

A large scale study of SVM based methods for
abstract screening in systematic review

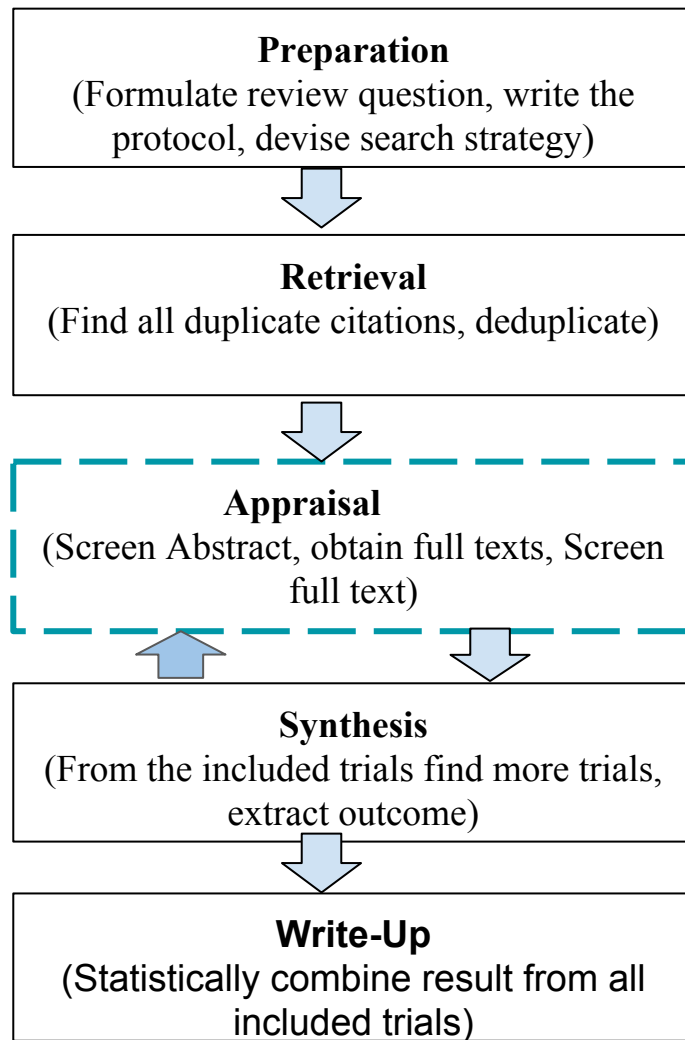
Systematic Review

The key characteristics of a systematic review are:

- a clearly stated set of objectives with predefined **eligibility criteria** for studies;
- an explicit, reproducible methodology;
- a systematic **search** that attempts to identify all studies that would meet the **eligibility criteria**;
- an **assessment** of the validity of the findings of the **included** studies, for example through the assessment of risk of bias; and
- a systematic presentation, and synthesis, of the characteristics and findings of the **included** studies.

* Many systematic reviews contain meta-analyses. It uses statistical methods to summarize the results of independent studies. By combining information from **all relevant** studies, meta-analyses can provide more precise estimates of the effects of health care than those derived from the individual studies included within a review. They also facilitate investigations of the consistency of evidence across studies, and the exploration of differences across studies.

Steps in Systematic Review Preparation



Can be
repeated
multiple times

Systematic Review (An Example)

