Kaiser Permanente Northwest (KPNW) data (2015–2019).

when designing studies that involve rare outcomes such as anaphylaxis.

This study had several strengths, including the validation of anaphylaxis events using established clinical criteria and rigorous methods, and reporting of reliable cross-validated performance metrics in the source population (75). Comparing the AUC of the BART2-Retain-All model on the external validation data (0.67) with the cross-validated AUC on the original data (0.70) rather than the empirical AUC (0.89) provided a valid demonstration of the model's ability to generalize well within the Kaiser network.

Limitations of this study should also be noted. Although our model development site included data from a diverse collection of inpatient facilities, we externally validated our models in a single, integrated care setting, Kaiser Permanente Northwest, where access to records facilitating NLP was assured. To the extent that there are substantial differences in documentation of anaphylaxis in other healthcare systems, the generalizability of our findings to those other settings remains unknown. Our decision to manually curate NLP covariates limits the algorithm's scalability. Using paper charts converted by optical character recognition introduced errors that may have had a negative impact on NLP performance. Another important limitation of this study was the modest sample size, a common challenge in drug safety surveillance when events of interest are rare. NLP and machine learning are "data hungry" methods that generally perform best in big data sets. Large amounts of data give both technologies opportunities to discover signals among rich sets of covariates and model their potentially complex relationships with the outcome of interest. The relatively small sample size may have constrained algorithm performance by limiting the number of covariates that could be considered. Applying these same methods in larger samples may yield better results, but increasing gold-standard samples increases costs. In this study, minority class size (i.e., the 85 KPWA nonevents) constrained model development even more than the overall sample size. The approximately 2:1 ratio of actual anaphylaxis events to nonevents (Table 1) means that 3 additional, costly chart reviews are required to add a single (minority class) nonevent to the sample. Our modest sample size also influenced our decision not to use gold-standard training data during NLP covariate engineering, which may have had a negative impact on algorithm performance.

Other factors that may limit algorithm performance—for anaphylaxis and other health outcomes of interest—are ambiguity in clinical presentation and incompleteness of EHR documentation. In prior work applying NLP to anaphylaxis cases in Sentinel, reasons for misclassification included errors in identifying timing, severity, or presence of competing diagnoses (e.g., angioedema, asthma, chronic obstructive pulmonary disease, serious infection), and the presence of language consistent with anaphylaxis but also

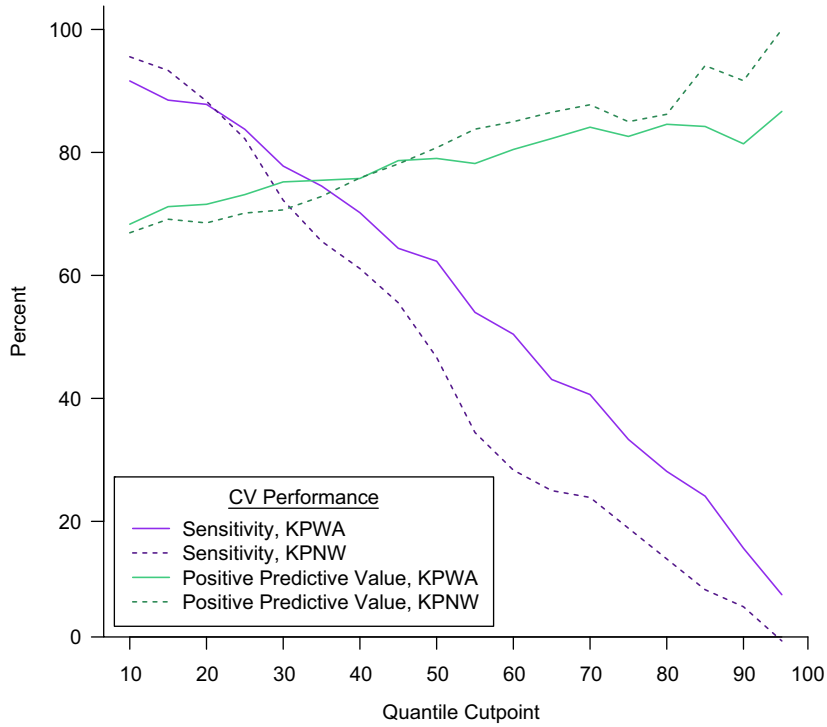

Figure 3. Model performance for classifying actual anaphylaxis events in Kaiser Permanente Washington (KPWA) data and Kaiser Permanente Northwest (KPNW) data, 2015–2019 . The model, developed in KPWA data, was a Bayesian additive regression trees (BART) model 2 retaining all covariates. Plotted values are cross-validated (CV) positive predictive value and sensitivity for increments of the predicted risk between the 10th and 95th percentiles.

