Natural Language Processing tool for extraction of patient-reported outcomes from a national multi-EHR registry

Marie Humbert-Droz[1], Zara Izadi[2], Gabriela Schmajuk[2], Jinoos Yazdany[2] and Suzanne Tamang[1]

[1]Stanford University School of Medicine, Stanford, California
[2]University of California, San Francisco, San Francisco, California

Abstract
Patient-reported outcomes (PRO) are important measures of quality of life for chronically ill patients. These indicators are often difficult to access because they are documented in the clinical notes. We propose a natural language processing pipeline to extract PROs developed using data from a national multi-electronic health record (EHR) registry. Our approach is rule-based and general enough to successfully extract PROs from notes coming from more than 100 practices using more than 20 different EHR brands, demonstrating a great generalizability and potential for transportability.

Keywords
Patient-reported outcome, Clinical natural language processing, Health information systems,

1. Introduction

Patient-reported outcomes (PROs) are important measures of quality of life for chronically ill patients, population health management or clinical decision support for example. These measures are found in the clinical notes, making their access for disease tracking, population-health management and quality reporting challenging.

PROs are especially important in rheumatoid arthritis (RA), a leading cause of disability in the United States and many other countries. Functional status assessments are important outcome measure in RA clinical trials and have been shown to be strong predictors of future disability and mortality[1]. In addition, PROs have been successfully used to inform evidence-based treatment decisions and engage the patient in decision-making, facilitating more patient-centered care[2, 3].

We are using data from the Rheumatology Informatics System for Effectiveness (RISE). RISE is a national EHR-enabled registry that passively collects data from EHRs of participating practices. It has been developed by the American College of Rheumatologists (ACR) to build digital research infrastructure nationally and facilitate quality reporting. The registry passively extracts data from practices, aggregates and analyses it centrally and feeds the information back to clinicians in the form of a dashboard, summarizing performance. The passive extraction of data is key here, as it allows practices to contribute EHR data that is collected during the course of routine clinical care, thereby minimizing impact on their workflow[4, 5].

The data collected by RISE allows the calculation of performance on specific quality measures, such as assessment of disease activity (DA) and functional status (FS), for rheumatoid arthritis patients. These two quality measures are often not available from coded data resources and obtained using standardized and validated questionnaires about their quality of life. These questionnaires are referred to as patient reported outcome (PRO) instruments, and can be for general health or specific to a disease or condition. In the case of RA patients, there is a multitude of different instruments measuring disease activity or functional status. Two commonly used measures of disease activity are the Routine Assessment of Patient Index-3 (RAPID-3) and Clinical Disease Activity Index (CDAI) questionnaires[6]. Functional status assessment is obtained with the Health Assessment Questionnaire (HAQ) or one of its versions[1]. Each instrument has a number of components, that are then added to yield the total score. For example, the CDAI instrument has four components: swollen joints counts, tender joints counts, patient global assessment and physician global assessment. The sum of each score yields the total CDAI score (see Figure 3c. for an example).

Management of patients with RA focus on early intervention and a treat-to-target goal of remission or low disease activity[7]. This strategy has been shown effective in reducing disease activity and joint damage but does not account for the patient's well-being improvement[8]. The use of patient reported outcomes (PROs) allow to

AI4Function 2021
mhdroz@stanford.edu (M. Humbert-Droz); zara.izadi@ucsf.edu (Z. Izadi); gabriela.schmajuk@ucsf.edu (G. Schmajuk); jinoos.yazdany@ucsf.edu (J. Yazdany); stamang@stanford.edu (S. Tamang)

