Computers in Biology and Medicine 170 (2024) 108014

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Modest performance of text mining to extract health outcomes may be almost sufficient for high-quality prognostic model development

Zwierd Grotenhuis [a,b], Pablo J. Mosteiro [a], Artuur M. Leeuwenberg [b,*]

[a] Department of Information and Computing Sciences, Utrecht University, The Netherlands
[b] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, The Netherlands

ARTICLE INFO

Dataset link: https://github.com/zwierd99/ClinicalTM, https://physionet.org

Keywords:
Text mining
Prognostic prediction modeling
In-hospital mortality
Performance evaluation

ABSTRACT

Background: Across medicine, prognostic models are used to estimate patient risk of certain future health outcomes (e.g., cardiovascular or mortality risk). To develop (or train) prognostic models, historic patient-level training data is needed containing both the predictive factors (i.e., features) and the relevant health outcomes (i.e., labels). Sometimes, when the health outcomes are not recorded in structured data, these are first extracted from textual notes using text mining techniques. Because there exist many studies utilizing text mining to obtain outcome data for prognostic model development, our aim is to study the impact of the text mining quality on downstream prognostic model performance.

Methods: We conducted a simulation study charting the relationship between text mining quality and prognostic model performance using an illustrative case study about in-hospital mortality prediction in intensive care unit patients. We repeatedly developed and evaluated a prognostic model for in-hospital mortality, using outcome data extracted by multiple text mining models of varying quality.

Results: Interestingly, we found in our case study that a relatively low-quality text mining model (F1 score ≈ 0.50) could already be used to train a prognostic model with quite good discrimination (area under the receiver operating characteristic curve of around 0.80). The calibration of the risks estimated by the prognostic model seemed unreliable across the majority of settings, even when text mining models were of relatively high quality (F1 ≈ 0.80).

Discussion: Developing prognostic models on text-extracted outcomes using imperfect text mining models seems promising. However, it is likely that prognostic models developed using this approach may not produce well-calibrated risk estimates, and require recalibration in (possibly a smaller amount of) manually extracted outcome data.

1. Introduction

Prognostic prediction models (and risk scoring rules) have been built and are used in all areas of medicine to estimate patient risk on certain future health outcomes [1–6]. Two well-known models to assess in-hospital mortality are the EuroSCORE [7,8], for patients undergoing cardiac surgery, and the APACHE score system [9–12], to assess the risk for intensive care unit patients at their admission. General guidelines for diagnostic and prognostic prediction model development are clear about the fact that developing high-quality prognostic models – which are able to sufficiently discriminate between low and high-risk patients, and provide well-calibrated predicted risks that correspond to observed risks – requires having representative patient-level data about all important predictive factors for the outcome to be predicted, as well as high-quality data about the predicted outcome itself [13–16]. With the increasing availability of electronic medical records, often representative of daily clinical practice populations, researchers have increasingly started using text mining to extract relevant study variables from the clinical notes in those records that were not recorded in structured form initially [17–21]. This includes the extraction of data on patient's health outcomes from clinical notes, which are then subsequently used downstream for prognostic model development (schematically shown in the yellow pathway in Fig. 1). This use of text mining – to extract outcome data for downstream prognostic model development – opens up the use of large amounts of readily available medical record data to develop prognostic models for outcomes that were not structurally collected. Using this approach, prognostic models have for example been developed for the prediction of (text-mined) problematic opioid use after chronic opioid therapy [22], prediction of (text-mined) falls among

* Corresponding author.
E-mail address: a.m.leeuwenberg-15@umcutrecht.nl (A.M. Leeuwenberg).

https://doi.org/10.1016/j.compbiomed.2024.108014
Received 26 July 2023; Received in revised form 3 January 2024; Accepted 19 January 2024
Available online 23 January 2024
0010-4825/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

elderly [23], and the prediction of (text-mined) parkinsonism side-effects of antipsychotic polypharmacy prescribed in secondary mental healthcare [24]. However, when using text mining to extract outcomes for subsequent prognostic model development, potential mistakes by the text mining model could impact the predictive performance of the downstream prognostic models, possibly limiting their clinical utility. While recent guidance on the use of natural language processing in observational medical research recommends reporting the text mining extraction performance (e.g., via precision, recall, and F1 score) [25], to our best knowledge, there has been no methodological research structurally investigating the relation between common text mining performance measures and the quality of the downstream prognostic models developed on text-mined outcome data.

Objective

The main objective of this study is to methodically investigate and demonstrate how the performance of text mining models to extract health outcomes (assessed by calculating precision, recall, and F1 score) relates to the performance of a downstream prognostic prediction model (assessed via discrimination and calibration measures), developed on the text-mined outcome data. For this, we used an illustrative case study on the development of a prognostic model to predict in-hospital mortality for intensive care unit (ICU) patients.

2. Methods

2.1. Study design

An overview of our study design is shown in Fig. 1. First, a prognostic prediction model is developed using well-recorded structured outcome data (the top blue pathway in the figure), acting as the reference prognostic model in our study. Then, we simulated the envisioned target setting, in which the true outcome data are unavailable in structured form, and text mining is first needed to extract the outcomes, to then be used for prognostic model development (the yellow part of the figure). The text-mining based prognostic model development process is repeated for different text mining models, with different levels of extraction performance. Finally, the different prognostic models are evaluated and compared in terms of their predictive performance in new unseen patients, and related to the text mining quality during their development. Further details about the study data, the prognostic modeling, the text mining architecture, and how text mining performance was varied are described in the following sections.

