Detecting adverse events in clinical care using natural language processing

Dorthe Klein (✉ dorthe.klein@mumc.nl)

Maastricht Universitair Medisch Centrum+ <https://orcid.org/0000-0003-0182-9569>

R.J.M.W. Rennenberg

Maastricht Universitair Medisch Centrum+

F.A.G. van den Heuvel

Praktijk index

R.P. Koopmans

Maastricht Universitair Medisch Centrum+

M.H. Prins

Universiteit Maastricht

Research article

Keywords: adverse events, hospital, natural language processing, medical record review

Posted Date: July 17th, 2019

DOI: <https://doi.org/10.21203/rs.2.11633/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Current methods for retrospective review of medical records require both time- and cost-wise a substantial effort. Therefore, we wanted to find the best method (based on natural language processing (NLP)) to select cases out of the medical records for further investigation in search for a (potentially preventable) adverse event (AE) to decrease this effort. **Methods** The basic dataset consisted of 2987 medical records of patients who died during their hospitalization. To gain insight into the signal to noise ratio of the various resources, several subsets of our basic dataset were tested. Thereafter, we tested the scalability. After the best subset was chosen, several NLP algorithms were tested to select the best performing algorithm for the detecting of AEs. In the last experiment we tested the performance of the computer algorithms to predict potentially preventable AEs. The results of the NLP were compared with the outcome of the original retrospective medical record review. **Results** The dataset which contained the last three letters of the medical record showed the biggest potential. The scalability experiment showed that more data leads to a better performance of the algorithm. The best performing algorithm in the third test was the one based on support vector machine (SVM), with a precision of 79%, a negative predictive value (NPV) of 95% and a specificity of 85%. The results of the preventability experiment showed that the performance of the algorithms was almost equal to the results of the AEs. **Conclusions** In this study, we have shown that the SVM algorithm generates the most accurate results for the selection of cases for further investigation in the search for a (potentially preventable) AE. The sensitivity of the algorithms was around 75%. However, the SVM algorithm selected fewer cases to be examined for AEs compared to the original method. Consequently, this would lead to a lower workload for the committee. At the same time, there are a substantial number of cases, with potentially preventable AEs, not detected by machine learning.

Background

Current methods for retrospective review of medical records require both time- and cost-wise a substantial effort. Although trigger systems aim to decrease the burden, they still select many cases without adverse events (AEs). To optimize the process, sometimes cases are preselected leading to the investigation of a specific subset of cases. Examples are: deceased patients, patients of a particular department or patient with a longer length of stay. Results are therefore not generalizable to all hospital patients. Another issue is the relatively high inter-rater variation and low reproducibility of trigger results and subsequently AE detection with current methods.¹⁻⁷ This suggests that these methods generate results that make translation to the clinical practice difficult.

To increase the efficacy, computer assisted detection of triggers, or even better AEs, in medical records has been developed. The aim is to reduce the labor-intensive manual record review and to increase the reproducibility.⁸ Thus far, text mining software is available for the detection of triggers and a limited selection of AEs.⁹⁻¹⁷ The software has to be taught what triggers or AEs are, preferably with a large database of well characterized cases with their triggers and AEs. Until now, the most optimal dataset with

patient details to be investigated is not known. For example, incorporating all information about the patients might influence the signal to noise ratio negatively and thus also the reliability of the results. Therefore, the optimal dataset has to be determined. Thereafter, different natural language processing (NLP) algorithms can be tested to detect the algorithm with the highest sensitivity and specificity for detecting triggers or AEs.

In this study, we wanted to find the most optimal method to select cases for further investigation in search for a (potentially preventable) AE. We describe our stepwise approach for developing a computer algorithm to search reliably for AEs in medical records. First, we used an NLP algorithm with excellent computing power restrictions to identify the optimal dataset selected from the medical record in our database with well characterized cases of deceased patients. Second, we evaluated the influence of the amount of records to be included for finding optimal results concerning agreement (true positives + true negatives divided by the total group). Third, several experiments were performed with different NLP algorithms in this optimal dataset to find the algorithm with the best performance for finding AEs and their preventability. Then, the performance of the best computer algorithm was validated in a different part of the data.

