MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231031

Automatic Extraction of Skin and Soft Tissue Infection Status from Clinical Notes

Jamie L.W. Rhoads[a,b,1], Lee Christensen[c], Skylar Westerdahl[a], Vanessa Stevens[b,d], Wendy W. Chapman[e], Mike Conway[e]
[a] *Dept. Dermatology, University of Utah, Salt Lake City, UT, USA*
[b] *Informatics, Decision-Enhancement and Analytic Sciences (IDEAS) Center of Innovation, VA Salt Lake City Health Care System, Salt Lake City, UT, USA*
[c] *Dept. Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*
[d] *Div. Epidemiology, University of Utah, Salt Lake City, UT, USA*
[e] *Centre for Digital Transformation of Health, University of Melbourne, VIC, Australia*
ORCiD ID: Jamie L.W. Rhoads https://orcid.org/0000-0002-2859-4168

Abstract. The reliable identification of skin and soft tissue infections (SSTIs) from electronic health records is important for a number of applications, including quality improvement, clinical guideline construction, and epidemiological analysis. However, in the United States, *types* of SSTIs (e.g. is the infection *purulent* or *non-purulent*?) are not captured reliably in structured clinical data. With this work, we trained and evaluated a rule-based clinical natural language processing system using 6,576 manually annotated clinical notes derived from the United States Veterans Health Administration (VA) with the goal of automatically extracting and classifying SSTI subtypes from clinical notes. The trained system achieved mention- and document-level performance metrics of the range 0.39 to 0.80 for mention level classification and 0.49 to 0.98 for document level classification.

Keywords. Natural Language Processing; Skin and Soft Tissue Infections; Electronic Health Records

1. Introduction

Skin and soft tissue infections (SSTIs) are one of the most common infectious diseases treated in the US ambulatory setting [1]. Evidence has recently emerged that is inconsistent with the current established guidelines regarding both *when* and *how* adjuvant antibiotics should be prescribed with respect to *purulent* (i.e. contain a collection of pus) and *non-purulent* (i.e. not containing pus) SSTIs [2–4]. A series of studies suggest that providers are not adhering to SSTI prescribing guidelines regarding overuse of adjuvant antibiotics, choosing inappropriate antibiotics, and underutilizing incision and drainage (I&D), a procedure to drain a focus of pus [4–7].

A key challenge in conducting research on SSTIs using electronic health records (EHRs) is the inadequacy of International Classification of Disease (ICD) -9 and -10 codes for clinical classification because SSTI sub-types are grouped under the same code. Similarly, there is currently limited evidence available regarding the accuracy of Current

[1] Corresponding author: Jamie L.W. Rhoads - jamie.rhoads@hsc.utah.edu

Procedure Terminology codes for identifying I&D procedures. As a result, the structured component of EHRs cannot be used to decisively sub-type SSTIs.

The utility of natural language processing (NLP) methods for extracting clinical information has been recognized by the dermatology research community [8] for various applications [9]. While machine learning currently dominates in the wider NLP research community, rule-based NLP algorithms have continued to play an important role in clinical NLP due to the reduced cost associated with developing a dataset [10], the relative ease of adaptation to new contexts [11], and the reduced need for resource-intensive hardware [10].

Our objective was to develop and evaluate a rule-based algorithm for the automatic extraction of SSTI subtype (*purulent*/*non-purulent*) and I&D procedures at the document (patient note) level for the purpose of assessing SSTI management at the United States Veterans Health Administration (VA). The VA EHR provides a unique national dataset derived from 1,255 healthcare facilities and over 9 million patients [12] facilitating investigations into SSTI treatment practices. We utilized the rule-based Moonstone clinical NLP system, a system initially developed for the automatic identification of social risk factors associated with readmission [11]. Contributions of this work include (1) evidence that an existing rule-based NLP system can be adapted to a new, unrelated task; and (2) the development of an NLP tool that can be utilized for clinical, operational, and epidemiological goals related to automated SSTI type detection.

