April – November 2022

Utrecht University *Catharina Ziekenhuis Eindhoven*

Internal validity of a text mining algorithm to identify Adverse Drug Reactions in free-text entries from electronic health records of geriatrics and orthopedics patients

Student: *Loes van Dijck 6142036*
Daily supervisor: *Britt van de Burgt MSc.*
Referee: *dr. René Grouls*
Examiner: *prof. dr. Toine Egberts*

Abstract

Background: Structured registration of adverse drug reactions (ADRs) in the electronic health record (EHR) is vital in preventing recurrence of ADRs, but in practice, ADRs are often saved as free-text only. Using a text mining tool could be useful in identifying these ADRs.

Aim: To determine the internal validity of a previously developed ADR-identifying text mining algorithm at the geriatrics and orthopedics department at Catharina Hospital Eindhoven.

Methods: One year of EHR data from 15 orthopedics patients and 6 geriatrics patients were manually reviewed for ADRs, creating a gold standard. MedDRA and SNOMED-CT terminology was used to identify symptoms and the Dutch G-standard database was used to identify offending medications. The same data was reviewed by the algorithm and its output was compared to the gold standard.

Results: A total of 100 unique ADRs were identified in the gold standard, 20 of which were potentially serious. 14 ADRs were also found by the algorithm (true positives); 86 ADRs were marked as false negatives. The algorithm also returned 49 false positives. Overall, the algorithm reached a 22% PPV (positive predictive value), 14% sensitivity and an F-measure of 0.17. At the geriatrics department the PPV was 28%, as opposed to 15% at the orthopedics department. For serious ADRs, the algorithm reached an overall sensitivity of 20%.

Conclusion: In this preliminary analysis, the algorithm did not meet our goals. This study needs to be finished in order to draw valid conclusions. Future research into this algorithm is required for further improvements and evaluation of its performance in different settings.

Introduction

A useful strategy to prevent Adverse Drug Reactions (ADRs) from recurring to a patient is systematic registration of the ADR as structured information in a designated field in the electronic health record (EHR) [1,2], but in clinical practice, this registration of ADRs is poorly performed [1,3-5]. Inadequate IT systems, inadequate support from colleagues and professional organizations, lack of knowledge recognizing ADRs, failure to recognize the importance of registering ADRs, interruption of the normal workflow and time constraints are at the root of this problem [5]. Instead of registering an ADR in as structured data, ADRs are often recorded using free-text entries (e.g. in clinical notes or reports) [6]. Free-text information can presently not be used by clinical decision support systems (CDSSs) to alert healthcare professionals if a problematic drug is represcribed. Information on ADRs therefore goes unnoticed, but since ADRs contribute significantly to morbidity, mortality and expenses [7,8], it is vital to prevent them from recurring as much as possible. Additional strategies are therefore necessary to make free-text ADRs accessible for use in clinical practice.

A solution to this problem lies in text mining (TM), also known as natural language processing (NLP). Text mining tools have shown to be able to screen unstructured free-text and transform the retrieved data into computer-readable knowledge that can help in clinical decision-making [9-11]. In the past, some TM tools have been developed for the purpose of tracing ADRs, such as the tool by Honingman et al [12], which screened primary care records for ADRs. Our previous work centers around the development of a TM tool to identify ADRs in all free-text of the EHR at Catharina Hospital Eindhoven. This Dutch-language algorithm initially achieved an overall sensitivity of 57% and a positive predictive value (PPV) of 32% [6]. The algorithm has since been developed further. The first results from a follow-up study suggest a 93% sensitivity and an 11% PPV on the same data set [13].

This algorithm was developed and tested at the departments of geriatrics, oncology and internal medicine at Catharina Hospital exclusively, which are all non-surgical departments. It is currently unknown how this algorithm performs in different departments, particularly surgical departments, where both the patient population and conventions in reporting are likely to be different. It is also unknown how the algorithm performs in a different hospital.

The objective of this study is therefore two-fold. Firstly, this study aims to determine the internal validity of the algorithm by evaluating its performance at a department at Catharina Hospital that has not been studied in this context yet. Secondly, this study aims to assess the external validity of the algorithm by evaluating its performance in a different hospital. This paper only concerns the first objective.

Methods

Design and setting

This retrospective study was performed at Catharina Hospital Eindhoven, the Netherlands, a 660 beds teaching center. We conducted this research at the departments of geriatrics (non-surgical) and orthopedics (surgical). A visual representation of the full methods can be found in figure 1.

