Open Globe Injury Patient Identification in Warfare Clinical Notes[1]

Emilia Apostolova, PhD[1,3], Helen A. White, MBA[2], Patty A. Morris[2], David A. Eliason, MD[2], Tom Velez, PhD[3]
[1]Language.ai, Chicago, IL; [2]DoD/VA Vision Center of Excellence, Falls Church, VA; [3]Computer Technology Associates, Cardiff, CA

Abstract

The aim of this study is to utilize the Defense and Veterans Eye Injury and Vision Registry clinical data derived from DoD and VA medical systems which include documentation of care while in combat, and develop methods for comprehensive and reliable Open Globe Injury (OGI) patient identification. In particular, we focus on the use of free-form clinical notes, since structured data, such as diagnoses or procedure codes, as found in early post-trauma clinical records, may not be a comprehensive and reliable indicator of OGIs. The challenges of the task include low incidence rate (few positive examples), idiosyncratic military ophthalmology vocabulary, extreme brevity of notes, specialized abbreviations, typos and misspellings. We modeled the problem as a text classification task and utilized a combination of supervised learning (SVMs) and word embeddings learnt in a unsupervised manner, achieving a precision of 92.50% and a recall of 89.83%. The described techniques are applicable to patient cohort identification with limited training data and low incidence rate.

Introduction

Open globe injury (OGI) refers to full thickness wound of the eyeball. OGIs range from small and self-sealing penetrations or lacerations to globe rupture with prolapse of intraocular contents or obliteration of the whole globe. The Birmingham Eye Trauma Terminology (BETT)[1] defines and classifies OGIs as shown in Figure 1.

OGIs are considered a major cause of significant loss of vision, blindness and/or total loss of the eye. Given the possibility of such catastrophic outcomes, OGI treatment, outcomes, and risk factors are an active field of research, for example[2–9]. OGI clinical research poses challenges, as such injuries are typically rare. It has been reported that in the US, penetrating eye injury accounts for 3.81 injuries per 100 000 injuries annually[10]. In the context of military operations, however, the occurrence of OGI is more prevailing. Eyes injuries make up a significant portion of all casualties experienced in battle by our service members and veterans. OGIs are a significant number of these. Furthermore, unlike in peacetime where unilateral injuries are the rule, ocular war injuries are bilateral in 15 to 25% of cases[11].

In this study, we utilize the Defense and Veterans Eye Injury and Vision Registry (DVEIVR), a web-based application, to develop methods for comprehensive and reliable OGI patient cohort identification. Congressionally mandated, DVEIVR was established to longitudinally track diagnosis, surgical intervention or other procedures, other treatments and follow up of active Service members and Veterans with significant eye and vision injuries from point of injury through rehabilitation starting 9/11/2001 and forward[12]. The DVEIVR registry includes abstracted data derived from initial battlefield acute care and subsequent definitive care encounters. The goal of DVEIVR is to aid researchers and providers in discovery of post-injury care pathways designed to prevent vision related injuries, protect and preserve the visual system, and restore vision. The DVEIVR data provided for the study is de-identified and complies with the DoD Privacy Office requirements, therefore exempted from IRB review.

In the context of DVEIVR, structured patient data can be used to somewhat reliably identify OGI patients. For example, ICD9 Diagnosis Codes *871.1 (Ocular laceration with prolapse or exposure of intraocular tissue), 360.61 (Foreign body in anterior chamber)*, etc. or Exam Types *Globe - Anophthalmic, Globe - IOFB* are reliable indicators of OGI. However, structured data is not always a comprehensive indicator of OGIs[13]. Diagnosis Codes can be ambiguous

