Article published online: 2021-12-29

A Data-Driven Iterative Approach for Semi-automatically Assessing the Correctness of Medication Value Sets: A Proof of Concept Based on Opioids

Linyi (Sabrina) Li[1,2] Adela Grando[3] Abeed Sarker[4]

[1] Department of Computer Science, Emory University, Atlanta, Georgia, United States
[2] Language Technologies Institute, Carnegie Mellon University, Pennsylvania, United States
[3] College of Health Solutions, Arizona State University, Phoenix, Arizona, United States
[4] Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, Georgia, United States

Address for correspondence Abeed Sarker, PhD, 101 Woodruff Circle, Suite 4101, Atlanta, GA 30322, United States (e-mail: abeed@dbmi.emory.edu).

Methods Inf Med 2021;60:e111–e119.

Abstract

Background Value sets are lists of terms (e.g., opioid medication names) and their corresponding codes from standard clinical vocabularies (e.g., RxNorm) created with the intent of supporting health information exchange and research. Value sets are manually-created and often exhibit errors.

Objectives The aim of the study is to develop a semi-automatic, data-centric natural language processing (NLP) method to assess medication-related value set correctness and evaluate it on a set of opioid medication value sets.

Methods We developed an NLP algorithm that utilizes value sets containing mostly true positives and true negatives to learn lexical patterns associated with the true positives, and then employs these patterns to identify potential errors in unseen value sets. We evaluated the algorithm on a set of opioid medication value sets, using the recall, precision and F_1-score metrics. We applied the trained model to assess the correctness of unseen opioid value sets based on recall. To replicate the application of the algorithm in real-world settings, a domain expert manually conducted error analysis to identify potential system and value set errors.

Results Thirty-eight value sets were retrieved from the Value Set Authority Center, and six (two opioid, four non-opioid) were used to develop and evaluate the system. Average precision, recall, and F_1-score were 0.932, 0.904, and 0.909, respectively on uncorrected value sets; and 0.958, 0.953, and 0.953, respectively after manual correction of the same value sets. On 20 unseen opioid value sets, the algorithm obtained average recall of 0.89. Error analyses revealed that the main sources of system misclassifications were differences in how opioids were coded in the value sets—while the training value sets had generic names mostly, some of the unseen value sets had new trade names and ingredients.

Keywords

- prescription drugs
- natural language processing
- opioids
- medication value sets

received
September 2, 2021
accepted after revision
October 11, 2021
published online
December 24, 2021

DOI https://doi.org/10.1055/s-0041-1740358.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Conclusion The proposed approach is data-centric, reusable, customizable, and not resource intensive. It may help domain experts to easily validate value sets.

Background and Significance

Value sets (e.g., opioid medications) are lists of terms or descriptions (e.g., oxycodone hydrochloride 15-mg oral tablet) and corresponding codes from standard clinical vocabularies. The value set authority center (VSAC) was created by the National Library of Medicine, in collaboration with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare and Medicaid Services, as a repository for public, reusable value sets and an authoring tool. Value sets have grown in importance and ubiquity with the advent of common data models, distributed research networks, and the availability of higher-order, reusable analytic resources like electronic phenotypes and electronic clinical quality measures. A major difference between value sets and software implementations of analytic methods is that value sets do not need to be expressed in a specific programming or query language.

The process of creating and maintaining value sets is labor-intensive, involving domain experts working with terminologists to identify concepts for inclusion in a particular value set. Past studies have investigated automated methods to assist the authoring and quality assurance of value sets, making use of information such as the hierarchical structure of terminologies, the semantic types of the codes and inter-terminology maps.[1] A value set should contain all the relevant codes for a particular data element (completeness) and should contain only the relevant codes for a particular data element (correctness).[2] Multiple value sets with the same codes should be harmonized, to facilitate maintenance and prevent inconsistencies over time (non-redundancy). It was found when looking at 526 SNOMED CT value sets for diagnosis that only 271 (52%) of those value sets were complete, 65 (12%) were nearly complete, and 190 (36%) were missing a significant proportion of concepts.[3] It was also reported that value sets often exhibit errors and redundancy, illustrating the need for reliable methods and tools to address those issues.[4]

One possible mechanism to discover errors or inconsistencies in value sets is to develop automatic methods that apply natural language processing (NLP). Ideally, such methods should (1) be data-centric—so that it does not require re-programming or re-configuring when applying to distinct value sets, (2) be simple—so that it can be used by medical experts who may not have expertise in informatics, and (3) not require many external datasets—given a single or small number of value sets, it should be able to detect potential errors within them.

