Active learning aided systematic reviews within highly inclusive datasets

A simulation study into the performance of active learning aided systematic reviews in datasets with a high number of relevant records.

T.S. Klopper

Department of Methods & Statistics, Utrecht University

Bachelor thesis

Supervised by: L. Hofstee

Date: 28-06-2021

1. Background

The Cochrane handbook for Systematic Reviews of Interventions specifies the necessity of systematic reviews (Higgins et al., 2019). According to them, systematic reviews prevent bias of individual research, create a complete understanding of a certain research topic, and could identify gaps in the literature. This could ensure that scarce human effort is deployed in the right area of study. Furthermore, they emphasize that good systematic reviews should be updated every two years to stay relevant. Although this is important, it takes a tremendous amount of time and money to complete a systematic review. Borah et al. (2017) estimated that a full systematic review takes on average 67 weeks to complete. Additionally, manually screening is sensitive to human errors. Wang et al. (2020) found that there is on average an error rate of 10% while manually screening abstracts.

Active learning could provide a solution for these problems as it offers an approach which is faster and less prone to human errors than the traditional method. Active learning is defined as a subset of machine learning in which a researcher supervises the machine learning process (Settles, 2009; Miwa et al., 2014). Miwa et al. (2014) describe this process as training a machine through an iterative process until a certain stopping criteria is reached. In abstract screening, the phase in which active learning could be beneficial, this comes down to a researcher screening an abstract and then labelling it relevant or irrelevant. Afterwards, the machine trains with the newly acquired data and returns a record for labelling to the researcher (Ferdinands et al., 2020). To test the efficacy of this active learning cycle during the screening phase, O'Mara-Eves et al. (2015) conducted a systematic review of its benefits. They found that the use of active learning could decrease the average workload with 30% up to 70% and one study reported greater results which ranged up to a decrease in workload of 98% (Bekhuis et al., 2014). Additionally, this decrease in workload could decrease the error rate of systematic reviews (Wang et al., 2020). Current errors in the screening phase occur through researchers fatigue, distraction and systematic human biases (Bannach-Brown, 2019). Active learning could be a solution to these problems as well. A decreased workload could be beneficial for decreasing the researchers fatigue and distraction errors while software features like author removal could prevent systematic human biases (Wallace et al., 2010; van de Schoot et al., 2021).

To successfully complete the active learning cycle, machine learning models consist of: feature extraction techniques, classifiers, query strategies and balance strategies. First, the feature extraction technique extracts the most important features from a text with the goal to reduce noise and create a vector which is used for analysis later in the process (Guyon& Elisseeff, 2006). In abstract prioritization, the text which is transformed by the feature extraction technique is the title and abstract of a record. Secondly, classifiers are used to assign records to different classes based on the feature vectors (Colas & Brazdil, 2006). In the case of abstract screening these classes are 'relevant' or 'irrelevant'. For each record, a score is generated between 1 and 0 which predicts the relevancy of the record (Ferdinands et al., 2020). Subsequently, the query strategy determines which records are selected by the machine to present to the researcher. Certainty based sampling will show the researcher the record with the highest relevancy scores, or query probability (Fu & Lee, 2013), while uncertainty based sampling focusses on the relevancy score in the middle of the spectrum, ergo the papers of which the relevancy is the most uncertain (Van de Schoot et al., 2021). Finally, there are balance strategies to address class imbalance problems. These issues occur because the class distribution is often extremely skewed. As a result, classifiers tend to be biased towards the majority class and show worse performance on the minority class (Longadge & Dongre, 2013). The most common methods of dealing with this problem are oversampling and undersampling (Chawla et al., 2002). Oversampling focusses on resampling the minority group by adding copies of relevant papers and undersampling focusses on resampling the majority group by removing random papers (Chawla, 2009). Both balance strategies, and various variations, tend to have positive results towards the effectiveness of machine learning in imbalanced datasets (Ertekin et al., 2007).

Although the effectiveness of abstract prioritisation has been verified multiple times over the years (Jin & Yen, 2015; O'Mara-Eves et al., 2015; Howard et al., 2016; Olorisade et al., 2019), there are still gaps in the literature. Whilst the previously mentioned studies address the problem of class imbalance, all datasets have an inclusion percentage between the 2-35%. This is surprising, as studies with a high number of inclusions which use active learning do occur in the literature (van Lissa, 2021). Examples of these datasets with a high number of inclusions could be systematic reviews with a narrow search string or the creation of subsets from a dataset. However, little to no simulation studies have been conducted on the performance and the improvement of active learning in datasets with high inclusion percentages. To make active learning a more inclusive solution to the problem of systematic reviews, it is important to fill this gap in the literature and better examine the possibilities of systematic reviews for all types of datasets. Therefore, this research aims to provide the first insights into the performance of active learning aided systematic reviews in highly inclusive datasets.

To do so, this paper will try to answer three main research questions. Firstly, how does active learning perform in datasets with an high inclusion percentage? It is essential to measure the overall performance of active learning to test if highly inclusive datasets are suitable for practical research. It is expected that the performance of active learning in general becomes lower in datasets with a high percentage of relevant records. This is due to the measurements of performance. The ultimate goal of active learning is to decrease the workload of researchers. However the higher the percentage of inclusions, the lower the margin in which active learning can be beneficial. Therefore, it is expected that the performance of active learning is worse in datasets with a higher percentage of relevant records. Secondly, how do different machine learning models perform in datasets with a high inclusion percentage? This will be studied to examine if a certain model would outperform in highly inclusive datasets. Previous research into datasets with a low number of inclusions has shown that there is little deviation between models (Ferdinands et al., 2020). Therefore, it is expected that the performance of different active learning models will not differ significantly from each other. Finally, could the performance of active learning in systematic reviewing be increased by inverting inclusion labels? This inversion method could offer a solution to the benefit margin problem. By decreasing the inclusion percentage, the margin in which active learning can be beneficial increases. However, the excluded class tends to be more heterogenous as the only connection between papers is the original search query. This could create noise in the classifiers and result in lower relevancy scores. Since certainty based sampling is used, it is anticipated that the heterogeneous groups perform worse because certainty based sampling will suggest papers with high relevancy scores first. Therefore, it is expected that the inverted datasets show a small improvement in performance which is lower than performance reported in the literature.