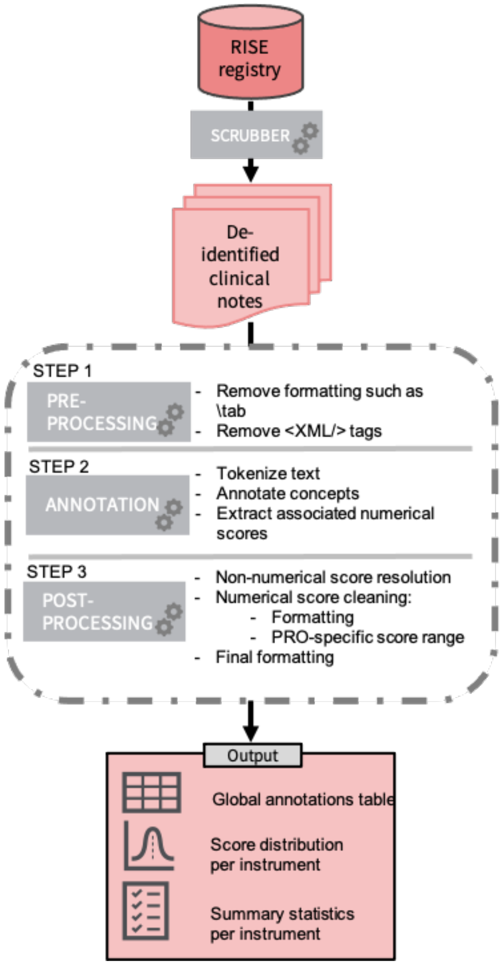

Figure 1: Pipeline description

account for the patient's perspective in the assessment of disease impact and disease activity, with the goal of providing patient-centered care, ultimately improving outcomes. Collecting PROs provides useful information about the impact of RA on a patient's quality of life and monitor the effects of interventions.

2. Materials and methods

We received de-identified notes from RISE, where mentions of personally identifiable information such as names, dates or ZIP codes for example, were automatically removed in compliance with the Health Insurance Portability and Accountability Act (HIPAA)[9]. We processed notes from January 1 2015 to December 30 2018, coming from 158 practices and 24 EHR brands. Earlier notes (January 1 2012 to December 31 2014) were used for the pipeline development. Over 34 million notes for 854,628 patients were processed through the pipeline.

2.1. Challenges for developing a rule-based NLP system with multiple EHR products

Due to the large number of different practices and EHR brands used for the development of this pipeline, one of the initial challenges was to harmonize the raw notes. Many raw notes contained formatting and XML tags, that introduce a significant amount of noise and artificial artefacts in the notes, making it difficult to develop a rule-based system to extract PRO instruments and their associated scores. Different practices may also have different styles and templates for their notes. Some may have a semi-structured template for reporting test results for example. Some may report PROs differently. For example, for composite assessments like CDAI, some providers may report the detailed results for each component before reporting the total score (Figure 3c.). Other providers may report the score interpretation with (Figure 3a.) or without (Figure 3e.) the numerical score.

Finally, the de-identification process also led to challenges for our NLP tool development. Some instrument names and/or scores as well as dates have been removed from the original notes.

2.2. Pipeline description

The PRO and associated scores extraction pipeline (Figure 1) is constituted of three steps. The first step of is a text cleaning step. The second step is the annotation step, where the concepts of interest and scores are annotated. Finally, the post-processing step is a succession of cleaning functions that format the mentions and performs the score resolution.

The pre-processing of the raw text to harmonize the notes and clean out the text of any XML or formatting markup (see Figure 2 for an illustrative example). The formatting markup is cleaned using regular expressions and the text is extracted from the XML markup using the BeautifulSoup package[10]. Next, the note is tokenized using a spacy pipeline [11] (using the en_core_sci_md language model by scispacy [12]). A terminology curated by domain experts is used as the source for a string matching module to annotate PRO mentions and non-numerical scores. Sometimes, the score resulting from the questionnaire is not documented in the note. Instead, the clinical interpretation of the disease activity or functional status is. This interpretation is what we refer to