other conditions (19). The rapid onset and responsiveness of anaphylaxis to epinephrine treatment, often administered in the field, further limit clinicians' opportunities to directly observe and document symptoms and progression critical to accurate diagnosis. Such diagnostic challenges were evident during our gold-standard creation, where 20% of KPWA charts were assigned discordant case status or were subjectively judged "difficult to adjudicate" by one or both independent physician reviewers. This raises questions regarding the potential upper limits of algorithm performance and may account for the modest algorithm performance reported by most others (20, 34, 35).

Several lessons from this study are relevant to broader efforts to improve surveillance methods for rare outcomes using EHR data and machine learning. First, this work illustrated the strengths of applying data-adaptive machine learning to a rich, high-dimensional set of covariates. A traditional parametric modeling approach incorporating a limited number of preselected covariates should no longer be the sole method under consideration. Second, future work should consider using automated NLP development approaches (47, 76–78). Such approaches may reduce burdens and operator-dependencies of manual NLP engineering and be replicated more easily in multiple settings thereby addressing site-specific textual heterogeneity and facilitating larger sample sizes by combining NLP data from multiple sites (79). Third, whenever possible, use of paper charts should be avoided; converting paper charts via optical character recognition is burdensome and introduces error in a corpus.

As the quantity of EHR data available for research and public health surveillance expands, other more sophisticated but data-hungry machine-learning methods may be useful for identifying rare, acute health outcomes (such as anaphylaxis) without burdensome manual feature engineering. Long/short-term memory (LSTM) neural networks and bidirectional encoder representations from transformers (BERT) have been shown to improve classification models when applied to data sets containing tens of thousands of the health outcome of interest (80, 81).

In planned future work using our existing data, we will we explore whether modeling approaches based on distant supervision (77, 78) may be useful for shifting some of the burden of generating strong NLP-derived predictors can be shifted from humans to machines. We will also combine KPWA and KPNW data in an attempt to improve sensitivity and PPV, develop second-stage models designed to reduce misclassifications, and investigate whether reducing hypothesized heterogeneity in the covariate-outcome associations by restricting to an adults-only study population can improve performance. As we reported elsewhere, the distribution of causes of anaphylaxis are different in children and adults (25), and others have reported age-related differences in anaphylaxis symptoms and comorbidities (82).

Table 3. Cross-Validated Performance Metrics at Cutpoints of Predicted Risk from the Bayesian Additive Regression Trees, Retain-All Model Developed Using Kaiser Permanente Washington Structured and All Natural Language Processing Data. Kaiser Permanente, 2015–2019

Cutpoint of Predicted Risk, %	KPWA Data						KPNW Data					
	Sensitivity	Specificity	PPV	NPV	F1	F0.5	Sensitivity	Specificity	PPV	NPV	F1	F0.5
10	92.9	17.5	67.2	57.7	78.0	67.3	95.6	12.4	66.9	60.0	78.7	67.0
15	89.5	26.8	69.0	58.4	77.9	69.0	93.3	22.7	69.1	64.7	79.4	69.2
20	84.5	30.9	69.0	52.2	76.0	69.0	88.3	24.7	68.5	53.3	77.2	68.6
25	83.9	43.8	73.1	59.9	78.1	73.1	82.2	35.1	70.1	51.5	75.7	70.2
30	78.0	48.9	73.5	55.0	75.7	73.6	72.2	44.3	70.7	46.2	71.4	70.7
35	76.3	59.8	77.5	58.0	76.9	77.5	65.6	54.6	72.8	46.1	69.0	72.8
40	70.9	63.4	77.9	54.5	74.2	77.9	61.1	63.9	75.9	47.0	67.7	75.8
45	65.8	67.5	78.7	52.0	71.7	78.6	55.6	71.1	78.1	46.3	64.9	78.1
50	60.7	70.6	79.0	49.7	68.6	78.9	46.7	79.4	80.8	44.5	59.2	80.6
55	56.1	76.8	81.5	49.0	66.5	81.4	34.4	87.6	83.8	41.9	48.8	83.5
60	49.6	78.4	80.7	46.1	61.4	80.5	28.3	90.7	85.0	40.6	42.5	84.6
65	44.5	83.0	82.7	45.1	57.9	82.5	25.0	92.8	86.5	40.0	38.8	86.0
70	38.0	86.1	83.3	43.3	52.2	83.0	23.9	93.8	87.8	39.9	37.6	87.2
75	30.7	87.1	81.3	40.8	44.5	80.9	18.9	93.8	85.0	38.4	30.9	84.3
80	25.3	90.8	83.3	40.0	38.8	82.8	13.9	95.9	86.2	37.5	23.9	85.1
85	18.2	90.8	78.1	37.9	29.5	77.5	8.9	99.0	94.1	36.9	16.2	91.9
90	13.6	96.9	89.0	38.1	23.6	87.8	6.1	99.0	91.7	36.2	11.5	88.6
95	7.1	100.0	100.0	37.2	13.3	96.8	0.6	100.0	100.0	35.1	1.1	69.1