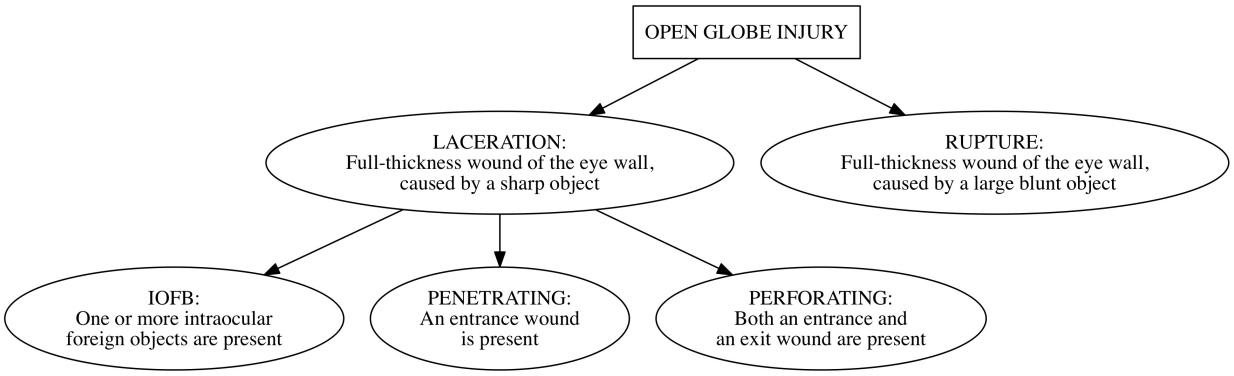

Figure 1: The BETT Classification of Open Globe Injuries.

and used to describe both OGI and closed globe injuries. Furthermore, codes and structured data in early post-trauma clinical records could be inaccurate for a number of practical reasons. We have observed a number of OGI incidents identified only in free-form text (clinical notes). The following brief notes are all clear indicators of the presence of an OGI:

HISTORY OF PRESENT ILLNESS: ruptured right globe in 2004 with a metal FB going through nasal right cornea through lens and into the vitreous. Pt underwent vitrectomy and phaco iol surgery right eye in 2006

CHIEF COMPLAINT TEXT: EYE TRAUMA INTRAOCULAR FOREIGN BODY

ENCOUNTER COMMENT: SPECIFY: Other (GLOBE EXPLORATION); FINDINGS:EBL ¡0.1 CCNO GLOBE WALL PENETRATION AND DISRUPTION NOTED WITH THROUGH EXPOSURE AND EXPLORATION

In such cases, structured data could be missing (e.g. no ICD9 Diagnosis Code), the ICD9 Code used could include eye injuries referring to both OGI and non-OGI instances, e.g. 871.4 (Unspecified laceration of eye), 369.9 (Unspecified visual loss), or the code may have been incorrectly selected (e.g. a foreign body coded as *intraocular* with an associated free text treatment comment that the foreign body was removed using a *q-tip*, inferring the wound was actually superficial).

Given the importance of identifying all available OGI patients, we explored the use of Natural Language Processing and Machine Learning (ML) to identify OGI patients from clinical notes. The nature of the notes poses a number of challenges: the notes are typically very short, abundant in domain-specific concepts (ophthalmology in the context of warfare and veteran care), with a high rate of abbreviations and typos. The described techniques are applicable, in general, to patient cohort identification with challenging datasets, limited training data, and low incidence rates.

Dataset

The dataset consists of de-identified patient data from the DVEIVR database exported in May 2016. It comprises of both structured and unstructured fields describing diagnoses, treatments, procedures, visits, and exams. The data contains 26,131 unique patients and 76,809 unique encounters.

We utilized the available structured OGI information to semi-automatically generate a training data set. Tables 1 and 2 list structured information values found to be reliable indicators of OGI patients. All free-form text fields associated with structured OGI information were identified and labeled as positive examples. A randomized sample of the remaining records was used as negative examples. As descriptive free-form text associated with OGI structured data does not always indicate OGI by itself, the positive dataset was manually reviewed and false positive examples were discarded. This resulted in a training dataset consisting of 1,645 free-form text notes indicating OGI (positive examples) and 10,732 negative examples.

The characteristics of the training dataset poses a number of challenges. First, the number of positive examples is small for the purposes of text classifications. In addition, individual examples are typically very short. The median

Table 1: List of ICD9 Diagnosis Codes used as reliable indicators of Open Globe Injuries.

ICD9 Diagnosis Code	ICD9 Diagnosis Code Description
360.5	Retained (old) intraocular foreign body magnetic
360.50	Foreign body, magnetic, intraocular, unspecified
360.51	Foreign body, magnetic, in anterior chamber of eye
360.52	Foreign body, magnetic, in iris or ciliary body
360.53	Foreign body, magnetic, in lens
360.54	Foreign body, magnetic, in vitreous
360.55	Foreign body, magnetic, in posterior wall
360.59	Intraocular foreign body, magnetic, in other or multiple sites
360.6	Retained (old) intraocular foreign body nonmagnetic
360.60	Foreign body, intraocular, unspecified
360.61	Foreign body in anterior chamber
360.62	Foreign body in iris or ciliary body
360.63	Foreign body in lens
360.64	Foreign body in vitreous
360.65	Foreign body in posterior wall of eye
360.69	Intraocular foreign body in other or multiple sites
871	Open wound of eyeball
871.0	Ocular laceration without prolapse of intraocular tissue
871.1	Ocular laceration with prolapse or exposure of intraocular tissue
871.2	Rupture of eye with partial loss of intraocular tissue
871.3	Avulsion of eye
871.5	Penetration of eyeball with magnetic foreign body
871.6	Penetration of eyeball with (nonmagnetic) foreign body
871.7	Unspecified ocular penetration
871.9	Unspecified open wound of eyeball
940.5	Burn with resulting rupture and destruction of eyeball