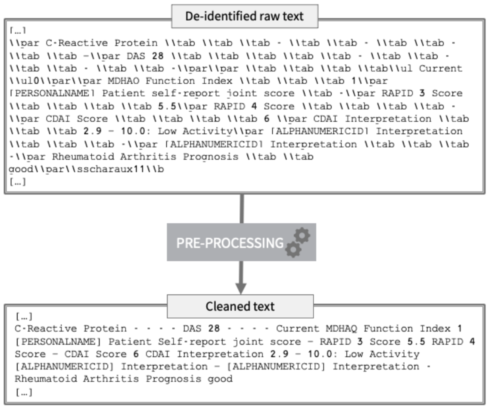

Figure 2: Snippet examples for the pre-processing step

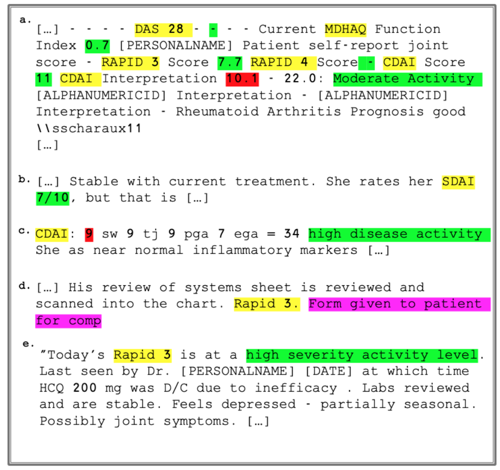

Figure 3: Snippet examples for some challenging extractions. Instruments names are over-lined in yellow. When the score is picked up correctly, it is over-lined in green, in red when incorrect. Finally, the purple over-line shows an example where there is mention that a form was given to the patient for completion, but no score was recorded.

as non-numerical score (see Figure 3e. for an example). Finally, a set of rules to identify numbers in the proximity of the PRO mention is used to resolve the score associated with the mention. The rules are fairly simple and identify numbers after the instrument mention, up to the end of the sentence.

Finally, the extractions are cleaned out in the post-processing step. This last step has three major components. First, non-numerical scores that were detected by the matching module are associated with their respective PRO mention. This part is done associating PRO mentions and non-numerical scores in the same sentence, in close distance to each other. Then, scores that are documented as a fraction, representing SCORE/SCALE (see Figure 3b. for an example) are resolved. Finally, some minor formatting errors are resolved and scores outside the PRO-specific range are discarded.

The final output contains global annotation tables, ready for post-analysis, score distribution plots and a summary statistics for numerical scores only.

3. Results

Figure 3 shows some examples of the annotation output. Instrument names are in yellow, the associated scores are in green (when correct) or red (when incorrect). In Figure 3a., most scores are picked up correctly, except for the second CDAI mention where the lower bound of the range is picked up. Another challenging part of extracting scores associated with a PRO instrument is exemplified in Figure 3c. where the score picked up is from a component of the CDAI instrument (swollen joints).

3.1. Performance assessment

100 mentions from 48 randomly selected notes were used for manual chart review of disease activity and functional status assessment. The performance was evaluated at the mention level (for score extraction). Precision, sensitivity, specificity were evaluated for each mention-score pair.

Out of 59 mentions associated with a score, 55 were correct and 33 out of 41 mentions without associated score in the note were correctly extracted. This results to a precision of 87.30%, recall of 93.22% and F score of 90.16%. The overall accuracy of the system is 88.00%.

We are continuing to extend our manual chart review to 100 documents.

3.2. Statistics of NLP extractions

After processing through the extraction pipeline, notes from 95 practices yielded PRO extractions. The raw notes came from 14 different EHR vendors. Table 1 shows a summary statistics of all the different instruments extracted from the notes. RAPID-3 and CDAI were the most extracted disease activity instruments and MDHAQ the most extracted functional status assessment instrument, reflecting the general popularity of these measures. The *result range* column of the table shows the range of the extracted scores, providing a sanity check with respect to the validity of the score resolution for each instrument.

Table 1
Extraction statistics per instrument

Instrument	# assessments	# patients	# practice	Date range	Result range	Theoretical range
			Disease Activity			
RAPID-3	285,391	29,997	54	2015-01-02 - 2018-12-28	0.0 - 30.0	0.0 - 30.0
CDAI	165,181	20,922	39	2015-01-02 - 2018-07-10	0.0 - 75.0	0.0 - 76.0
DAS28	11,985	2,242	21	2015-01-02 - 2018-06-28	0.0 - 10.0	0.0 - 9.4
SDAI	2,620	619	4	2015-01-05 - 2018-06-27	0.0 - 83.1	0.0 - 86.0
			Functional Status			
MDHAQ	31,821	4,413	8	2015-09-25 - 2018-06-29	0.0 - 3.0	0.0 - 3.0
HAQ	9,609	4,015	15	2015-01-02 - 2018-06-28	0.0 - 3.0	0.0 - 3.0
MHAQ	18,812	3,328	5	2015-01-05 - 2018-12-13	0.0 - 3.0	0.0 - 3.0
HAQ II	455	189	4	2017-02-02 - 2018-05-30	0.0 - 2.7	0.0 - 3.0