Here, we describe the development of such a system. From a high-level perspective, the proposed system utilizes the correct entries within a given value set to automatically generate lexical patterns via regular expressions that can detect entries that should and should not be in the value set. A regular expression is a sequence of characters that specifies a search pattern that can be used to detect segments from text. They have been widely used in information extraction and text classification tasks. Past studies have employed regular expressions in either a top-down or bottom-up manner to automatically find patterns from training set samples. The top-down approach requires a task-specific manual construction of seed patterns followed by modifications and adaptations of the regular expressions to meet a predetermined performance threshold. The bottom-up approach builds regular expression patterns based on the similarities within clusters of the training set samples. The method proposed in this paper relies on an automatic bottom-up NLP approach where the system automatically learns the rules under the assumption that most of the entries in the value set are correct. Manual error analysis involves domain experts whose engagement is limited to correcting system misclassifications or datasets errors.

Objectives

The objectives of this study were to: (1) propose a semi-automatic method based on a data-driven iterative pattern generation NLP algorithm to assess value set correctness; and (2) demonstrate the methodology in the context of quality assurance of VSAC value sets for opioid medications.

We chose opioids because there are many opioid value sets and prescription opioids have received considerable research and public health attention in the recent past. Opioid use and overdose have emerged as an alarming crisis in the United States (US). In 2019, an estimated 10.1 million people aged 12 or older reported misuse of opioids in the past year.[5] Prescription drug monitoring programs (PDMPs) have been implemented throughout the US as a decision support for prescribers, pharmacists, and regulators.[6] PDMPs are electronic databases that collect and analyze patient prescription data. Providers, such as prescribers and pharmacists, are required to check the PDMPs before they prescribe controlled substances, such as opioids. PDMPs need to rely on up-to-date, correct, and complete opioid datasets to inform prescribers about concurrent prescriptions and expose drug misuse at the time of prescribing. While the overarching goal behind the development of the method is to make it reusable on any value set, opioid-related value sets represent targets for potential immediate application.

Methods

Datasets

We queried VSAC using the keyword "opioid" and chose "RxNorm," and "SNOMED CT" as clinical vocabularies. VSAC returned 38 value sets. After manual inspection, we identified

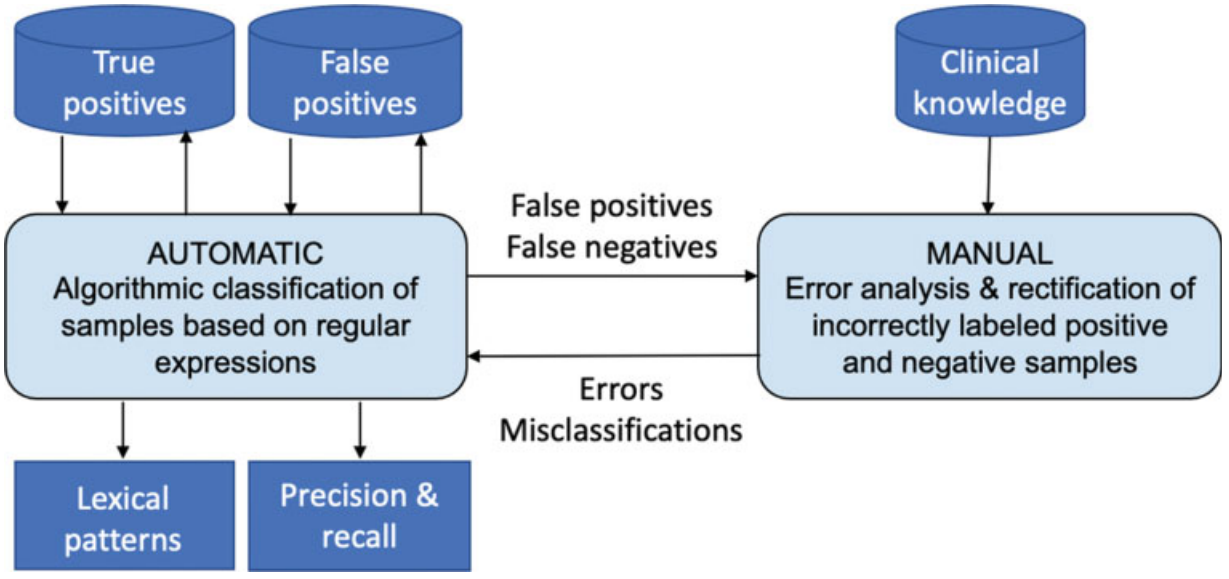

Fig. 1 Flowchart illustrating the application of our method in a practical setting, involving an automatic and a manual component to assess the correctness of value sets.

relevant value sets containing opioid (*e.g.,* long-acting opioids) and non-opioid (*e.g.,* non-opioid pain relievers) medications. Then, we conducted an initial assessment of value sets' completeness, correctness and redundancy through the analysis of dataset intersections, and concept uniqueness and repetition.