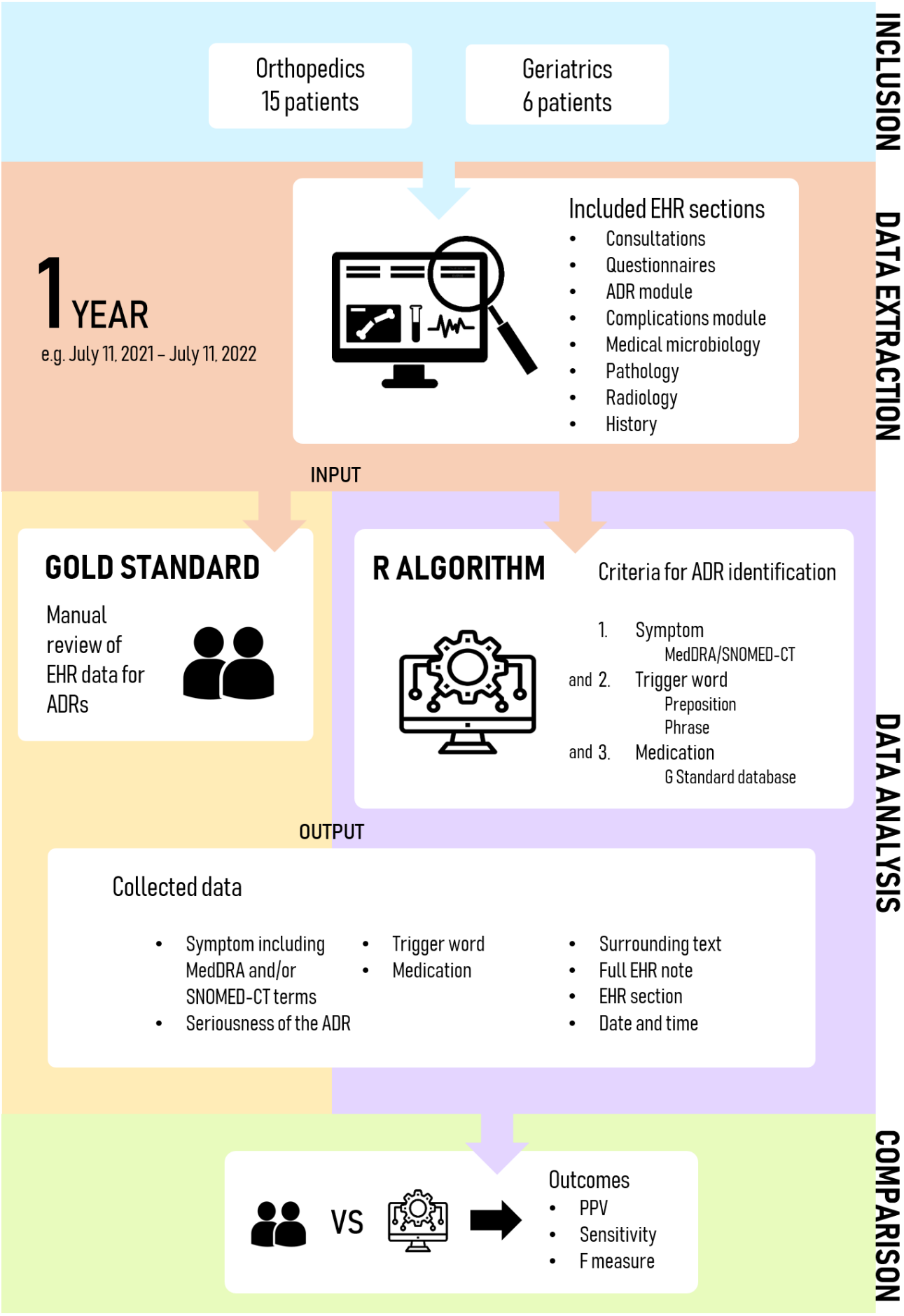

Figure 1: Overview of the methods of this study. EHR: electronic health record, ADR: adverse drug reaction, MedDRA: Medical Dictionary for Regulatory Activities, SNOMED-CT Systematized Nomenclature of Medicine Clinical Terms, PPV: positive predictive value.

Inclusion

The study population consisted of patients hospitalized for at least 24 hours at either department. In our first study [6], a sample size of 15 patients per department was determined. Considering that the sensitivity of the algorithm has increased since then, we aimed for a sample size of 15 patients per department in this study too.

Physicians selected and approached competent, eligible patients during their hospitalization. When a patient expressed interest, additional written and oral information was given by the research team. Written informed consent was obtained from all participants. The study was declared not subject to Research Involving Human Subjects Act (non-WMO) by the medical ethics committee MEC-U (Nieuwegein, The Netherlands).

Systems

Catharina Hospital uses HiX® (version 6.1, ChipSoft B.V., Amsterdam, The Netherlands), a comprehensive EHR that supports a wide array of functionalities, including medication management and some decision support [14,15]. Research Manager® (Cloud9, Deventer, The Netherlands) was used to record, encrypt and save the data. The algorithm was programmed in R (The R Foundation, Auckland, New Zealand).

Data extraction

We extracted all free-text data that was entered into a participant's EHR in the year preceding their most recent discharge, i.e. if someone was discharged August 1st, 2022, we extracted the data between August 1st 2021 and August 1st 2022.

We used extracts of all free-text sections of HiX® 6.1: consultations (which contained physician and physiotherapy reports), questionnaires (which included nursing reports), microbiology reports, pathology reports, radiology reports and medical history. We also used the ADR module and complications module to see how many ADRs were (also) registered as structured data.