number of words per example is only 7. Furthermore, as in most clinical texts, abbreviations, spelling variations, typos, and misspellings are quite common. All these factors contribute to relatively large vocabulary size, while at the same time both the number of positive examples and the number of words per example are relatively small. Lastly, the vocabulary of the military ophthalmology domain is distinctive from general-domain clinical notes and doesn't quite fit general purpose medical ontologies and terminologies.

Method and Results

We modeled the task as a binary classification of free-form clinical notes. It has been shown that Support Vector Machines[14] achieve superior results in most text classifications tasks, including short texts (e.g.[15,16]). SVMs were selected as a sensible first choice. The individual clinical notes were represented as a bag of words (1-grams)[2]. The tokens were all converted to lower case and non-alphabetic characters were discarded. In addition, tokens that are present in more than 50% of all documents or less than 3 times across all documents were also discarded. This resulted in a dictionary of size 4,032 unique tokens. Tokens were weighted using the term-frequency / inverse-document-frequency scheme (tf-idf). We performed 10-fold cross-validation using linear kernel SVMs. We increased the positive class weight to address the unbalanced dataset (fewer positive examples). Best results achieved with positive class weight of 2. This resulted in a precision of 90.25 and a recall of 87.78 as shown in Table 3.

While the available positive training data is quite limited, it seemed sensible to attempt to utilize all available free-form text notes from all 76,809 patient encounters. Word embeddings[18,19] have gained popularity as a way of capturing semantic knowledge of words without relying on available training data. Unsupervised algorithms have also been

[2]Including bi-grams slightly deteriorated the model performance, a behaviour not atypical for text classification tasks[17].

Table 2: List of Exam Detail Types used as reliable indicators of Open Globe Injuries.

Eye Laceration	Globe - Anophthalmic
Globe - IOFB	Globe - IOFB - Earth/Mud
Globe - IOFB - Glass	Globe - IOFB - Metallic
Globe - IOFB - Metallic-Magnetic	Globe - IOFB - Metallic-Non Magnetic
Globe - IOFB - NFS	Globe - IOFB - Other
Globe - IOFB - Stone/Sand	Globe - IOFB - Wood
Globe - Intraocular Foreign Body (IOFB)	Globe - Intraocular Foreign Body (IOFB) - Metallic
Globe - Open, Blunt Trauma, Ruptured	Globe - Open, Penetrating
Globe - Open, Penetrating - Cornea Only	Globe - Open, Penetrating - Cornealscleral
Globe - Open, Penetrating - NFS	Globe - Open, Penetrating - Other
Globe - Open, Penetrating - Post Equatorial	Globe - Open, Penetrating - Sclera Only
Globe - Open, Penetrating - With Uveal Prolapse	Globe - Open, Perforating
Globe - Open, Perforating - NFS	Globe - Open, Perforating - Other
Globe - Open, Perforating - With Uveal Prolapse	Globe - Ruptured
Retina - Macular Foreign Body - Glass	Retina - Macular Foreign Body - Other
Retina - Macular Foreign Body - Stone/Sand	Vitreous - Foreign Body
Vitreous - Foreign Body - Metallic	Vitreous - Foreign Body - Metallic-Magnetic
Vitreous - Foreign Body - Metallic-Non Magnetic	Vitreous - Foreign Body - NFS
Vitreous - Foreign Body - Other	Vitreous - Foreign Body - Stone/Sand
Socket - Status of Globe - Enucleation	Socket - Prosthetic Fit - NFS
Socket - Status of Globe - Exenteration	Anterior Chamber - Foreign Body
Anterior Chamber - Foreign Body - Metallic	Cornea - Laceration Full Thickness (Perforating)
Cornea - Laceration Full Thickness (Perforating) - Linear, Corneal Scleral	Cornea - Laceration Full Thickness (Perforating) - NFS
Cornea - Laceration Full Thickness (Perforating) - Stellate, Corneal Scleral - with Tissue Loss	Cornea - Laceration Partial Thickness (Penetrating)
Iris - Iridectomy	Iris - Iridotomy - Other
Macula - Foreign Body	Retina - Choroidal Rupture
Retina - Foreign Body	Sclera - Laceration
Sclera - Perforation Posterior - Superior Nasal	Socket - Prosthesis
Socket - Prosthesis - Conformer	Socket - Prosthesis - Prosthetic
Socket - Prosthesis - Scleral Shell	Socket - Prosthetic Fit - Good
Socket - Prosthetic Motility - Fair	Socket - Prosthetic Motility - Good
Socket - Prosthetic Motility - Poor	Socket - Status of Globe - Enucleation

