Machine learning for identifying relevant publications in updates of systematic reviews of diagnostic test studies

Toni Lange | Guido Schwarzer | Thomas Datzmann | Harald Binder

1Center for Evidence-based Healthcare, University Hospital Carl Gustav Carus and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Germany
2Faculty of Medicine and Medical Center, Institute of Medical Biometry and Statistics, University of Freiburg, Freiburg, Germany

Correspondence
Toni Lange, Center for Evidence-based Healthcare, University Hospital Carl Gustav Carus and Medical Faculty Carl Gustav Carus, TU Dresden, Fetscherstr. 74, 01307 Dresden, Germany.
Email: toni.lange@uniklinikum-dresden.de

Abstract
Updating systematic reviews is often a time-consuming process that involves a lot of human effort and is therefore not conducted as often as it should be. The aim of our research project was to explore the potential of machine learning methods to reduce human workload. Furthermore, we evaluated the performance of deep learning methods in comparison to more established machine learning methods. We used three available reviews of diagnostic test studies as the data set. In order to identify relevant publications, we used typical text preprocessing methods. The reference standard for the evaluation was the human-consensus based on binary classification (inclusion, exclusion). For the evaluation of the models, various scenarios were generated using a grid of combinations of data preprocessing steps. Moreover, we evaluated each machine learning approach with an approach-specific predefined grid of tuning parameters using the Brier score metric. The best performance was obtained with an ensemble method for two of the reviews, and by a deep learning approach for the other review. Yet, the final performance of approaches strongly depends on data preparation. Overall, machine learning methods provided reasonable classification. It seems possible to reduce human workload in updating systematic reviews by using machine learning methods. Yet, as the influence of data preprocessing on the final performance seems to be at least as important as choosing the specific machine learning approach, users should not blindly expect a good performance by solely using approaches from a popular class, such as deep learning.

Keywords
machine learning, deep learning, text mining, preprocessing, systematic reviews

1 INTRODUCTION
In patient-centered medicine, the integration of external evidence is a crucial component in deciding on the use of medical services. Clinicians, researchers and health policy makers have to deal with a multitude of publications within their field of expertise. A systematic review summarizes the empirical evidence according to a priori defined inclusion criteria and serves as an external evidence basis for clinical decisions. An increasing number

© 2021 The Authors. Research Synthesis Methods published by John Wiley & Sons Ltd.
of studies such as seen in the field of medicine in recent years, also leads to the need for a higher frequency of systematic review updates. Updating systematic reviews is a very resource-intensive process, which creates a major barrier for up-to-date evidence syntheses. Intelligent technical support systems, for example, based on machine learning techniques, are a promising approach for reducing human effort in this updating process. In particular, deep learning techniques, which have recently become quite popular, are expected to considerably advance this area. We specifically considered the setting of systematic reviews of diagnostic test studies and compared the performance of several machine learning techniques, also including deep learning approaches.

In the context of systematic reviews and living reviews, that is, continuous updates, various computer-assisted approaches were introduced to increase the efficiency. Currently, however, living reviews are not widely used as human effort for continuous updating within a few months is very substantial. Despite existing efforts to automate sub-processes of systematic/living reviews, these are only rarely applied in research practice. Initially, studies investigated the performance of established machine learning methods, such as Support Vector Machines or Random Forests. The results obtained with these methods have already highlighted the potential of such support systems. Recently, the application of deep learning methods for natural language processing has been growing in addition to conventional machine learning methods. Machine learning and in particular deep learning methods have mostly been successful in big data applications. In this research study, we investigated to what extent such techniques can be useful in settings with relatively small data sizes.

To the best of our knowledge, only one study by Marshall et al. compared convolutional neural networks (CNN) with other machine learning methods in the context of systematic reviews. Furthermore, current research has mainly focused on the classification of publications for which detailed reporting standards have long been available like randomized controlled trials (RCTs).

We specifically evaluated the performance of machine learning methods for identifying relevant publications in the context of updating a systematic review of diagnostic studies. Since they are characterized by a rather low standardization of reporting, classification is made more difficult. In addition to comparing different machine learning techniques, we also focused in particular on the effects of data preprocessing and tuning parameter selection on classification performance, as these might have considerable influence.