2. Methods

2.1 Data collection

To test the performance of active learning in highly inclusive datasets, two datasets were acquired. First, the Bayesian dataset which was collected for a systematic review of Bayesian psychology articles which were published between 1990 and 2015 (Van de Schoot et al., 2017). The search was performed through Scopus, which resulted in 1669 records in the dataset before pre-processing (see Appendix A, Table 1). Second, the PBPK dataset is data which was collected by the Radboud UMC for a study on Physiologically Based Pharmacokinetic (or PBPK) modelling. The data collection of this study was completed at the end of 2020. The results of the study or yet to be published. Both datasets contain pre-labelled data. This labelling was performed by the researchers of each study. For the purpose of this study, these labels were assumed to be correct.

2.2 Outcome of interest

The outcome of interest of this study was the relevance of the records which the machine presented to the researcher. Thus, if the paper which was presented had a relevant or irrelevant label.

2.3 Simulation models

Various models were created with all possible combinations of the Logistic Regression, Naïve bayes, Random forest, and Support Vector Machines classifiers and the Doc2Vec, Sentence BERT, and Term Frequency – Inverse Document Frequency feature extraction techniques. Both the Doc2Vec and the sBERT feature extraction technique cannot be combined with Naïve Bayes classifier because they both generate negative values which the Naïve Bayes classifier cannot deal with (van de Schoot et al., 2021). The query strategy which is used is certainty based sampling. The software refers to this as the 'max' strategy. The balance strategy which is used as a method of oversampling is called dynamic resampling. The software refers to this as 'double'. Both strategies are chosen because they are the default settings of the software. Additionally, these settings are kept as constant over all models. For additional information about all the models, Appendix A, Table 1 can be consulted.

2.4 Simulation design

The simulation makes use of four datasets. The two original datasets (see 'data collection') and copies of these datasets in which the inclusion labels have been inverted. All ten models from Appendix A, Table 1 were used in a simulation with three runs per model. A run was defined as a full cycle of the simulation in which the model was trained to find all relevant records. This simulation was executed for every dataset resulting in 120 runs overall. Every model was trained with one relevant record and one irrelevant record, this is hereafter mentioned as prior knowledge. The prior knowledge changes between runs of a single model, but are identical for all different models within the dataset. Only the first run of each model has been used for the results. The goal of the other runs was to indicate the stability and reliability of the models over different runs with different prior knowledge.

2.5 Performance metrics

To answer the research questions, this research focussed on two main performance metricises: Work Saved over Sampling (WSS) and Relevant References Found (RRF).

The WSS@95% is a widely used metric for measuring the performance of screening prioritisation models (Cohen et al, 2006; O'Mara-Eves et al., 2015). This metric indicates how much time would have been saved when 95% of the relevant records have been found, while using abstract prioritisation compared to random sampling. The WSS is measured at 95% to resemble a more realistic situation than at 100%.

The RRF is a metric which represents the amount of relevant references found compared to the percentage of screened papers. This gives a representation of how effective screening prioritisation is compared to random screening. The RRF is measured at two instances: @10 and @90 to represent the first and last segments of the simulation. These values were chosen because the RRF@10 is more commonly used in the literature and because they are more informative when poor performances are expected because they are further from the edge values. The RRF is complementary to the WSS because the WSS compares to the percentage of relevant records found while the RRF compares to the percentage of evaluated records.

Across all performance metrices, the change caused by the inversion of inclusion labels is observed. This is hereafter mentioned as Delta (or Δ). Additionally, the mean and the median of all models is calculated to give insights into the overall performance of active learning across all models in a dataset.

2.6 Data treatment

Before the inversion, both the PBPK and the Bayesian datasets were prepared for simulation. Duplicate records were removed to prevent that certain terms receive disproportional weights in the models. Additionally, non-English records were removed due to problems with term frequencies because the software is unable to translate and match these terms. Inaccessible records were also removed from the data.

The original Bayesian dataset was not designed for active learning purposes. Therefore, it did not contain titles nor abstracts of all records, only references. The abstracts were automatically extracted from online databases using DOIs and references (n = 1591). The remaining missing abstracts were manually inserted (n= 78) to create the raw dataset. After this, duplicates (n = 5), non-English records (n = 21), and inaccessible records (n = 4) were removed. The titles of the records were extracted from the references column using Regular Expressions to increase the performance of ASReview. Additionally, inclusion labels, which were in text notation, were replaced with binary in which ‘1’ represents inclusion (n = 1579) and ‘0’ represents exclusion (n = 60). Afterwards, inclusion labels were inverted which resulted in two datasets which were ready for simulation; one dataset with an inclusion rate of 96.3% and an inverted dataset with an inclusion rate of 3.7%.

For the PBPK dataset, duplicates were removed (n = 5) and missing abstracts were manually inserted (n = 24). The dataset did not contain any non-English or inaccessible records. There were records which had no abstract because of the nature of the publication, such as responses or editorial notes (n = 42). These records are not removed to mimic the behaviour of the original researchers. Subsequently, the dataset was split into two data frames, from which one had inverted the inclusion labels. This resulted in two datasets in which the original has a 48.8% inclusion rate (n= 1047) and the inverted has a 51.2% inclusion rate (n = 1100). Additional statistics of the data treatment for all datasets can be found in Appendix A, Table 2.

2.7 Software

The study was performed using the simulation mode of ASReview (version 0.17) and Python (version 3.9.4). ASReview is active learning software specifically designed to aid researchers in their systematic review. The simulation mode is a method in which the performance of a labelled dataset can be tested. For additional information about ASReview, the article of Van de Schoot et al. (2021) can be consulted. Additional information about the reproducibility and set-up of ASReview can be found in the GitHub repository of this paper.