used to represent variable pieces of texts such as sentences and paragraphs as fixed-length feature representations (Paragraph Vectors)[20]. Studies have shown that Paragraph Vectors outperform bag-of-words models on a number of text classification tasks[20,21]. We used the free-text notes from all patient encounters to train Paragraph Vectors. The paragraph vectors the training notes were then fed to a logistic regression classifier and a neural network. In both cases results significantly under-performed the bag-of-words model. The underperformance of Paragraph Vectors is possibly due to the idiosyncrasies of the data: short text snippets, occasionally providing information on several, not necessarily related topics.

However, word embeddings revealed to be a powerful source of identifying semantically related abnormalities and procedures, variant spellings, abbreviations, and typos. Given the rare occurrence of OGIs, the intuition was that identifying all possible tokens expressing key OGI concepts would improve our model performance. Using all available free-form text, we generated word vectors for all words in the vocabulary[3]. The words in the training set were then clustered using Agglomerative Clustering. The average cosine distance between word vectors was used as a measure of similarity between clusters. The automatically derived clusters revealed to accurately capture related vocabulary words. Table 4 shows samples of word clusters derived with the described approach. The samples show that word embeddings accurately captured semantically related concepts (e.g. Cluster 7 lists various antibiotics); abbreviations

[3]We used word skip-gram model, vectors of size 400, and window of size 4.

Table 3: Classification Results. SVM^{b}=baseline, SVM^{wc+ne}=word vector clusters + negation

	Precision	**Recall**	**F1-score**	**Vocabulary Size**
SVM^{b}	90.25	87.78	88.99	4,032
SVM^{wc+ne}	92.50	89.83	91.14	2,706

and typos (e.g. Cluster 6 lists 7 different spellings of *fluorescein*); and many domain-specific semantic relations (e.g. Cluster 10 lists various military vehicles, while Cluster 3 lists procedures and conditions relevant to the concepts of *enucleation*). We utilized the generated clusters to reduce the vocabulary size by treating all words in a cluster as equivalent.

Table 4: Sample word clusters generated using agglomerative clustering and the cosine distance of word embeddings.

1	detonated, exploded, attack
2	eccymosis, brusing, ecchymoses, echymosis, bruising, ecchymosis
3	enucleated, exenteration, anophthalmic, enucleation, orbilotomy, enculeation, evisceration, nlp, enuc, phthisis
4	eyelids, lids, eyelid, palpebral, lid
5	flipped, everted, inversion, flipping, inverted, eversion
6	fluorescein, flouroscein, fluoroscein, florescein, fluorescin, flourescein, fluoro
7	gentamicin, mycin, emycin, erythromycin, antibiotic, abx, bacitracin, antibiotics, polytrim, tobradex, vigamox, ilotycin
8	glasses, spec, specs, gls, srx, eyeglasses, eyeglass
9	ha, confusion, lightheadedness, vomiting, headache, dizziness, headaches, nausea, phonophobia
10	humvee, vehicle, truck, mrap
11	lasix, lasic, lasik, prk
12	schrapnel, gunshot, scrapnel, gsw, shrapnel
13	swab, tip, cotton, cta
14	washed, wash, irrigated, flushed, irrigation, lavage, rinse, flushing, flush

Lastly, we addressed the issue of negation (e.g. *no globe rupture*) as negated conditions and procedures are quite common in the dataset. We ran a customized version of the NegEx[22] algorithm to identify negated words and phrases. Encoding negated tokens in a bag-of-words model as separate token (e.g. *rupture* vs *no_rupture*) is a common approach to treating negation. However, given the limited dataset and the large vocabulary size, we instead simply discarded the negated tokens.