2.2. Data

We used data from MIMIC-III (v1.4) [26], a publicly available deidentified database of medical record data from 53,423 ICU admissions (of 38,597 uniqie patients) from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA between 2001 and 2012. It includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, diagnostic codes, survival data, length of stay, among other factors. In MIMIC-III, the in-hospital mortality (the outcome) is 11.5%. For our methodological study, we split these data into four non-overlapping subsets: one to develop the text mining model (TM-train), one to evaluate the text mining model (TM-test), one to develop the downstream prognostic model (PM-train), and one set to evaluate the prognostic model (PM-test). To realistically simulate the prognostic model development process (described in more detail in the next paragraph) we carefully followed the pipeline by Harutyunyan et al. [27], who developed a prognostic in-hospital mortality model in the same setting, and publicly released the code to reproduce their results. Our eligibility criteria are identical to theirs,[1] except for two aspects (needed for our study):

1. A pre-selection is made based on availability of clinical text notes.
2. Two extra subsets were made for text mining development and evaluation.

Important to note is that we use the same 85%–15% train-test split as Harutyunyan et al. [27], resulting in the same PM-test set, facilitating direct comparison with their reported performance (see Fig. 2).

Table 1
Table with predictive factors used in the reference prognostic model, following those used by Harutyunyan et al. [27] (CC BY 4.0).

Predictive factor	MIMIC-III table	Modeled as
Capillary refill rate	Chartevents	Categorical
Diastolic blood pressure	Chartevents	Continuous
Fraction inspired oxygen	Chartevents	Continuous
Glascow coma scale eye opening	Chartevents	Categorical
Glascow coma scale motor response	Chartevents	Categorical
Glascow coma scale total	Chartevents	Categorical
Glascow coma scale verbal response	Chartevents	Categorical
Glucose	Chartevents, labevents	Continuous
Heart rate	Chartevents	Continuous
Height	Chartevents	Continuous
Mean blood pressure	Chartevents	Continuous
Oxygen saturation	Chartevents, labevents	Continuous
Respiratory rate	Chartevents	Continuous
Systolic blood pressure	Chartevents	Continuous
Temperature	Chartevents	Continuous
Weight	Chartevents	Continuous
pH	Chartevents, labevents	Continuous

2.3. Prognostic model development

The model's predictive input factors (i.e., features or predictors) are obtained within the first 48 h of admission, and are listed in Table 1. For each factor, multiple variants are included: the first value, the final value, the minimum, maximum, mean, standard deviation, and skew for the first 10%, 25% and 50% as well as the last 10%, 25% and 50% of time and the full time period. This results in a total of $17 \times 7 \times 6 = 714$ predictive factors used as input for the prognostic model. Following Harutyunyan et al. [27], all predictive factors are normalized and missing values are replaced with their mean value in the training set.

The outcome of our prognostic model is in-hospital mortality, which is – in our methodological study – available for all patients, and is defined as whether the patient died within the given hospitalization or survived until discharge. These structured outcome data are used to develop our reference prognostic model, while text-mined outcomes are used to develop all text-mining-based prognostic models (as we simulate the setting where this outcome is not available in structured data).

The aforementioned predictive factors and outcome variable are used to train a prognostic model to predict in-hospital mortality. Following Harutyunyan et al. [27], we used an L2/Ridge-penalized logistic regression model (with inverse penalization factor C=0.001, the optimal value in their study). To assess whether our conclusions generalize to other prognostic modeling techniques, we additionally conducted all our experiments also using a feed-forward neural network (FFNN) instead of a logistic regression. The details and results on these experiments can be found in Appendix A.

[1] Focusing on first ICU stays of adult patients (due to the substantial differences between adult and pediatric physiology), for whom their length of stay exceeds 48 h (to ensure sufficient observations are available).

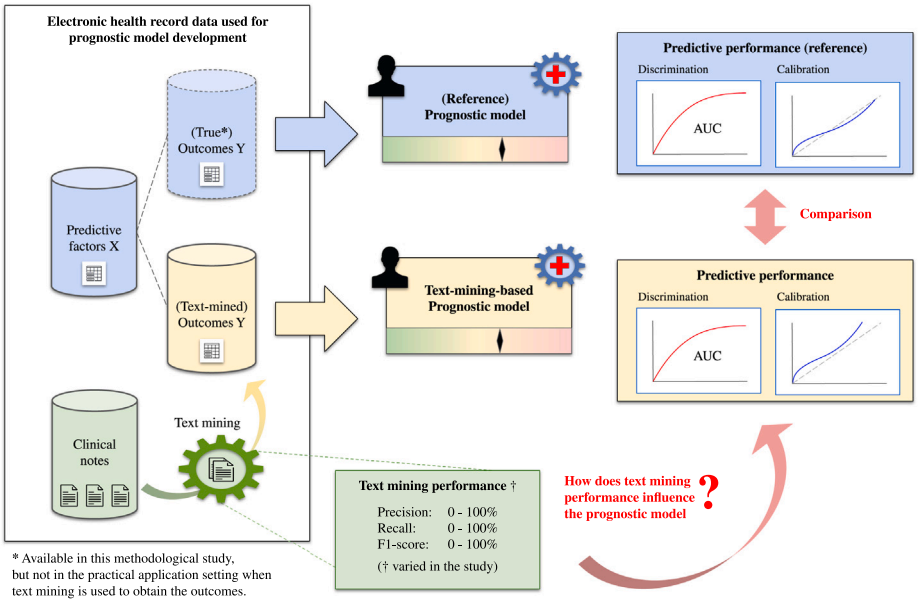


Fig. 1. Diagram of how text mining has been used to extract outcome data from text to be used in prognostic model development.