### What is already known

- Updating systematic reviews is a resource- and time-intensive process that requires a lot of human effort.
- Machine learning algorithms were successfully tested in the automatic classification of studies that had high standardized report quality, such as randomized controlled trials (RCTs).

### What is new

- This study shows the potential of different machine learning algorithms to reduce human workload in updating systematic reviews of studies with a low standardized reporting quality.

### Potential impact for Research Synthesis Methods: Readers outside the authors’ field

- Data preprocessing was the main influencing factor for a successful classification of studies. Thus, researchers should be aware of the importance of directly integrating text preprocessing procedures into model training as tuning parameters.
- Existing training and evaluation metrics encounter problems in class imbalance situations. A metric for probability models with a weighting factor including an adjustment for prevalence is required.

## 2 | MATERIALS AND METHODS

### 2.1 | Data sets

We used three reviews to assess the performance of various machine learning methods in the context of systematic reviews (Table 1). These are data from the screening process of two already published reviews (R-1, R-3) of diagnostic studies in the field of orthopedics and physical therapy. The review R-2 contains data of an unpublished Cochrane review. The sizes are typical for diagnostic test systematic reviews in the field of medicine. The investigated systematic reviews aimed to evaluate the validity and reliability of a physical examination...
test for structural and functional testing of the anterior and posterior cruciate ligament and shoulder pathologies. Included were studies which (a) examined the desired physical target structure and (b) the quality of the physical tests. The original queries were carried out in the following databases: MEDLINE via PubMed/Ovid, EMBASE via Ovid, AMED via Ovid, CINAHL via EBSCOhost, DARE via CRDWeb, MEDION, and ARIF via its own websites (www.mediondatabase.nl, www.arif.bham.ac.uk/).

### 2.2 Study procedure

Figure 1 illustrates the entire study process containing text preprocessing, scenario preparation, training of machine learning methods and evaluation. In order to identify relevant publications, the title and abstract of each publication were available as training data. The reference standard for the evaluation of the machine learning methods was the human-consensus based classification (inclusion, exclusion) of the studies based on title and abstract screening. The various phases of the study are described in more detail below.

#### 2.3 Text preprocessing

After the export from the literature databases, the titles and abstracts highly differed in terms of format. Abstracts, for example, varied in structure and sometimes additional unnecessary information was reported and had to be removed. Otherwise, it would have distorted the classification. Therefore, in the first processing step, we preprocessed the texts with the aim of creating a consistent record format with optimized information content.

We used various text cleaning procedures and removed additional content by using regular expressions in an iterative process (e.g., journal or author information). Furthermore, we applied standardized text procedures (punctuation handling and upper/lower case, replacement of numbers and symbols as well as removal of stop words) to all titles and abstracts. Two different
types of text reduction techniques were compared to each other: (a) Stemming - individual text words were reduced to the undeclined root word to identify words with the same content as the same token. (b) Lemmatization - inflected words were grouped together into their original form.

### 2.4 | Data scenario preparation

After text preprocessing, all titles and abstracts were prepared for model training and various scenarios were generated using a mixture of data preprocessing steps. Figure 2 outlines this procedure. Our aim was to ensure a fixed allocation of publications into training as well as test data and to provide various data preparation scenarios for model training in a compact way. We manipulated the sampling ratio of relevant versus irrelevant publications (inclusion/exclusion) to investigate a possible effect of the balancing on model training.

Two different types of data representation were required for different machine learning methods. On the one hand, (a) conventional machine learning methods use sparse vectors/document-term-matrices as input data. On the other hand, (b) deep learning methods with word embedding require the input data represented as dense vectors. For both types of the data representation, we varied (a) both types of text reduction techniques, and (b) the maximum number of initial tokens (original frequency, fifteen, ten as well as one thousand and five hundred). Furthermore, for the document-term-matrix, we used (a) term-frequency (TF) and (b) term-frequency—inverse document frequency (TF-idf) as weighting scheme for tokens and varied the number of n-grams between one and three words/tokens. This resulted in 150 different data scenarios per review for the conventional machine learning methods and 50 data scenarios for the deep learning approaches.