2. Methods

We conducted a retrospective cohort study of Veterans diagnosed with SSTIs in VA ambulatory clinics and emergency departments, which has been described in a prior publication [13]. To identify our cohort, we utilized the VA Informatics & Computing Infrastructure to identify ambulatory VA encounters associated with an ICD-9 or -10 code for an SSTI between 1/1/2005 and 6/30/2018 [13]. We randomly sampled 2,000 of these SSTI events, which yielded 6,576 clinical notes. To limit the clinical notes to the prescribing provider for these encounters, the provider types were limited to physicians, physician assistants, advance nurse practitioners, podiatry and dental providers.

An annotation scheme and associated guidelines were iteratively developed using five batches of 80 notes (i.e. 400 notes), with the annotation scheme refined as a result of discussion between two domain-expert annotators (a medical student and a pharmacist) and an adjudicator (a board-certified dermatologist – see Supplementary Materials for annotation guidelines[2]). These 400 notes were double annotated at the mention and document-level creating a reference standard corpus with agreement calculated using f-measure. Mention level annotations focused on identifying phrases in the notes indicating that the patient exhibited an active SSTI, an exclusionary diagnosis (see Supplementary Materials) or had recently undergone an I&D procedure. Finally, annotators assigned a document-level annotation representing the provider's diagnosis from one of the following: *purulent SSTI*, *non-purulent SSTI*, *non-specific SSTI*, *exclusionary diagnosis*, or *not applicable* (i.e. a clinical note unrelated to the SSTI event). An I&D document-level assessment was automatically assigned as *performed* or *not performed* if there was one positive I&D mention in the note or not, respectively. The

[2] Supplementary Materials: https://maconway.github.io/medinfo_supp.zip

annotated corpus was then split into a training set (3,288 notes) and a test set (3,288 notes) to develop our rules-based NLP approach. 2,322 notes from the training set were used to train the NLP tool. 966 notes from the training set were used to validate the NLP performance prior to applying the trained system to the test set.

To develop our NLP algorithm we used the rule-based open-source Moonstone NLP system. Moonstone training involves using a bespoke tool to visually compare the sentence-level reference training-set standard annotations (i.e. the annotations generated by our domain experts) with annotations over those same documents generated by Moonstone and iteratively refining a semantic grammar to improve Moonstone's performance [11,14]. We manually created a document-level decision tree using the output of the mention-level classifier in conjunction with rules generated by our domain experts to determine the SSTI document-level classification (see Supplementary Materials). Once an acceptable performance in f-measure (0.8 or higher) was achieved for *purulent* and *non-purulent* mention-level and *purulent*, *non-purulent*, and *I&D* document-level, we applied the resulting classifier to the validation set. We then made further modifications before applying the final classifier to the held-out test set.

This study was approved by the University of Utah's institutional review board and the Research and Development Committee at the VA Salt Lake City Health Care System.

3. Results

Interannotator agreement between the two domain expert annotators proved to be high for both mention and document level annotation (F-score > 0.9) indicating the feasibility of the task. The mention-level classification performance of the trained algorithm for the *purulent* and *non-purulent* categories was relatively high (0.76 and 0.80 F-score, respectively), while performance for *non-specific* and *exclusion* were low (0.39 and 0.51 F-score, respectively) perhaps due to their relatively low frequency in our corpus. The document-level classifier generally performed well, especially for our core task of predicting whether a note made references to a non-negated *purulent* or *non-pururlent* SSTI (0.82 and 0.84 F-score, respectively). Similar to our mention-level classifier, *non-specific* SSTIs and *exclusionary diagnosis* did not perform as well. Regarding document-level annotations specifically for I&D, the algorithm was relatively successful at identifying documents positive for I&D procedures. The "all class" document level classifier for I&D performed very well (0.96 F-score) due to the relative rarity of reported affirmed or negated I&D procedures present in the corpus. Note that for document-level *exclusionary diagnosis* classification, performance was moderate (0.63 F-score).