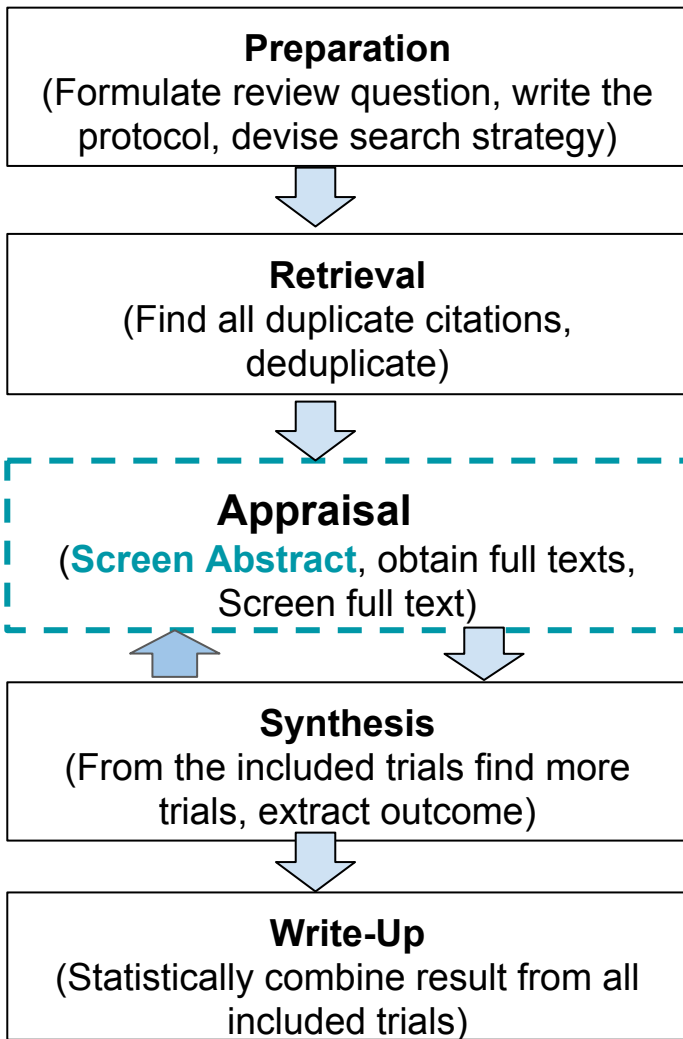
Review Title: Factors influencing falls after lower limb total joint arthroplasty: a systematic review and meta-analysis

Review question(s)

To summarize the evidence regarding factors that are related to post-TKA or post-THA falls in the hospital and beyond.

Abstract Screening

- First step in Appraisal
- Filters out irrelevant citations from relevant ones based on titles and abstracts
- Needs to download **full texts** only for **relevant** citations
- An instance of bipartite ranking problem (relevant citation should be ranked higher than irrelevant citation) [**Presentation point of view, Rayyan**]
- Also an instance of binary classification problem (Predicting the relevancy of a particular citation)



Why Important ?

- 27 million abstracts
- Two new abstracts every minute
- Adds over one million every year

The Problem (Total Recall => 100% Recall)

- Vanity search: find out **everything** about *me*
- Fandom: find out **everything** about *my hero*
- Research: find out **everything** about *my PhD topic*
- Investigation: find out **everything** about *something or some activity*
- **Systematic review:** find **all published studies** evaluating some method or effect
- Patent search: find **all prior art**
- Electronic discovery: find **all** documents *responsive to a request for production* in a legal matter
- Creating archival collections: label **all relevant** documents, for posterity, future IR evaluation, etc.

Expectation from Systematic Review App designer perspective (Rayyan's perspective)

- Feature Extraction should be fast and cacheable
- Features should be readily available
- The learning and prediction algorithm should be very efficient
- The algorithm (method/model) should be able to handle extreme data imbalance problem

Problems with existing studies

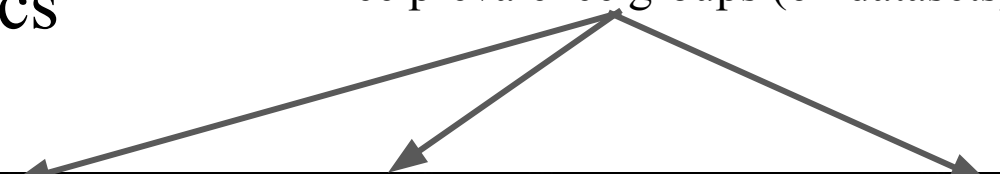
- Usage of small set of reviews (Unavailability of such kind of data)
- Usage of Non-overlapping metrics for the evaluation (Prioritizing over a specific metric)
- Does not perform variability analysis of metrics (Needs huge number of experiment and computation)
- No solid statistical testing or equivalence grouping of methods (No widely accepted method in the area)
- Does not take into account an app designers' perspective mentioned in previous slide