3. Results

First of all, the WSS@95 values are examined. The PBPK dataset saves 23.7% of time using active learning over random sampling (M = 23.7, Mdn = 24.8). The classifiers of the PBPK dataset show no difference in performances. For the feature extraction techniques, the WWS@95 of sentence BERT (WSS@95$_{\text{sBERT}}$ = [14.5 – 19.8]) is noticeably lower than for Doc2Vec (WSS@95$_{\text{D2V}}$ = [24.2 – 26.9]) and TF-IDF (WSS@95$_{\text{TF-IDF}}$ = [25.3 –29.6]). The LR + TF-IDF model performs best (WSS@95 = 29.6) while the SVM + sBERT model performs the worst (WSS@95 = 14.5). For the inverted PBPK datasets ($M_{\text{ Inverted}}$ = 16.4, $Mdn_{\text{ Inverted}}$ = 17.6), the WSS@95 does not show an increase over the normal dataset (M_{Δ} = -7.3, Mdn_{Δ} = -7.4). The WSS@95 $_{\Delta}$ values of all models in the inverted PBPK dataset range from -1.4 to -14.7. The LR + TF-IDF model has the best performance (WSS@95$_{\text{Inverted}}$ = 20.0, WSS@95 $_{\Delta}$ = -9.6). The Naïve Bayes classifier, and therefore the NB + TF-IDF model, underperforms compared to the other models (WSS@95 $_{\text{Inverted}}$ = 10.7, WSS@95 $_{\Delta}$ = -14.7). For the feature extraction techniques, the WSS@95 values of sBERT range from 12.4 up to 14.9 which is generally lower than other feature extraction techniques. The Bayesian dataset shows minimal change compared to random sampling (M = 0.7, Mdn = 0.6). There are no noteworthy differences across classifiers, feature extraction techniques and models. The inversion of the inclusion labels resulted in an average growth across all models of 10.8%. ($M_{\text{ Inverted}}$ = 11.5, $Mdn_{\text{ Inverted}}$ = 11.5). All feature extraction techniques report performances which are similar to each other. The NB + TF-IDF model severely underperforms with a 3% increase over random sampling (WSS@95$_{\text{Inverted}}$ = 3.0, WSS@95$_{\Delta}$ = 2.0). With a 19.2% increase, the RF + TF-IDF reported the highest increase of inverted Bayesian dataset (WSS@95$_{\text{Inverted}}$ = 20.1, WSS@95$_{\Delta}$ = 19.2). For additional information on the WSS@95 of all models, Appendix A, Table 3 can be consulted.

Secondly, the RRF@10 values are observed. For the PBPK dataset, active learning contributed to a 6.7% increase of relevant references found over random sampling when 10% of the total number of papers is screened (M = 16.7, Mdn = 16.9). The results show no differences in performance of classifiers or feature extraction techniques. However, the SVM + sBERT model (RRF@10 = 13.9) is slightly underperforming. The RRF@10 values of all models, excluding the SVM + sBERT model, show little variation across models (RRF@10 = [16.3 – 18.0]). Inversion of inclusion labels shows a similar performance as the PBPK dataset represented by a delta of 0.4% (RRF@10 $_{\text{Inverted}}$ = 17.2, $Mdn_{\text{ Inverted}}$ = 17.3). The NB + TF-IDF (RRF@10$_{\Delta}$ = -2.4), RF + D2V (RRF@10$_{\Delta}$ = -0.7), and LR + TF-IDF (RRF@10$_{\Delta}$ = -0.1) models have lower scores in the inverted dataset than before label inversion. The Bayesian dataset (M = 10.1, Mdn = 10.1) show minimal deviation across all classifiers, feature extraction techniques, and models with RRF@10 values ranging from 9.8 to 10.4. The inverted Bayesian dataset ($M_{\text{ Inverted}}$ = 29.2, $Mdn_{\text{ Inverted}}$ = 28.2) experienced positive effects of the label inversion (M_{Δ} = 19.0, Mdn_{Δ} = 18.7). Noteworthy models are the RF + TF-IDF (RRF@10$_{\text{Inverted}}$ = 44.1, RRF@10$_{\Delta}$ = 33.9) and LR + TF-IDF (RRF@10$_{\text{Inverted}}$ = 40.7, RRF@10$_{\Delta}$ = 30.3) models which perform far above the mean. Additionally, the NB + TF-IDF (RRF@10$_{\text{Inverted}}$ = 16.9, RRF@10$_{\Delta}$ = 6.6) and RF + D2V (RRF@10$_{\text{Inverted}}$

= 11.9, RRF@10$_{\Delta}$ = 1.9) models perform severely under the mean. For additional information on the RRF@10 metrics, Appendix A, Table 4 can be consulted.

Thirdly, the RRF@90 is analysed. The PBPK dataset shows little deviation across models as all values are ranging from 98.9 to 100% (*M* = 99.6, *Mdn* = 99.7). It is noteworthy that the NB + TF-IDF model is the only model which finished simulation (RRF@90 = 100.0). The inversion of the PBPK dataset (M_{Inverted} = 98.3, Mdn_{Inverted} = 98.4) does not result in an improvement of performance (M_{Δ} = -1.3, Mdn_{Δ} = -1.1). Additionally, no differences in classifier or feature extraction technique are observed. The NB + TF-IDF (RRF@90$_{\text{Inverted}}$ = 97.2, RRF@90$_{\Delta}$ = -2.8) is the only model which shows a minor deviation from the mean. The Bayesian dataset reports a 0.4% to 1.3% increase of performance over random sampling (*M* = 90.8, *Mdn* = 90.7). In this dataset, there appears to be no distinct differences in performance between classifiers, feature extraction techniques, and models. The inverted Bayesian dataset (M_{Inverted} = 98.0, Mdn_{Inverted} = 98.3) shows a small improvement of performance over the Bayesian dataset (M_{Δ} = 7.2, Mdn_{Δ} = 7.4). Two out of three models with Support Vector Machines classifiers finished simulation when 90% of all records were screened (RRF@90$_{\text{SVM+D2V}}$ = 100.0, RRF@90$_{\text{SVM+sBERT}}$ = 100.0). the use of different feature extraction techniques did not result in different performance across the models. Additional information about the RRF@90 can be consulted in Appendix A, Table 5.

Finally, the recall plots (see Appendix B) are examined. Figure 1 shows the recall plot of the PBPK dataset. This figure shows two clustered groups of lines. The models in the worst performing group all use the sBERT feature extraction technique. This pattern continues in the inverted PBPK dataset (see Figure 2). Additionally, the NB + TF-IDF shows worse performance after label inversion. The recall plot of the Bayesian dataset (see Figure 3) is identical to the random sampling diagonal. The inverted Bayesian dataset (see Figure 4) shows improvement over the random sampling diagonal. It is noteworthy that plot steepness of the plot peaks in the first 10% of the run. Afterwards, the lines flatten and show a somewhat linear trend towards completion of the simulation. The RF + D2V model shows an irregularity with a starting peak around 20%. Figure 5 shows the recall plots across all runs of the simulation. This shows that the trends are similar over all runs. The irregular gradient of the RF + D2V model seems to be an outlier as the other runs do not mimic the same behaviour.