Table 1. Mention-level performance for training, validation, and test sets

Class	Training			Validation			Testing		
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
Purulent	0.77	0.86	0.81	0.71	0.85	0.78	0.7	0.85	0.76
Non-purulent	0.82	0.80	0.81	0.79	0.80	0.80	0.79	0.80	0.80
Non-specific	0.49	0.30	0.38	0.46	0.35	0.39	0.49	0.32	0.39
Exclusion	0.67	0.61	0.64	0.46	0.30	0.62	0.44	0.62	0.51
I&D	0.75	0.7	0.72	0.75	0.70	0.72	0.72	0.67	0.64
All classes	0.78	0.51	0.61	0.75	0.48	0.58	0.74	0.50	0.60

Table 2. Document-level performance for training, validation, and test sets

	Attribute	Training			Validation			Testing		
		Pre	Rec	F	Pre	Rec	F	Pre	Rec	F
SSTI	purulent	0.79	0.85	0.82	0.75	0.88	0.81	0.75	0.90	0.82
	non-purulent	0.91	0.74	0.81	0.89	0.71	0.79	0.91	0.77	0.84
	non-specific SSTI	0.23	0.49	0.32	0.28	0.62	0.39	0.47	0.52	0.49
	N/A	0.92	0.88	0.9	0.92	0.89	0.91	0.93	0.91	0.92
	all classes	0.83	0.83	0.83	0.84	0.84	0.84	0.87	0.87	0.87
I&D	performed	0.77	0.83	0.8	0.70	0.86	0.78	0.68	0.84	0.75
	not performed	0.98	0.98	0.98	0.99	0.98	0.99	0.99	0.97	0.98
	all classes	0.97	0.97	0.97	0.98	0.98	0.98	0.96	0.96	0.96

4. Discussion

We successfully developed a rule-based NLP system capable of distinguishing *purulent* and *non-purulent* SSTIs and identifying I&D procedures at the document level. Sub-typing SSTIs and having a complete understanding of SSTI treatment (i.e. frequency of I&D procedures) are necessary for an epidemiologic analysis of practice variation. Identification of clinical variables associated with practice outliers allow for development of decision support tools that encourage clinicians to follow SSTI treatment guidelines and practice antibiotic stewardship. Extracting SSTI sub-type from clinical notes is a complex task. Simply identifying *purulent* and *non-purulent* SSTIs is well suited to a rule-based approach given the relatively limited number of terms and phrases associated with the documentation of common SSTI types. There was greater difficulty distinguishing exclusionary diagnoses due to their relatively low frequency in our corpus. SSTIs where the sub-type was not clearly described by providers were also challenging to characterize. Given the known difficulties in porting clinical NLP algorithms between hospital systems [11], we cannot be confident that our system would achieve similar performance standards on non-VA data. In addition to the general issue of NLP algorithm portability, the VA patient cohort has a number of characteristics that may impede algorithm generalizability, including gender, age and socio-economic distributions that do not necessarily reflect the wider US population. A further important limitation of this work is that we did not evaluate our trained classifier against a machine-learning based NLP baseline. We made this decision due to the feasibility of generating sufficient labelled data to effectively train and evaluate a machine learning algorithm.

5. Conclusions

This work demonstrates the feasibility of creating mention and document level NLP algorithms to identify *purulent* and *non-purulent* SSTIs and I&D procedures. We achieved good performance at the mention (0.76 and 0.80 F-score) and document level (0.82 and 0.84 F-score) for *purulent* and *non-purulent* SSTIs and excellent performance for I&D at the document level (0.96 F-score). Finally, this paper demonstrates the continuing utility of rule-based methods to perform complex NLP tasks using limited annotated data.

Acknowledgements

The authors thank Garrett G. Cole, a data scientist at the University of Utah Division of Epidemiology for preparing the annotation support. May he rest in peace. Research reported in this publication was supported by VA Health Services Research and Development [PPO 17-016] and utilized the resources of the VA Health Care System. The views expressed in this article are those of the authors and do not necessarily reflect the position of the Department of Veterans Affairs or the United States government.

References

[1] L. May, P. Mullins, and J. Pines, Demographic and treatment patterns for infections in ambulatory settings in the United States, 2006-2010, *Acad Emerg Med.* **21** (2014) 17–24. doi:10.1111/acem.12287.