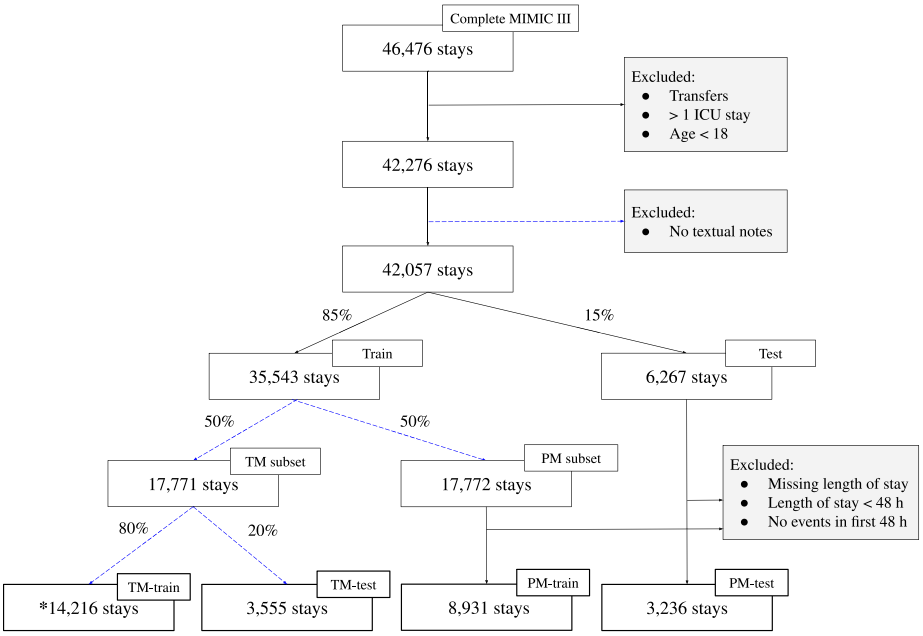


Fig. 2. Data flow diagram, following Harutyunyan et al. [27]. Any deviation from their data selection is indicated by dashed arrows (in blue). *The number indicated is the total number of stays in the TM-train dataset. In our experiments, varying fractions of this number will be using for training. TM: text mining, PM: prognostic model.

2.4. Text mining model development

The text mining models extract the mortality labels via document level classification. The classification model is based on tf-idf[2] features combined with a regularized logistic regression. Based on tf-idf we selected the 2000 most important words (with the highest tf-idf) and represent each document as a vector containing the tf-idf values for these 2000 words. The logistic regression used L2-regularization, with C = 1.[3] Common text preprocessing steps, such as lemmatization, number removal, lowercasing, stop word removal and punctuation removal were performed to increase the performance of our tf-idf model.

2.5. Variation in text mining performance

We aimed to obtain text mining models with a wide variety in extraction performance in order to see how their quality affects the

[2] Tf-idf; term frequency–inverse document frequency. The (automatic) tf-idf heuristic is commonly used for text classification and identifies frequent but discriminative words within a certain document collection by prioritizing words that occur frequently in certain documents (term frequency), but occur less frequently across documents (inverse of the document frequency).

[3] This is determined based on a grid search with parameters C ∈ {0.001, 0.01, 0.1, 1, 10, 100, 1000} and {L1, L2}, using a small validation set (a subset of training).

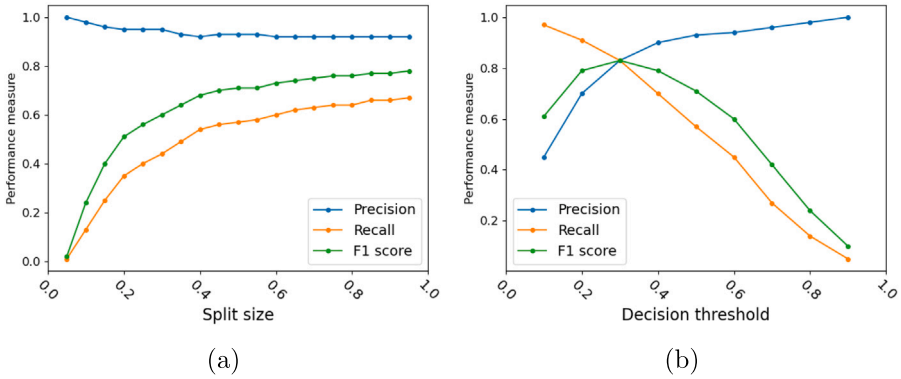


Fig. 3. The precision, recall and F1 score: (a) for different split sizes with a fixed decision threshold of 0.5, (b) for different decision thresholds with an equal split size of 0.5.

downstream prognostic models. We used two methods to vary the performance of the text mining model.

The first method to manipulate the performance of the text mining model was by using a smaller amount of training data for its development. Instead of using all samples from TM-train, to widely vary text mining performance, we used training dataset proportions in the range of 0.05 up to 0.95 with increments of 0.05 (in total 19).

The second method to change the performance of the text mining model was by changing the threshold value for the predicted output label probability to assign a positive outcome label. For binary classification settings this is usually set at 0.5 by default (i.e., assigning the label which is given the highest probability by the model). Increasing the decision threshold increases precision but reduces recall, while reducing the decision threshold reduces precision but increases recall. Changing the decision threshold also influences F1 score. With these two adaptations, we generated a wide range of precision and recall values for our text mining model. We used decision thresholds 0.1 up to 0.9 in increments of 0.1 (in total 9).

Combining both methods resulted in $19 \times 9 = 171$ text mining models.

2.6. Evaluation

Following relevant guidance for the evaluation of clinical prognostic prediction models [15], for each prognostic model, we calculated the area under the ROC curve (AUROC), the calibration slope, and the calibration intercept in PM-test. The AUROC measures discrimination: how well the model differentiates between patients with and without the in-hospital mortality outcome, but does not reflect how close the predicted probabilities are to actually observed probabilities. This component is captured by calibration measures [28]. The calibration slope (CS) indicates whether the probabilities are well spread out or too extreme (too close to 0 or 1): CS < 1 indicates the probabilities are too extreme, whereas a CS > 1 indicates predicted probabilities are too conservative (CS = 1 is the optimal value). The calibration intercept (Intercept) measures whether the mean predicted probability is in line with the observed outcome prevalence: Intercept > 0 indicates underestimation, and Intercept < 0 indicates overestimation (Intercept=0 is the optimal value).