Abbreviations: F1, the harmonic mean of PPV and sensitivity; F0.5, the harmonic mean of PPV and 0.5 × sensitivity; KPNW, Kaiser Permanente Northwest; KPWA, Kaiser Permanente Washington; NPV, negative predictive value; PPV, positive predictive value.

CONCLUSIONS

Identification of anaphylaxis, a rare and clinically complex health outcome of interest to clinicians, epidemiologists, and postmarketing medical product safety surveillance, is improved by the addition of NLP-derived EHR data and machine-learning approaches. Future anaphylaxis modeling work should attempt to assemble larger study samples, exclusively use native EHR text, and automate some or all aspects of NLP development to enhance scalability and reduce development costs.

ACKNOWLEDGMENTS

Author affiliations: Kaiser Permanente Washington Health Research Institute, Seattle, Washington, United States (David S. Carrell, Kara L. Cushing-Haugen, Ron L. Johnson, Vina Graham, David J. Cronkite, Jennifer C. Nelson); Department of Biomedical Informatics and Medical Education, UW Medicine, University of Washington, Seattle, Washington, United States (David S. Carrell); Putnam Data Sciences, Boston, Massachusetts, United States (Susan Gruber); Department of Medicine, School of Medicine, University of Washington, Seattle, Washington, United States (James S. Floyd, Maralyssa A. Bann); Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, United States (James S. Floyd); Center for Health Research, Kaiser Permanente Northwest, Portland, Oregon, United States (Brian L. Hazlehurst, Andrew H. Felcher); Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee, United States (Cosmin A. Bejan); Harvard Pilgrim Health Care Institute, Boston, Massachusetts, United States (Adee Kennedy, Mayura U. Shinde); and Food and Drug Administration, Silver Spring, Maryland, United States (Sara Karami, Yong Ma, Danijela Stojanovic, Yueqin Zhao, Robert Ball).

D.S.C. and S.G. contributed equally to this work.

This work was supported by the Food and Drug Administration (contract HHSF223201400030I).

The data sets used in this study are not available for use outside the 2 health-care institutions from which all study data were derived. Programming resources and documentation are available on GitHub at: https://github.com/kpwhri/Sentinel-Anaphylaxis.

The views expressed in this paper represent those of the authors and do not necessarily represent the official views of the US Food and Drug Administration.

R.B. is an author on US Patent 9,075,796, "Text mining for large medical text datasets and corresponding medical text classification using informative feature selection." At

present this patent is not licensed and does not generate royalties. The other authors report no conflicts.

