Results

2022-01-19

Results

First session

The initial search query for the example topic 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))

selecting only results between 2010 and 2020 (included). Results were collected from Pubmed, WOS, IEEE, EMBASE and SCOPUS, using the framework tools as described in Methods and Supplemental Material S1.

The first search session returned a total of 27600 records, specifically 12719 (71.6% of the total) records from the EMBASE database, followed by 9546 (53.8%) from Pubmed, 3175 (17.9%) from SCOPUS, 2100 (11.8%) from WOS, and 60 (0.34%) from IEEE (Table 1). There were various degrees of overlapping between sources, 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 data set was composed by 17755 unique records. The first 250 records (based on "simple query ordering") were manually labeled. Of these 43 (17.2%) were labeled as positive, and 207 (82.8%) as negative.

The labeled records were used to train the Bayesian classification model used to label the remaining records. After seven classification and review (CR) iterations (three resulting in new positive matches and four extra replications to account for stochastic variability), a total of 101 positives matches were found, requiring manual review of 766 records (13.2% positivity rate).
It is possible to observe how the number of records that required manual review dropped rapidly between iterations (Table 2), indicating that the engine was converging while the uncertainties were resolved.
This phenomenon is better depicted in Fig. 1 of the Supplemental Material S2, showing the mixture distribution of the PPDs of the records, 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 shrinked (they concentrate in a shorter probability range) and shifted toward the negative zone as more positive matches are found and reviewed.

We extracted the 15 term more relevant for the classification model, described as: Term (citation part): Inclusion Rate (Inclusion Stability) [linear Relative Risk, Statistic].

Patient Transport (Keyword): 61.2 (3.77) [99.1, 21.3], Transfer (Abstract): 57 (3.93) [22.5, 15.4], Network (Title): 56.5 (2.91) [18, 14.2], Network & Patient (Abstract): 54.2 (4.66) [26.3, 15.2], Donker T (Author): 53.5 (4.56) [159, 16.5], Worker (Abstract): 50 (3.33) [0.421, -1.21], Hospitals (Keyword): 49.8 (4.31) [27.8, 16.5], Movement (Abstract): 47.8 (2.7) [27.2, 15], Spread (Title): 46.6 (2.25) [16.2, 12.1], Facility (Abstract): 45 (2.22) [19.6, 14.8], Orange County (Keyword): 44.3 (3.19) [199, 17.2], Conduct (Abstract): 42.6 (3.7) [0.221, -2.57], Patient (Abstract): 42 (3.61) [27.6, 7.23], Perform (Abstract): 41.9 (2.38) [0.342, -2.55], Hospital (Title): 39 (1.95) [12.5, 12.5].

The "&" indicates nc-ngrams, i.e., terms strongly co-occurrent in the documents.
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

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. All session shows records after joining and deduplication of the Session 1 and Session 2 data set.

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 Session 2, the engine uses the labels at the end of Session 1 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

this study (Donker T.) as well as the region of interest ("Orange County") of another research group active on the topic of healthcare asociated pathogen spreading over hospital networks. Some terms were considered highly relevant for the BART models (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.); these are terms which are relevant only in conjunction with other terms but not by themselves, highlighting the extra

predictive power brought by a non-linear model like BART.
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, 17070 (20.7%) from SCOPUS, 12956 (15.7%) from WOS, and 61 (0.074%) from IEEE; compared with the first session, the relative weight of EMBASE and Pubmed 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. The first CR iteration required the user to review a substantial number of records (1,273), but just labelling 275 of them (the canonical 250 plus 25 that were already labelled during the framework hyperparameter tuning) was sufficient to drop this number to just 190 in the subsequent round. An evaluation of the convergence (Figure 1, Supplemental Material 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 skeptical and asks to review even some records labelled as positive in the initial training set generated during 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 inclusion rate of the terms showed that the engine was quite capable of internalizing the concepts behind the research topic A subsample of the terms is reported in Table 2 of Supplemental Material 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 validation set of 1200 completely manually labelled records. 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). The results of the hyperparameter tuning are reported in Table 3 of Supplemental Material S2. Figure 2 in Supplemental Material S2 demonstrates that the positive record oversampling rate, the number of ensemble models and the size of the initial training set were the parameters that most impact performance.

