Window Classifiers and Conditional Random Fields for Medical Report De-Identification

Viviana Cotik[1,2], Franco M. Luque[3,4], and Juan Manuel Pérez[1,2] *

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Instituto de Ciencias de la Computación, CONICET, Argentina
{vcotik, jmperez}@dc.uba.ar
[3] FAMAF, Universidad Nacional de Córdoba, Argentina
[4] CONICET, Argentina
francolq@famaf.unc.edu.ar

Abstract. Information extraction of medical reports is key in order to improve timely discoveries of findings and as an aid to improve decisions about medical treatments and budget. In order to develop information extraction methods, medical data has to be available. Since this data is extremely sensitive due to the presence of personal information, report de-identification is needed. We present two methods, a window classifier and an implementation of conditional random fields (CRF) in order to de-identify personal information of Spanish medical records provided by the MEDDOCAN challenge. CRF obtained the best results with a F1-measure of 0.897 for named entity recognition with exact match (subtask 1), and 0.930 and 0.940 for inexact match (subtask 2 strict and merged respectively).

Keywords: report anonymization · report de-identification · named entity recognition · BioNLP · Spanish medical reports

1 Introduction

Last years' exponential growth of available biomedical texts and the contribution of structured information to the triggering of automatic alerts about urgent situations, to performing better diagnosis, and as input to clinical decision support systems among others, has led to the need of automatically extracting information through the use of disciplines such as natural language processing (NLP) and information extraction (IE). The area of study that deals with information extraction from biomedical texts is called BioNLP or biomedical text mining.

Medical data is of sensitive nature, due to the presence of personal information. Therefore, in order to process medical reports, they have to be anonymized.

In the clinical domain, anonymization or de-identification is the process of removing from medical records all information that could identify a patient or the physician performing the study or diagnosis. In some cases, names and patients

[*] All authors contributed equally to the work.

identifications -patient ids- (identification codes from a knowledge base) have to be removed, in others, even the diseases have to be changed by others, since otherwise, they could help identify the patient.

Institutions and countries might have data sharing policies and legislation, stating to which degree information has to be anonymized in order to be shared. For that reason, de-identification is a very important task in BioNLP. Some examples of data sharing policies are US HIPAA,[5] the former European Data Protection Directive 95/46/EC[6] and Argentinian laws 26529, 17132 and 25326 [2].[7]

Many factors have to be taken into account when anonymizing medical records. See [6, 8] to read about perturbative and non-perturbative methods. Besides, the concept of data anonymization is ambiguous, as explained in [6], where three forms of sharing data (*public*, *quasi-public*, *non-public*) and different ways of data perturbation with de-identification goals and their effects with regards to the possibility of a meaningful analysis are presented.

Some de-identification challenges have been organized in the past, for example i2b2 *2006 De-identification and Smoking Challenge* [20] and *2014 De-identification and Heart Disease Risk Factors Challenge*.[8]

In this paper we present two methods we have developed for MEDDOCAN [13],[9] a medical document anonymization task of the IberLEF 2019 (Iberian Languages Evaluation Forum).[10] MEDDOCAN is specifically devoted to the anonymization of medical documents in Spanish.

The MEDDOCAN task was structured into two sub-tasks: 1) NER offset and entity type classification, and 2) sensitive token detection.

A synthetic corpus of 1000 clinical case studies augmented with protected health information (PHI) from discharge summaries and medical genetics clinical records was provided. Additionally, a number of linguistic resources were supplied:[11]

- **AbreMES-DB**,[12] a Spanish medical abbreviation database, that contains abbreviations and their potential definitions automatically extracted from the metadata of titles and abstracts of biomedical publications written in Spanish,

[5] https://www.hhs.gov/hipaa/index.html

[6] https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

[7] http://servicios.infoleg.gob.ar/infolegInternet/anexos/160000-164999/160432/texact.htm, http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=19429,http://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm

[8] https://www.i2b2.org/NLP/HeartDisease/

[9] MEDDOCAN: http://temu.bsc.es/meddocan/.

[10] IBERLEF 2019: https://sites.google.com/view/iberlef-2019/.

[11] http://temu.bsc.es/meddocan/index.php/resources/

[12] https://github.com/PlanTL-SANIDAD/AbreMES-DB

- **MEDDOCAN-Gazetteer**, composed of MEDDOCAN related entities. It includes names, surnames, addresses, hospitals, professions, and different types of locations (provinces, cities and towns, among others),
- **SPACCC_POS-TAGGER**,[13] a Part-of-Speech tagger for Spanish texts in the medical domain based on FreeLing [15], and a
- **Sentence-split test-set** computed using the SPACCC_POS-TAGGER.