REFERENCES

1. Yu JE, Lin RY. The epidemiology of anaphylaxis. *Clin Rev Allergy Immunol.* 2018;54(3):366–374.
2. Lieberman P, Camargo CA Jr, Bohlke K, et al. Epidemiology of anaphylaxis: findings of the American College of Allergy, Asthma and Immunology Epidemiology of Anaphylaxis Working Group. *Ann Allergy Asthma Immunol.* 2006;97(5): 596–602.
3. Rudders SA, Arias SA, Camargo CA Jr. Trends in hospitalizations for food-induced anaphylaxis in US children, 2000–2009. *J Allergy Clin Immunol.* 2014;134(4):960–2 e3.
4. Shrestha P, Dhital R, Poudel D, et al. Trends in hospitalizations related to anaphylaxis, angioedema, and urticaria in the United States. *Ann Allergy Asthma Immunol.* 2019;122(4):401–406.e2.
5. Mulla ZD, Lin RY, Simon MR. Perspectives on anaphylaxis epidemiology in the United States with new data and analyses. *Curr Allergy Asthma Rep.* 2011;11(1):37–44.
6. Lin RY, Anderson AS, Shah SN, et al. Increasing anaphylaxis hospitalizations in the first 2 decades of life: New York state, 1990–2006. *Ann Allergy Asthma Immunol.* 2008;101(4): 387–393.
7. Decker WW, Campbell RL, Manivannan V, et al. The etiology and incidence of anaphylaxis in Rochester, Minnesota: a report from the Rochester Epidemiology Project. *J Allergy Clin Immunol.* 2008;122(6):1161–1165.
8. Turner PJ, Jerschow E, Umasunthar T, et al. Fatal anaphylaxis: mortality rate and risk factors. *J Allergy Clin Immunol Pract.* 2017;5(5):1169–1178.
9. Takazawa T, Oshima K, Saito S. Drug-induced anaphylaxis in the emergency room. *Acute Med Surg.* 2017;4(3):235–245.
10. Lee JK, Vadas P. Anaphylaxis: mechanisms and management. *Clin Exp Allergy.* 2011;41(7):923–938.
11. Platt R, Wilson M, Chan KA, et al. The new Sentinel Network—improving the evidence of medical-product safety. *N Engl J Med.* 2009;361(7):645–647.
12. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf.* 2012;21(suppl 1): 1–8.
13. Platt R, Brown JS, Robb M, et al. The FDA Sentinel Initiative—an evolving national resource. *N Engl J Med.* 2018;379(22):2091–2093.
14. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf.* 2012; 21(suppl 1):23–31.
15. Connolly JG, Wang SV, Fuller CC, et al. Development and application of two semi-automated tools for targeted medical product surveillance in a distributed data network. *Curr Epidemiol Rep.* 2017;4(4):298–306.
16. Lanes S, Brown JS, Haynes K, et al. Identifying health outcomes in healthcare databases. *Pharmacoepidemiol Drug Saf.* 2015;24(10):1009–1016.
17. Brown JS, Maro JC, Nguyen M, et al. Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the Food and Drug Administration's Sentinel system. *J Am Med Inform Assoc.* 2020;27(5):793–797.
18. Duke-Margolis Center for Health Policy. *Discussion Guide: Improving the Efficiency of Outcome Validation in the Sentinel System*. Washington DC: Duke-Margolis Center for Health Policy; 2018. https://healthpolicy.duke.edu/sites/default/files/2020-03/discussion_guide.pdf. Accessed July 20, 2021.
19. Sampson HA, Munoz-Furlong A, Campbell RL, et al. Second symposium on the definition and management of anaphylaxis: summary report—second National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network symposium. *J Allergy Clin Immunol.* 2006;117(2):391–397.
20. Schneider G, Kachroo S, Jones N, et al. A systematic review of validated methods for identifying anaphylaxis, including anaphylactic shock and angioneurotic edema, using administrative and claims data. *Pharmacoepidemiol Drug Saf.* 2012;21(suppl 1):240–247.
21. Beachler DC, Taylor DH, Anthony MS, et al. Development and validation of a predictive model algorithm to identify anaphylaxis in adults with type 2 diabetes in U.S. administrative claims data. *Pharmacoepidemiol Drug Saf.* 2021;30(7):918–926.
22. Walsh KE, Cutrona SL, Foy S, et al. Validation of anaphylaxis in the Food and Drug Administration's Mini-Sentinel. *Pharmacoepidemiol Drug Saf.* 2013;22(11):1205–1213.
23. Tuttle KL, Wickner P. Capturing anaphylaxis through medical records: are ICD and CPT codes sufficient? *Ann Allergy Asthma Immunol.* 2020;124(2):150–155.
24. Bohlke K, Davis RL, DeStefano F, et al. Epidemiology of anaphylaxis among children and adolescents enrolled in a health maintenance organization. *J Allergy Clin Immunol.* 2004;113(3):536–542.
25. Bann MA, Carrell DS, Gruber S, et al. Identification and validation of anaphylaxis using electronic health data in a population-based setting. *Epidemiology.* 2021;32(3): 439–443.
26. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform.* 2017;73:14–29.
27. Yim WW, Yetisgen M, Harris WP, et al. Natural language processing in oncology: a review. *JAMA Oncol.* 2016;2(6): 797–804.
28. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol.* 2020;145(2):463–469.
29. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA.* 2018;320(11): 1107–1108.
30. Naylor CD. On the prospects for a (deep) learning health care system. *JAMA.* 2018;320(11):1099–1100.
31. Hinton G. Deep learning-a technology with the potential to transform health care. *JAMA.* 2018;320(11):1101–1102.
32. Bi Q, Goodman KE, Kaminsky J, et al. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol.* 2019;188(12):2222–2239.
33. Floyd JS, Heckbert SR, Weiss NS, et al. Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. *JAMA.* 2012;307(15):1580–1582.
34. Ball R, Toh S, Nolan J, et al. Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System. *Pharmacoepidemiol Drug Saf.* 2018;27(10): 1077–1084.
35. Yu W, Zheng C, Xie F, et al. The use of natural language processing to identify vaccine-related anaphylaxis at five