Our Contribution

- We use a large sample of reviews (61)
- We evaluate 18 different methods and report on 11 different metrics
- We perform a 500 x 2 cross validation We apply a 2-factor ANOVA analysis with a paired t-test and group the equivalent methods
- We present an ensemble method that present prediction results through a 5-star rating method

Dataset Statistics

Three prevalence groups (61 datasets)



Prevalence [0.22% – 5.92%]				Prevalence [6.79% – 13.07%]				Prevalence [13.45% – 40.08%]			
Review	Total	Pos	Prev (%)	Review	Total	Pos	Prev (%)	Review	Total	Pos	Prev (%)
P18	2241	5	0.22	P39	1149	78	6.79	P41	3250	437	13.45
P31	3034	16	0.53	P16	484	34	7.02	P22	1352	193	14.28
C12	1643	9	0.55	P32	895	63	7.04	P21	1352	193	14.28
P5	8812	60	0.68	P19	541	39	7.21	P36	449	66	14.70
C9	1914	15	0.78	P44	643	50	7.78	P12	820	121	14.76
P30	1864	19	1.02	P23	565	49	8.67	P23	910	137	15.05
P35	2601	33	1.27	P9	257	23	8.95	P20	2703	410	15.17
C1	2544	41	1.61	C6	1113	100	8.98	P17	1704	266	15.61
P7	417	8	1.92	P27	1243	114	9.17	P38	1386	221	15.95
C5	1965	42	2.14	P6	2539	235	9.26	C4	899	146	16.24
C2	845	20	2.37	P28	1242	115	9.26	P1	906	150	16.56
C13	3377	85	2.52	P43	996	95	9.54	P24	2488	419	16.84
P11	1580	41	2.59	P15	954	95	9.96	P42	4019	715	17.79
C14	660	24	3.64	P10	616	63	10.23	P2	1484	265	17.86
P26	351	13	3.70	P33	640	68	10.63	P8	498	100	20.08
C11	1330	51	3.83	P46	551	59	10.71	C7	368	80	21.74
P4	1187	56	4.72	P34	1728	200	11.57	P45	957	230	24.03
C3	296	16	5.41	C8	343	41	11.95	C10	503	136	27.03
P37	730	40	5.48	P40	850	110	12.94	P3	1487	404	27.17
P25	338	20	5.92	C15	306	40	13.07	P14	822	256	31.14
								P13	819	328	40.08

Metrics

Considered

Depends on
Threshold
settings

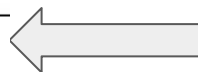
Threshold
Independent

Generally Used with
Active Learning
setting

Metric	Definition	Formula
Recall (Sensitivity)	Ratio of correctly predicted relevant citations to all relevant ones.	$\frac{TP}{TP + FN}$
Precision	Ratio of correctly identified relevant citations to all of those predicted as relevant.	$\frac{TP}{TP + FP}$
F-Measure	Combines Precision and Recall values. It corresponds to the harmonic mean of Precision and Recall for $\beta = 1$.	$\frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$
Accuracy	Ratio of relevant and irrelevant citations predicted correctly to all citations.	$\frac{TP + TN}{TP + TN + FP + FN}$
ROC (AUC)	Area under the curve obtained by graphing the true positive rate against the false positive rate; 1.0 is a perfect score and 0.5 is equivalent to a random ordering.	
AUPRC	Area under precision recall curve.	
AM ERROR	Arithmetic mean of the loss in Recall of the relevant (L_{R_p}) and irrelevant class (L_{R_n})	$\frac{\frac{FN}{TP + FN} + \frac{FP}{FP + TN}}{2}$
QUADMEAN ERROR	Quadratic mean, aka. root mean square, measures the magnitude of varying quantities. It is defined as the square root of the arithmetic mean of the squares of the loss in Recall of the relevant (L_{R_p}) and the irrelevant class (L_{R_n}).	$\sqrt{\frac{\frac{FN}{TP + FN}^2 + \frac{FP}{FP + TN}^2}{2}}$
Burden	The fraction of the total number of citations that a human must screen.	$\frac{TP^L + TN^L + TP^U + FP^U}{N}$
Yield	The fraction of citations that are identified by a given screening approach.	$\frac{TP^L + TP^U}{TP^L + TP^U + FN^U}$
Utility	Utility is a weighted sum of Yield and Burden. Here, β is a constant. It represents the relative importance of Yield, in comparison to Burden. We use $\beta = 19$ in our experimental evaluations.	$\frac{\beta \cdot \text{Yield} + (1 - \text{Burden})}{\beta + 1}$