4. Discussion

4.1 Main findings

The PBPK dataset, which had 48.8% inclusions, performed moderately well. Active learning contributed to a 24.8% Work Saved over Sampling. The Bayesian dataset, on the other hand, performed poorly and showed no evidence that active learning performs better than random sampling. The RRF@10 and the RRF@90 showed a similar trends in which the PBPK dataset performed slightly better than random sampling and the Bayesian dataset showed no improvement over random sampling. This is in line with the predictions of the hypothesis that it is expected that the performance of active learning in general becomes lower in datasets with a high percentage of relevant records.

It was hypothesized that the performance of active learning models would not differ between models. It was found that all classifiers behave similar throughout the datasets. The sBERT feature extraction technique performed less well based on the WSS@95 for the PBPK dataset. The lowered performance is not observed in the RRF values. However, when observing the recall plots (see Figure B1,2 &4), it becomes clear that the sBERT models are always underperforming compared to the D2V and TF-IDF feature extraction techniques. Finally, the NB + TF-IDF model was underperforming for both inverted datasets across all performance metrics. When observing the inverted PBPK plot (see Appendix B, Figure 3), the plot showed less divergence from the random sampling diagonal, which suggests worse performance. The performance of the inverted Bayesian dataset (see Appendix B, Figure 4) seemed similar in the plot. It is noteworthy that the incremental peaks of the NB + TF-IDF line are steeper and longer. This would suggest that the Naïve Bayes classifier is more sensitive to clusters of similar papers. To conclude, the findings regarding the stability of different models of this research were not in line with the expectations due to the questionable performance of the sBERT feature extraction technique and the performance of the NB + TF-IDF model in inverted datasets.

The efficacy of inclusion label inversion was dependent on the distribution of relevant and irrelevant papers. It was found that the inverted Bayesian dataset, which had 3.7% inclusions, showed a small improvement over random sampling. When observing the plot, all models seemed to find the first part of relevant records quickly. Hereafter, a linear trend similar to the random sampling strategy is observed. Unlike random sampling, these lines displayed small stepwise increases in which multiple relevant records are found. A possible explanation for this could be that there are clusters of similar papers within the heterogenous class. This would explain both the peaks and the similarity to the random sampling strategy after the initial peak. All in all, the performance of inverted datasets does not match the results seen in previous research (O’Mara-Eves et al., 2015; Ferdinands et al., 2020) which reported a higher performance. This is in line with the hypothesis regarding inverting the inclusion labels. Therefore, these results show first evidence that there are differences in homogeneity between the relevance classes.

4.2 Applicability of the research

This research provides evidence that the use of active learning is not efficient in highly inclusive datasets. The inversion of inclusion labels does not yet provide an efficient solution. Therefore, it is not recommended to use active learning when a high number of inclusions is expected.

4.3 Limitations and future research

Although this study is carefully constructed, the results are based on only one run and one dataset per inclusion distribution. Because of this, the generalizability of the research should be questioned as there is no evidence of the influence of prior knowledge and the internal deviation of models over different runs. Therefore, future research should use more runs per model to test the deviation of the models between different runs and minimalize the influence of prior knowledge. Additionally, future research should contain more datasets with different inclusion distributions to conclusively proof the relationship between inclusion percentages and performance of active learning.

One of the possible explanations for the poor performance of the inverted models could be the heterogeneity of the included group. This is based on two assumptions. First, the assumption that the group which was originally exclusions is less homogeneous than the inclusions group. This is assumed because the inclusions group has the query string and the topic of research in common, while the exclusions are only connected through the query string of the research. Secondly, the assumption that models perform less in heterogenous groups than in homogeneous groups. This would be suspected because the terms and vectors within homogeneous papers are more likely to be similar and are therefore receive higher weights. This is, however, only speculation as the effects of homogeneity of a dataset have not, to the best of my knowledge, been tested to this date. Therefore, future research should focus on the effects of homogeneity and heterogeneity within datasets on the performance of active learning.

Finally, the results may be influenced by the effects of the query and balance strategies. The ‘max’ query strategy could result in an overperformance of homogeneous groups because the strategy suggests papers with high relevancy scores first. Additionally, dynamic resampling could result in a overfitted model. Because of this and the lack of runs of this research, the reliability and validity of this research should be questioned.

5. Conclusion

The main goal of this research was to explore the performance of active learning within highly inclusive datasets. The overall performance of active learning becomes worse if the inclusion percentage of a dataset increases. Inversion of the dataset only slightly increased the performance. However, this subject needs more study into the optimisation of models to become viable in practice. Finally, evidence is found to support that the NB + TF-IDF model underperforms in inverted datasets.

6. Acknowledgements

First of all, I would like to thank the Radboud UMC, Joyce van der Heijden, Jolien Freriksen, and Saskia de Wildt for letting me use their collected data on PBPK models. I would like to thank Bianca Kramer for her assistance in the abstract retrieval process. Furthermore, I would like to thank Jonathan de Bruin and Rens van de Schoot for sharing their expertise and providing technical assistance when necessary. Last but not least, I would like to thank Laura Hofstee to whom I am very grateful for all the helpful meetings, feedback, words of encouragement, and enthusiasm throughout the entire research process. Thank you all.

7. Data availability

All datasets and scripts to replicate this study are available on the GitHub repository of this paper: https://github.com/asreview/thesis-asreview-performance-in-highly-inclusive-datasets

8. Abbreviations

D2V – Doc2Vec

LR – Logistic Regression

NB – Naïve Bayes

RF – Random Forest

RRF – Relevant References Found

sBERT – Sentence BERT

SVM – Support Vector Machines

TF-IDF –Term Frequency – Inverse Document Frequency

WSS – Work Saved over Sampling

9. References

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, *8*(1), 1-12. https://doi.org/10.1101/255760

Bekhuis, T., Tseytlin, E., Mitchell, K. J., & Demner-Fushman, D. (2014). Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PloS one*, *9*(1), e86277. https://doi.org/10.1371/journal.pone.0086277

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open*, *7*(2), e012545. https://doi.org/10.1136/bmjopen-2016-012545

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875-886. https://doi.org/10.1007/0-387-25465-x_40

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. https://doi.org/10.1613/jair.953

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, *13*(2), 206-219. https://doi.org/10.1197/jamia.m1929

Colas, F., & Brazdil, P. (2006, August). Comparison of SVM and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice* (pp. 169-178). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-34747-9_18

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007, November). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 127-136). https://doi.org/10.1145/1321440.1321461

Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D., Tummers, L., & van de Schoot, R. (2020). Active learning for screening prioritization in systematic reviews. https://doi.org/10.31219/osf.io/w6qbg

Fu, J., & Lee, S. (2013). Certainty-based active learning for sampling imbalanced datasets. https://doi.org/10.1016/j.neucom.2013.03.023

Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In *Feature extraction* (pp. 1-25). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-35488-8_1

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). Cochrane handbook for systematic reviews of interventions. *John Wiley & Sons*, 3-12.