health care systems in the Vaccine Safety Datalink. *Pharmacoepidemiol Drug Saf.* 2020;29(2):182–188.
36. Segura-Bedmar I, Colon-Ruiz C, Tejedor-Alonso MA, et al. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *J Biomed Inform.* 2018;87:50–59.
37. Fung KW, Richesson R, Smerek M, et al. Preparing for the ICD-10-CM transition: automated methods for translating ICD codes in clinical phenotype definitions. *EGEMS (Wash DC).* 2016;4(1):1211.
38. United States Food and Drug Administration. Sentinel distributed database and common data model. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model. Accessed July 21, 2017.
39. Botsis T, Ball R. Automating case definitions using literature-based reasoning. *Appl Clin Inform.* 2013;4(4):515–527.
40. U.S. National Library of Medicine. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/. Accessed November 1, 2017.
41. Wikipedia. Anaphylaxis. https://en.wikipedia.org/wiki/Anaphylaxis. Accessed August 5, 2021.
42. Mustafa SS. Anaphylaxis. https://emedicine.medscape.com/article/135065-print. Accessed September 25, 2021.
43. Delves PJ. Anaphylaxis. https://www.merckmanuals.com/professional/immunology-allergic-disorders/allergic,-autoimmune,-and-other-hypersensitivity-disorders/anaphylaxis. Accessed September 25, 2021.
44. Mayo Clinic. Anaphylaxis. https://www.mayoclinic.org/diseases-conditions/anaphylaxis/symptoms-causes/syc-20351468?p=1. Accessed September 25, 2021.
45. National Library of Medicine. Anaphylaxis. https://medlineplus.gov/ency/article/000844.htm. Accessed September 25, 2021.
46. Campbell RA, Kelso JM. Anaphylaxis: acute diagnosis. https://www.uptodate.com/contents/anaphylaxis-acute-diagnosis?search=anaphylaxis&source=search_result&selectedTitle=5~150&usage_type=default&display_rank=5. Accessed September 25, 2021.
47. Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc.* 2015;22(5):993–1000.
48. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(database issue):D267–D270.
49. Palmer RE, Carrell DS, Cronkite D, et al. The prevalence of problem opioid use in patients receiving chronic opioid therapy: computer-assisted review of electronic health record clinical notes. *Pain.* 2015;156(7):1208–1214.
50. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc.* 2010;17(5):507–513.
51. The Apache Software Foundation. Apache cTAKES Examples. https://ctakes.apache.org/examples.html. Accessed June 1, 2022.
52. Harkema H, Dowling JN, Thornblade T, et al. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform.* 2009;42(5):839–851.
53. Source GO. Tesseract OCR: An optical character recognition (OCR) engine. https://opensource.google/projects/tesseract. Accessed October 5, 2021.
54. Kernighan MD, Church KW, Gale WA. A spelling correction program based on a noisy channel model.*COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*. 1990:205–210.
55. Mays E, Damerau FJ, Mercer RL. Context based spelling correction. *Inf Process Manag.* 1991;27(5):517–522.
56. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc.* 2015;22(6):1220–1230.
57. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc Sep-Oct.* 2011;18(5):544–551.
58. Savova GK, Deleger L, Solti I, et al. Natural language processing: applications in pediatric research. In: Hutton JJ, ed. *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*. New York, NY: Springer; 2012:173–192.
59. Pestian JP, Deleger L, Savova GK, et al. Natural language processing—the basics. In: Hutton JJ, ed. *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*. New York, NY: Springer; 2012:149–172.
60. Carrell DJ. Anaphylaxis-Runrex. GitHub, 2020. https://github.com/kpwhri/anaphylaxis-runrex#about-the-project. Accessed October 20, 2021.
61. R Core Team. What is R? R Foundation for Statistical Computing, https://www.r-project.org/about.html. Accessed June 1, 2022.
62. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22.
63. Kaufman L, Rousseeuw PJ. Clustering by means of medoids. In: Dodge Y, ed. *Statistical Data Analysis Based on the L1–Norm and Related Methods*. Amsterdam, the Netherlands: North-Holland/Elsevier; 1987:405–416.
64. Chen T, He T, Benesty M, et al. xgboost: Extreme Gradient Boosting, R package, version 0.81.0.1. https://cran.r-project.org/web/packages/xgboost/index.html. Accessed September 15, 2021.
65. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat.* 2010;4(1):266–298.
66. Dorie V. dbarts: Discrete Bayesian Additive Regression Trees Sampler, R package, version 0.9-17, 2020. https://CRAN.R-project.org/package=dbarts. Accessed October 10, 2021.
67. Fritsch S, Guenther F, Wright MN, Suling M, Mueller SM. neuralnet: Training of Neural Networks, R package, version 1.44.2, 2020. https://cran.r-project.org/web/packages/neuralnet/index.html. Accessed October 10, 2021.
68. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge, United Kingdom: Cambridge University Press; 1996.
69. Polley EC. SuperLearner: Super Learner Prediction, R package, version 2.0-19, 2020. http://CRAN.R-project.org/package=SuperLearner. Accessed August 15, 2021.
70. Polley EC, van der Laan MJ. *Super Learner in Prediction*, Technical Report 200. Berkeley, CA: UC Berkeley: UC Berkeley Division of Biostatistics Working Paper Series, No. 266; 2010.
71. Office for Human Research Protections (OHRP). 45 CFR 46. U.S. Department of Health & Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html. Updated March 10, 2021. Accessed November 5, 2021.
72. Rosati K, Jorgensen N, Soliz M, Evans B. *HIPAA and Common Rule Compliance in the Sentinel Initiative, White Paper. Sentinel Initiative Principles and Policies*. Houston,