The rest of the paper is organized as follows. Section 2 presents previous work of anonymization in the biomedical domain. Section 3 presents the data set used, the implemented methods and the performance evaluation metrics used. Section 4 presents the results of our methods and its analysis. Finally, Section 5 presents conclusions and future work.

2 Previous work

In this section we describe the previous work in anonymization in the biomedical domain.

Emam [5] wrote a technical report telling how patient health data is being anonymized in Canada in year 2006 and discusses the adequacy of this practices. Some de-identification challenges have been organized in the past, for example i2b2 *2006 De-identification and Smoking Challenge* [20] and *2014 De-identification and Heart Disease Risk Factors Challenge*.[14] The annotation process performed for the later is explained in Stubbs and Özlem Uzuner [17]. The i2b2 effort was focused on documents in English and covered characteristics of US-healthcare data providers. Gkoulalas-Divanis and Loukides [8] describe algorithms to anonymize while preserving patient demographics (non-perturbative anonymization) and to anonymize diagnosis codes.

A tutorial on Privacy Challenges and Solutions for Medical Data Sharing was organized by IBM Research Zurich and Cardiff University in 2011.[15] Gkoulalas-Divanis et al. [9] present a survey of more than 45 algorithms that have been proposed for publishing data of EHRs preserving patient's privacy

Emam et al. [6] mention the ambiguity of the concept of anonymous data and different ways of data perturbation with de-identification goals and their effects with regards to the possibility of a meaningful analysis.

Many other analysis and surveys on the subject [12, 20, 23] and anonymization implementations [4, 7, 18, 21, 22] have been published.

Rules and regular expressions have been used, eg. for medical reports in Spanish [3]. Token and character level conditional random fields and long short-term memory networks are methods with good results in the de-identification of medical records [10, 11].

[13] https://github.com/PlanTL-SANIDAD/SPACCC_POS-TAGGER

[14] https://www.i2b2.org/NLP/HeartDisease/

[15] Tutorial on Privacy Challenges and Solutions for Medical Data Sharing. Slides. https://www.zurich.ibm.com/medical-privacy-tutorial/.

3 Methods

This section presents the datasets used, the evaluation metrics, the preprocessing of data and the two proposed methods. A flow chart figure for the used processing pipeline can be seen in Figure 1.

3.1 Data

We used the MEDDOCAN corpus, a synthetic corpus composed of 1,000 clinical cases manually selected by a practicing physician and enriched with PHI expressions extracted from discharge summaries and genetics clinical records. It has around 33,000 sentences, with an average of around 33 sentences per clinical case, summing up 495,000 words and an average of 494 words per clinical case.

The corpus was divided into three subsets. The training set comprises 500 clinical cases, and the development and test set 250 clinical cases each. The test set was not provided until the end of the challenge.

An additional test set, that includes the actual test set, the training, the development set, and a background set -of 2,000 documents-, composed of 3,751 clinical cases was provided in order to test the submissions.[16] The actual test set (250 clinical cases) is called *test set with Gold Standard Annotations* and the test set provided before the end of the challenge (3,751 clinical cases) is called *test set (including background set)*.

A conversion script between BRAT standoff annotation format[17] and i2b2 annotation format was provided.[18]

The annotation schema used by the challenge organizers to create the dataset defines 29 entities and was inspired by the annotation schema used for the i2b2 de-identification tracks, adapting it to the specificities of the MEDDOCAN document collection. An iterative process was followed until the final guidelines were obtained. The inter-annotation agreement (IAA) calculated on a set of 50 double-annotated records is reported as 98% in exact match.[19] The dataset annotation guidelines can be downloaded.[20]

3.2 Evaluation metrics

Given that different uses might require different balance in terms of precision and recall, two different sub-tasks have been proposed: sub-task 1 (named entity recognition offset and entity type classification) and sub-task 2 (sensitive span detection). Sub-task 1 consists in exact match (ie. prediction and gold standard

[16] Datasets can be found in `http://temu.bsc.es/meddocan/index.php/data/`.

[17] `https://brat.nlplab.org/standoff.html`

[18] `https://github.com/PlanTL-SANIDAD/MEDDOCAN-Format-Converter-Script`

[19] An abstract of the annotation process and schema can be seen in `http://temu.bsc.es/meddocan/index.php/annotation-guidelines/`.

[20] http://temu.bsc.es/meddocan/wp-content/uploads/2019/02/guADas-de-anotacin-de-informacin-de-salud-protegida.pdf

have to begin and end in the same offset and also have the same classification). Sub-task 2 allows inexact match and possibly wrong entity types, evaluating whether spans belonging to sensitive phrases are detected correctly. Sub-task 2 has two evaluations: strict and merged.