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., ... & Thayer, K. (2016). SWIFT-Review: a text-mining workbench for systematic review. *Systematic reviews*, *5*(1), 1-16. https://doi.org/10.1186/s13643-016-0263-z

Ji, X., & Yen, P. Y. (2015). Using MEDLINE elemental similarity to assist in the article screening process for systematic reviews. *JMIR medical informatics*, *3*(3), e28. https://doi.org/10.2196/medinform.3982

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1305/1305.1707.pdf

Miwa, M., Thomas, J., O’Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, *51*, 242-253. https://doi.org/10.1016/j.jbi.2014.06.005

O’Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 1-22. https://doi.org/10.1186/2046-4053-4-5

Olorisade, B. K., Brereton, P., & Andras, P. (2019). The use of bibliography enriched features for automatic citation screening. *Journal of biomedical informatics*, *94*, 103202. https://doi.org/10.1016/j.jbi.2019.103202

Settles, B. (2009). Active learning literature survey. Retrieved from: http://digital.library.wisc.edu/1793/60660

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217. https://doi.org/10.1037/met0000100

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., ... & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125-133. https://doi.org/10.1038/s42256-020-00287-7

van Lissa, C. J. (2021). Mapping Phenomena Relevant to Adolescent Emotion Regulation: A Text-Mining Systematic Review. *Adolescent research review*, 1-13. https://doi.org/10.1007/s40894-021-00160-7

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010, July). Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 173-182). https://doi.org/10.1145/1835804.1835829

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PloS one*, *15*(1), e0227742. https://doi.org/10.1371/journal.pone.0227742

Appendix A – Tables

Table 1

The descriptive statistics of all pre-processed datasets used for simulation.

	Number of records before pre-processing	Number of records after pre-processing	Number of relevant labelled records	Inclusion rate (in %)
PBPK	2152	2147	1047	48.8
PBPK (inverted)	2152	2147	1100	51.2
Bayesian	1669	1639	1579	96.3
Bayesian (inverted)	1669	1639	60	3.7

Note. Inverted datasets contain the same records as the original dataset. However, inclusion labels are inverted.

Table 2

All used model combinations with their respective classifiers, feature extraction techniques, query strategies, and balance strategies.

Name model	Classifier	Feature extraction technique	Query Strategy	Balance strategy
LR + D2V	Logistic Regression	Doc2Vec	Max	Double
LR + sBERT	Logistic Regression	Sentence BERT	Max	Double
LR + TF-IDF	Logistic Regression	Term Frequency – Inverse Document Frequency	Max	Double
NB + TF-IDF	Naïve Bayes	Term Frequency – Inverse Document Frequency	Max	Double
RF + D2V	Random Forest	Doc2Vec	Max	Double
RF + sBERT	Random Forest	Sentence BERT	Max	Double
RF + TF-IDF	Random Forest	Term Frequency – Inverse Document Frequency	Max	Double
SVM + D2V	Support Vector Machines	Doc2Vec	Max	Double
SVM + sBERT	Support Vector Machines	Sentence BERT	Max	Double
SVM + TFIDF	Support Vector Machines	Term Frequency – Inverse Document Frequency	Max	Double

Note. The ‘*Max’* query strategy and the ‘*Double’* balance strategy are both default settings in ASReview.

Table 3

The Worked Saved over Sampling when 95% of the relevant records are found (in percentages).

	PBPK	PBPK (inverted)	Δ PBPK	Bayesian	Bayesian (inverted)	Δ Bayesian
LR + D2V	24.3	19.1	-5.2	0.6	9.0	8.5
LR + sBERT	19.8	14.9	-4.9	0.5	7.5	7.0
LR + TF-IDF	29.6	20.0	-9.6	1.2	12.4	11.2
NB + TF-IDF	25.3	10.7	-14.7	1.0	3.0	2.0
RF + D2V	26.9	19.7	-7.2	0.7	11.8	11.0
RF + sBERT	18.7	12.4	-6.2	0.3	14.6	14.3
RF + TF-IDF	27.5	19.0	-8.4	0.9	20.1	19.2
SVM + D2V	24.2	16.6	-7.6	0.5	11.3	10.8
SVM + sBERT	14.5	13.1	-1.4	0.3	14.3	14.0
SVM + TFIDF	26.1	18.5	-7.6	1.0	10.8	9.7
Mean	23.7	16.4	-7.3	0.7	11.5	10.8
Median	24.8	17.6	-7.4	0.6	11.5	10.9

Note. Delta (Δ) represents the difference in WSS@95 caused by inverting the inclusion labels of the dataset.

Table 4

The percentage of Relevant References Found after screening 10% of all records.

	PBPK	PBPK (inverted)	Δ PBPK	Bayesian	Bayesian (inverted)	Δ Bayesian
LR + D2V	16.9	18.3	1.4	9.8	37.3	27.5
LR + sBERT	16.3	17.8	1.5	10.0	30.5	20.5
LR + TF-IDF	17.3	17.2	-0.1	10.3	40.7	30.3
NB + TF-IDF	16.7	14.3	-2.4	10.3	16.9	6.6
RF + D2V	17.8	17.1	-0.7	10.0	11.9	1.9
RF + sBERT	16.6	16.7	0.0	10.4	27.1	16.7
RF + TF-IDF	18.0	17.9	0.0	10.1	44.1	33.9
SVM + D2V	16.9	17.8	0.9	10.0	27.1	17.1
SVM + sBERT	13.9	17.5	3.6	10.0	25.4	15.4
SVM + TFIDF	17.0	17.2	0.2	10.2	30.5	20.3
Mean	16.7	17.2	0.4	10.1	29.2	19.0
Median	16.9	17.3	0.1	10.1	28.8	18.7

Note. Delta (Δ) represents the difference in RRF@10 caused by inverting the inclusion labels of the dataset.

Table 5

The percentage of Relevant References Found after screening 90% of all records.