Evaluation of existing SVM based Models (Models evaluated)

Feature Space	Algorithm, Parameter & Loss Function	Method Id
UniBi	SVM^{perf} (B0, AUC)	1
	— (B1, AUC)	2
	— (B1, KLD)	3
	— (B1, QuadMean)	4
	SVM (Default)	5
	SVM^{cost} (J, B0)	6
	— (J, B1)	7
<hr/>		
	SVM TRANSDUCTION	11
<hr/>		
WORD2VEC ROW	SVM^{perf} (B1, AUC)	21
	— (B1, KLD)	22
	— (B1, QuadMean)	23
	SVM^{cost} (J, B0)	24
	— (J, B1)	25
<hr/>		
WORD2VEC COL	SVM^{perf} (B1, AUC)	31
	— (B1, KLD)	32
	— (B1, QuadMean)	33
	SVM^{cost} (J, B0)	34
	— (J, B1)	35



In PX2 evaluation (50%-50%) settings, it already knows how many positive to predict. We still evaluate the case to show the best performance. In extreme imbalance cases, we may need to set the -p parameter, i.e. number of examples to be predicted as positive.

Results from WORD2VEC Model

Query	Result
model.similarity ('liver', 'cirrhosis')	0.63
model.similarity('breast','cancer')	0.46
model.most_similar(positive=['liver'], top-k with k=10)	hepatic (0.67), cirrhosis (0.63), cholestatic (0.51), spleen (0.51), steatosis (0.5), kidney (0.50), steatohepatitis (0.50), extrahepatic (0.46), pancreas (0.45), cirrhotic (0.45)
model.most_similar(positive=['cancer','cirrhosis'], negative = ['breast'], top-k with k=5)	cirrhotic (0.46), hcc (0.46), liver (0.45), hepatocellular (0.43), metavir (0.43)

Statistical Testing Procedure

1. METRIC \sim DATA + METHOD (Fit the Model)
2. Perform a 2-factor (DATA, METHOD) analysis of variance
3. Helps us to identify whether there is any statistically significant difference among the methods, the datasets, and the interactions between the methods and the data
4. If the test succeeds we do the following:
 - a. We find the best method based on average value in a certain metric
 - b. All the methods which are not statistically significant with the best method falls in the same group
 - c. We repeat the process in Step a and b for rest of the methods until all the methods have a rank group id

Evaluation of existing Models (Results)