Approach

We propose a semi-automatic approach (►**Fig. 1**) for the overall task of value set correctness assessment that takes as input manually-created sets of potential positive (true positives) and negative (false positives) examples. The system development, training and intrinsic evaluation process starts with a data-driven pattern generation algorithm that automatically learns lexical patterns from the given training sets, resulting in the computation of precision and recall scores, and identification of potential false positives in a target value set; the process continues with a manual error analysis involving expert domain knowledge that leads to the correction of system misclassifications or dataset errors. We also conduct an extrinsic evaluation during which the lexical patterns learned are automatically used to assess the correctness of unseen datasets.

Algorithm

We modeled the problem of automatic detection of medications belonging to a specific class as a binary classification problem, focusing on the category of opioids as our use case. To ensure that the approach can separate potential true and false positives accurately and at scale, we combined the opioids and non-opioids from the value sets described above into a single dataset. As such, the initial modeled problem during system development is actually harder from the system's perspective than the target problem of finding potential errors from a single value set. The negative examples included represent a diverse set of medications, ensuring that the system is capable of handling a large set of false positive patterns. Specifically, the objective of our method and intrinsic evaluation is to accurately separate a set of opioid medications from a set of non-opioid medications, some of which may be semantically similar to the opioids.

We developed a data-driven iterative pattern generation algorithm that commences the pattern generation process with the negative and positive training sets. The algorithm first generates character-level *n*-grams (*i.e.*, the contiguous sequence of ***n*** characters within the medical description) with a predetermined *n*-gram range from both the positive and negative instances of the training set. We determined the *n*-gram range empirically via trial and error using the training set. The *n*-gram range we experimented with ranged from *3* to *10* characters. We extracted the character *n*-grams from the positive set (*i.e.*, opioids) and only kept those that did not appear in the negative set. This gave us a list of character *n*-gram sequences that were uniquely associated with the positive set.

Following the curation of the character *n*-grams, we iterated through them from high to low frequency on the positive set. We generated a regular expression pattern based on each *n*-gram pattern and updated the regular expressions at each iteration. The updated regular expression set was then tested against the entire training set, the precision and recall scores were calculated, and used as a criterion for the stopping condition. If both the precision and recall score met a predetermined cutoff or the number of iterations met a predetermined maximum number, the procedure stopped. The number of iterations we experimented with ranged from *100* to *500*. To classify each sample in the training set, we applied each pattern in the regular expression set and checked if it matched the medication's lexical description (name or ingredients (IN) list, as presented in the value set). The lexical description was marked as a positive sample as long as there was at least one regular expression pattern that matched with the description. All

```
T = all samples in the training set
PS = positive samples in the training set labeled as opioids medications
NS = negative samples in the training set labeled as non-opioids medications
MP = minimum acceptable precision score for regular expression generation
MR = minimum acceptable recall score for regular expression generation
ME = maximum number of epochs

Regular_Expression_Generation (PS, NS, MP, MR, ME):
        regex_set = {}                          /* initialize the regular expression set */
        P = generate_ngram(PS)                  /* generate ngram patterns from the positive samples */
        N = generate_ngram(NS)                  /* generate ngram patterns from the negative samples */
        P' = filter(P, N)                       /* filter the ngram patterns generated from the positive samples */

iterations = 0                                  /* initialize the number of iterations */
for each p in P'
        iterations  = iterations + 1
        R = generate_regex(p)                   /* generate a regular expression pattern from the ngram pattern */
        regex_set.add(R)                        /* add the expression to the set */
        precision, recall = apply_regex(T, regex_set)       /* apply the regular expression set to classify */
        if_break = check(precision, MP, recall, MR, epochs, ME) /* check for stopping condition */
```

Fig. 2 Pseudo-code of the regular expression generation algorithm described in the paper.

patterns above the cutoff threshold were kept while the rest were discarded. The pseudo-code for the algorithm is presented in ►Fig. 2.