Micro-averaged precision, recall, and F1 score are used to evaluate both subtasks (see definition below). Micro-balanced metrics calculate each metric over all the classes together.

$$\begin{aligned} Precision(P) &= \frac{TP}{TP+FP} \\ Recall(R) &= \frac{TP}{TP+FN} \\ F1 &= \frac{2PR}{P+R} \\ Leaks &= \frac{FN}{\#sentences} \end{aligned}$$

where TP are the number of true positives, FP the number of false positives, and FN the number of false negatives.

Two additional metrics were computed to evaluate the best methods of the tasks: *leaks* for sub-task 1 and *merged* for sub-task2. The first computes the proportion of PHIs that were not detected. The merged results of subtask 2 merges the spans of PHIs connected by non-alphanumerical characters before evaluating the results.

3.3 Data Preparation

We first use standard text preprocessing techniques such as sentence segmentation and word tokenization. Then, we use the BIO encoding, a standard technique that models Named Entity Recognition as a sequence tagging problem. In the BIO encoding, each entity type `<T>` has labels `B-<T>` and `I-<T>` to mark, respectively, those tokens as the beginning or the continuation of an entity. The label `O` is used to tag "outside" tokens, this is, tokens that do not belong to any entity.

We also do Part-of-Speech tagging using our own classifier-based PoS tagger.[21] The tagger was trained using the AnCora 3.0.1 Spanish corpus [19], that uses the EAGLES tagset. Instead of using the full tagset, we used a simplified 85-tag version as done in the Stanford CoreNLP tagger.[22]

3.4 Window Classifier

The window classifier consists in the use of a classifier to tag each token in the text based on information of the token itself and of tokens in a fixed window

[21] Unpublished work.

[22] `https://stanfordnlp.github.io/CoreNLP/`

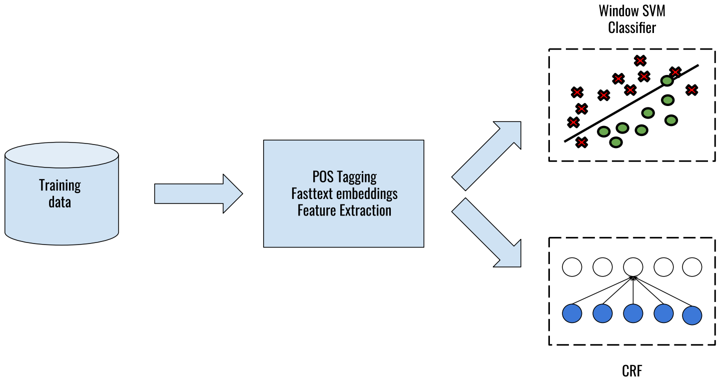

Fig. 1. Chart of the processing pipeline.

surrounding it. Our current system uses the two previous tokens and the next token, this is, a window of size four with the center word at the third position.

Features for all tokens include the lowercased token, a prefix and a suffix of length 2, the PoS tag, and boolean features indicating if the word is in upper case and if it is only composed of digits. For the center word, a word embedding is also included as a feature. Here we use 300-dimension fastText embeddings,[23] pre-trained on Spanish Wikipedia [1].

For the classifier, in this work, we use a Support Vector Machine (SVM) with a linear kernel, as provided by *scikit-learn* [16]. Each token is tagged independently by the classifier, leading to a very fast system.

As tags are independent to each other, the tagger may output sequences inconsistent with the BIO representation, such as an `I-<T>` tag after an `O` tag. To fix this, we introduce a post-processing step, that converts all inconsistent `I-<T>` tags to `B-<T>` tags.

3.5 Conditional Random Fields

Conditional Random Fields (CRFs) are probabilistic models used to predict sequences of labels, based on sequences of input samples.

Each token has an associated vector of features, such as the words' part of speech tag and the words' suffix of a given length. The input of a CRF is the sequence of tokens of the text. The features of a token and the pattern of labels assigned to previous words are used to determine the most likely label for the current token. In a linear chain CRF, only the label of the previous token is used.

As features we used the word in lower case, a prefix of length 3, a suffix of length 4, the truth value of whether the word is all written as upper case if all

[23] `https://fasttext.cc/`

Table 1. Results of de-identification methods sub-task 1, sub-task2 strict and sub-task2 merged.

Sub-Task	Method	Leak	Precision	Recall	F1
1	Window	0.073	0.866	0.903	0.884
	CRF	0.090	0.914	0.880	0.897
2 strict	Window		0.872	0.910	0.891
	CRF		0.948	0.912	0.930
2 merged	Window		0.887	0.921	0.904
	CRF		0.956	0.924	0.940

are digits, PoS tag and reduced PoS tag of length 2. Also, one word before and one word after were considered.

