Results

2021-10-11

Results

First session

The input search query was:

((model OR models OR modeling OR network OR networks) AND (dissemination OR transmission OR spread OR diffusion) AND (nosocomial OR hospital OR "long-term-care" OR "long term care" OR "longterm care" OR "long-term care" OR "healthcare associated") AND (infection OR resistance OR resistant))

subsetting results between 2010 and 2020 (included). The automatic search tools were used for MEDLINE, WOS and IEEE, while a manual search and importation was necessary for EMBASE and SCOPUS. The search was repeated manually also for MEDLINE, since, as reported in the methods, the web interface may return different results from the API.

The first search session returned a total of 27600 unique records, specifically 12719 (71.6% of the total) records from the *EMBASE* database, followed by 9546 (53.8%) from *Pubmed/Medline*, 3175 (17.9%) from *Scopus*, 2100 (11.8%) from *Web of Science*, and 60 (0.34%) from *IEEE* (Table 1). There were various degrees of overlapping with the 38.4% of records being present in more than one database and *EMBASE* and *IEEE* being the databases with the higher uniqueness ratios. The final dataset was composed by 17755 unique records.
The first 250 records were manually labeled. Of these 43 (17.2%) were labeled as positive, and 207 (82.8%) as negative.

After the first manual classification, 7 automatic classification and manual review rounds (CR) were performed (Table 2). It is possible to observe how the number of records that required manual review dropped fast between iterations, indicating that the engine was converging while the uncertainties were resolved.
This phenomenon is better depicted in figure 1 in the Suppl. Mat. S2. It shows the mixture distribution of the predicted probabilities for each record, specifically for the reviewed positive and negative records, and for records that need manual review after the classification step: it can be noticed how the distribution of the uncertain records shrinks (they concentrate in a shorter probability range) and shifts toward the negative zone as more positive matches are found and reviewed. Accordingly (Table 2), the number of new positives found drops after the first couple of iterations. As per default, the classification process was stopped after four rounds with no new positive matches; a total number of 101 positives were found over 766 manually reviewed records (13.2% positivity rate).

By evaluating how often the BART algorithm decided to use a term for classification, it is possible to list a subset of the variables most relevant for the classification of the documents: Patient Transport (Keyword): 61.2 [21.3], Transfer (Abstract): 57 [15.4], Network (Title): 56.5 [14.2], Network & Patient (Abstract): 54.2 [15.2], Donker T (Author): 53.5 [16.5], Worker (Abstract): 50 [-1.21], Hospitals (Keyword): 49.8 [16.5], Movement (Abstract): 47.8 [15], Spread (Title): 46.6 [12.1], Facility (Abstract): 45 [14.8], Orange County (Keyword): 44.3 [17.2], Conduct (Abstract): 42.6 [-2.57], Patient (Abstract): 42 [7.23], Perform (Abstract): 41.9 [-2.55], Hospital (Title): 39 [12.5]. In parenthesis, it is reported the part of the record in which the term was used, while the numeric values indicate respectively the rate of posterior trees (over 10000 trees) in which a term was used, and the number of standard errors of the association between a term and the probability of a positive match according to a simple logistic linear model (see methods). The "&" indicates that two terms

Table 1. Distribution of retrieved records by source and session. For each source it is reported the number of records, percentage over the session total (after removing duplicates)' and, number or records specific for a source as absolute value and as percentage over the source total. SAll session shows records after joining and deduplication of the Session1 and Session2 dataset.