TX: University of Houston; 2018. https://www.sentinelinitiative.org/sites/default/files/communications/publications-presentations/HIPAA-Common-Rule-Compliance-in-Sentinel-Initiative.pdf. Accessed November 5, 2021.
73. Federal Policy for the Protection of Human Subjects, 82 Federal Register 12. 2017.
74. Gruber S, Carrell DS, Floyd JS, et al. Letter to the editor re Beachler, et al, 2021. *Pharmacoepidemiol Drug Saf.* 2021; 30(12):1735–1736.
75. Moons KG, Altman DG, Reitsma JB, et al. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. *Adv Anat Pathol.* 2015;22(5):303–305.
76. Yu S, Chakrabortty A, Liao KP, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *J Am Med Inform Assoc.* 2017;24(e1):e143–e149.
77. Yu S, Ma Y, Gronsbell J, et al. Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc.* 2018;25(1): 54–60.
78. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc.* 2019; 14(12):3426–3444.
79. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc.* 2017;24(5):986–991.
80. Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* 2021;4(1):86.
81. Mallya S, Overhage JM, Srivastava N, et al. Effectiveness of LSTMs in predicting congestive heart failure onset [preprint]. *arXiv.* 2019. https://arxiv.org/abs/1902.02443. Accessed November 9, 2022.
82. Braganza SC, Acworth JP, McKinnon DR, et al. Paediatric emergency department anaphylaxis: different patterns from adults. *Arch Dis Child.* 2006;91(2):159–163.