### 2.5 | Word embedding

We generated the word embeddings by a skip-gram model with the preprocessed three reviews (R-1, R-2, R-3) as inputs. In total, two word embeddings were used based on both text reduction techniques (a) stemming and (b) lemmatization. We used the following parameters for the word embeddings: Loss function: binary cross-entropy, epochs: 10, batch size 1, learning rate 0.01, optimizer: Adam (decay: 0, $\beta_1$: 0.9, $\beta_2$: 0.999), dimensions: 300, skip window: 5, negative sample: 1.

### 2.6 | Model training

We trained different models on each specific data scenario (Figure 2). Overall, we investigated six conventional machine learning methods, two deep learning methods, and finally, two ensemble methods.

For conventional machine learning methods: (SL-1) logistic regression with elastic net regularization, (SL-2) Naive Bayes, (SL-3) Support Vector Machines, (SL-4) K-Nearest Neighbors, (SL-5) C5.0 trees, and (SL-6) a Random Forest algorithm were used (Table S1).

Further, we investigated the performance of different deep learning methods: (DL-1) Multilayer Perceptron (MLP)—Bag of Words approach, (DL-2) MLP—with word embedding, (DL-3) Convolutional Neural Networks...
We evaluated two different ensemble methods, which comprise conventional machine learning methods. (E-1) Ensemble stack with a Random Forest algorithm as meta learner and the (E-2) ensemble soft voting averaging the prediction probabilities of the base learners. Both Ensembles (E-1, E-2) used the scenario specific best-trained base learner (B-1) logistic regression with elastic net regularization, or (B-2) Support Vector Machines, as well as (B-3) a Random Forest algorithm. In addition, we used a soft voting ensemble to combine (E-3) deep learning methods.

The pseudo code in Figure 3 illustrates the entire model training process. For each model, a 10-fold cross-validation with three repetitions was used to identify the optimal tuning parameter set for each data scenario. To enable a fair comparison, it was ensured that resampling was identical across all models during cross-validation. The optimal data scenario for each algorithm with its specific optimal tuning parameters was determined by averaging across hold-out predictions while training and choosing the best performing scenario.

2.7 Model evaluation

The evaluation of the methods was carried out step by step. As outlined in Figure 3, (a) we selected the optimal tuning parameter constellation per scenario, (b) we identified the optimal scenario per method, and then (c) we used the optimal tuning parameters per method and scenario to determine the predictive value of the respective method on independent test data.

1. The Brier score (Formula 1) was primarily used as selection criterion for the optimal tuning parameters. This score describes the precision of predictions by determining the distance of predicted probabilities from the reference standard.\(^{15}\)

2. To enable a fair comparison between scenarios with differently balanced proportions, the Brier score for each scenario was compared to the scenario-specific null model. We call this “Brier comparison score” and used this score to determine the optimal scenario (Formula 2). As a null model, we trained an intercept only model.

3. After the models were trained with the optimal tuning parameter set on the optimal scenario, we determined final prediction probabilities on a corresponding independent test data set. The Brier comparison score was primarily used as an evaluation metric. Other metrics such as sensitivity, specificity, as well as positive and negative predictive values, were also calculated.

Formula 1: Brier score
Formula 2: Brier comparison score

\[
BS_{\text{comp}} = BS_{\text{null}} - BS_{\text{mod}}
\]

2.8 | Sensitivity analysis

To compensate for the high class imbalance (few inclusions/many exclusions) that is inherent in all title and abstract screenings, we lowered the default threshold value (50:50) for binary classification to evaluate its influence on the prediction success. We reduced the threshold from 0.5 to 0.2 to improve the classification sensitivity of the models (i.e., lower the false negative rate).

All statistical analyses were performed using the statistical software R. For training and systematic evaluation, the caret package was used, which streamlines the process for creating predictive models including implementation of self-developed algorithms, loss functions, and evaluation metrics. This package also enables the evaluation of models in direct comparison in a sufficient manner.