Evaluations

We performed two levels of evaluations of the system—intrinsic and extrinsic. For the intrinsic evaluations, which were performed iteratively during development, we used two opioid value sets combined with four non-opioid value sets. Using these, we evaluated the model via stratified fivefold and Monte Carlo cross-validations.[7] We computed the precision, recall, and F_1-score for the system for each fold of cross-validation. Since the disagreements between the system predictions and the value set classification were often due to errors in the value set (e.g., opioids appearing in the non-opioid value sets), we performed training and evaluation before and after correcting all errors in these datasets.

For extrinsic evaluation, we used 20 opioid value sets, which were not part of the training/development data. Our goal was to assess how well the system performed when applied to unseen, real-world opioid value sets. Since these value sets are not supposed to contain negative samples, assuming they are 100% correct, we evaluated the system based on recall and manually reviewed entries that were not detected by the system. This evaluation, thus, mirrored the application of the system in a real-world setting. The initial training and development sets were manually corrected before performing the extrinsic evaluations.

Error Analysis

We conducted error analyses in multiple rounds—for the intrinsic and the extrinsic evaluations. These manual analyses were conducted to (1) study the causes of classification errors, and (2) verify whether the false positive and false negative cases were system misclassifications or dataset errors. For (1) we manually reviewed a sample of the misclassified medication names/descriptions to identify potential patterns of repeated errors. For (2) we scanned through the system outputs and manually assessed if an entry undetected by the system was actually an opioid.

Results

Datasets

For the system development and intrinsic evaluation, we used two opioid value sets with 1,532 entries and four non-opioid value sets with 3,006 entries in total. We found 214 repeated concepts in the opioid value sets and 57 repeated concepts in the non-opioid value sets, indicating value set redundancy. All value sets except for one (#3 in ►Table 1) contain clinical dose forms (e.g., acetaminophen 325 mg/oxycodone hydrochloride 10 mg oral tablet; acetaminophen 325 mg/oxycodone hydrochloride 5 mg oral tablet). Importantly, we found 29 entries that occurred in both the opioid and non-opioid value sets, indicating potential value set incorrectness.

Classification Results

►Fig. 3 shows the recall and precision obtained by the system in each cycle of the fivefold cross validation. The system performance over the fivefold cross validations was fairly consistent: precision consistently higher than recall, average F_1-score of 0.912 and average precision, recall, and overall accuracy of 0.934, 0.906, and 0.906, respectively. For

Table 1 System performance over 20 opioid value sets. The table shows the number of samples in the dataset and the system recall (assuming the value set is completely correct)

Dataset	Entries	Prediction recall
1	67	1.00
2	17	1.00
3	182	0.995
4	106	0.981
5	96	0.980
6	198	0.975
7	392	0.960
8	688	0.956
9	688	0.956
10	933	0.955
11	252	0.952
12	714	0.948
13	343	0.930
14	476	0.887
15	471	0.885
16	320	0.869
17	1579	0.855
18	577	0.827
19	352	0.478
20	42	0.452

50 iterations of Monte Carlo cross validation, the average precision, average recall, average F_1-score, and average overall accuracy were 0.932, 0.904, 0.909, and 0.904, respectively. ►**Table 2** presents some examples of system-generated *n*-gram patterns.

Since some of the classification errors made by the system were not actually system errors, but dataset errors (see: error analysis results), we repeated the intrinsic evaluations after manually correcting errors in the opioid and non-opioid value sets. The system performance improved considerably following the value set correction: average precision, recall, F_1-score, and overall accuracy for fivefold cross validation increased to 0.956, 0.953, 0.953, and 0.953. The corresponding average precision, average recall, average F_1-score, and average overall accuracy for 50 iterations Monte Carlo cross validation increased to 0.958, 0.953, 0.953, and 0.953.

Extrinsic Evaluation Results

We used 20 opioid value sets with a total of 8,063 entries for the extrinsic evaluation, the largest had 1,579 entries and the smallest had 17 entries. The average recall of the classification performance was 0.89 with, ranging between 0.452 and 1.0. Only value sets #19 and #20 had a recall lower than 0.85. Absence of clinical dose forms (#3) did not appear to affect performance. Further details about the value sets are provided in the ►**Supplementary Tables S1** and **S2, available in the online version only.**