We also collected the median number of hospitalizations, the number of ambulatory visits, the length of the most recent hospital stay, the cumulative length of all hospital stays and the Charlson Comorbidity Index. The patients' age and sex were recorded as well. Lastly, the number of notes, words and characters in the EHR extracts were documented.

Data analysis – gold standard

To create a gold standard, a pharmacist and a pharmacist in training manually reviewed the EHR extracts for potential ADRs independently. The two assessors' results were matched. Duplicate ADRs (i.e. if the same ADR was phrased in different ways, e.g. *hemorrhage due to rivaroxaban* and *bleeding after Xarelto*) were removed until a set of unique potential ADRs was left. Any unlikely ADRs were discarded and discrepancies between the assessors were discussed until consensus on the final gold standard was reached.

ADRs were identified using MedDRA (Medical Dictionary for Regulatory Activities, version 25) and SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms, version January 2022). The ADRs were categorized according to MedDRA's System Organ Classes (SOCs). The seriousness of the ADRs was classified using the European Medicines Agency's (EMA) Important Medical Events list (version 25) [16]. Additionally, the medication associated with the ADR was recorded with the Anatomical Therapeutic Chemical (ATC) code system. If a single drug or drug class was associated with multiple symptoms (e.g. nausea and constipation related to opiates), this was considered two ADRs. If a symptom was associated with two drugs (e.g. hypotension because of metoprolol/hydrochlorothiazide), this was considered one ADR. We also collected the EHR section where the ADR was found.

Data analysis – R algorithm

Parallel to the creation of the gold standard, the same EHR extracts were reviewed by the algorithm.

The algorithm was initially coded in Gaston Pharma® (Gaston Medical, Eindhoven, The Netherlands), a rule-based clinical decision support system, and first tested on 45 patients in the study by Wasylewicz et al. [6] In the second study by Van de Burgt et al. (forthcoming) [13], the algorithm was rewritten in R and developed to be able to identify and categorize ADRs using the same data set. SNOMED-CT was added to the catalog for even greater coverage of possible ADRs. The resulting algorithm was used in this study.

The algorithm identified an ADR when it encountered a combination of a symptom, a (type of) medication and a trigger word. MedDRA and SNOMED-CT terminology was used to identify a symptom. The EMA Important Medical Events list was used to categorize the seriousness of this symptom [16]. Trigger words included prepositions and phrases such as *after, as a result of* or *due to*. Thirdly, medications were specified by the Dutch G-standard database (version September 2020), which contains generic and trade names of all medications registered in the Netherlands. The algorithm could also detect some manually pre-specified terms, such as synonyms or abbreviations. The algorithm ran in two phases. In phase 1, it searched the full EHR extract for medications in combination with a trigger word. The output from phase 1 then served as the input for phase 2, in which it searched for a symptom within a 38-character distance of the medication. The algorithm then removed duplicate ADRs. Additional output of the algorithm included a snippet of text surrounding the ADR, the full EHR note, the EHR section and the time and date.

Comparison

To assess the performance of the algorithm, its output was compared and matched to the gold standard. Correspondence in the gist of an ADR was enough to qualify as a match; we considered perfect overlap in their phrasing and context unnecessary from a clinical viewpoint (e.g. *hives on administration of penicillin* and *urticaria after penicillin* were considered a match). If no match was found, we evaluated what went wrong.

Statistical analysis

The performance of the algorithm was expressed with the metrics sensitivity, positive predictive value (PPV) and the F measure. True positives (TP) were defined as ADRs identified by both the gold standard and the algorithm (a match). False positives (FP) were entries found by the algorithm that were not included in the gold standard. False negatives (FN) were ADRs from the gold standard that the algorithm failed to identify. Sensitivity was calculated as TPs/(TPs + FNs). The PPV was calculated as TPs/(TPs + FPs). The F measure (the harmonic mean of precision and sensitivity; a measure of accuracy) was calculated as 2((precision x sensitivity)/(precision + sensitivity)). The algorithm's ability to identify causative medication was expressed as a separate PPV. Causative medication means that a single, specific drug was associated with an ADRs, not a class of drugs. The performance of the algorithm at the two departments was compared by assessing the confidence intervals. We aimed for a sensitivity of >80% and a PPV of >50%.

Results

Characteristics of patients and EHRs

At the time of this analysis, 15 patients had been included in the orthopedics group and 6 in the geriatrics group. Patients in the geriatrics group were older, had more comorbidities and were hospitalized for longer periods of time compared to the orthopedics patients. The EHR extracts from geriatrics patients were also longer than those from orthopedics patients. Details can be found in table 1.

Table 1. Characteristics of included patients and EHRs at the orthopedics and geriatrics department.