To evaluate the text mining models, common evaluation measures are calculated [29] in the TM-test data: precision (also called positive predictive value), recall (also called sensitivity), and their harmonic mean (i.e., F1 score).

3. Results

3.1. Text mining model adjustments and performance

In Fig. 3(a) we can see the effect of changing the training data split size on the F1 score of the positive label. Increasing the training data size increases the F1 score. The relation seems asymptotic. The first 20% of training data accounts for 60% of the increase in F1 score. Fig. 3(b) shows the effect of a change in decision threshold on the performance metrics of the positive label. Interestingly, the optimal decision threshold seems to be below the default threshold of 0.5, at around 0.3. The further the threshold from that point, the lower the F1 score. Increasing the threshold increases precision at the cost of a lower recall, and decreasing the threshold means that the recall increases at the cost of precision. An optimal F1 score emerges at the intersection of precision and recall.

3.2. Text mining F1 score and prognostic model discrimination

In Fig. 4 we can see the effect of the F1 score of the text mining model on the AUROC of the clinical prognostic model, with a line indicating the models with a 0.5 decision threshold. A higher F1 score of the text mining model leads to a higher AUROC of the prognostic model. A relatively low F1 score of around 0.5 already yields an AUROC above 0.8, which is generally considered good [30,31]. We can also see that a high enough F1 score (in this case 0.8) can approximate the AUROC of the model using the true labels. We obtain similar results for the FFNN (Appendix A).

3.3. Text mining F1 score and prognostic model calibration

The same trend, where a higher F1 score of the text mining model leads to a prognostic model that more closely resembles the reference model, continues for the calibration metrics as seen in Fig. 5. In contrast to discrimination, for both calibration slope and calibration intercept there are multiple text mining models with higher F1 but worse calibration scores compared to the reference model, indicating that calibration is more sensitive to the quality of the text-extracted outcomes. Similar results were found for the FFNN (Appendix A).

4. Discussion

In our study, we observed that a higher text mining F1 score was associated with better prognostic prediction model discrimination and calibration, closer to the reference model.

We found that model calibration was more sensitive to imperfectly text-mined outcomes compared to model discrimination. This finding is consistent with other studies investigating measurement error in predictors and outcomes for (prognostic) prediction models [32–34]. For quite poor text mining models (F1 score ≈ 0.50), requiring only about 40% of the text mining training data, the downstream prognostic model obtained still quite good discrimination (AUROC ≈ 0.80). In contrast, even for good text mining models (F1 score ≈ 0.80) the calibration was not always in line with the reference model. This indicates that if calibration is important, which it generally is for prognostic models [5,

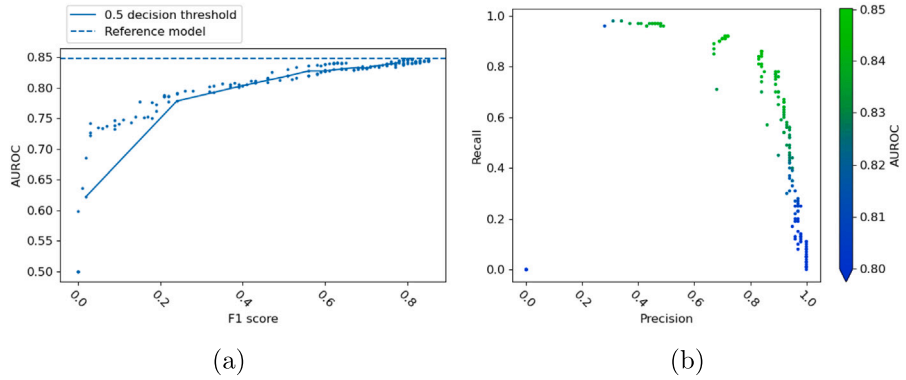


Fig. 4. The AUROC of prognostic model against (a) the F1 scores of the text mining models, and (b) the precision and recall of the text mining model. Each point on the plots represents a single choice of training set size and classification threshold. The continuous line in (a) represents the models that have a classification threshold of 0.5, for different training set sizes. The dotted line in (a) indicates the reference model trained on the ground truth labels.