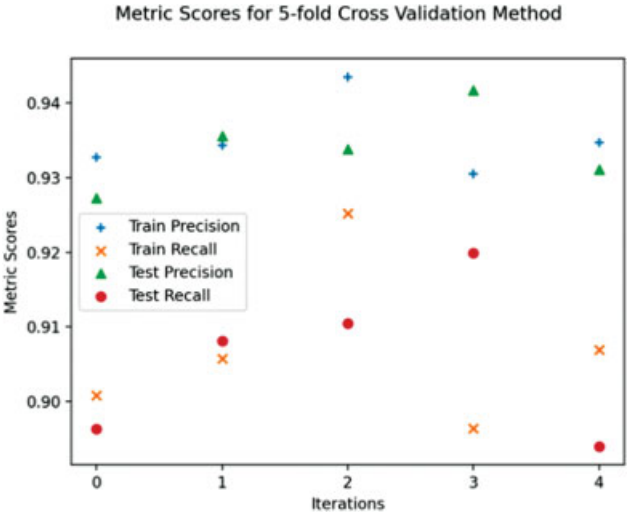

Fig. 3 Performance of the classification algorithm with fivefold cross-validation in terms of train and test precision and recall.

Error Analysis Results

Intrinsic Evaluations

During the development and intrinsic evaluation of the system, we collected the false positive and false negative samples from each of the 50 iterations of Monte Carlo cross validation results and randomly selected one sample for error analysis. We used an expert knowledge base created by the American Society of Addiction Medicine[8] (List S1 of ►**Supplementary Material**, available in the online version only) and advise from a physician to manually identify "evidence of opioids" in the false negative samples and "evidence of non-opioids" in the false positive samples. Ninety-four false negatives were actually opioid medications and four false positives non-opioid medications.

According to our classification algorithm, a medical description is only classified as an opioid medication as long as there is at least one regular expression pattern generated from the opioid *n*-grams that matches with the description. Similarly, a medical description is classified as a non-opioid medication only if it does not contain any of the *n*-gram patterns generated from the positive training samples. Therefore, contamination of the initial dataset with incorrect negative examples (*i.e.*, opioids appearing in the non-opioid datasets) caused the system to reject the patterns emerging from them. For example, in one sample of the intrinsic error analyses, we found that 53% of the false negative samples contained the keyword "Hydrocodone" (*e.g.*, hydrocodone/ibuprofen oral tablet [Ibudone]), meaning that the algorithm learned the pattern from the negative samples. We thus hypothesized that some positive samples were incorrectly labeled as negative because of value set errors. We validated our hypothesis by checking if the negative samples in the training set contained any of the "evidence of opioids" from the list above. Out of all 2,928 negative samples in the dataset, we identified 209 samples that were incorrectly labeled as non-opioid medications, resulting in the system errors.

Table 2 Example of character n-grams and the opioid medications that they identified according to the result of intrinsic evaluation

N-grams	Opioid medications that they identified
morp	Duramorph morphine oral liquid product morphine sulfate 60 mg extended release oral capsule Abuse-Deterrent 12 hours morphine sulfate 30 mg extended release oral tablet [Morphabond] hydromorphone hydrochloride 8 mg oral tablet [Dilaudid]
odon	Hydrocodone Roxicodone pill acetaminophen/hydrocodone oral solution oxycodone hydrochloride 5 mg oral capsule Abuse-deterrent 12 HR oxycodone 36 mg extended release oral capsule [Xtampza]
fentany	Fentanyl fentanyl 0.2 mg oral lozenge fentanyl 0.2 mg oral lozenge [Actiq] 72 HR fentanyl 0.0625 MG/HR transdermal system {2 (fentanyl 0.8 MG/ACTUAT mucosal spray [Subsys]) } pack [Subsys 1600 mcg]
orpho	Oxymorphone oral tablet hydromorphone oral liquid product 1 mL hydromorphone hydrochloride 10 mg/mL injection 24 HR hydromorphone hydrochloride 12 mg extended release oral tablet [Exalgo] hydromorphone hydrochloride 3 mg rectal suppository
enorp	Buprenorphine 0.7 mg 1 mL buprenorphine 0.3 mg/mL cartridge 0.5 mL buprenorphine 200 mg/mL prefilled syringe [Sublocade] 168 HR buprenorphine 0.005 mg/HR transdermal system

Table 3 The composition of value sets #19 and #20

Dataset	Entries	Brand name (BN)	Ingredient (IN)	Multiple ingredient (MIN)	Precise ingredient (PI)	Semantic branded drug form (SBDF)	Semantic branded dose form group (SBDG)	Semantic clinical drug (SCD)
19	352	58	42	192	51	4	4	1
20	42	26	15	0	1	0	0	0

Note: Examples of each ingredient type are provided in ►**Supplementary Table S3**, available in the online version only.