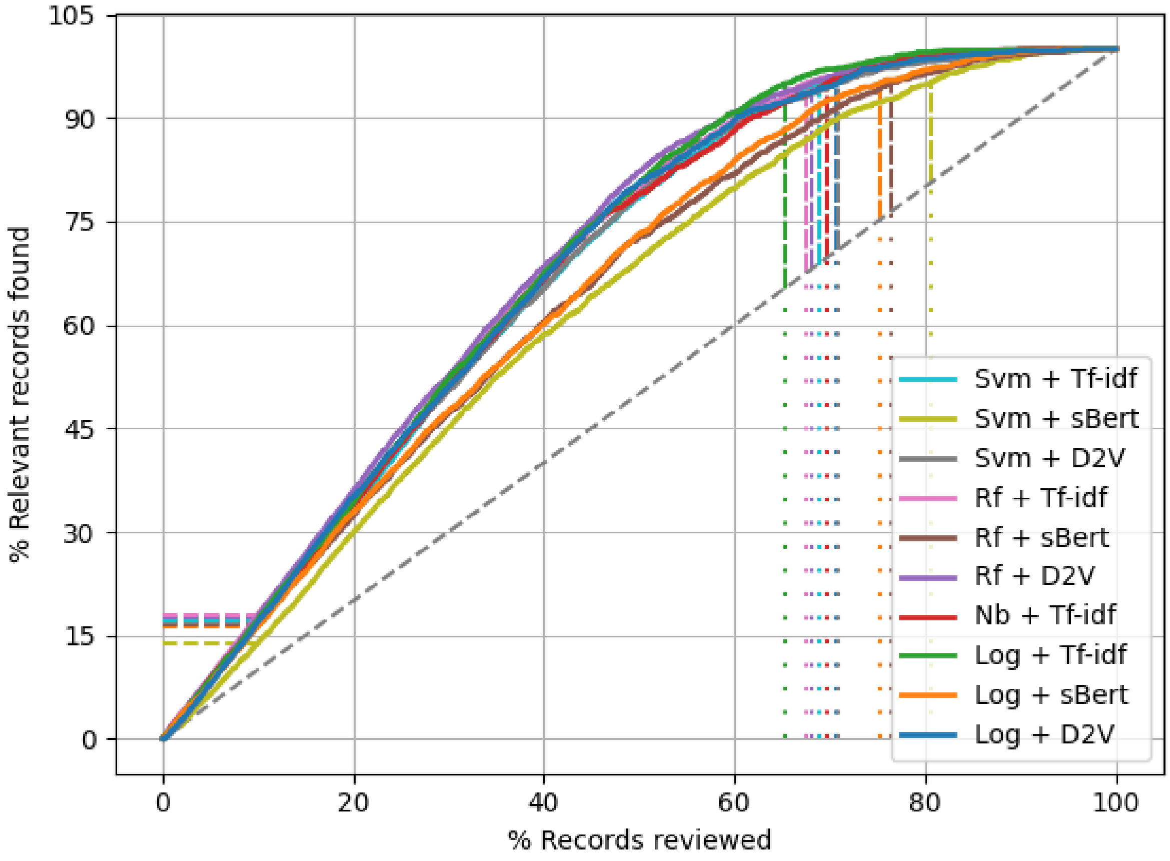
	PBPK	PBPK (inverted)	Δ PBPK	Bayesian	Bayesian (inverted)	Δ Bayesian
LR + D2V	99.7	98.6	-1.1	90.6	96.6	6.0
LR + sBERT	99.5	98.4	-1.2	90.5	98.3	7.8
LR + TF-IDF	99.9	98.4	-1.5	91.3	96.6	5.4
NB + TF-IDF	100.0	97.2	-2.8	91.1	94.9	3.8
RF + D2V	99.8	98.5	-1.4	90.7	98.3	7.6
RF + sBERT	99.1	97.7	-1.4	90.4	98.3	7.9
RF + TF-IDF	99.9	98.8	-1.1	91.2	98.3	7.1
SVM + D2V	99.4	98.7	-0.7	90.4	100.0	9.6
SVM + sBERT	98.9	97.9	-0.9	90.4	100.0	9.6
SVM + TFIDF	99.7	98.7	-1.0	91.1	98.3	7.2
Mean	99.6	98.3	-1.3	90.8	98.0	7.2
Median	99.7	98.4	-1.1	90.7	98.3	7.4

Note. Delta (Δ) represents the difference in RRF@90 caused by inverting the inclusion labels of the dataset.

Appendix B – Figures

Figure 1

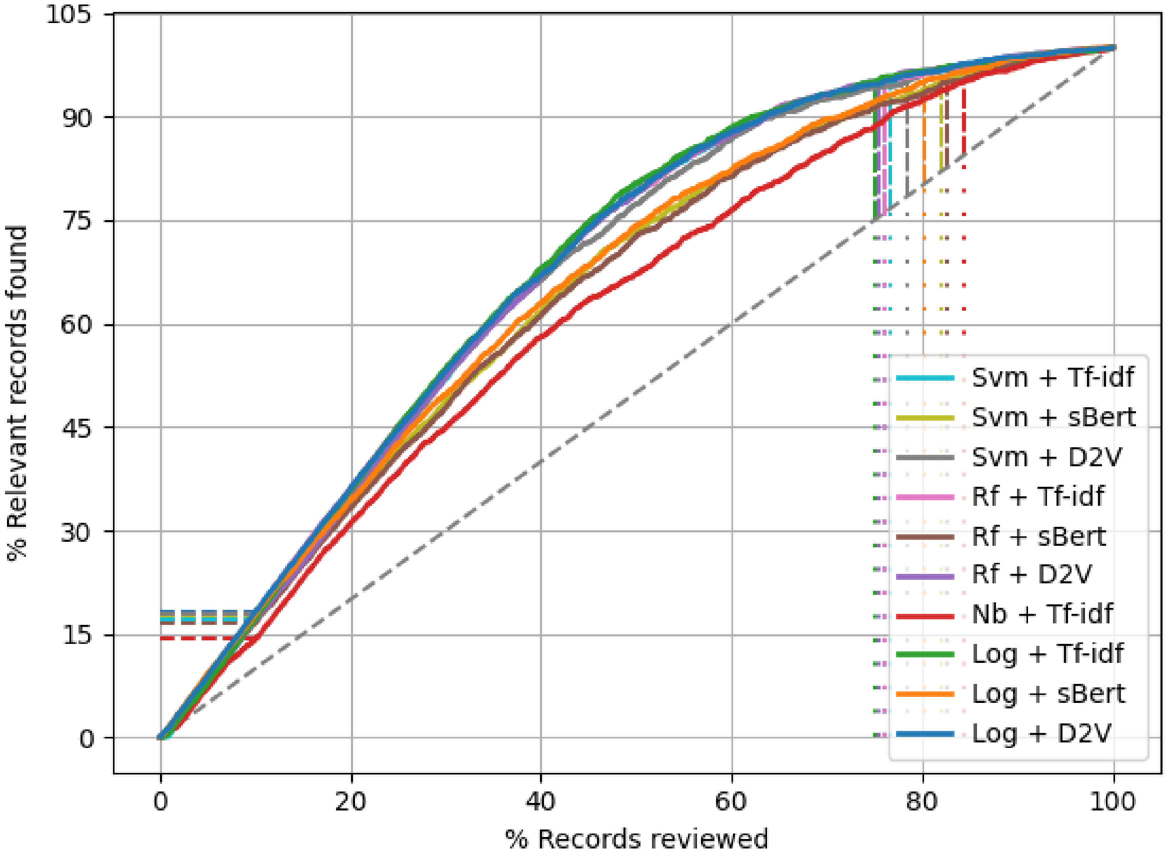
Graph representing the percentage of Relevant Records Found against the total percentage of papers screened for the PBPK dataset.