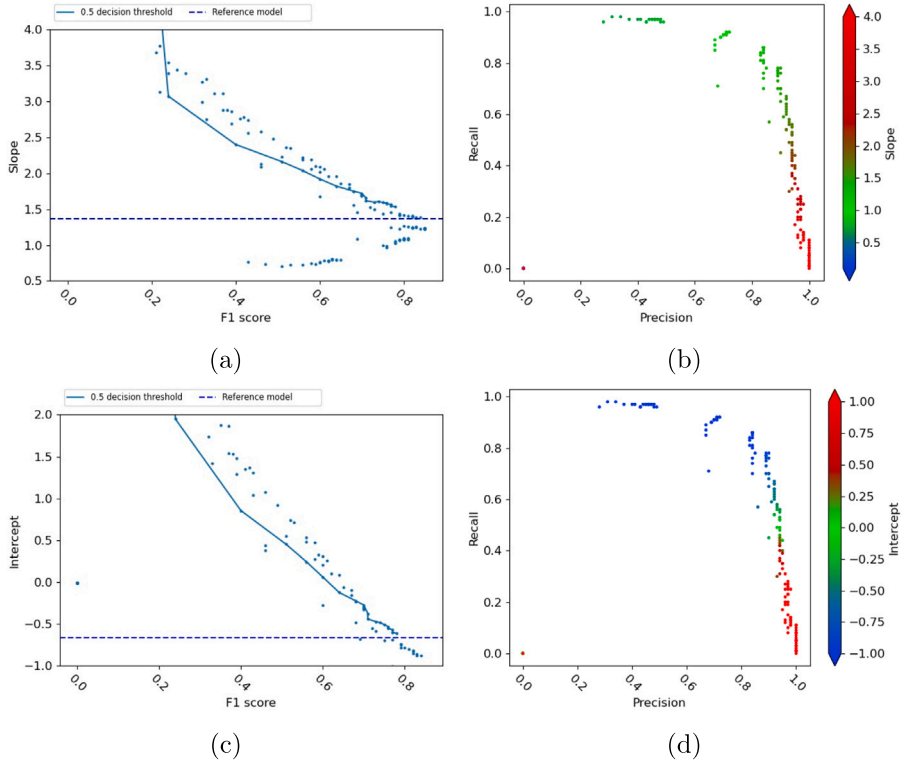


Fig. 5. The calibration slope of the prognostic models against (a) the F1 scores, and (b) the precision and recall (right) of the text mining model. And, similarly, the calibration intercept of the prognostic models against (c) the F1 scores, and (d) the precision and recall (right) of the text mining model.

28,35,36], (a smaller amount of) manually labeled data might still be needed to recalibrate a prognostic model developed on text-mined outcomes [37]. To place this potential requirement in perspective, it is important to consider that local evaluation and potential recalibration is also recommended for prognostic models that are not specifically developed using text-extracted outcomes [15,38]. So a potential need for recalibration, and the collection of (a smaller amount of) high quality outcome data may not per se impose extra work compared to the 'normal' prognostic model deployment process.

There is no general consensus on what is a sufficient minimum quality (i.e., minimum F1 score) of a text mining tool to extract outcomes sufficiently accurately to be used for prognostic model development. A general threshold of having at least precision of 90% and recall of 90% was used by van Laar et al. [39] for extraction of inclusion criteria, predictors, and outcomes when performing survival analysis (estimating Kaplan Meier curves). They found only small differences in results compared to manual extraction of the study variables. Regarding extraction of prognostic outcomes, this is in line with our study results about model discrimination.

To fully appreciate the results of our study, a few points should be considered. First of all, to further confirm the generalizability of the identified relation between text mining performance and prognostic model performance our findings should be confirmed in other data, and for other outcomes. While we varied the prognostic modeling strategy in our study (using logistic regression and neural networks), other prognostic modeling techniques, choices in study design (e.g., missing data or measurement error handling), and different text mining models should be further explored. Another consideration is that in our study, the data used to develop the text mining models and prognostic models were collected in a single institution (with likely similar patient

characteristics). To expand our findings, it would be interesting to see our analysis performed on a separate external data set from another medical institution. This situation, where a text mining model is created by an organization and adopted by another for usage to develop their own prognostic model, would pave the way for broader text mining usage in prognosis and is a valuable next step to consider. Finally, we did not analyze whether the samples misclassified by the text mining model were completely at random or systematic. If the outcomes are systematically misclassified for a certain group of patients, it may bias the performance of the prognostic model for certain groups of patients, which may not be visible when inspecting overall performance measures [40].

5. Conclusions

We concluded that – in our study – a relatively low quality text mining model (F1 score ≈ 0.50) could still be used to extract outcomes to train a prognostic prediction model with good discrimination (AUROC ≈ 0.80, close to using ground truth outcomes). The prognostic model's discrimination ability could be increased by slightly decreasing the decision threshold of the text mining model (below 0.5), resulting in an increase in the number of outcome events, contributing to the prognostic model's discriminative ability. In the vast majority of the experimental settings, the prognostic model's calibration was off, even for quite good text mining models (F1 score ≈ 0.80). Two possible ways to resolve this in practice are to (1) further improve the used text mining models if possible, or (2) recalibrate the prognostic model in a (possibly smaller manually annotated) dataset with high quality outcome labels.

Future work should focus on using available manually labeled data more efficiently, developing strategies to distribute labeled data effectively over text mining model development, and validating or recalibrating prognostic models. This would reduce the time-effort and domain expertise required to manually label clinical data.

Summary table

What was already known on the topic:

- Text mining is increasingly being used to extract relevant health outcomes from clinical notes in the development of prognostic models.
- Text mining systems rarely extract information without making errors.

What this study added to our knowledge:

- Even if text mining is of moderate quality, it can be used to extract outcomes to develop prognostic models with good discriminative ability.
- Even when text mining is of good quality, prognostic models developed on the text-extracted outcomes may still produce poorly calibrated risks.

CRediT authorship contribution statement

Zwierd Grotenhuis: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Pablo J. Mosteiro:** Methodology, Supervision, Writing – review & editing. **Artuur M. Leeuwenberg:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code used for this study is publicly available at https://github.com/zwierd99/ClinicalTM. The MIMIC-III database can be obtained with a data sharing agreement at https://physionet.org.

Acknowledgments

This research was funded by the Dutch Research Council, as part of the project "RAISE: Responsible AI Science Explorations" (Grant Number NWA.1418.22.008), and the Utrecht University Applied Data Science Fund. The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Appendix A. Feed-forward neural network details and results

A.1. Model architecture

The feed-forward neural network (FFNN) experiments, we used the same input and outcome data as for the logistic regression. After minimal tuning of the network's structure (splitting PM-train in a small train-validation split, maximizing for AUC), we arrived at a single densely connected layer with 16 units (from tuning grid {512, 256, 128, 16}), ReLU activation for the intermediate layer, and a logistic (softmax) activation for the final layer, to obtain probabilistic outputs. Dropout regularization was then added and set to 0.5. Finally, two output nodes with softmax activation were used. The loss function was binary cross-entropy.

A.2. FFNN results

See Figs. A.6–A.8.