Extrinsic Evaluations

To analyze the system errors in the extrinsic evaluation phase, we focused on the two value sets that yielded low recall (#19 and #20 from ►**Table 1**). See ►**Table 3** for details on the composition of these value sets. For value set #19, the system failed to detect 185 entries and we verified that 159 of these were opioids. Within the false negatives, there were 80 multiple ingredients (MINs), 49 brand names (BNs), 19 precise ingredients (PIs), and 11 INs. For #20 value set, out of the 23 false negative samples, there were 22 BNs, and one IN.

The error analyses revealed that the errors in the system were primarily caused by two description types: BNs and MINs. This is not surprising since the training datasets did not contain a comprehensive list of BNs, as shown in ►**Fig. 4**. MINs also rarely appeared (less than 2%) in the other opioid value sets. Since 54% of the samples in the second dataset were labeled as MIN, meaning that two or more INs appeared together in a single drug preparation, this list of opioids was considerably different in focus and broader in scope than the training dataset.

Discussion

Significance

The proposed approach has the potential to assist with quality assurance of value sets. In terms of correctness, it reduces subject matter expert engagement; only false negatives concepts need to be reviewed. Standardized, up-to-date, reusable, correct, complete, and non-redundant clinical definitions are needed to support HIT initiatives. For instance, reliable opioid value sets could help with the implementation of PDMPs, used to detect “doctor shopping” and target the opioid epidemic.

The design of the system has several advantages that are lucrative for real-life use. The system is data-centric; it learns from a given value set without requiring human experts to

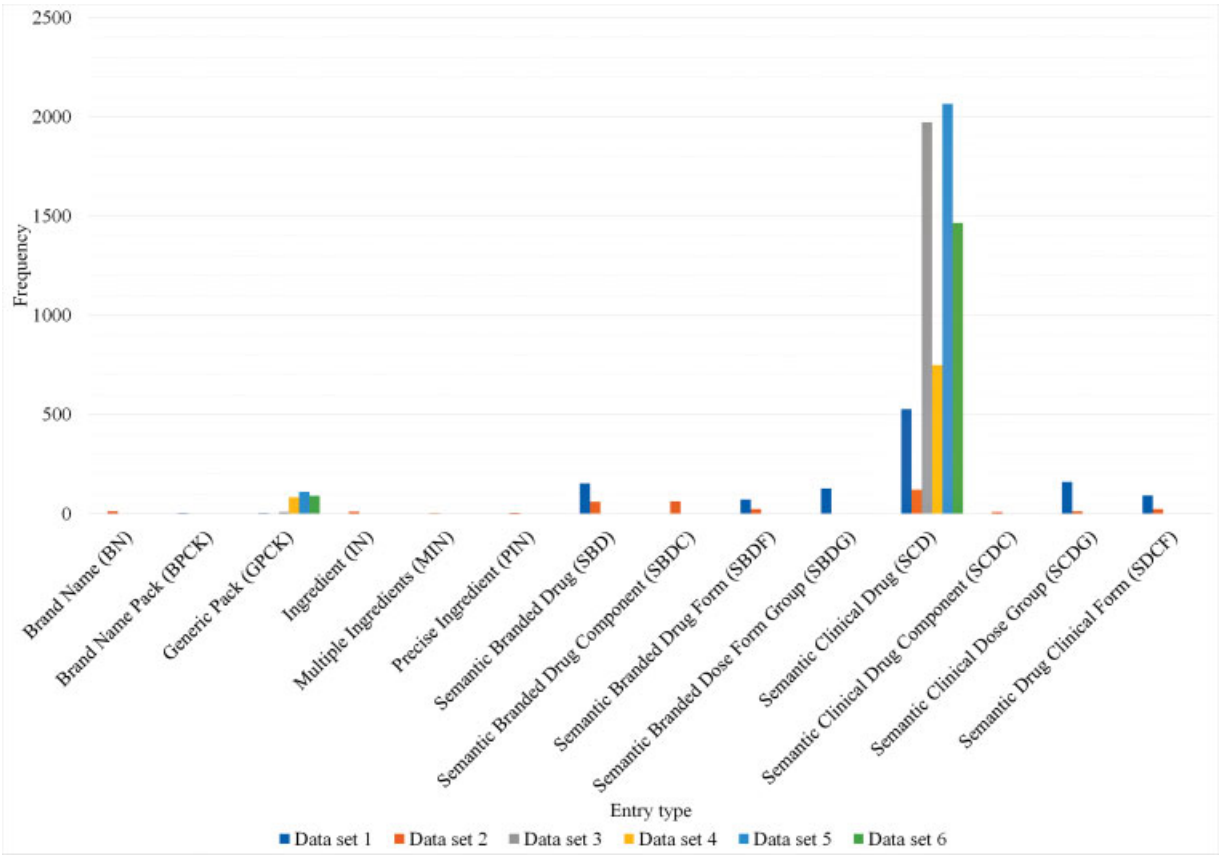

Fig. 4 Distribution of entity types in the training dataset. The vast majority were SCDs with few to no BNs and MINs.