Note. The horizontally dotted lines represent the RRF@10 for each line. The vertically dotted lines represent the WSS@95 for each line.

Figure 2

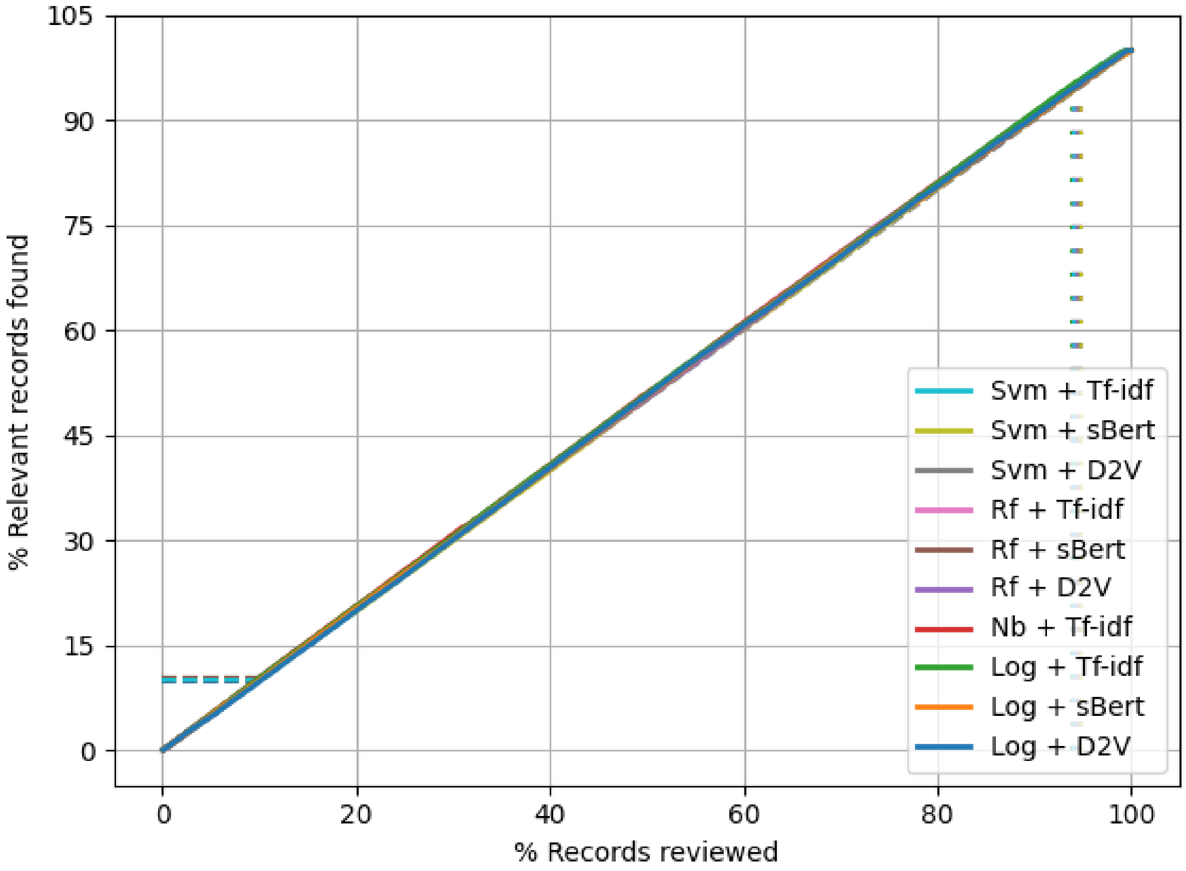
Graph representing the percentage of Relevant Records found against the total percentage of papers screened for the inverted PBPK dataset.



Note. The horizontally dotted lines represent the RRF@10 for each line. The vertically dotted lines represent the WSS@95 for each line.