Appendix B. Decision threshold results

B.1. Text mining decision thresholds and prediction model discrimination

Fig. B.9 shows how changing the decision threshold of the text mining model changes the AUROC of the prediction model for an equal split size of 0.5. Since we know from Fig. 3 that a decision threshold of around 0.3 leads to a higher F1 score and a higher F1 score leads to a higher AUROC, this figure is somewhat implied, but it is still interesting to confirm that a decision threshold of 0.3 led to the model with the highest AUROC for both FFNN and logistic regression.

B.2. Text mining decision thresholds and prediction model calibration

The influence of the text mining decision threshold, as seen in Figs. B.10–B.12, seems quite uniform across the intercept and calibration in the large for both prediction models, but the slope increases with a an increase of the decision threshold for the logistic regression model, while for the FFNN model the opposite is true. This means that the logistic regression model becomes more moderate (predicted probabilities closer to the average) the fewer samples it has to learn from, whereas the FFNN becomes more extreme (predicted probabilities closer to 0 and 1). The logistic regression has the best slope of around 1 (optimal is 1) at decision threshold 0.1, while the FFNN has the best slope it can achieve at 0.2. The intercept increases for both models as the decision threshold increases, and both models see an intercept closest to optimal (0) for a decision threshold around 0.5 or 0.6. The calibration in the large also increases for both models as the decision threshold shifts to the right, but the FFNN crosses the optimal CITL (0) at a decision threshold of 0.3, the logistic regression model only gets closer but never reaches it.

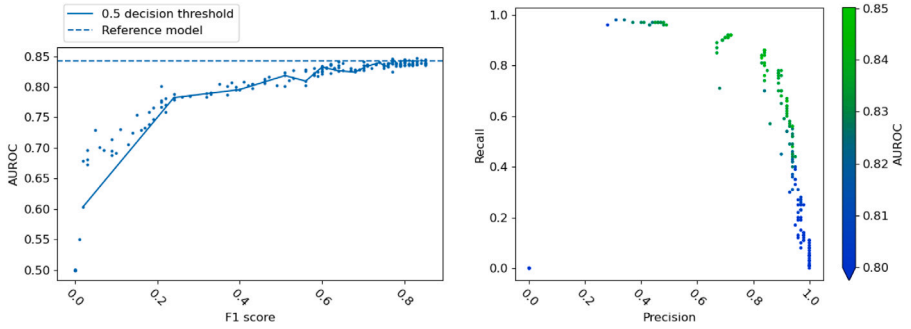


Fig. A.6. Left: The AUROC of the FFNN prediction model against the F1 scores of the text mining models. The dotted line indicates the reference model trained on the ground truth labels. Right: Precision and recall of the text mining model against the AUROC of the FFNN prediction model.

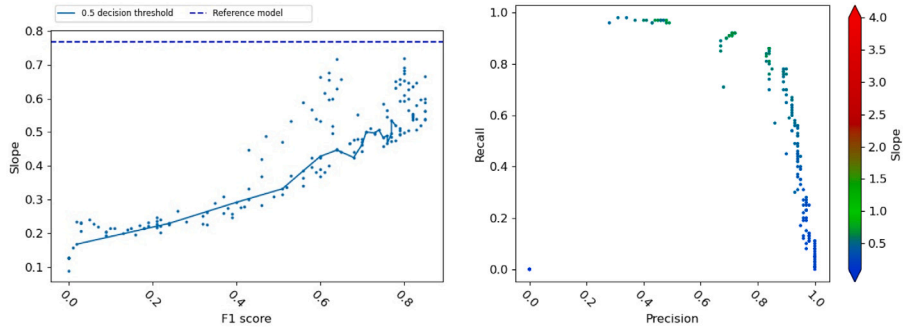


Fig. A.7. The calibration slope of the FFNN prediction models for the F1 scores (left) and the precision and recall (right) of the text mining models.

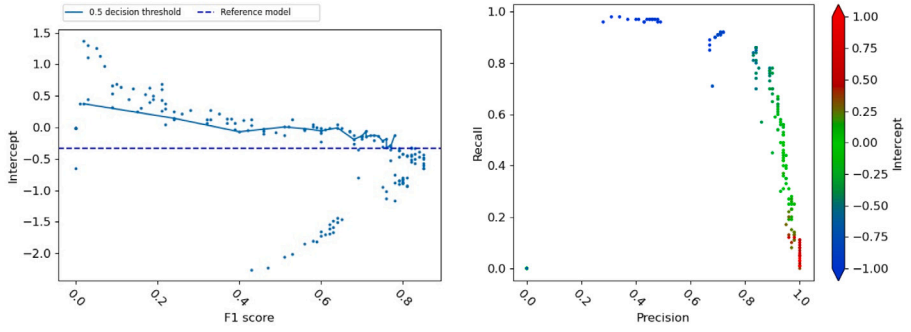
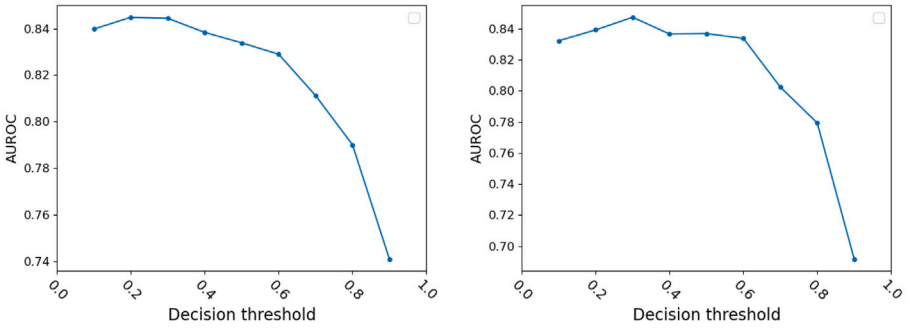
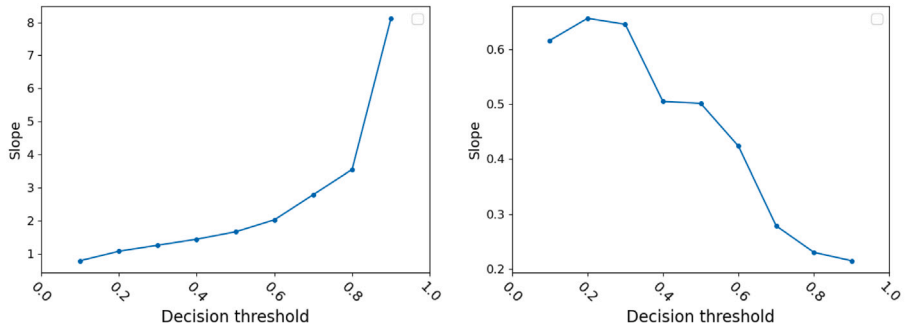