Methods

The usual procedure of the medical record review (our “gold standard”)

Since 2008, a team of trained (according to the EMGO/NIVEL standards¹⁸) nurses screened the medical records of all deceased inpatients (approximately 700 annually) for the presence of (one of the fifteen) triggers (figure 1). Triggers are clues which alert screeners for potential AEs (for example “unplanned transfer to the intensive care unit”). To accommodate the process, a database facilitating the necessary steps in this procedure was introduced in 2010 (Medirede®, Clinical File Search; Mediround BV). We used the triggers originally proposed by the Harvard medical practice study (HMPS) in 1991¹⁹ with a slight adjustment to fit the group of deceased patients. Therefore, the triggers regarding transfer to another acute care hospital and unplanned inappropriate discharge to home were omitted as they have no relevance in deceased patients.

The medical records with at least one trigger were redirected to the review committee. This committee consisted of both active and retired medical specialists with considerable clinical experience in the field of quality and safety in healthcare and medical record review. After a thorough review by a member specialised in the field of medicine related to the main diagnosis of the case (e.g. a surgeon investigates surgical patients etc.), this case was presented to the other members of the review committee in a regular meeting. A first conclusion on the potential presence of an AE was then established. Subsequently, after

consulting the involved specialists, the committee finally decided on the presence of an AE and the potential preventability of this AE. Since 2012 there is a stable formation of the review committee.

Previous research showed that the average time nurses needed for the manual screening of triggers was 38 minutes (they had no time restrictions), for the reviewers this was on average 60 minutes (excluding the time needed for the discussion in a meeting). Thus, it takes approximately 1.5 hours for the total review of a single medical record.

Figure 1: Presentation of the regular medical record review procedure

Data

The total dataset (originated from our “gold standard” procedure) consisted of 2987 medical records of patients who died during their hospitalization. All records between 2011 and 2016 were included. Of these records 59% contained one or more triggers after the screening by the nurses. In 742 of these medical records (42%) with a trigger, one or more AEs were detected by the review committee. 208 of these AEs were classified as potentially preventable. For these records, there was full access to surgical reports, discharge letters, patient records, nursing reports, use of medication, radiology reports, lab results and the medical history.

Definitions

An AE was defined as an unintended outcome arising from the (non)-action of a caregiver and/or the health care system with damage to the patient resulting in temporary or permanent disability or death of the patient.²⁰

The patient file was defined as the document including all reports, letters, lab results, scans, reports and medical history.

Modified data

From the basic dataset several parts of data were selected. Machine learning is based on NLP. This is the ability of a computer program to understand human language as it is spoken. It is a component of artificial intelligence. In machine learning, it is important to make a selection out of data with a high signal-to-noise ratio. Preferably you would like to find the “signal” in the data, rather than fitting the noise. The signal was in this case the useful information in the medical record that is pointing towards the AE

and the noise is the information in the medical record that is not helpful in locating/finding the AE. To gain insight into the signal to noise ratio of the various resources, several subsets of our basic dataset were tested (A-F).

A: Last general letter to the general practitioner (GP)

This letter describes the last communication from the hospital to the GP of the patient.

In this selection, all records without a last general GP letter were excluded. There were 476 medical records without this letter, leaving 2511 (84%) for inclusion in the analysis. We have chosen the last GP letter because we assumed this contained the most useful information regarding the hospitalization of the patient.

B: Last letter

In this selection, the last general GP letter was used, but for the 476 cases in which this letter was missing, the last written document was used instead. Therefore, in this analysis, the original 2987 records were included.

C: Last three letters

In this analysis, the last three letters of all medical records were included.

D: Patient file

For this selection, the full patient file was used.

E: General GP letter combined with patient file

In this selection, the general GP letter was combined with the patient file, for every record with a general GP letter.

F: Last (general GP) letter combined with the edited patient file

For this dataset, the patient record was electronically edited, leaving only 20% of the rarest words in the patient record. After the editing, the patient record was combined with dataset B. This was executed by a preprocessing script. First the whole text was evaluated and then the rare words were filtered out. After the editing, the patient record was combined with the last letter (dataset B).

Outcome measures

The following outcome measures were determined;

- Accuracy (agreement): calculated as the sum of true positives and true negatives divided by the total population.
- Precision: calculated as the sum of true positives divided by the number of predicted condition positives (in this case AE present)
- Recall: calculated as the number of true positives divided by the total number of medical records which were identified as containing an AE.
- Negative predictive value (NPV): calculated as the number of true negatives divided by the number of predicted condition negative (in this case no AE present).
- Specificity: calculated as the number of true negatives divided by the total number of medical records which were identified as not containing an AE.