3 | RESULTS

Overall, three different but thematically similar systematic medical reviews were analyzed. Specifically, titles and abstracts of all individual publications contained in these reviews were used, however, no full texts. An abstract contains on average 225.2 ± 117.4 words (median = 226; min/max = 2/708). After data preprocessing (stemming and lemmatization), the maximum overall number of tokens (words) per review varied within a range of: R-1: 16.807/19.162; R-2: 12.722/14.684; R-3: 5.924/6.905.

The data preprocessing grid (shown in Figure 2) resulted in 1038 different data scenarios in total.
3.1 | Evaluation of data scenarios

Figure 4 gives a comparison of models during the training step (cross validation) for one review (R-1) as example. Each data point represents a prediction for a specific data preparation scenario as measured by the Brier score. On the y-axis, the Brier comparison score is shown. Positive values indicate a better model performance than the scenario specific intercept only null model. Note that some methods tend to yield rather poor predictions with negative Brier comparison scores (e.g., K-NN, Naïve Bayes). Furthermore, variation between models within one algorithm is higher than the variation between different algorithms. Depending on the specific data preparation procedures, a method might either perform well or rather poor. Yet, there is no clear relationship between specific data preparation steps and the performance of the methods. The algorithms perform differently on the same data preparation scenarios but with no distinguishable patterns. It seems that each algorithm needs its own optimal preprocessing steps, also strongly depending on the input data. Two different data preprocessing methods that were designed to reduce dimensionality of the feature space showed an impact on model performance in many cases. One method is token reduction, which aimed to increase the signal to noise ratio. The other method is word embedding, which was specifically used for the deep neural networks. The latter takes proximity of words into account and generated a lower dimensional vector out of a high dimensional matrix. We used the word embeddings to pre-weight input data of our deep neural network models. The above findings remained consistent across all reviews examined.

3.2 | Evaluation of test data performance

The final evaluation of model performance (i.e., generalizability) was based on independent test data. Figure 5 shows a heatmap with Brier comparison scores for all investigated reviews (R1-3). Cells with a light gray background (values above zero) point to a better performance of the corresponding model compared to the scenario specific null model, while dark cells (values below zero) indicate a worse performance. The best performing algorithm varied between the reviews examined (R1: Ensemble (Soft-Voting), R2: MLP (20 epochs), R3: Ensemble (Soft-Voting)). Across all reviews, there was no clear winner and performance varies between reviews. Review R3 was most challenging during test data performance.
prediction. In review R2, deep learning outperformed most conventional machine learning methods, while in review R1 and R3 ensembles of conventional machine learning methods were superior.

Note that model tuning, especially the use of thresholding or generalized word embeddings, has not been fully explored systematically for deep learning methods. Depending on the evaluation metric, the best performing algorithm changed within a review. In the present case of updates of systematic reviews, there is always a strong class imbalance with the minority class being of particular interest. Under this scenario, a stable evaluation metric is needed incorporating this high class imbalance. This would certainly improve the quality of results. Ensemble methods, incorporating more than one algorithm, may lead to better results here compared to single algorithms since multi-model ensembles increased the sensitivity of the analysis purely by incorporating more than one algorithm, especially by aggregating with plurality or soft voting.

### 3.3 Sensitivity analysis

For sensitivity analysis, we lowered the classification threshold from 0.5 (equal probabilities on both sides) to 0.2 inducing a more sensitive and less specific classification. This clearly led to better results since a more sensitive classification was superior in this task. Not missing a potentially relevant manuscript and therefore increasing the inclusion rate (lower the false negative rate) was of advantage here (Table S3–S5).

### 4 DISCUSSION

We investigated the performance of machine learning methods for systematic reviews in the challenging setting of diagnostic test studies. Besides the general level of performance, we were interested in whether deep learning approaches would have a considerable advantage over conventional machine learning methods. We found that deep neural network approaches were among the top performers but did not consistently or considerably outperform conventional machine learning methods such as Random Forest algorithm or ensembles. This might be due to the rather low number of publications, and small amounts of text available from titles and abstracts.