For the implementation, CRFsuite [14] and the *sklearn-crfsuite* wrapper were used.

4 Results

A single run was submitted for each method. Results for both methods can be seen in Table 1. Our results are compared to the results obtained by other MEDDOCAN participating teams in [13].

Table 2 displays classes with lowest F1 and a minimum of 50 instances and Figure 2 presents the confusion matrix of a subset of the named entities.

We can observe some common errors: e.g, *TERRITORIO* (territory) is tagged frequently as *CALLE* (street) or as some of the0 *ID* entities (postal codes are annotated as *TERRITORIO*). Another type of common error is to misclassify *NOMBRE_SUJETO_ASISTENCIA* as *NOMBRE_PERSONAL_SANITARIO* (and viceversa); this is probably due to the fact that both refer to proper nouns. A model with more features (and more sophisticated, such as word embeddings) and eventually a different window size would probably overcome these problems.

We also noticed some annotation errors –such as *Madrid* and *Mendieta Espinosa*– that were not annotated as entities. This kind of errors can influence negatively the performance of the learning algorithm. Moreover, entities correctly discovered are erroneously computed as errors.

Table 2. Labels with lowest F1 among those with more than 50 occurrences in the development set.

Tag	Precision	Recall	F1
ID_SUJETO	0.75	0.87	0.81
TERRITORIO	0.93	0.57	0.71
INSTITUCION	0.79	0.58	0.67
FAMILIARES_SUJETO	0.77	0.63	0.69
INSTITUCION	0.83	0.49	0.61
NUMERO_TELEFONO	0.85	0.88	0.86

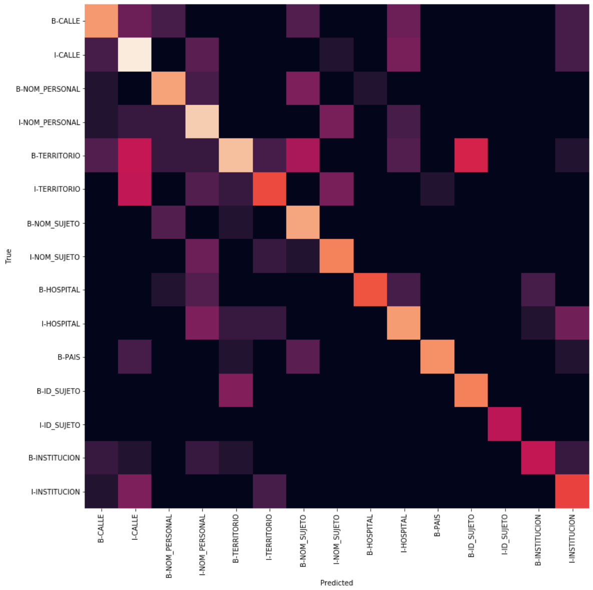

Fig. 2. Confusion Matrix for CRF Classifier. Vertical axes represent gold standard labels, horizontal axes represent predicted labels, and brighter colors mean more occurrences of instances for that pair of actual-predicted label.

5 Conclusions and Future work

As future work we plan to include in our models some of the linguistic resources provided by the challenge: namely, AbreMES-DB, a Spanish medical abbreviation database; the MEDDOCAN gazetteer, a gazetteer that includes names, surnames, addresses, hospitals, professions, and different types of locations, such as provinces, cities and towns. We tried to integrate them to our system, but it resulted in degraded performance. We also plan to include the use of regular expressions for detecting e-mails and phone numbers. More feature engineering in CRF should also be performed.

Another line of work to tackle NER could be the use of recurrent neural networks (RNN); in particular, bidirectional RNNs. We trained a biLSTM but, due to the small size of the dataset, the model failed to converge.

Since we did not use the Spanish abbreviation database nor the MEDDOCAN gazetteer the only components that takes language into account are our PoS tagger (trained with the AnCora Spanish corpus) and the word embeddings. Therefore, we think that our methods could be easily adapted to other languages. Regarding Spanish, although there are many differences in different countries (e.g. the Rioplatense Spanish of Argentina and Uruguay) the entities of interest in this task, that were recognized with our methods should have similar results with Spanish of different regions of the world.