We again performed 10-fold cross-validation using linear kernel SVM and positive class weight set to 2. Both precision and recall increased from 90.25 to 92.50 and from 87.78 to 89.83, respectively, while the vocabulary size was reduced significantly.

Related Work

There is a vast amount of literature on the use of free-form text and clinical notes for the purposes of patient cohort identification. A review of approaches to identifying patient phenotype cohorts using electronic health records[23] describes a number of studies using clinical notes, most often in combination with additional structured information, such as diagnosis codes. The study asserts that clinical notes are often the only source of information from which to infer important phenotypic characteristics.

Most studies map textual elements to standard vocabularies, such as the Unified Medical Language System (UMLS). For example, Bejan et al.[24] describe a system for pneumonia identification from narrative reports using n-grams, UMLS concepts, and assertion status. Similarly, Elkin et al.[25] encoded radiology reports using the SNOMED CT Ontology and developed a set of rules to identify pneumonia cases. Carroll et al.[26] used as features a combination of ICD-9 codes, UMLS concepts, and medication names and build an SVM classifier for the identification of rheumatoid arthritis patients.

A number of studies use rules, such as keywords and regular expressions, to identify the phenotype of interest. Friedlin et al.[27] compare methods for identifying pancreatic cancer patients and utilize a tool that uses a combination of regular expressions and algorithms to detect keywords and their context. Xu et al.[28] develop a system to detect patients with colorectal cancer that uses a combination of heuristic rule-based approach and the MedLEE[29] system to extract relevant concepts. Sohn and Savova[30] improve the Mayo Clinic Smoking Status Classification System by introducing a rule-based component for patient-level smoking status assignments.

A growing number of studies utilize the use of Machine Learning (ML) on the task of patient cohort identification from clinical narratives. The i2b2 (Informatics for Integrating Biology to the Bedside) project organized a challenge with the task of determining the patient smoking status from free-form discharge notes[31]. The 11 participating teams utilized a variety of rule-based and ML approaches, including Nave Bayes, Decision Trees, AdaBoost classifiers, logistic regression, and neural networks. The most common ML algorithm was SVMs, as reported by 5 participating teams. The ShARe/CLEF eHealth 2013 task was organized to evaluate the state of the art in disorder recognition and normalization of the clinical narrative[32]. Most of the participating systems employed hybrid approaches by supplementing features to a machine-learning algorithm and the best performing system utilized Conditional Random Fields and Structured SVMs. Lin et al.[33] develop a system for automatic prediction of rheumatoid arthritis disease activity from the Electronic Medical Records (clinical narratives and lab values). They report that the best performing combination was linear kernel SVMS with UMLS concepts, feature selection, and lab values. Rumshisky et al.[34] utilize Latent Dirichlet Allocation and SVMs to predict early psychiatric readmission from narrative discharge summaries.

Conclusion

We demonstrated an approach to identifying rare patient phenotypes (OGIs) utilizing free-form text in a challenging domain: military ophthalmology. The characteristics of the clinical narratives include extreme brevity, idiosyncratic terms and abbreviations, typos and misspellings. Using available structured data, we were able to construct a free-text training dataset with minimal manual review. We then employed supervised Machine Learning methods for classification of clinical notes. Best results were achieved utilizing clustering of the vocabulary (unigrams) using word embeddings and performing SVM classification by increasing the weight of the positive class to account for the imbalanced dataset. We were able to achieve an overall F1-score of 91.14. The described methods are applicable to short clinical text classification in narrowly specialized domains.

Acknowledgements

Research reported in this Open Globe Injury Patient Identification in Warfare Clinical Notes was supported by the DoD/VA Vision Center of Excellence under award number GS00Q09BGD0031. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

References

1. F Kuhn, R Morris, CD Witherspoon, and V Mester. The birmingham eye trauma terminology system (bett). *Journal Français d'Ophtalmologie*, 27(2):206–210, 2004.

2. Dante J Pieramici, Mathew W MacCumber, Michael U Humayun, Marta J Marsh, and Eugene de Juan. Open-globe injury: update on types of injuries and visual results. *Ophthalmology*, 103(11):1798–1803, 1996.

3. GW Schmidt, AT Broman, Holly B Hindman, and Michael P Grant. Vision survival after open globe injury predicted by classification and regression tree analysis. *Ophthalmology*, 115(1):202–209, 2008.