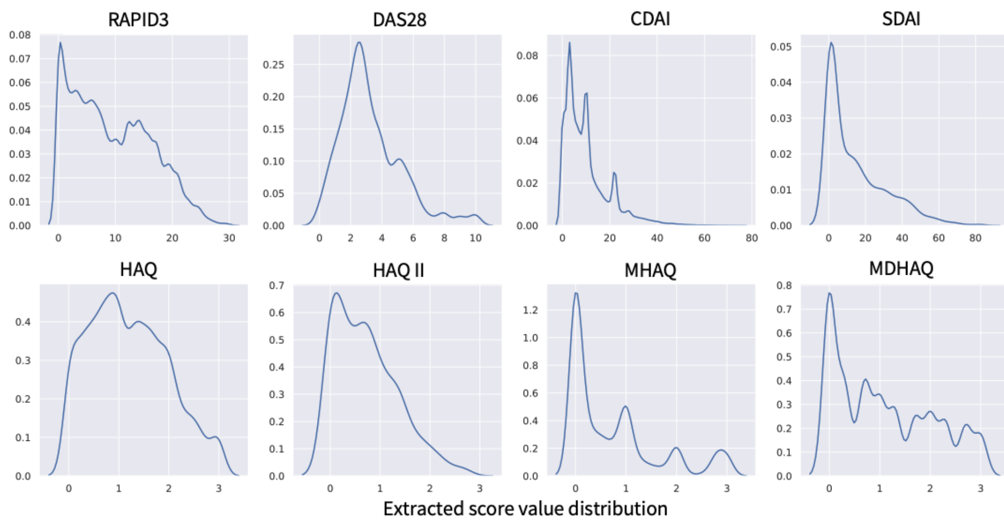

Figure 4: Score distributions for all considered instruments

Indeed, the extracted ranges fall within normal range for each instrument. Figure 4 shows the distribution of the scores for each instrument. The scores extracted are within the theoretical range for each measure and the distribution is as expected. Both Table 1 and Figure 4 demonstrate the feasibility of extracting PRO instruments and scores with high fidelity.

4. Discussion

Our work demonstrated a framework for developing a new tool to extract information related to PRO measurement that are often absent form healthcare data in EMRs. For example, our system extracts self-reported disease activity measures, like the RAPID-3 questionnaire, which provides disease progression information and highlights patients' pain and overall perception of their condition. Functional status assessment, like the HAQ for example, can also be self-administered and measures the patients' physical functioning through 3 patient-centered domains (disability, pain, global health). The questionnaire is very responsive to change and comorbidities, healthcare resource use or need for social support measures can be predicted by the HAQ score. Such information can be used for healthcare performance and quality measurement. Paired with a predictive layer, it may also have the ability to be used proactively within clinical decision support systems.

Our system achieved good fidelity for PRO measurement and score extraction. The documenting of disease activity or functional status assessment is usually done in the clinical notes, thus in a format not readily available for analysis. Our tool that tags notes for the mention of various PRO instruments and extracts both the mention of the instrument and its score (if present). Our tool is rule-based and generic enough to successfully extract mentions and scores throughout a variety of practices (> 100) and EHR brands (> 20). Although more sophisticated algorithms may also be used to achieve the same goal, the simplicity of our approach is attractive. It also provides an interpretable baseline approach for comparison with more sophisticated machine and deep learning methods that may have the potential to improve performance.

We found key types of systematic errors that were attributed to our system, such as the extraction of the lower bound of the interpretation scale (Figure 3a.) or the extraction of the instrument component instead of the total score (Figure 3c.) could be removed. The variation of documentation styles among the different practices might have contributed to the occurrence of such type of errors. Since the pipeline is based off of expert-curated terminology, the extent of the PROs extracted can be easily extended. Finally, because the tool was developed to extract information successfully from multiple sources, it is likely to generalize to other practices or inpatient setting for example.