Variable	**Orthopedics (n=15)**	**Geriatrics (n=6)**
Mean age in years (range)	64 (33-83)	87 (79-91)
Sex (% female)	53	50
Variable	**Median (range)**	
Charlson Comorbidity Index at last hospitalization[1]	3 (0-9)	5 (3-6)
Duration of most recent hospitalization in days	3 (2-17)	9 (6-27)
Cumulative duration of hospitalizations	6 (2-17)	11 (6-27)
Hospitalizations[2]	1 (1-2)	1 (1-2)
Ambulatory visits[3]	4 (0-53)	5 (0-22)
Free-text EHR notes	163 (88-578)	265 (139-482)
Words[4]	6096 (1920-39915)	15965 (9424–38629)
Characters[5]	37270 (12665-230203)	98889 (55546-220697)

[1] Only the available data (1 year) was used to calculate the Charlson Comorbidity Index. This was possible because many notes contain a summarized medical history.
[2] Hospitalizations were >24 hours; hospitalizations <24 hours were considered ambulatory visits.
[3] Ambulatory visits also included telephone and video consultations.
[4] Only free-text was included, structured data within notes (e.g. dates, timestamps, weight, height) was disregarded.
[5] This number excludes spaces.

Creation of the gold standard

The two assessors identified 151 potential ADRs. 114 unique ADRs remained after exclusion of duplicates. After removal of 14 unlikely ADRs a 100 ADRs were included in the final gold standard. This process is presented in figure 2.

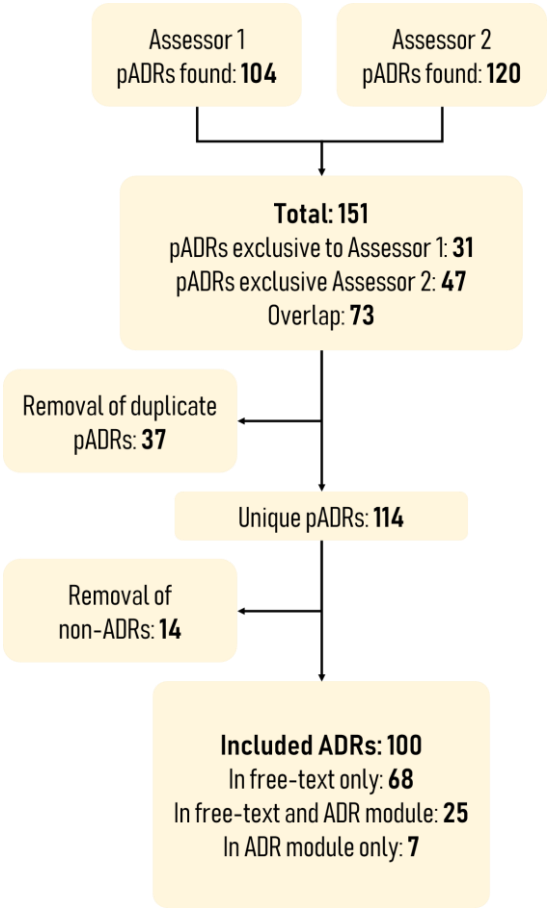

Figure 2: Overview of the creation of the gold standard, inclusion and exclusion of potential ADRs. pADR: potential ADR.

In the geriatrics group, 55 ADRs were found, of which 16% (n=9) was serious. Patients had an median of 8.5 unique ADRs per person (range 4-11). We identified 45 ADRs in the orthopedics group, of which 24% (n=11) was serious. These serious ADRs all belonged to the same patient. In total, 20 ADRs (20%) were serious. A median of 1.5 (range 0-25) unique ADRs per person was found in the orthopedics group. 40% (n=6) of orthopedics patients did not have any ADRs in their data, whereas each geriatrics patient contributed at least six unique ADRs to the gold standard. A single causative medication was identified 58 times (58%). In all other cases a drug class was mentioned, e.g. *diuretic*, or simply *medication*. A full overview of serious ADRs and ADRs per SOC can be found in the supplementary tables.

The consultations section contained the majority of the ADRs (74%, n=74), followed by the questionnaires section (n=18, 18%). The ADR module contained seven (7%) additional unique ADRs. 32% of ADRs in the gold standard were (also) registered as structured data in the ADR module. At the orthopedics department this applied to 12 out of 45 (27%), whereas this number was 20 out of 55 (36%) at the geriatrics department.

Review by the R algorithm

The algorithm identified 63 potential ADRs in phase 2, 14 of which were TPs. The other 49 entries were considered FPs. Moreover, the algorithm failed to identify the remaining 86 ADRs (FNs). Data per department can be found in figure 3. This results in an overall PPV of 22% (95%CI 13-35%), a sensitivity of 14% (95%CI 8.1-23%) and an F measure of 0.17. At the orthopedics department, the algorithm reached a PPV of 15% (95%CI 4.9-35%). At the geriatrics department, the PPV was 28% (95%CI 15-45%). The algorithm correctly identified 4 serious ADRs, reaching an overall 20% sensitivity (95%CI 6.6-44%) for identifying serious ADRs. The sensitivity to identify serious ADRs was 17% (95%CI 2.9-49%) at the orthopedics department; 25% (95%CI 4.4-64%) at the geriatrics department. Table 2 provides an overview of all outcomes.

Table 2: overview of all statistical outcomes by department and overall. PPV: positive predictive value.