Computer NLP algorithms

We tested different computer algorithms and explored the feasibility of this software (Open Mines Platform supplied by “the Praktijk Index”).

The following NLP algorithms have been used:

- Naive Bayes (NB) with n-gram input and term frequency–inverse document frequency (tf-idf) scores;
- Fast-text (FT) 2-layer neural network with hierarchical softmax output
- Linear Support Vector Machine (SVM)
- Convolutional neural network (CNN) based on pre-trained word vectors^{21,22}

Experiments

As a first step, the 6 datasets (described in section datasets) were provided to the NB algorithm to select the dataset that provided the highest performance in predicting an AE. This selected dataset was then used for the next experiments. To correct for the variation of the initialization, this experiment was repeated 28 times. Due to time and computing power restrictions, we chose to test the fast NB algorithm for all selections. In the second experiment (scalability) was tested whether the performance would decrease if a smaller training set was available. In the second experiment, we attempted to predict the preventability of an AE. AEs were therefore categorized as “probably not preventable” or “potentially preventable”. The results of these test were compared with the outcome of the gold standard (committee review)

In the third experiment all four algorithms were trained, optimized and tested for AE prediction with use of the best-fitted dataset, which resulted from the first experiment (see box 1).

Box 1: Explanation of training, validation and test

1. *Training set: this dataset consisted of examples used for the NLP to learn. In this case the NLP was taught which cases contained an AE and which didn't. learning, that is to fit the parameters (e.g., weights) of, for example, a [classifier](#).^{[7][8]}*
2. *Validation set: a set of examples used to tune the parameters of a classifier. It is sometimes also called the development set or the "dev set".*
3. *Test set : A test dataset is a [dataset](#) that is [independent](#) of the training dataset, but that follows the same [probability distribution](#) as the training dataset. A test set is therefore a set of examples used only to assess the performance (i.e. generalization) of a fully specified classifier.^{[7][8]}*

60% of the data was used for training, 20% for validation and 20% for testing.

Privacy

To guarantee patients privacy, information such as names, addresses, and other personal information were deleted from the data. After the anonymization, the data was checked by the privacy officer and researcher DK.

Results

Experiment 1: Dataset selection

In table 1 the results are presented for the individual datasets. The positive predictive value varied widely for the different datasets. The accuracy was for all datasets around 75% and the specificity was close to 100% for every dataset.

Dataset C, which contained the last three letters, showed the biggest potential and was therefore used in the next experiment to test the performance of the different algorithms. For this first experiment, the NB algorithm was used.

Table 1: Performance of the six datasets

Dataset	Accuracy	Precision	Recall	Specificity
A	74.9	66.8	75.1	98.8
B	75.3	72.1	75.2	99.2
C	76.1	78.3	76.1	99.7
D	75.8	57.0	76.0	100
E	75.7	57.0	76.0	100
F	74.9	67.6	74.9	99.0

Experiment 2: Scalability

Figure 1 below shows the influence of the amount of data plotted against the accuracy. This is shown for the NB algorithm. In this figure can be seen that more data leads to a better performance of the algorithm.

Figure 1: Influence of the amount of data against the accuracy of the algorithm

The size of the dataset is plotted against the accuracy of the NB algorithm. The resulting line is flattened and shown with corresponding trend line.

Experiment 3: Algorithm performances with dataset C

For accuracy, sensitivity and specificity the results were close together for the four algorithms (shown in table 2).

The best performing algorithm was the SVM algorithm, with a precision of 79%, an accuracy of 84% and a specificity of 85%. Figure 2 shows the precision presented against the recall/sensitivity for the SVM algorithm. This shows, if we would create a set with 82% precision, this would result in a recall/sensitivity of 40% (percentage AEs with respect to all AEs). In figure 3, the ROC curve of the SVM algorithm is presented. Table 2 shows the performance of the four algorithms separately. These results are based on 12 repetitions of the experiment, with the same settings for the algorithm but with a different distribution for the training set and the test set.