manually specify rules. Therefore, while our use case was for opioids, the system can potentially be reused without modifications on other classes of medications (e.g., benzodiazepines). This design increases the reusability of the system and ensures that it is domain agnostic. The optimization of the system is done by regulating the threshold during training and validation. This adds further customizability to the system as it can be fine-tuned for precision or recall, as required, and also fine-tuned for diverse value sets.

In addition to being reusable and customizable, the system is intentionally simple and not resource intensive. While recent progress in NLP has been driven largely by advances in deep neural networks, such approaches are extremely resource intensive, requiring substantial expertise to configure and execute. In contrast, our proposed method requires very little system resources and can essentially run on any computer with minimal processing power. In addition, the simplicity of the system means that it can easily be configured, if required, without requiring expert programming knowledge.

Related Work

While our approach to value set correctness determination is novel, there has been a considerable amount of related past work in the broader field of biomedical NLP. Broadly speaking, our work falls under the umbrella of text classification, where the objective is to correctly classify instances occurring in a value set in a binary manner—*correct* or *incorrect*. To the best of our knowledge, no prior work has attempted to automate this task, but related works have attempted to generate patterns in a similar manner from text.

Numerous studies have used NLP-based approaches to detect patterns in texts, such as for computational phenotyping from electronic health records,[9] detecting and disambiguating geographical entities,[10] finding measurements in radiology narratives[11] and detecting demographic information such as age and gender from patient notes.[12] While many studies employed manually-crafted patterns, others focused on creating specialized lexicons for entity recognition and extraction from noisy free texts, such as those from electronic health records and social media.[13–15] A major drawback of lexicon-based approaches is that lexicons are static and new lexicons need to be built for every new problem domain. Meanwhile, for a task such as ours, a traditional machine learning-based sequence labeling approach, such as conditional random fields[16] or recurrent neural networks,[17] is not suitable since the character-level patterns need to be detected from within short text segments that do not have any additional context. Plus, the number of example instances in a value set may also be small.

Perhaps the approach most similar to ours was proposed by Bui and Zeng-Treitler,[18] who developed an algorithm that followed a bottom-up approach. They first identified key tokens using text alignments between pairs of phrases and built regular expression patterns based on the key tokens with distance controls between the phrases. Each newly

generated regular expression was tested against the rest of the training set and filtered based on the precision threshold. This study itself built on earlier similar studies that employed regular expression-based pattern matching.[19–22]

Limitations and Future Work

While promising, the reported performance of the system is specific for opioids. Performance may be reduced for certain categories of medications that do not necessarily have learnable patterns, although significant performance drops are unlikely. The system also relies on the assumptions that there are value sets with true positive (opioids) and false negatives (non-opioids) available and that the value sets are (1) mostly accurate and (2) contain sufficient numbers of correct examples. While that is true for most value sets, the system performances are likely to drop for noisy value sets and value sets with few examples. The system is also limited to curated value sets only, and its performance is likely to be reduced if applied to informal datasets (e.g., free text sources that contain spelling variants or misspellings). The system may also underperform if applied to value sets containing trade names or BNs only, since trade names may not follow distinguishable patterns.

Our immediate future work will include further testing the proposed approach with existing VSAC value sets corresponding to other medication categories (e.g., benzodiazepines). Additionally, further improvements to the system could be achieved if patterns could be learned using lower numbers of examples, and we will explore possible mechanisms to add such functionality. Finally, the effectiveness of the data-centric approach will be compared with existing terminology-based techniques currently used for value set quality assurance.[1]

Conclusion

We proposed a novel and promising data-centric, reusable, customizable, simple and not resource intensive NLP approach to improve value set quality assurance. When we evaluated the system with opioid value sets, it helped to automate error analysis, assess value set correctness, and reduce domain expert engagement. Future work will involve testing with other medication value sets.