References

[1] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), ISSN 2307-387X

[2] Cotik, V.: Information extraction from Spanish radiology reports. In: PhD Thesis (2018)

[3] Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., Xu, F.: Annotation of entities and relations in spanish radiology reports. In: RANLP, pp. 177–184 (2017)

[4] Dalianis, H., Velupillai, S.: De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. J. Biomedical Semantics **1**, 6 (2010), https://doi.org/10.1186/2041-1480-1-6, URL https://doi.org/10.1186/2041-1480-1-6

[5] Emam, K.E.: Data Anonymization Practices in Clinical Research. A descriptive study. , University of Ottawa (2006), URL http://www.ehealthinformation.ca/wp-content/uploads/2014/07/2006-Data-Anonymization-Practices.pdf

[6] Emam, K.E., Rodgers, S., Malin, B.: Anonymising and sharing individual patient data. BMJ **350** (2015), URL http://www.bmj.com/content/350/bmj.h1139

[7] Gentili, M., Hajian, S., Castillo, C.: A case study of anonymization of medical surveys. In: Proceedings of the 2017 International Conference on Digital Health, London, United Kingdom, July 2-5, 2017, pp. 77–81 (2017), https://doi.org/10.1145/3079452.3079490, URL http://doi.acm.org/10.1145/3079452.3079490

[8] Gkoulalas-Divanis, A., Loukides, G.: Overview of Patient Data Anonymization, pp. 9–30. Springer New York, New York, NY (2013), ISBN 978-1-4614-5668-1

[9] Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of Biomedical Informatics **50**, 4 – 19 (2014), ISSN 1532-0464, https://doi.org/https://doi.org/10.1016/j.jbi.2014.06.002, URL http://www.sciencedirect.com/science/article/pii/S1532046414001403, special Issue on Informatics Methods in Medical Privacy

[10] Jiang, Z., Zhao, C., He, B., Guan, Y., Jiang, J.: De-identification of medical records using conditional random fields and long short-term memory networks. Journal of biomedical informatics **75**, S43–S53 (2017)

[11] Liu, Z., Chen, Y., Tang, B., Wang, X., Chen, Q., Li, H., Wang, J., Deng, Q., Zhu, S.: Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. Journal of biomedical informatics **58**, S47–S52 (2015)

[12] Malin, B., Karp, D., Scheuermann, R.H.: Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational

research. Journal of Investigative Medicine **58**(1), 11–18 (2010), ISSN 1081-5589, https://doi.org/10.2310/JIM.0b013e3181c9b2ea, URL http://jim.bmj.com/content/58/1/11

[13] Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodrguez, H., Lopez Martin, J.A., Villegas, M., Krallinger, M.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), vol. TBA, p. TBA, CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain (Sep 2019), URL TBA

[14] Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007), URL http://www.chokkan.org/software/crfsuite/

[15] Padr, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey (May 2012)

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[17] Stubbs, A., Özlem Uzuner: Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics **58**, 20 – 29 (2015), ISSN 1532-0464, https://doi.org/https://doi.org/10.1016/j.jbi.2015.07.020, URL http://www.sciencedirect.com/science/article/pii/S1532046415001823, proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data

[18] Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. In: J Am Med Inform Assoc., vol. 14, pp. 574–580 (2007), URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1975791/

[19] Taulé, M., Martí, M.A., Recasens, M.: AnCora: Multilevel annotated corpora for catalan and spanish. In: Proceedings of the International Conference on Language Resources and Evaluation, European Language Resources Association, Marrakech, Morocco (2008)

[20] Uzuner, Ö., Luo, Y., Szolovits, P.: Viewpoint paper: Evaluating the state-of-the-art in automatic de-identification. JAMIA **14**(5), 550–563 (2007), https://doi.org/10.1197/jamia.M2444, URL https://doi.org/10.1197/jamia.M2444

[21] Uzuner, Ö., Sibanda, T.C., Luo, Y., Szolovits, P.: A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine **42**(1), 13–35 (2008), https://doi.org/10.1016/j.artmed.2007.10.001, URL https://doi.org/10.1016/j.artmed.2007.10.001

[22] Velupillai, S., Dalianis, H., Hassel, M., Nilsson, G.H.: Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computer-

ized annotation trial. I. J. Medical Informatics **78**(12), 19–26 (2009), https://doi.org/10.1016/j.ijmedinf.2009.04.005, URL `https://doi.org/10.1016/j.ijmedinf.2009.04.005`

[23] Wellner, B., Huyck, M., Mardis, S.A., Aberdeen, J.S., Morgan, A.A., Peshkin, L., Yeh, A.S., Hitzeman, J., Hirschman, L.: Research paper: Rapidly retargetable approaches to de-identification in medical records. JAMIA **14**(5), 564–573 (2007), https://doi.org/10.1197/jamia.M2435, URL `https://doi.org/10.1197/jamia.M2435`