Figure 3

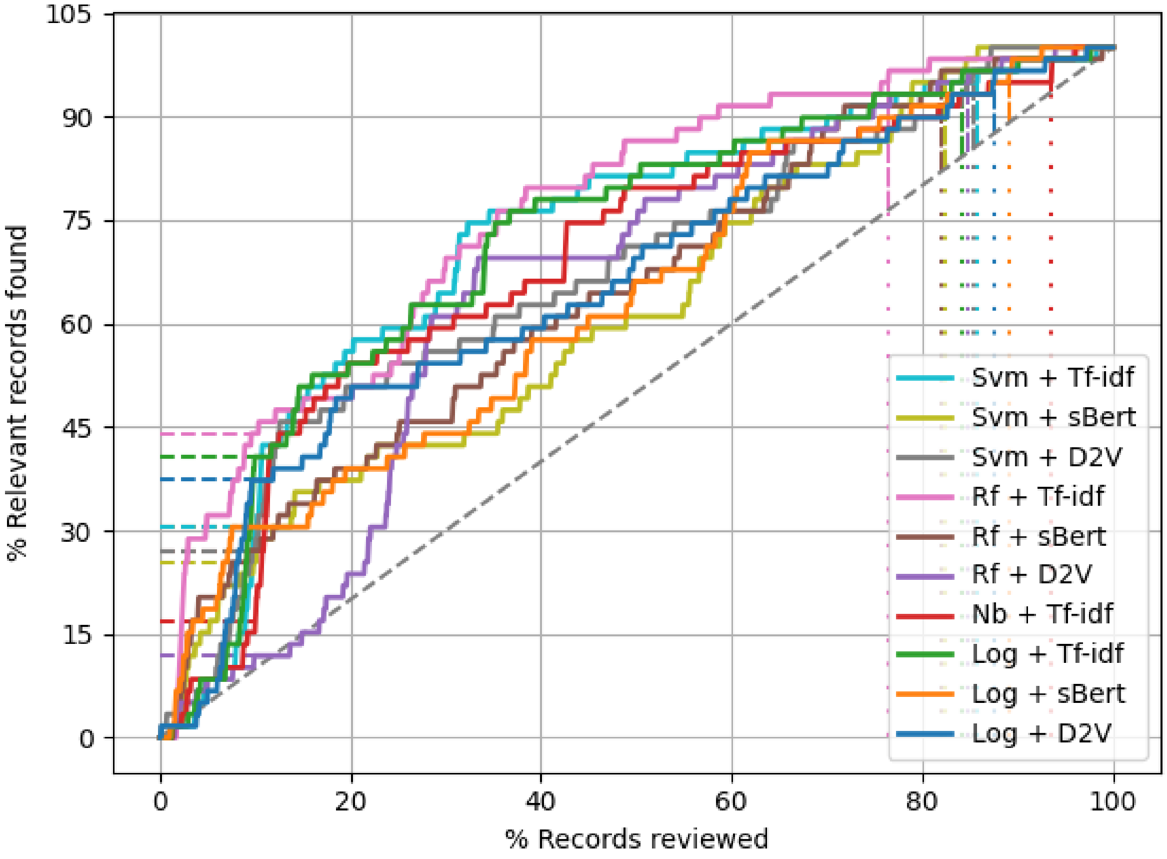
Graph representing the percentage of Relevant Records found against the total percentage of papers screened for the Bayesian dataset.



Note. The horizontally dotted lines represent the RRF@10 for each line. The vertically dotted lines represent the WSS@95 for each line.

Figure 4

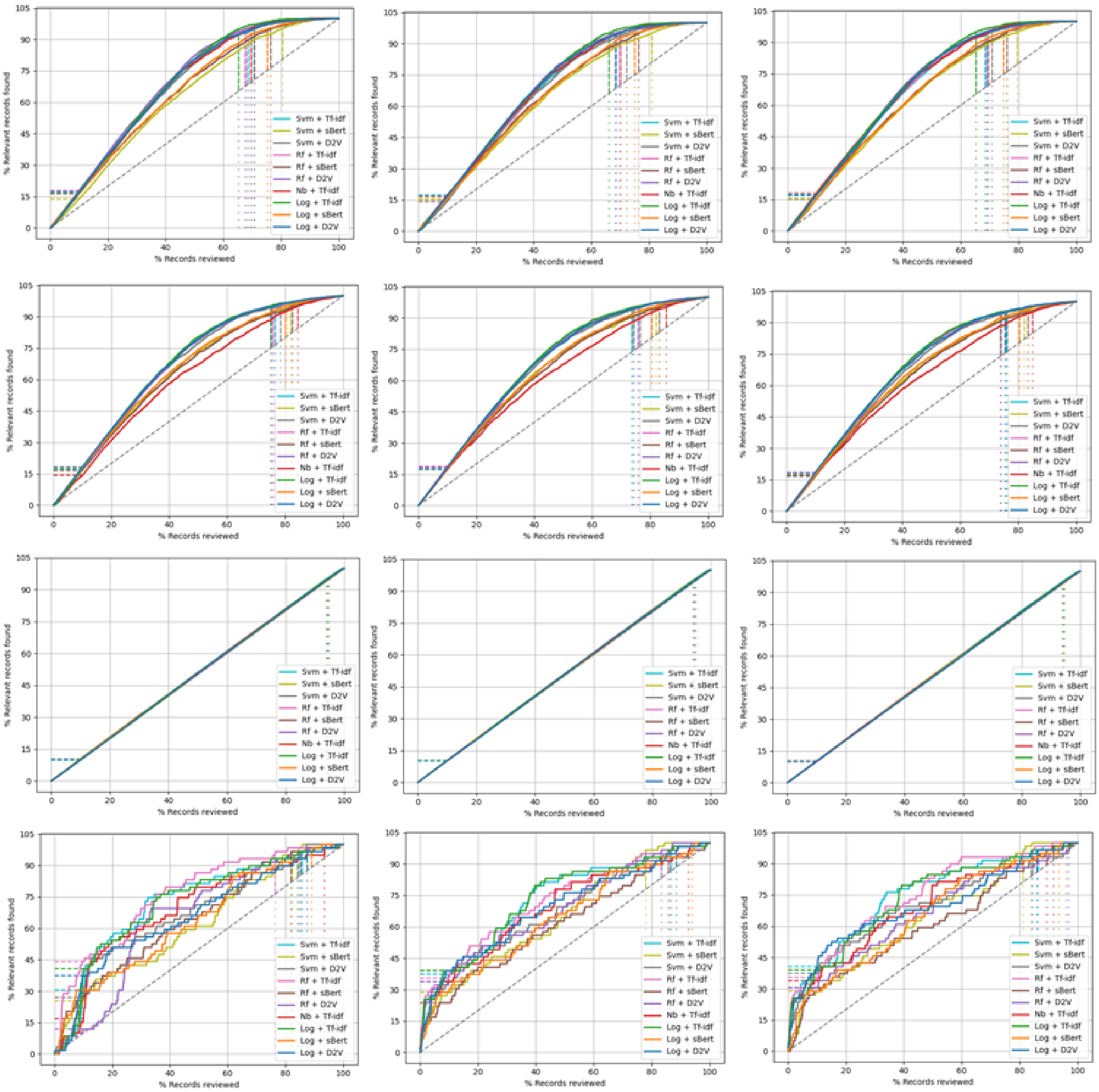
Graph representing the percentage of Relevant Records found against the total percentage of papers screened for the inverted Bayesian dataset.



Note. The horizontally dotted lines represent the RRF@10 for each line. The vertically dotted lines represent the WSS@95 for each line.

Figure 5

Recall plots collected over three runs representing the percentage of Relevant Records found against the total percentage of papers screened for the inverted Bayesian dataset.



Note. The order of datasets, from top to bottom, is: PBPK, inverted PBPK, Bayesian, and inverted Bayesian.