Fig. A.8. The calibration intercept of the FFNN prediction models against the F1 scores (left), and the precision and recall (right) of the text mining model.



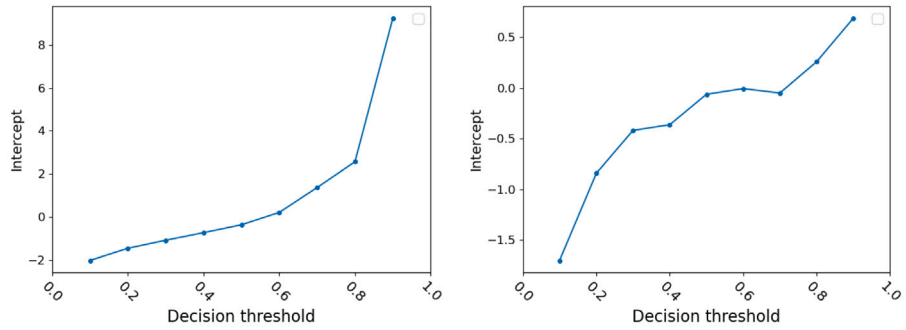
(a) Logistic regression (b) FFNN

Fig. B.9. The AUROC of the logistic regression model and the FFNN model for the different decision thresholds of the text mining model, with an equal split size of 0.5.



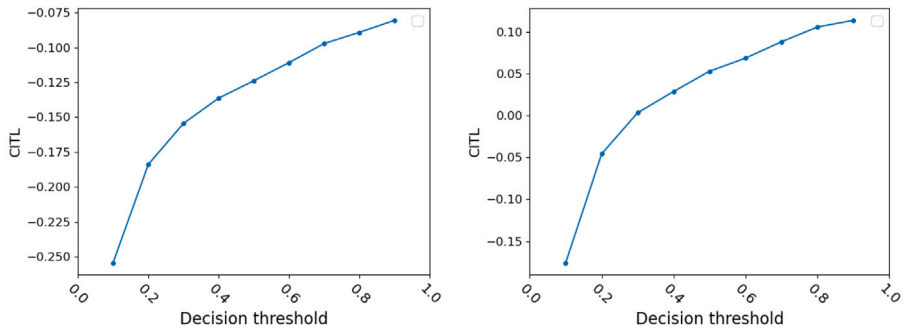
(a) Logistic regression (b) FFNN

Fig. B.10. The calibration slope of the prediction models for the different decision thresholds of the text mining model, with an equal split size of 0.5.



(a) Logistic regression (b) FFNN

Fig. B.11. The calibration intercept of the prediction models for the different decision thresholds of the text mining model, with an equal split size of 0.5.



(a) Logistic regression (b) FFNN

Fig. B.12. The calibration in the large of the prediction models for the different decision thresholds of the text mining model, with an equal split size of 0.5.

References

[1] A. Laupacis, N. Sekar, et al., Clinical prediction rules: a review and suggested modifications of methodological standards, JAMA 277 (6) (1997) 488–494.

[2] W. Bouwmeester, N.P. Zuithoff, S. Mallett, M.I. Geerlings, Y. Vergouwe, E.W. Steyerberg, D.G. Altman, K.G. Moons, Reporting and methods in clinical prediction research: a systematic review, PLoS Med. 9 (5) (2012) e1001221.

[3] B.E. Keuning, T. Kaufmann, R. Wiersema, A. Granholm, V. Pettilä, M.H. Møller, C.F. Christiansen, J. Castela Forte, H. Snieder, F. Keus, et al., Mortality prediction models in the adult critically ill: A scoping review, Acta Anaesthesiol. Scand. 64 (4) (2020) 424–442.

[4] L. Wynants, B. Van Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, M.M. Bonten, D.L. Dahly, J.A. Damen, T.P. Debray, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, Bmj 369 (2020).

[5] K.G. Moons, P. Royston, Y. Vergouwe, D.E. Grobbee, D.G. Altman, Prognosis and prognostic research: what, why, and how? Bmj 338 (2009).

[6] B.M. Reilly, A.T. Evans, Translating clinical research into clinical practice: impact of using prediction rules to make decisions, Ann. Intern. Med. 144 (3) (2006) 201–209.

[7] F. Roques, P. Michel, A. Goldstone, S. Nashef, The logistic euroscore, Eur. Heart J. 24 (9) (2003) 882–883.

[8] S.A. Nashef, F. Roques, L.D. Sharples, J. Nilsson, C. Smith, A.R. Goldstone, U. Lockowandt, Euroscore ii, Eur. J. Cardio-Thorac. Surg. 41 (4) (2012) 734–745.

[9] W.A. Knaus, J.E. Zimmerman, D.P. Wagner, E.A. Draper, D.E. Lawrence, APACHE—acute physiology and chronic health evaluation: a physiologically based classification system, Crit. Care Med. 9 (8) (1981) 591–597.

[10] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, APACHE II: a severity of disease classification system., Crit. Care Med. 13 (10) (1985) 818–829.