Session	Source	Records	% over total	Source specific records	% over source total
Session1	Total	17755			
	Embase	12719	71.6%	6683	52.5%
	Pubmed	9546	53.8%	3457	36.2%
	Scopus	3175	17.9%	298	9.39%
	WOS	2100	11.8%	473	22.5%
	IEEE	60	0.34%	29	48.3%
Session2	Total	82579			
	Embase	48396	58.6%	40826	84.4%
	Pubmed	28811	34.9%	18021	62.5%
	Scopus	17070	20.7%	4908	28.8%
	WOS	12956	15.7%	2817	21.7%
	IEEE	61	0.074%	22	36.1%
All Sessions	Total	98371			
	Embase	59604	60.6%	46942	78.8%
	Pubmed	37278	37.9%	21371	57.3%
	Scopus	19353	19.7%	5181	26.8%
	WOS	14367	14.6%	3175	22.1%
	IEEE	108	0.11%	48	44.4%

Table 2. Results of the automatic classifaction and manual review rounds. For each iteration, the cumulative number of positives and negative records and their sum (Total labelled) and percentage over total are shown. Also, the number of changes after review and their description is reported. "Unlab." indicates unlabelled records marked for review. For each Iteration, also the number of features used by the engine is reported. The first row reports the results of the initial manual labelling of records, which acted as input for the automatic classification in Iteration 1. In Session2, the engine uses the labels at the end of Session1 to classify the newly added records.

Session	Iteration	Positives	Negatives	Total labelled (%)	Unlab. -> y	Unlab. -> n	Unlab. -> *	n -> y	Changes	N. features
Session1 (n = 17755)	Initial labelling	43	207	250 (1.41%)	43	207	0	0	250	2289
	1	93	529	622 (3.5%)	50	322	0	0	372	2289
	2	100	614	714 (4.02%)	6	86	0	1	93	3750
	3	101	625	726 (4.09%)	1	11	0	0	12	3834
	4	101	648	749 (4.22%)	0	23	0	0	23	3856
	5	101	651	752 (4.24%)	0	3	0	0	3	3856
	6	101	660	761 (4.29%)	0	9	0	0	9	3856
	7	101	665	766 (4.31%)	0	5	0	0	5	3856
Session2 (n = 98371)	1	106	934	1040 (1.06%)	5	270	998	0	1273	4729
	2	107	1123	1230 (1.25%)	1	189	0	0	190	4729
	3	107	1176	1283 (1.3%)	0	53	0	0	53	4733
	4	107	1200	1307 (1.33%)	0	24	0	0	24	4729
	5	107	1209	1316 (1.34%)	0	9	0	0	9	4729
	6	107	1226	1333 (1.36%)	0	17	0	0	17	4729

are present together in a document but not close to each other.
The engine was able to pick up the central concept of the research (i.e., "patient transport" or "transfer" through a "network" of "facility"ies that facilitates the "spread" of infections) and even one of the authors of the current study (Donker T.) or the region of interest ("Orange County") of another research group active on the topic. It is interesting to see that some terms were considered highly relevant (e.g., "Worker" in 6th

position out of more than 3800 terms considered) although in a simpler linear model, their effect would be hardly significant (statistic: -1.21 s.e.), and this highlight the extra predictive power brought by a highly non-linear model.
A more extensive set of terms is presented in Table 1 of Supplemental Material S2.

Second session

The results of the first classification session were used to create a second, data-driven query with the purpose of performing a more large-spectrum search to find records which may have escape the first search session. The resulting query was the following:

(((Donker T) NOT (bacterium isolate)) OR ((network patient) AND (resistant staphylococcus aureus) NOT (monte carlo) NOT isolation) OR (facility AND (network patient) AND regional NOT hospitals NOT increase NOT (patient transport) NOT (control infection use)) OR ((patient transport) NOT (Donker T) NOT worker) OR (hospitals AND (network patient) NOT (patient transport) NOT regional NOT clinical) OR (facility AND (network patient) NOT hospitals NOT (patient transport) NOT regional NOT prevention NOT medical) OR ((healthcare facility) NOT (Donker T) NOT worker NOT positive) OR (hospitals NOT (network patient) NOT medical NOT environmental NOT outcome NOT global) OR ((network patient) NOT facility NOT hospitals NOT (patient transport) NOT therapy NOT global)) AND ((antimicrobial resistance) OR (healthcare infection))

The final piece *AND ((antimicrobial resistance) OR (healthcare infection)* was added manually to better define the search domain, since the algorithm was trained on documents that were all more or less related to these topics.
The generated query also provides a more nuanced understanding of the engine's internal classification logic, and this is helpful to spot possible biases in the model.