Author Contributions

All authors made substantial contributions to manuscript revisions and approved the final version. A.S. and A.G. contributed to the design of the study. S.L. conducted data analysis under the mentorship of A.S.. A.S. and A.G. supervised the conception, design, and revision of the manuscript. All authors also agree to be accountable for the accuracy and integrity of the work presented here

Funding

A.G.'s efforts were funded by the National Institute of Mental Health through the "*My Data Choices, evaluation of effective consent strategies for patients with behavioral health conditions*" (R01 MH108992) grant. A.S.'s efforts were funded by the National Institute on Drug Abuse through the "*Mining Social Media Big Data for Toxicovigilance: Automating the Monitoring of Prescription Medication Abuse via Natural Language Processing and Machine Learning Methods*" (R01 DA046619) grant.

Conflict of Interest

None declared.

References

1 Fung KW, Xu J, Gold S. The Use of Inter-terminology Maps for the Creation and Maintenance of Value Sets. Paper presented at: AMIA Annu Symp Proc. Vol 2019. American Medical Informatics Association; Accessed June 8, 2021. /pmc/articles/PMC71531322019: 438–447
2 Winnenburg R, Bodenreider O. Metrics for assessing the quality of value sets in clinical quality measures. Paper presented at: AMIA Annu Symp Proc. Vol 2013. American Medical Informatics Association; 2013:1497–1505
3 Bodenreider O. Title: Criteria and Metrics for Assessing the Quality of SNOMED CT Value Sets in Clinical Quality Measures. Vol. 2012; 2012. Accessed June 8, 2021 at: http://www.ihtsdo.org/fileadmin/user_upload/doc/showcase/show13/
4 Winnenburg R, Bodenreider O. Issues in creating and maintaining value sets for clinical quality measures. Paper presented at: AMIA Annu Symp Proc. Vol 2012. American Medical Informatics Association; Accessed June 8, 2021. /pmc/articles/PMC35405852012: 988–996
5 National Survey on Drug Use and Health (NSDUH). 2019. Results from the 2018 National Survey on Drug Use and Health. Accessed November 21, 2021 at: https://www.samhsa.gov/data/report/
6 Ponnapalli A, Grando A, Murcko A, Wertheim P. Systematic Literature Review of Prescription Drug Monitoring Programs. Paper presented at: AMIA Annu Symp Proc. Vol 2018. American Medical Informatics Association; Accessed June 8, 2021. /pmc/articles/PMC63712702018:1478–1487
7 Xu QS, Liang YZ. Monte Carlo cross validation. Chemom Intell Lab Syst 2001;56(01):1–11
8 American Society of Addiction Medicine (ASAM) Opioids: Brand names, generic names & street names. Published August 28, 2017. Accessed October 10, 2021 at: https://www.asam.org/docs/default-source/education-docs/opioid-names_generic-brand-street_it-matttrs_8-28-17.pdf?sfvrsn=7b0640c2_2
9 Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Trans Comput Biol Bioinformatics 2019;16(01):139–153
10 Nesi P, Pantaleo G, Tenti M. Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering. Eng Appl Artif Intell 2016;51:202–211
11 Sevenster M, Buurman J, Liu P, Peters JF, Chang PJ. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2015;6(03):600–110
12 Sarker A, Klein AZ, Mee J, Harik P, Gonzalez-Hernandez G. An interpretable natural language processing system for written medical examination assessment. J Biomed Inform 2019;98:103268
13 Arnoux-Guenegou A, Girardeau Y, Chen X, et al. The adverse drug reactions from patient reports in social media project: protocol for an evaluation against a gold standard. JMIR Res Protoc 2019;8 (05):e11448
14 Pérez-Pérez M, Pérez-Rodríguez G, Fdez-Riverola F, Lourenço A. Using twitter to understand the human bowel disease community: exploratory analysis of key topics. J Med Internet Res 2019;21 (08):e12610

15 Hostetter J, Wang K, Siegel E, Durack J, Morrison JJ. Using standardized lexicons for report template validation with LexMap, a web-based application. J Digit Imaging 2015;28(03): 309–314
16 Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Dep Pap. Published online June 28, 2001. Accessed May 28, 2020 at: https://repository.upenn.edu/cis_papers/159
17 Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–2681
18 Bui DDA, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc 2014;21(05):850–857
19 Frenz CM. Deafness mutation mining using regular expression based pattern matching. BMC Med Inform Decis Mak 2007;7:32
20 Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(05):301–310
21 Brauer F, Rieger R, Mocan A, Barczynski WM. Enabling information extraction by inference of regular expressions from sample entities. Paper presented at: Int Conf Inf Knowl Manag Proc; 2011
22 Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish HV. Regular Expression Learning for Information Extraction. Paper presented at: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing; 2008:21–30