[11] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, A. Damiano, et al., The APACHE

III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults, Chest 100 (6) (1991) 1619–1636.
[12] J.E. Zimmerman, A.A. Kramer, D.S. McNair, F.M. Malila, Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients, Crit. Care Med. 34 (5) (2006) 1297–1310.
[13] E.W. Steyerberg, Applications of prediction models, in: Clinical Prediction Models, Springer, 2009, pp. 11–31.
[14] E.W. Steyerberg, K.G. Moons, D.A. van der Windt, J.A. Hayden, P. Perel, S. Schroter, R.D. Riley, H. Hemingway, D.G. Altman, P. Group, Prognosis research strategy (PROGRESS) 3: prognostic model research, PLoS Med. 10 (2) (2013) e1001381.
[15] A.A. de Hond, A.M. Leeuwenberg, L. Hooft, I.M. Kant, S.W. Nijman, H.J. van Os, J.J. Aardoom, T. Debray, E. Schuit, M. van Smeden, et al., Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review, npj Digit. Med. 5 (1) (2022) 1–13.
[16] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, J. Br. Surg. 102 (3) (2015) 148–158.
[17] B.A. Goldstein, A.M. Navar, M.J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (1) (2017) 198–208.
[18] B. Percha, Modern clinical text mining: a guide and review, 2021.
[19] H. Dalianis, Clinical Text Mining: Secondary Use of Electronic Patient Records, Springer Nature, 2018.
[20] R. Zhu, X. Tu, J.X. Huang, Utilizing BERT for biomedical and clinical text mining, in: Data Analytics in Biomedical Engineering and Healthcare, Elsevier, 2021, pp. 73–103.
[21] T.M. Seinen, E. Fridgeirsson, S. Ioannou, D. Jeannetot, L.H. John, J.A. Kors, A.F. Markus, V. Pera, A. Rekkas, R.D. Williams, et al., Use of unstructured text in prognostic clinical prediction models: a systematic review, 2022, medRxiv.
[22] T.R. Hylan, M. Von Korff, K. Saunders, E. Masters, R.E. Palmer, D. Carrell, D. Cronkite, J. Mardekian, D. Gross, Automated prediction of risk for problem opioid use in a primary care setting, J. Pain 16 (4) (2015) 380–387.
[23] N. Dormosh, M.C. Schut, M.W. Heymans, N. van der Velde, A. Abu-Hanna, Development and internal validation of a risk prediction model for falls among older people using primary care electronic health records, J. Gerontol. (2021).
[24] G. Kadra, A. Spiros, H. Shetty, E. Iqbal, R.D. Hayes, R. Stewart, H. Geerts, Predicting parkinsonism side-effects of antipsychotic polypharmacy prescribed in secondary mental healthcare, J. Psychopharmacol. 32 (11) (2018) 1191–1196.
[25] S. Fu, L. Wang, S. Moon, N. Zong, H. He, V. Pejaver, R. Relevo, A. Walden, M. Haendel, C.G. Chute, et al., Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review, Clin. Transl. Sci. (2022).
[26] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.
[27] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (1) (2019) 1–18.
[28] B. Van Calster, D.J. McLernon, M. Van Smeden, L. Wynants, E.W. Steyerberg, Calibration: the achilles heel of predictive analytics, BMC Med. 17 (1) (2019) 1–7.
[29] S. Fu, L. Wang, S. Moon, N. Zong, H. He, V. Pejaver, R. Relevo, A. Walden, M. Haendel, C.G. Chute, et al., Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review, Clin. Transl. Sci. 16 (3) (2023) 398–411.
[30] D.E. Shapiro, The interpretation of diagnostic tests, Stat. Methods Med. Res. 8 (2) (1999) 113–134.
[31] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, vol. 398, John Wiley & Sons, 2013.
[32] J.J. Lahoz-Monfort, G. Guillera-Arroita, B.A. Wintle, Imperfect detection impacts the performance of species distribution models, Glob. Ecol. Biogeography 23 (4) (2014) 504–515.
[33] L.C. Rosella, P. Corey, T.A. Stukel, C. Mustard, J. Hux, D.G. Manuel, The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model, Popul. Health Metr. 10 (1) (2012) 1–11.
[34] K. Luijken, R.H. Groenwold, B. Van Calster, E.W. Steyerberg, M. van Smeden, Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective, Stat. Med. 38 (18) (2019) 3444–3459.
[35] T.H. Kappen, W.A. van Klei, L. van Wolfswinkel, C.J. Kalkman, Y. Vergouwe, K.G. Moons, Evaluating the impact of prediction models: lessons learned, challenges, and recommendations, Diagn. Progn. Res. 2 (1) (2018) 1–11.
[36] K.G. Moons, A.P. Kengne, D.E. Grobbee, P. Royston, Y. Vergouwe, D.G. Altman, M. Woodward, Risk prediction models: II. External validation, model updating, and impact assessment, Heart 98 (9) (2012) 691–698.
[37] Y. Vergouwe, D. Nieboer, R. Oostenbrink, T.P. Debray, G.D. Murray, M.W. Kattan, H. Koffijberg, K.G. Moons, E.W. Steyerberg, A closed testing procedure to select an appropriate method for updating prediction models, Stat. Med. 36 (28) (2017) 4529–4539.
[38] B. Van Calster, E.W. Steyerberg, L. Wynants, M. van Smeden, There is no such thing as a validated prediction model, BMC Med. 21 (1) (2023) 70.
[39] S.A. van Laar, K.B. Gombert-Handoko, H.-J. Guchelaar, J. Zwaveling, An electronic health record text mining tool to collect real-world drug treatment outcomes: a validation study in patients with metastatic renal cell carcinoma, Clin. Pharmacol. Ther. 108 (3) (2020) 644–652.
[40] S.A. Frost, E. Alexandrou, Misclassification and measurement error–planning a study and interpreting results, Nurse Res. 30 (3) (2022).