	Orthopedics			**Geriatrics**			**Overall**		
	PPV (95% CI)	*Sensitivity (95% CI)*	*F measure (95% CI)*	*PPV (95% CI)*	*Sensitivity (95% CI)*	*F measure (95% CI)*	*PPV (95% CI)*	*Sensitivity (95% CI)*	*F measure (95% CI)*
Non-serious ADRs	0.10 (0.018-0.33)	0.06 (0.011-0.22)	0.075	0.32 (0.16-0.54)	0.17 (0.081-0.31)	0.22	0.22 (0.12-0.37)	0.125 (0.065-0.22)	0.16
Serious ADRs	0.29 (0.051-0.70)	0.17 (0.029-0.49)	0.21	0.18 (0.032-0.52)	0.25 (0.044-0.64)	0.21	0.22 (0.074-0.48)	0.20 (0.066-0.44)	0.21
Overall	0.15 (0.049-0.35)	0.09 (0.028-0.22)	0.11	0.28 (0.15-0.45)	0.22 (0.095-0.31)	0.18	0.22 (0.13-0.35)	0.14 (0.081-0.23)	0.17

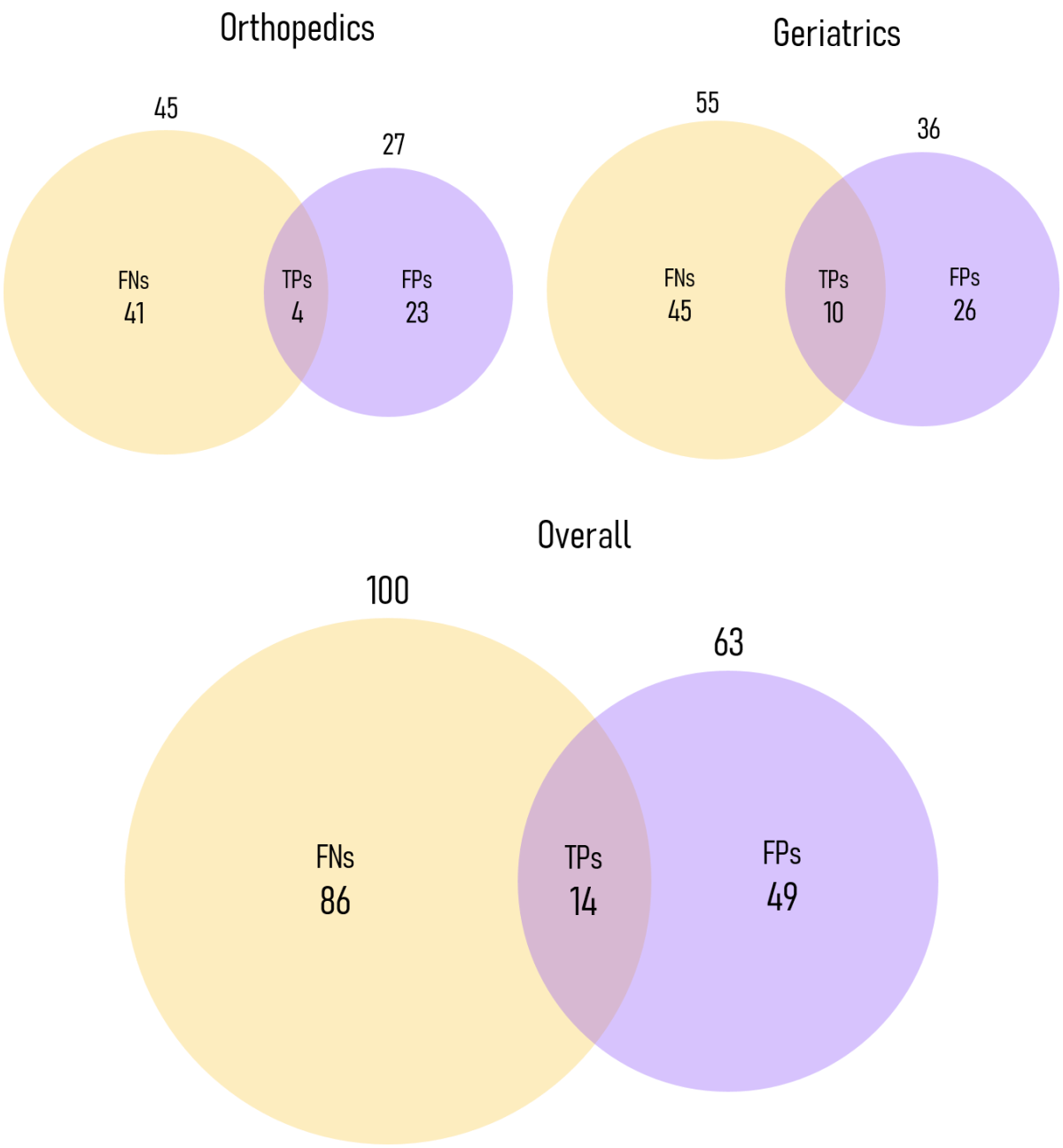

Figure 3: Venn diagrams presenting unique ADRs at the orthopedics department, the geriatrics department and overall. The yellow circles represent the unique ADRs found in the gold standard. The purple circles represent unique ADRs identified by the algorithm. Overlap between the circles represents true positives, i.e. correctly identified ADRs. TPs: true positives, FNs: false negatives, FPs: false positives.