Figure 2: SVM algorithm with corresponding precision and accuracy

Figure 3: ROC curve of the SVM algorithm

Table 2: Performance for every algorithm separately for dataset C

Algorithm	Accuracy	Precision	NPV	Recall	Specificity
NB	0.76	0.67	0.77	0.11	0.98
SVM	0.84	0.79	0.95	0.51	0.85
FT	0.80	0.61	0.85	0.53	0.89
CNN	0.78	0.78	0.79	0.99	0.15

Experiment 4: Preventability

Table 3: Performance for the four algorithms separately for dataset C

Algorithm	Accuracy	Precision	Recall	Specificity
NB	0.762	0.568	0.075	0.99
SVM	0.822	0.755	0.428	0.955
FT	0.798	0.610	0.513	0.89
CNN	0.778	0.730	0.173	0.98

The results of this experiment shows that the performance of the algorithms for the preventability is almost equal to the results of the AEs. However, it must be considered that these results are based on a small dataset. This experiment was repeated five times.

Discussion And Conclusion

In this study, we have shown that the SVM algorithm generates the most accurate results for the selection of cases for further investigation in the search for a (potentially preventable) AE. The sensitivity of the algorithms was around 75%. However, the SVM algorithm selected about 50% fewer cases to be examined for AEs compared to the original method. Consequently, this would lead to a lower workload for the committee.⁷ At the same time, there are a substantial number of cases, with potentially preventable AEs, not detected by machine learning (20% for CNN and FT algorithm, 18% for SVM algorithm, 24% for NB algorithm). If we look at the best accuracy, then the SVM algorithm performs best. However, for precision, the algorithms don't differ that much. The CNN and the SVM algorithm used the letters and the

patient record. By using more structured information, which is already generated by hospitals, such as age and length of stay, the algorithms could be improved. Another option is using more data. The number of examples of which a algorithm will learn has an influence on the performance of the algorithm.

Because of an unfavorable signal to noise ratio in the total record the performance of the algorithms is dependent on the selected dataset. When different dataset selections are compared, it was indeed shown that the use of the unedited total patient record doesn't improve the output and can even have a negative effect on the results. Different datasets show different results for sensitivity and specificity. This is probably caused by the length of a patient record in itself and especially by long sentences.

The use of discharge letters, in particular, the letters for the GP, appeared to be useful because of a more favorable signal to noise ratio. These letters contain brief information regarding the hospitalization of the patient in contrast to the extensive patient record which contains a lot of "noise". The scalability experiment shows that by increasing the dataset, the precision of the algorithming will improve.

Finally, the results of the preventability experiment showed that the preventability can be adequately estimated despite scarce records in the training set.

Until now, the literature concerning the automatic detection of AEs focused on a short list of pre-specified AEs or only the triggers for which the data were searched.^{9,11,14,16,17} Forster et al²³ and Murff et al (2003) screened discharge letters in two small patient samples for the presence of terms that could indicate an AE but the actual evaluation was performed by physician reviewers.^{23,24} Murff showed a sensitivity of 69%, a specificity of 48% and a precision of 52%. For Forster, the results were a sensitivity of 23%, specificity of 92%, precision of 41% and NPV of 83%. In general, our results suggest a more optimal performance compared to these studies.

The strengths of our study were the careful step by step analysis with NLP to find the best algorithm that improved the precision for a potentially preventable AE compared with the current trigger system. Several, currently common NLP algorithms were tested. Experiments were repeated until saturation was reached. Also, we tested different dataset to find the most useful one with the most valuable information.

Still, some questions remain because of some weaknesses in our study. First, it is not clear what an optimal cut off point for sensitivity or specificity could be. This has to be determined by further research to see if the data generated with this study make policymakers draw the same conclusions as with conventional methods (in this study our "gold standard"). Second, the number of well-characterized records can be considered as rather small for NLP programs to be optimized, especially when we take the potential preventability into account. Finally, this was a selection of deceased patients and it might be even more interesting to use this method in patients that were alive at discharge from the hospital. Therefore, we think these data are not generalizable to living patients.

We think if the algorithms' performance is considered acceptable it should be tested prospectively and compared to conventional methods. Considering that the conventional method is our "gold standard" the

agreement will never be 100%. However, this is not an issue if the conclusions that can be drawn from the final results are comparable to the original ones. Still, we think the word of the committee cannot be fully replaced because the communication with the involved departments is crucial in the success of this quality and safety instrument.