Performance evaluation

```
## Running MCMC with 8 parallel chains...
##
## Chain 2 finished in 1.0 seconds.
## Chain 3 finished in 1.0 seconds.
## Chain 4 finished in 1.0 seconds.
## Chain 1 finished in 1.2 seconds.
## Chain 5 finished in 1.1 seconds.
## Chain 6 finished in 1.1 seconds.
## Chain 7 finished in 1.1 seconds.
## Chain 8 finished in 1.5 seconds.
##
## All 8 chains finished successfully.
## Mean chain execution time: 1.1 seconds.
## Total execution time: 1.8 seconds.
## Running MCMC with 8 parallel chains...
##
## Chain 1 finished in 1.5 seconds.
## Chain 3 finished in 1.5 seconds.
## Chain 4 finished in 1.6 seconds.
## Chain 7 finished in 1.6 seconds.
## Chain 2 finished in 1.6 seconds.
## Chain 5 finished in 1.7 seconds.
## Chain 6 finished in 1.7 seconds.
## Chain 8 finished in 1.7 seconds.
##
## All 8 chains finished successfully.
## Mean chain execution time: 1.6 seconds.
## Total execution time: 2.0 seconds.
```

To evaluate the theoretical performance of the engine, a surrogate Bayesian logistic regression model was trained on the manually reviewed labels using only the lower bound of the record PPDs as predictor (see Methods for details). The surrogate model show the high predictive power of the scores produced by the classification model (Bayesian R2: 98.1% [97.4%, 98.3%] for session 1 and 98.2% [97.6%, 98.3%] for session 2).

Figure 2 presents the actual and predicted (from the surrogate model) cumulative number of positive matches, ordered by the initial simple ordering query: the median of surrogate models' cumulative predictive distributions matches quite well the actual number of positive records found. It is striking how many more records would have needed to be evaluated manually to find the same number of positive matches without using a smart classification tool; some relevant record was found even close to the end of the heuristically ordered list of records.

Table 3 shows various performance indexes for both sessions, both descriptive (Total records, Reviewed records, Observed positive matches) and estimated through the surrogate model (Expected efficiency, Predicted positive matches, Expected sensitivity, R^2).
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 95.6% [95.3%, 95.7%]. In session 2 we observed a drop in the expected sensitivity, especially in the lower margin (97.3% [72.8%, 100%]), due to the fact that as the number of records grows, even a small probability can translate, in the worst case scenario, into a relevant number of missed positive matches (147 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 additional positive matches.

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 "Efficiency" (as WSoR) are reported, with their 90% PrI truncated to the observed realization in the dataset [trunc. PrI] (see. methods). Finally the median Bayesian R^2 [90% CrI] of the logistic models is reported. PrI: Predictive Intervals; CrI: Credibility Intervals.

Indicator	Session 1	Session 2
Total records	17755	98371
Reviewed records (% over total records)	766 (4.31%)	1333 (1.36%)
Expected efficiency (over random) [trunc. 90% PrI]	95.6% [95.3%, 95.7%]	98.6% [98.1%, 98.6%]
Observed positive matches (% over total records)	101 (0.57%)	107 (0.11%)
Predicted positive matches [trunc. 90% PrI]	101 [101, 108]	110 [107, 147]
Expected sensitivity [trunc. 90% PrI]	100% [93.5%, 100%]	97.3% [72.8%, 100%]
Simple Model R^2 [90% CrI]	98.1% [97.4%, 98.3%]	98.2% [97.6%, 98.3%]

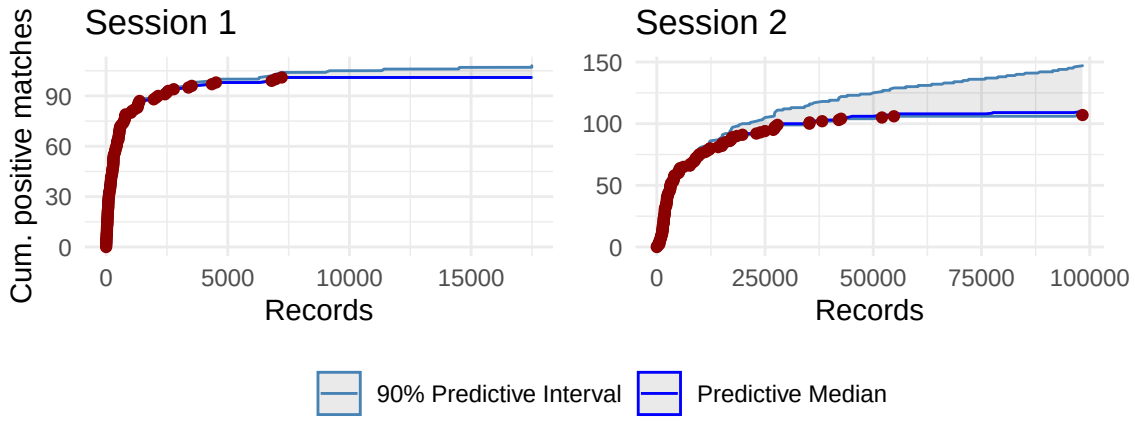

Figure 2. Observed cumulative number of positive matches (red dots) sorted by simple query ordering. The [trunc. 90% PrI] of the cumulative positive matches estimated by the logistic Bayesian model is shown as shaded area delimited by the 95% quantile of the PrI and by the observed number of positive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