When the deduplication and distance (n=38) functions were disabled, the algorithm returned 5 additional TPs, increasing the sensitivity to 19%. This, conversely, also added 64 FPs, dropping the PPV to 14%.

After phase 1, the overall sensitivity reached 71%. The algorithm was able to correctly identify a single causative agent in 10 instances (PPV 24%).

False negatives analysis

The analysis and categorization of false negatives is shown in table 3. For some ADRs, more than one reason could be identified, which is why the sum exceeds the total number of FNs. The most common reason for a FN (n=36) was absence of a familiar MedDRA or SNOMED-CT term. In 23 cases the distance between the medication and symptom exceeded 38 characters. Absence of a familiar trigger word accounted for 17 FNs, whereas an unfamiliar, alternative or colloquial drug name (e.g. *water pill, painkiller)* led to 16 FNs. Three times a typo in the medication resulted in a FN. Finally, no category could be determined for 13 ADRs, 15% of all FNs.

Table 3: Analysis of false negatives, prevalence of reasons why the ADRs was not identified by the algorithm.

Reason for FN	n
Unfamiliar MedDRA of SNOMED-CT term	36
Distance >38 characters	23
No trigger word	17
Unfamiliar medication	16
Typo	3
Unknown	13

Discussion

This manuscript presents preliminary data to evaluate the internal validity of a previously developed text mining algorithm to identify ADRs in free-text sections of the EHR. The algorithm reached an overall PPV of 22% and an overall sensitivity of 14% in phase 2, where we aimed for a PPV of 50% and 80% sensitivity.

A sensitivity of 71% was found in phase 1. This is a decrease compared to 93% in our previous study [13]. At that time, the algorithm had been tested on data from 45 patients and improvements were based on findings from that dataset specifically, increasing the sensitivity to 93% for that specific dataset. It seems these improvements did not fully apply to the data in the present study, likely causing this decrease to 71%.

Geriatrics patients were older, suffered from more comorbidities and were hospitalized longer than the participants in the orthopedics group, which led to more and longer notes. The majority of patients in the orthopedics group were hospitalized briefly for elective surgery. This is in line with our expectations for a non-surgical and surgical department.

We noticed a practice among physicians, especially non-surgical specialists, to copy and paste large sections of notes, particularly the medical history. This way clinically relevant ADRs from the past were transferred to more recent notes. This finding is in line with our expectations; we hypothesized that using one year of data would cover most of a patient's ADRs, while also keeping a manual review feasible. We also observed the use of colloquial drug names in the EHR free-text, which could lead to a misinterpretation of the offending drug. Transdermal patches of fentanyl or buprenorphine, for example, are referred to as *morphine patch*. The algorithm will logically identify an ADR related to morphine here, even though morphine is not available in patch form. Finally, we noticed a tendency to overgeneralize ADRs, i.e. when a drug class is accused of causing an ADR, when a single offending agent had been identified earlier. This is especially the case when previous notes are summarized. One patient, for example, experienced a rash, possibly caused by cefuroxime, which was later described as *rash from antibiotics.* If this ADR

were registered as such in a structured way, this would unnecessarily limit treatment options for this patient.

There were also some remarkable aspects to the R algorithm itself. Firstly, there were discrepancies in the seriousness assessment between physicians and the algorithm. The MedDRA term *allergy*, for example, is non-serious according to the EMA list. The algorithm therefore marks all allergies as non-serious, whereas some patients had allergies that had been marked as serious by their doctor. Secondly, the algorithm is not always able to distinguish a ADR-related symptom from an indication, e.g. when it returns *rivaroxaban after pulmonary embolism*. Thirdly, the false negatives analysis pointed out that the algorithm missed 13 ADRs for no apparent reason. The cause for this remains unclear. Lastly, the runtime of the R algorithm is quite long: the full review of the EHR data from the 21 participants took two days. We hoped that including only one year worth of data would mitigate this issue, but unfortunately this was not the case. This is a common issue. In Van de Burgt's forthcoming scoping review of studies combining text mining tools with clinical decision support systems [17], only 8% of studies implemented a real-time tool due to performance issues.

Our previous studies [6,13] indicated that its high prevalence of ADRs made the geriatrics department an ideal candidate for subsequent research. Consistent with this, we found more ADRs in the geriatrics group, both relatively and absolutely: the geriatrics group contributed more ADRs (n=55) to the gold standard than the orthopedics group (n=45), despite a smaller sample size. This is in agreement with our expectations for these departments and populations. Overall, the algorithm seems to do marginally better in the geriatrics department, but the confidence intervals are too wide to make a valid claim about any difference in performance.

Just like in the first study, a fifth of all ADRs were classified as serious. The algorithm identifies serious ADRs with a higher sensitivity than non-serious ADRs in both departments, which is good news from a clinical viewpoint. This outcome could be improved by adding the output from phase 1, but this would inevitably lead to disproportionately more false positives as well. In addition, we noticed that the vast majority (99%) of all ADRs in the gold standard were registered in either the consultations section, the questionnaires section or the ADR module. It might be worth only including the consultations section, questionnaires section and ADR module in future research. This will likely also decrease the runtime of the algorithm.