[2] D.A. Talan, W.R. Mower, A. Krishnadasan, F.M. Abrahamian, F. Lovecchio, D.J. Karras, M.T. Steele, R.E. Rothman, R. Hoagland, and G.J. Moran, Trimethoprim-sulfamethoxazole versus placebo for uncomplicated skin abscess, *N Engl J Med.* **374** (2016) 823–832. doi:10.1056/NEJMoa1507476.

[3] L.G. Miller, R.S. Daum, C.B. Creech, D. Young, M.D. Downing, S.J. Eells, S. Pettibone, R.J. Hoagland, H.F. Chambers, and DMID 07-0051 Team, Clindamycin versus trimethoprim-sulfamethoxazole for uncomplicated skin infections, *N Engl J Med.* **372** (2015) 1093–1103. doi:10.1056/NEJMoa1403789.

[4] K.J. Brinsley-Rainisch, R.L. Cochran, and M.L. Pearson, Dermatologists' perceptions and practices related to community-associated methicillin-resistant Staphylococcus aureus infections, *Am J Infect Control.* **36** (2008) 668–671. doi:10.1016/j.ajic.2008.02.010.

[5] R. Mistry, D. Shapiro, M. Goyal, T. Zaoutis, J. Gerber, C. Liu, and A. Hersh, Clinical management of skin and soft tissue infections in the U.S. emergency departments, *The Western Journal of Emergency Medicine.* **15** (2014) 491–8. doi:10.5811/westjem.2014.4.20583.

[6] D.J. Pallin, C.A. Camargo, and J.D. Schuur, Skin infections and antibiotic stewardship: analysis of emergency department prescribing practices, 2007-2010, *West J Emerg Med.* **15** (2014) 282–289. doi:10.5811/westjem.2013.8.18040.

[7] R.S. Kamath, D. Sudhakar, J.G. Gardner, V. Hemmige, H. Safar, and D.M. Musher, Guidelines vs actual management of skin and soft tissue infections in the emergency department, *Open Forum Infect Dis.* **5** (2018) ofx188. doi:10.1093/ofid/ofx188.

[8] A.J. Park, G.S. Weintraub, and M.M. Asgari, Leveraging the electronic health record to improve dermatologic care delivery: The importance of finding structure in data, *J Am Acad Dermatol.* **82** (2020) 773–775. doi:10.1016/j.jaad.2019.10.064.

[9] B.T. Bucher, J. Shi, J.P. Ferraro, D.E. Skarda, M.H. Samore, J.F. Hurdle, A.V. Gundlapalli, W.W. Chapman, and S.R.G. Finlayson, Portable automated surveillance of surgical site infections using natural language processing: development and validation, *Ann Surg.* **272** (2020) 629–636. doi:10.1097/SLA.0000000000004133.

[10] I. Spasic, and G. Nenadic, Clinical Text Data in Machine Learning: Systematic Review, *JMIR Med Inform.* **8** (2020) e17984. doi:10.2196/17984.

[11] R.M. Reeves, L. Christensen, J.R. Brown, M. Conway, M. Levis, G.T. Gobbel, R.U. Shah, C. Goodrich, I. Ricket, F. Minter, A. Bohm, B.E. Bray, M.E. Matheny, and W. Chapman, Adaptation of an NLP system to a new healthcare environment to identify social determinants of health, *J Biomed Inform.* **120** (2021) 103851. doi:10.1016/j.jbi.2021.103851.

[12] Veterans Health Administration, About VHA - Veterans Health Administration, (2022). https://www.va.gov/health/aboutvha.asp (accessed November 1, 2022).

[13] J.L.W. Rhoads, T.M. Willson, J.D. Sutton, E.S. Spivak, M.H. Samore, and V.W. Stevens, Epidemiology, disposition, and treatment of ambulatory veterans with skin and soft tissue infections, *Clin Infect Dis.* **72** (2021) 675–681. doi:10.1093/cid/ciaa133.

[14] M. Conway, S. Keyhani, L. Christensen, B.R. South, M. Vali, L.C. Walter, D.L. Mowery, S. Abdelrahman, and W.W. Chapman, Moonstone: a novel natural language processing system for inferring social risk from clinical narratives, *Journal of Biomedical Semantics.* **10** (2019) 6. doi:10.1186/s13326-019-0198-0.