The search was done with the same year filter and procedures of the first session.

The new search produced 107294 records (Table 1), of which 48396 (58.6%) from the *EMBASE*, followed by 28811 (34.9%) from *Pubmed/Medline*, 17070 (20.7%) from *Scopus*, 12956 (15.7%) from *Web of Science*, and 61 (0.074%) from *IEEE*; compared to the first session, the relative weight of *EMBASE* and *Pubmed* over the total was decreased, while the amount of content specificity was greatly increased, as it was for *Scopus*. After removal of duplicates, 82579 unique records were obtained. Once joined with the session 1 records and duplicates removed, we obtained 98371 unique records, with just 1963 shared records between searches, that is the 2%. The percentage of records shared by two or more source dropped to 22%.

Six CR rounds were necessary to complete the second session classification, with just 6 new positive found after reviewing 568 extra records. It is interesting to notice that the first CR round required the user to review a substantial number of records (1273), but just labelling 275 of them (the canonical 250 plus 25 that were already labelled during the framework performance evaluation) was sufficient to drop this number to just 190 in the subsequent round. An evaluation of the convergence (Figure 1, Suppl. Mat. S2) showed that, in addition to the dynamics already observed in session 1 (shrinkage and negative shift), a second mode appeared in the mixture distribution of the records to be reviewed, centred in a highly positive zone. The interpretation is that as the number of negative training records increases, the engine gets more and more sceptical and asks to review even some records labelled as positive in the initial labelling at the beginning of session 1. This behaviour can be useful to spot classification errors and inconsistencies. Considering both sessions, 1333 records were reviewed and 107 (8.03%) were found.

Again, the evaluation of the relative importance of the terms showed that the engine was quite capable of internalizing the concepts behind the research topic. A subsample of these terms is reported in Table 2 of Suppl. Mat. S2.

Hyperparameter selection

As described in the methods, the selection of hyperparameters was achieved via evaluation of sensibility and efficiency through a grid search on a subset of 1200 completely manually labelled records (validation set). The best set of parameters suggested an initial input of 250 labelled records with 10x positive matches oversampling, an averaged ensemble of 10 models, no bootstrapping and an uncertainty zone defined by the 98% predictive interval. On the validation set, this combination of parameters reached a sensitivity of 98.8% (81 / 82 positive matches found) and efficiency of 61.5% (462 / 1200 records evaluated). A summary of the results of the grid search is reported in Table 3 in Suppl. Mat. S2.

Performance evaluation

To evaluate the theoretical performance of the engine on the full datasets (i.e., session 1 and session 2 data), a Bayesian logistic model was trained on each session dataset to predict the label of the records from the probability estimated by the engine (see methods for details). The performance of such simple models is quite high (Bayesian R2: 98.1% [97.4%, 98.3%] for session 1 and 98.2% [97.6%, 98.3%] for session 2) and the median of their cumulative predictive distribution matches quite well the actual number of cumulative positive records found. The predicted cumulative number of positive matches was used to evaluate the performance in the non-reviewed records (Table 3).

Figure 2 shows the actual and predicted (from the logistic model) cumulative number of positive matches, ordered by the initial simple ordering query. As confirmed by the high efficiency values reported in Table 3, it is striking how many more records would need to be evaluated manually to find all positive matches without using a smart search tool. The engine was able to find matches even close to the end of the heuristically ordered list of records. Specifically, in session 1 we observe an expected total number of positives of 101 [101, 108] for an estimated sensitivity of 100% [93.5%, 100%] and efficiency of 95.7% [95.3%, 95.7%].
In session 2 we observed a drop in the expected sensitivity, especially in the lower margin (97.3% [73.8%, 100%]), due to the fact that as the number of records grows very large, even a small probability can translate, in the worst scenario, into a relevant number of predicted positive matches (145 in this case). To ascertain that no evident positives were missed, we evaluated 100 more records between the unreviewed ones with the highest median predicted probability produced by the engine and found no actual positive matches.