32% of ADRs in the gold standard was (also) registered as structured data in the ADR module. This is a stark contrast to our previous studies, when this applied to only 2%. More awareness of structured ADR registration may have caused this positive development, possibly as a result of the extensive research into ADR prevention and registration by Van der Linden et al. [2-4], geriatrician at Catharina Hospital and our previous studies. In contrast to our previous studies, the patients' treating physicians selected eligible patients for informed consent. Their involvement in the study may have inspired more structured ADR registration as well.

Strengths and limitations

The most important asset of this study is the creation of a gold standard by two assessors, allowing us to determine the sensitivity and find ways to improve the algorithm based on the false negatives analysis. Another strength is the scope of this algorithm. Other works into text mining of EHRs have focused on specific domains, such as diagnosis of depression [18] or adverse drug events related to antidepressants and antipsychotics [19]. This algorithm, however, takes a more

generalized approach, trying to find any ADR caused by any medication in all free-text.

This is a preliminary analysis, which limits the validity of the outcomes. The required sample size of 15 patients had not been reached in the geriatrics group at the time of this analysis. More participants need to be included to draw valid conclusions.

Another important limitation is the fact that the causality of the ADRs in the gold standard was not assessed formally. We originally planned for an expert team of clinical pharmacologists to solve any discrepancies between the assessors, but due to time constraints, the assessors did this ourselves. We also planned to have the ADRs scored using the Naranjo scale by the Netherlands' Pharmacovigilance Center Lareb and exclude all ADRs with a Naranjo score <1, but this will only be done when inclusion is complete. The goal of this analysis, however, was to evaluate the internal validity of the algorithm, not the clinical relevance of the identified ADRs.

Future perspectives

The false negatives analysis revealed that there are still accessible opportunities to improve the algorithm by adding more synonyms, colloquialisms, abbreviations and alternative terminology for symptoms and medications. In addition, real-time testing and/or combining the algorithm with existing clinical decision support systems will also be a valuable source of information in the future.

Unfortunately, the current EHR systems do not allow for quick and easy registration of ADRs, which contributes to the tendency to register important ADRs as free-text only [5]. This algorithm could be a tool to detect these ADRs, but is not meant as a substitute for structured ADR registration by physicians, especially because ADRs are generally underreported already [1]. It should also not limit the incentive for software developers to make necessary changes to the design of the EHR systems to allow for easier reporting.

Conclusion

This preliminary analysis shows that this algorithm currently does not perform as desired in either department. More research is needed to fine-tune the algorithm for application in a broader setting. Even though the algorithm currently does not meet our goals, finishing the study will provide insights into ADR registration elsewhere. It will clarify whether local customization of the algorithm is needed or that a standard set of symptoms, trigger words and medication could lead to similar outcomes elsewhere.

Acknowledgment

Many thanks to Britt van de Burgt, René Grouls and Toine Egberts for their supervision, valuable feedback and advice, to Roel Verheijen, Loes Lammers and Björn Dullemond for providing the data, to all physicians who helped to include participants and to all other contributors to the project.