Future and ongoing work includes an extension of the chart review to 100 documents rather than 100 mentions, comparison of the NLP-extracted scores with structured data for the practices for which such information is available, as well as a an analysis of the performance per EHR brand and practice type. Finally, we plan to explore the feasibility of computing ACR-endorsed quality measures from the data extracted with our tool in the hopes of improving the current standard

References

[1] L. Maska, J. Anderson, K. Michaud, Measures of functional status and quality of life in rheumatoid arthritis: Health assessment questionnaire disability index (haq), modified health assessment questionnaire (mhaq), multidimensional health assessment questionnaire (mdhaq), health assessment questionnaire ii (haq-ii), improved health assessment questionnaire (improved haq), and rheumatoid arthritis quality of life (raqol), Arthritis Care & Research 63 (2011) S4–S13. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acr.20620. doi:https://doi.org/10.1002/acr.20620.

[2] O. Malysheva, A. Bedrich, J. Kuipers, H. Kleine, B. Wolff, C. Baerwald, Use of clinical scores to guide therapeutic de cisions in patients with rheumatoid arthritis in daily care., Clinical and experimental rheumatology 33 (2015) 255–258.

[3] E. D. Newman, V. Lerch, J. Billet, A. Berger, H. L. Kirchner, Improving the quality of care of patients with rheumatic disease using patient-centric electronic redesign software, Arthritis Care & Research 67 (2015) 546–553.

[4] J. Yazdany, N. Bansback, M. Clowse, D. Collier, K. Law, K. P. Liao, K. Michaud, E. M. Morgan, J. C. Oates, C. Orozco, A. Reimold, J. F. Simard, R. Myslinski, S. Kazi, Rheumatology informatics system for effectiveness: A national informatics-enabled registry for quality improvement, Arthritis Care & Research 68 (2016) 1866–1873. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acr.23089. doi:https://doi.org/10.1002/acr.23089.

[5] C. Tonner, G. Schmajuk, J. Yazdany, A new era of quality measurement in rheumatology: electronic clinical quality measures and national registries, Current Opinion in Rheumatology 29 (2017) 131–137.

[6] K. Anderson, Jaclyn, L. Zimmerman, L. Caplan, K. Michaud, Measures of rheumatoid arthritis disease activity patient (ptga) and provider (prga) global assessment of disease activity, disease activity score (das) and disease activity score with 28-joint counts (das28), simplified disease activity index (sdai), clinical disease activity index (cdai), patient activity score (pas) and patient activity score-ii (pasii), routine assessment of patient index data (rapid), rheumatoid arthritis disease activity index (radai) and rheumatoid arthritis disease activity index-5 (radai-5), chronic arthritis systemic index (casi), patient-based disease activity score with esr (pdas1) and patient-based disease activity score without esr (pdas2), and mean overall index for rheumatoid arthritis (moi-ra), Arthritis Care & Research 63 (2011) S14–S36.

[7] J. S. Smolen, D. Aletaha, J. W. J. Bijlsma, F. C. Breedveld, D. Boumpas, G. Burmester, B. Combe, M. Cutolo, M. de Wit, M. Dougados, P. Emery, A. Gibofsky, J. J. Gomez-Reino, B. Haraoui, J. Kalden, E. C. Keystone, T. K. Kvien, I. McInnes, E. Martin-Mola, C. Montecucco, M. Schoels, D. van der Heijde, Treating rheumatoid arthritis to target: recommendations of an international task force, Annals of the Rheumatic Diseases 69 (2010) 631–637. URL: https://ard.bmj.com/content/69/4/631. doi:10.1136/ard.2009.123919.

[8] B. Fautrel, R. Alten, B. Kirkham, I. de la Torre, F. Durand, J. Barry, T. Holzkaemper, W. Fakhouri, P. C. Taylor, Call for action: how to improve use of patient-reported outcomes to guide clinical decision making in rheumatoid arthritis, Rheumatology

Internaltional 38 (2018) 935–947.

[9] Centers for Medicare & Medicaid Services, The Health Insurance Portability and Accountability Act of 1996 (HIPAA), Online at http://www.cms.hhs.gov/hipaa/, 1996.

[10] L. Richardson, Beautiful soup documentation, April (2007).

[11] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL: https://doi.org/10.5281/zenodo.1212303. doi:10.5281/zenodo.1212303.

[12] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019.