\begin{table}[!h]

\caption{Table 3. Estimated performance summary. The table reports for each session, the number of reviewed records and the percentage over the total. Also, the posterior expected number of positive records, *Sensitivity* and *Work saved over random* (WSoR) are reported, with their 90% PrI truncated to a number of positive matches equal to the observed one. Finally the median Bayesian R^2 [90% CrI] of the logistic models is reported. [PrI] represents the 90% Predictive Interval while [trunc. PrI] indicate the residual PrI truncated at the observed realization (see. methods).}

Indicator	Session 1	Session 2
Total records	17755	98371
Reviewed records (% over total records)	761 (4.29%)	1316 (1.34%)
Expected efficiency (over random) [trunc. 90% CrI]	95.7% [95.3%, 95.7%]	98.6% [98.2%, 98.7%]
Observed positive matches (% over total records)	101 (0.57%)	107 (0.11%)
Predicted positive matches [trunc. 90% PrI]	101 [101, 108]	110 [107, 145]
Expected sensitivity [trunc. 90% PrI]	100% [93.5%, 100%]	97.3% [73.8%, 100%]
Simple Model R^2 [90% PrI]	98.1% [97.4%, 98.3%]	98.2% [97.6%, 98.3%]

\end{table}

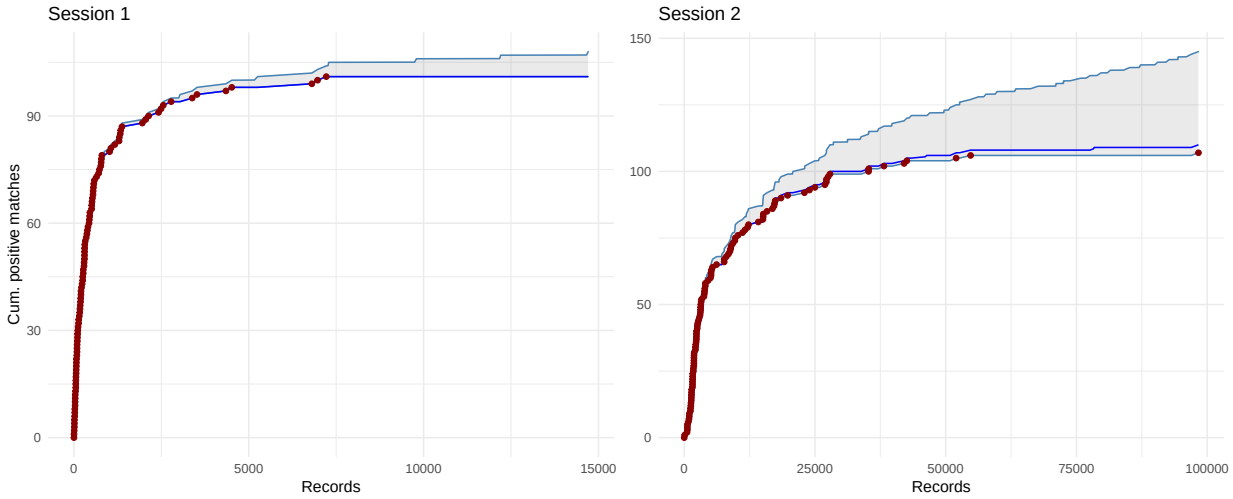

Figure 2. Observed cumulative number of positive matches (red dots) sorted by simple query ordering. The [trunc. 90% PrI] of the cumulative positive matches estimated by the logistic Bayesian model is shown as shaded area delimited by the 95% quantile of the PrI and by the observed number of positive matches (light blue lines). The median of the PrI is represented by a darker blue line.