In conclusion, the results of NLP algorithms to predict potentially preventable AEs in specific datasets from patient records is a promising tool to simplify record review. Further research is necessary, to investigate if the results of this method lead to the same overall conclusions from medical record review compared to the more expensive and labor-intensive conventional methods.

Abbreviations

AE adverse event

CNN convolutional neural network

FT fast text

GP general practitioner

NB naive bayes

NLP natural language processing

NPV negative predictive value

SVM support vector machine

Declarations

Ethics approval and consent to participate

The study protocol was evaluated by the ethics committee of our hospital.

Consent for publication

Not applicable

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

None

Funding

None

Authors' contributions

Concept and design of this study: DK, RR, MP. Acquisition of data: DK and FH. Analysis and interpretation of the data: DK, FH, RR. Drafting the manuscript: DK, FH, RR. Critical revision of the manuscript: DK, RR, MP, RK. All authors have read and approved the final version of the manuscript.

Acknowledgments

This study wouldn't have been executed without A. Witter, J. Hanhart, A. van der Veen and M. Aussems. Thank you for your time, effort and help!

References

1. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. *Health Serv Res*. 2011;46(2):654-78.
2. Mattsson TO, Knudsen JL, Lauritsen J, Brixen K, Herrstedt J. Assessment of the global trigger tool to measure, monitor and evaluate patient safety in cancer patients: reliability concerns are raised. *BMJ quality & safety*. 2013;22(7):571-9.
3. Hanskamp-Sebregts M, Zegers M, Vincent C, van Gorp PJ, de Vet HC, Wollersheim H. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. *BMJ Open*. 2016;6(8):e011078.

4. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. *J Clin Epidemiol*. 2010;63(1):94-102.
5. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. *Bmj Open*. 2015;5(7).
6. Klein DO, Rennenberg R, Koopmans RP, Prins MH. Adverse event detection by medical record review is reproducible, but the assessment of their preventability is not. *PloS one*. 2018;13(11):e0208087.
7. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. *Prev Med Rep*. 2017;8:250-5.
8. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128-44.
9. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association : JAMIA*. 2005;12(4):448-57.
10. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. *Journal of biomedical informatics*. 2007;40(2):174-82.
11. Rochefort CM, Buckeridge DL, Forster AJ. Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. *Implementation science : IS*. 2015;10:5.
12. Rochefort CM, Buckeridge DL, Tanguay A, Biron A, D'Aragon F, Wang S, et al. Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol. *BMC Health Serv Res*. 2017;17(1):147.
13. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306(8):848-55.
14. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association : JAMIA*. 2003;10(2):115-28.
15. Stockwell DC, Kirkendall E, Muething SE, Kloppenborg E, Vinodrao H, Jacobs BR. Automated adverse event detection collaborative: electronic adverse event identification, classification, and corrective actions across academic pediatric institutions. *Journal of patient safety*. 2013;9(4):203-10.
16. Gerdes LU, Hardahl C. Text mining electronic health records to identify hospital adverse events. *Studies in health technology and informatics*. 2013;192:1145.
17. Sammer C, Miller S, Jones C, Nelson A, Garrett P, Classen D, et al. Developing and Evaluating an Automated All-Cause Harm Trigger System. *Jt Comm J Qual Patient Saf*. 2017;43(4):155-65.

18. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, Waaijman R, van der Wal G. Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals. *BMC Health Serv Res*. 2007;7:27.
19. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med*. 1991;324(6):370-6.
20. Wagner C. Onbedoelde schade in ziekenhuizen: resultaten dossieronderzoek naar patiëntveiligheid. *Klachtenmanagement in de Zorg*. 2007;4(3-4):28-31.
21. J H. *Data Mining: Concepts and Techniques*: Elsevier Science & Technology; 2011.
22. CM B. *Pattern Recognition and Machine Learning*: Springer-Verlag New York Inc.; 2006.
23. Forster AJ, Andrade J, Van Walraven C. Validation of a discharge summary term search method to detect adverse events. *Journal of the American Medical Informatics Association*. 2005;12(2):200-6.
24. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. *Journal of biomedical informatics*. 2003;36(1-2):131-43.

Figures

Figure 1

Influence of the amount of data against the accuracy of the algorithm

Figure 2

SVM algorithm with corresponding precision and accuracy

Figure 3

ROC curve of the SVM algorithm