Reference List

1. McLachlan G, Broomfield A, Elliott R. Completeness and accuracy of adverse drug reaction documentation in electronic medical records at a tertiary care hospital in Australia. Health Inf Manag. 2021 Dec 21:18333583211057741.
2. Van der Linden C, Jansen P, van Geerenstein E, van Marum R, Grouls R, Egberts T. Reasons for discontinuation of medication during hospitalization and documentation thereof: a descriptive study of 400 geriatric and internal medicine patients. Arch Intern Med. 2010;170(12):1085–7.
3. van der Linden CM, Jansen PA, Grouls RJ, van Marum RJ, Verberne MA, Aussems LM, et al. Systems that prevent unwanted represcription of drugs withdrawn because of adverse drug events: a systematic review. Ther Adv Drug Saf. 2013 Apr;4(2):73-90.
4. Van der Linden CMJ, Jansen PAF, Van Marum RJ, Grouls RJE, Korsten EHM, Egberts ACG. Recurrence of adverse drug reactions following inappropriate re-prescription: Better documentation, availability of information and monitoring are needed. Drug Saf. 2010 Jul 1;33(7):535-8.
5. Geeven IPAC, Jessurun NT, Wasylewicz ATM, Drent M, Spuls PI, Hoentjen F, et al. Barriers and facilitators for systematically registering adverse drug reactions in electronic health records: a qualitative study with Dutch healthcare professionals. Expert Opin Drug Saf. 2022 May;21(5):699-706.
6. Wasylewicz A, van de Burgt B, Weterings A, Jessurun N, Korsten E, Egberts T. Identifying adverse drug reactions from free-text electronic hospital health record notes. Br J Clin Pharmacol. 2022 Mar;88(3):1235-1245.
7. Patton K, Borshoff DC. Adverse drug reactions. Anaesthesia. 2018 Jan;73 Suppl 1:76-84.
8. Eurpean Commission. Strengthening pharmacovigilance to reduce adverse effects of medicines. [Internet]. Available via: https://ec.europa.eu/commission/presscorner/detail/en/MEMO_08_782. [Accessed June 8 2022].
9. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: A review. J Healthc Eng. 2018 Apr 8;(2018):1–9.
10. Pereira L, Rijo R, Silva C, Martinho R. Text mining applied to electronic medical records: A literature review. Int J E-Health Med Commun. 2015 Jul 1;6(3):1–18.
11. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. J Heal Inf Manag. 2014;22(3):52–6.
12. Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, et al. Identifying Adverse Drug Events JAMIA Original Investigations Using Computerized Data to Identify Adverse Drug Events in Outpatients. J Am Med Inform Assoc. 2001 May-Jun;8(3):254-66.
13. Van de Burgt B et al. Identifying Adverse Drug Reactions from free-text Dutch EHR in hospitalized patients with the development of an algorithm (IADRESS). [Unpublished manuscript].
14. ChipSoft. Solutions. [Internet]. Available via: https://chipsoft.com/solutions/550. [Accessed June 8 2022].
15. ChipSoft. Health Care Logistics. [Internet]. Available via: https://www.chipsoft.com/solutions/536. [Accessed June 8 2022].
16. European Medicines Agency. Important medical event terms list (MedDRA version 25.0). [Internet]. Available via: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjO5cj485n7AhWw_7sIHbhGDJEQFnoECA4QAQ&url=https%3A%2F%2Fwww.ema.europa.eu%2Fen%2Fdocuments%2Fother%2Fmeddra-important-medical-event-terms-list-version-250_en.xlsx&usg=AOvVaw0J1fVqZ_7h01LdRfvrj3eB&cshid=1667749995810227. [Accessed 22 October 2022].
17. Van de Burgt BWM, Wasylewicz ATM, Dullemond B, Grouls BJE, Egberts ACG, Bouwman AG, et al. Integrating Text Mining with Clinical Decision Support in patient care: a scoping review. J Am Med Inform Assoc, forthcoming.
18. Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, et al. Identifying Patients with Depression Using Free-text Clinical Documents. Stud Health Technol Inform. 2015;216:629–33.
19. Iqbal E, Mallah R, Rhodes D, Wu H, Romero A, Chang N, et al. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. PLoS One. 2017 Nov 9;12(11): e0187121.

Supplements

Table S1: Unique ADRs per MedDRA system organ class (SOC) included in the gold standard.

System Organ Class (SOC) involved in the ADR[1]	Geriatrics Unique ADRs	Orthopedics Unique ADRs	Total
Blood and lymphatic system disorders	2	0	2
Cardiac disorders	3	2	5
Endocrine disorders	1	0	1
Gastrointestinal disorders	13	4	17
General disorders and administration site conditions	1	2	3
Immune system disorders	11	3	14
Injury, poisoning and procedural complications	1	3	4
Investigations	3	3	6
Metabolism and nutrition disorders	8	7	15
Nervous system disorders	2	7	9
Psychiatric disorders	1	0	1
Renal and urinary disorders	3	9	12
Respiratory, thoracic, and mediastinal disorders	1	3	4
Skin and subcutaneous tissue disorders	8	4	12
Vascular disorders	2	2	4

[1] *100 ADRs were included in the gold standard. Since some MedDRA terms have more than one SOC attached to them, the total number of ADRs in this table exceeds 100. The lower level MedDRA term confusion, for example, is in both the Psychiatric disorders and Nervous systems disorders categories.*

Table S3: Number of serious ADRs per MedDRA system organ class and related medication.

	Geriatrics		Orthopedics	
SOC and MedDRA PT of ADR	**Associated medication**	**Occurrences**	**Associated medication**	**Occurrences**
Gastrointestinal disorders				**1**
Gastrointestinal haemorrhage			Anticoagulants	1
Injury, poisoning and procedural complications				**2**
Nephropathy toxic			Medication	1
Nephropathy toxic			Flucloxacilline	1
Metabolism and nutrition disorders		**5**		**2**
Hypokalemia	Furosemide	1	Furosemide	1
Hypokalemia			Diuretics and flucloxacilline	1
Hypokalemia	Hydrochlorothiazide	1		
Hypokalemia	Metoprolol	1		
Hypokalemia	Irbesartan	1		
Hypokalemia	Medication	1		
Renal and urinary disorders		**4**		**6**
Tubulointerstitial nephritis	Ciproflocaxine	1		
Acute kidney injury	Ciprofloxacine	1		
Acute kidney injury	Medication	1		
Acute kidney injury			ACE-inhibitor and bumetanide	1
Acute kidney injury			Diuretics	1
Acute kidney injury			Deferasirox	
Urinary retention	Oxybutinine	1		
Renal impairment			Diuretica	1
Renal failure			Deferasirox	1
Nephropathy toxic			Medication	1
Nephropathy toxic			Flucloxacilline	1
Vascular disorders		**1**		**1**
Shock			Medication	1
Gastrointestinal haemorrhage	Anticoagulants	1		