Searching for the best algorithm for a task is only part of the picture. We did not intend to give a suggestion for a specific best method to be used in general based on our study including only three reviews. Instead, our results suggest that data preprocessing steps had much more influence on the final performance. With a bad choice of preprocessing steps, any approach could drop to even below the performance level of the scenario specific null model, which only uses an intercept as predictor variable and no words. Having encountered these difficulties in our three reviews, we are sure that these issues will also be problematic in other reviews. On the other hand, most algorithms could perform quite well on a specific combination of data preprocessing steps. Unfortunately, it seems difficult to predict in advance which preprocessing steps are best suited for which algorithm and review. Therefore, we strongly encourage considering data preprocessing besides other predictors as a tuning parameter for training of machine learning approaches. Yet, one consistent pattern is that techniques reducing the dimensional load of models are preferable. For example, token reduction, which increases the signal to noise ratio, is usually beneficial. The same trend is observed for embeddings, which reduce dimensions in deep neural network models. Our finding is that the choice of data preprocessing can have a dramatic effect, which can easily be overlooked when discussing different machine learning methods on a more general level. Thus, future research should in our opinion focus on investigating such effects also in other review question types.

In the process of updating systematic reviews, it is of utmost importance not to miss a relevant publication, thus high sensitivity was needed. In fact, in our medical context we aimed to achieve a sensitivity of one (zero missed studies). It is necessary to use all available evidence in medical decision-making. Accordingly, we required that a usable algorithm must have a sensitivity of one and a specificity that allows to reduce the number of publications to be examined manually at least by a factor of two. By lowering the binary classification threshold from 0.5 to 0.2, we were able to achieve a sensitivity of one. At the same time, the number of publications need to be screened manually was reduced by more than 50% in each review. We did not use a systematic search for the optimal threshold for each review, instead we deliberately chose a value of 0.2 for all sensitivity analyses. Cross-validation on a separate test data would in principle allow to find the optimal threshold for the specific problem.

### 4.1 Methodological considerations

A stable comparison metric in the setting of high class imbalance is desirable to clarify the question of which algorithm works best. In the presence of a high class imbalance, the null models tend to be extremely good in performance when measured by all used evaluation metrics, including the Brier comparison score. In fact, the
higher the imbalance, the better the null model. Another potential measure is the Net Benefit,\textsuperscript{18} which considers the class imbalance by weighting the difference in true and false positive rate. However, the advantage of the Brier score is the incorporation of the distance of the predicted probabilities. In contrast, the Net Benefit uses only the binary classification. Thus, an inclusion of a prevalence weighting factor for the Brier score could improve the performance in class imbalance problems.

A recent study showed that “generic design classifiers” (distinguishing between RCT and non-RCT) can achieve acceptable performance for use in practice.\textsuperscript{7} In contrast, our study focused on a specialized algorithm that simultaneously identifies the correct study design (diagnostic test studies) and topic (selection of studies investigating shoulder pathologies). Our results suggest that such a “review-specific classifier” that updates an existing review should additionally be the focus of further investigations to explore the potential of human workload reduction with an acceptable proportion of false negative predictions. Maybe, a combination of various meta data could increase the classification performance. In addition to title and abstract, further information such as full text, MeSH term indexing or hierarchical inclusion of keywords amongst others could be used. Thus, separate classifiers trained on the different input information could be combined for improved predictions e.g. with a majority-vote ensemble.

4.2 | Strengths and weaknesses

One of the main strengths of this work is the realization and implementation of all working processes in a unified development environment, which allow us to systematically generate and evaluate different models on different data sets with different data pre-processing steps. According to the review by O’Mara-Eves et al.,\textsuperscript{3} the Brier Score was not included in any recent scientific research of machine learning methods as evaluation metric in the context of updating systematic reviews. In favor of a balanced prevalence in training and test data sets, we did not assess the potential effect of changes in the quality of reporting over time. The tested machine learning and deep learning methods represent only a selection of methods potentially available for this classification task.

5 | CONCLUSIONS

Overall, the presented work shows that it is possible to reduce the human workload in updating scientific systematic reviews by machine learning methods even for studies with a low standardized reporting quality. There is no big performance difference between deep learning and other machine learning approaches. Instead, the influence of data preprocessing on the final model performance is quite strong. Optimizing the threshold assessment and evaluating additional, desirable stable, metrics in high class imbalance settings seems more urgent and should be the topic of future research.