4. Aaron Savar, Michael T Andreoli, Carolyn E Kloek, and Christopher M Andreoli. Enucleation for open globe injury. *American journal of ophthalmology*, 147(4):595–600, 2009.

5. C Yu Wai Man and D Steel. Visual outcome after open globe injury: a comparison of two prognostic modelsthe ocular trauma score and the classification and regression tree. *Eye*, 24(1):84–89, 2010.

6. Eric D Weichel, Marcus H Colyer, Spencer E Ludlow, Kraig S Bower, and Andrew S Eiseman. Combat ocular trauma visual outcomes during operations iraqi and enduring freedom. *Ophthalmology*, 115(12):2235–2245, 2008.

7. Y Ahmed, AM Schimel, A Pathengay, MH Colyer, and Harry W Flynn. Endophthalmitis following open-globe injuries. *Eye*, 26(2):212–217, 2012.

8. Marcus H Colyer, Eric D Weber, Eric D Weichel, John SB Dick, Kraig S Bower, Thomas P Ward, and Julia A Haller. Delayed intraocular foreign body removal without endophthalmitis during operations iraqi freedom and enduring freedom. *Ophthalmology*, 114(8):1439–1447, 2007.

9. Marcus H Colyer, Dal W Chun, Kraig S Bower, John SB Dick, and Eric D Weichel. Perforating globe injuries during operation iraqi freedom. *Ophthalmology*, 115(11):2087–2093, 2008.

10. CM Gilbert, HK Soong, and LW Hirst. A two-year prospective study of penetrating ocular trauma at the wilmer ophthalmological institute. *Annals of ophthalmology*, 19(3):104–106, 1987.

11. Robert Scott. The injured eye. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1562):251–260, 2011.

12. Vision registry. http://vce.health.mil/Vision-Registry. Accessed: 2016-12-27.

13. Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

14. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

15. Anthony Khoo, Yuval Marom, and David Albrecht. Experiments with sentence classification. In *Proceedings of the 2006 Australasian language technology workshop*, pages 18–25, 2006.

16. Atreya Basu, Christine Walters, and M Shepherd. Support vector machines for text categorization. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 7–pp. IEEE, 2003.

17. Ron Bekkerman and James Allan. Using bigrams in text categorization. Technical report, Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst, 2004.

18. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

19. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

20. Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

21. Andrew M Dai, Christopher Olah, and Quoc V Le. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015.

22. Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

23. Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, 2014.

24. Cosmin Adrian Bejan, Fei Xia, Lucy Vanderwende, Mark M Wurfel, and Meliha Yetisgen-Yildiz. Pneumonia identification using statistical feature selection. *Journal of the American Medical Informatics Association*, 19(5):817–823, 2012.

25. Peter L Elkin, David Froehling, Dietlind Wahner-Roedler, Brett E Trusko, Gail Welsh, Haobo Ma, Armen X Asatryan, Jerome I Tokars, S Trent Rosenbloom, and Steven H Brown. Nlp-based identification of pneumonia cases from free-text radiological reports. In *AMIA*, 2008.

26. Robert J Carroll, Anne E Eyler, and Joshua C Denny. Naive electronic health record phenotype identification for rheumatoid arthritis. In *AMIA Annu Symp Proc*, volume 2011, pages 189–196, 2011.

27. J Friedlin, M Overhage, MA Al-Haddad, JA Waters, JJR Aguilar-Saavedra, J Kesterson, and M Schmidt. Comparing methods for identifying pancreatic cancer patients using electronic data sources. In *AMIA*, page 237241, 2010.

28. Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B Peterson, Qingxia Chen, Subramani Mani, Mia A Levy, Qi Dai, and Josh C Denny. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. In *AMIA Annu Symp Proc*, volume 2011, pages 1564–1572, 2011.

29. Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

30. Sunghwan Sohn and Guergana K Savova. extensions and improvements. In *AMIA Annu Symp Proc*, 2009.

31. Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

32. Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154, 2015.

33. Chen Lin, Elizabeth W Karlson, Helena Canhao, Timothy A Miller, Dmitriy Dligach, Pei Jun Chen, Raul Natanael Guzman Perez, Yuanyan Shen, Michael E Weinblatt, Nancy A Shadick, et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS One*, 8(8):e69932, 2013.

34. A Rumshisky, M Ghassemi, T Naumann, P Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, 6(10):e921, 2016.