ACKNOWLEDGMENTS

The authors would like to thank Christian Kopkow and Alice Freiberg for providing their screening data for model training. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare they have no conflict of interest.

AUTHOR CONTRIBUTIONS

Toni Lange, Harald Binder, Guido Schwarzer, Thomas Datzmann: Conceived the study. Toni Lange: Performed the analysis. All authors drafted the paper and have critically revised it and approved the final version. A preprint was published on medRxiv (https://www.medrxiv.org/content/10.1101/2020.06.16.20132670v1).

DATA AVAILABILITY STATEMENT

The data sets used in the current study are available from the corresponding author on reasonable request. Example data set and R-Code is available at https://github.com/toni-ml/sys_review_ml

ORCID

Toni Lange https://orcid.org/0000-0002-8671-7496
Guido Schwarzer https://orcid.org/0000-0001-6214-9087
Thomas Datzmann https://orcid.org/0000-0003-0312-5901

REFERENCES

1. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6(7):e1000100. https://doi.org/10.1371/journal.pmed.1000100.
2. Fontelo P, Liu F. A review of recent publication trends from top publishing countries. Syst Rev. 2018;7(1):147. https://doi.org/10.1186/s13643-018-0819-1.
3. O’Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1):5. https://doi.org/10.1186/2046-4053-4-5.
4. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic review N. Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol*. 2017;91:31-37. https://doi.org/10.1016/j.jclinepi.2017.08.011.

5. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction-the why, what, when, and how. *J Clin Epidemiol*. 2017;11:1878-5921. https://doi.org/10.1016/j.jclinepi.2017.08.011.

6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. https://doi.org/10.1038/nature14539.

7. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying Randomized Controlled Trials: an evaluation and practitioner’s guide. *Res Synth Meth*. 2018;9(4):602-614. https://doi.org/10.1002/jrsm.1287.

8. Korevaar DA, van Enst WA, Spijker R, Bossuyt PMM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid Based Med*. 2014;19(2):47. https://doi.org/10.1136/eb-2013-101637.

9. Lange T, Matthijs O, Jain NB, Schmitt J, Lutzner J, Kopkow C. Reliability of specific physical examination tests for the diagnosis of shoulder pathologies: a systematic review and meta-analysis. *Br J Sports Med*. 2017;51(6):511-518. https://doi.org/10.1136/bjsports-2016-096558.

10. Lange T, Struyf F, Schmitt J, Lutzner J, Kopkow C. The reliability of physical examination tests for the clinical assessment of scapular dyskinesis in subjects with shoulder complaints: A systematic review. *Phys Therapy Sport*. 2017;26:64-89. https://doi.org/10.1016/j.ptsp.2016.10.006.

11. Kopkow CF, Kirschner A, Seidler S, Schmitt A. Physical examination tests for the diagnosis of posterior cruciate ligament rupture: a systematic review. *J Orthop Sports Phys Ther*. 2013;43(11):804-813. https://doi.org/10.2519/jospt.2013.4906.

12. Kopkow C, Lange T, Hoyer A, Lützner J, Schmitt J. Physical tests for diagnosing anterior cruciate ligament rupture. *Cochrane Database Syst Rev*. 2015. https://www.cochrane.org/CD011925/MUSKINJ_physical-tests-diagnosing-anterior-cruciate-ligament-rupture.

13. Goldberg YA. Primer on neural network models for natural language processing. *J Artif Intell Res*. 2016;57(1):345-420. https://doi.org/10.1613/jair.4992.

14. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *ICLR Workshop*. 2013.

15. Binder H, Graf E. *Brier score*. Vol 1. London, England: SAGE Publications; 2009.

16. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2018.

17. Kuhn M. *Caret: Classification and Regression Training*. Vol 6. R Package Version; 2018:1-80.

18. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagn Progn Res*. 2017;1(1):19. https://doi.org/10.1186/s41512-017-0020-3.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article**: Lange T, Schwarzer G, Datzmann T, Binder H. Machine learning for identifying relevant publications in updates of systematic reviews of diagnostic test studies. *Res Syn Meth*. 2021;12:506–515. https://doi.org/10.1002/jrsm.1486