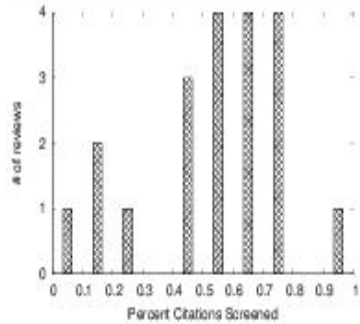
Metric	Prevalence [0.22% – 5.92%]			Prevalence [6.79% – 13.07%]			Prevalence [13.45% – 40.08%]		
	<i>rg</i> = 1	<i>rg</i> = 2	<i>rg</i> = 3	<i>rg</i> = 1	<i>rg</i> = 2	<i>rg</i> = 3	<i>rg</i> = 1	<i>rg</i> = 2	<i>rg</i> = 3
PREC	<u>7</u> , 11	3, <u>5</u> , 4, 6, 35		<u>5</u> , <u>7</u>	11	3, 6, 22, 25	<u>5</u>	<u>7</u>	11
REC	<u>21</u>	31	1, 2, 34	<u>21</u>	1, 2, 31, 34	22	<u>21</u>	1,2	31
F1	<u>11</u> , 25	24, 35	4, 22, 23	<u>11</u> , 25	4, 22, 23, 24, 35	3, 6, 33	7, <u>11</u> , 25	35	24
ACC	<u>7</u>	5	3, 6, 11	5, <u>7</u>	11	6	5, <u>7</u>	11	3, 4, 6
AUC	<u>21</u> , <u>25</u>	31, 35	1, 2, 5, 7, 23, 24	<u>21</u> , <u>25</u>	4, 24, 31, 35	1, 2, 5, 7, 22, 23	<u>21</u>	7, 25	(All Others)
AUPRC	<u>1</u> , <u>2</u> , 5, 7, 21	4, 25	11, 23, 24, 31	<u>1</u> , <u>2</u> , 4, 5	7, 21	25	<u>1</u> , <u>2</u>	(All Others)	-
AM ERROR	<u>4</u> , 31, 35, 24, <u>25</u>	1, 2, 11, 22, 23, 32, 33, 34	3, 6, 7, 21	<u>4</u> , <u>25</u>	24	31, 33, 35	<u>25</u>	24	4, 35
QUADMEAN ERROR	1, 2, 4, <u>24</u> , 25, 31, 32, 33	23, 34, 35	11, 22 4, 25, 33	<u>24</u>	23, 31, 35	25	<u>24</u>	4, 35	
YIELD	<u>21</u>	31	1, 2, 34	<u>21</u>	1, 2, 31, 34	32	<u>21</u>	1, 2	31
BURDEN UTILITY	<u>5</u> , <u>21</u>	7, 31	3, 6, 1, 2, 34	<u>5</u> , <u>21</u>	7, 1, 2, 31, 34	6, 32	<u>5</u> , 7, <u>21</u>	11, 1,2	6, 31
AUC (STD)	<u>21</u> , 22, 23, 25, 31, 35	All Others		<u>21</u> , 25	24, 31, 35	22, 23	<u>21</u>	25, 31	35
AUPRC (STD)	<u>32</u> , 34	3, 6, 22, 33	(All others)	No statistically significant difference among methods			No statistically significant difference among methods		

Observations from Evaluation

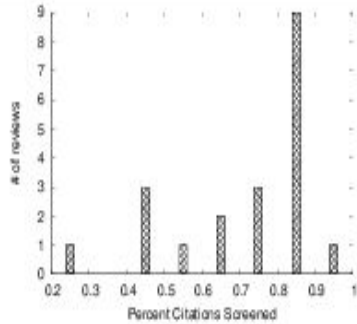
- There is almost always a method that ranks first in the three prevalence group
- Various methods perform well on different prevalence groups and for different metrics
- There is no “winner” or best method across all metrics
- Method 21 (Word2VEC ROW + SVM_Perf (AUC) seems to be a good choice, outperforming the other methods in five metrics

An Active Learning Experiment

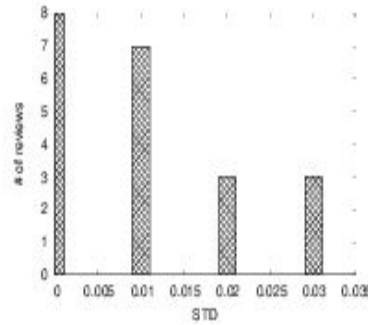
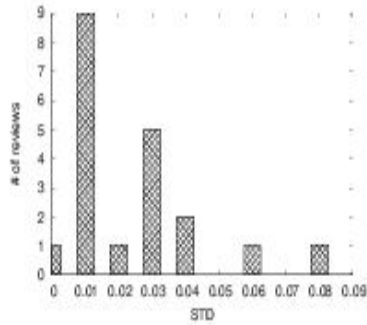
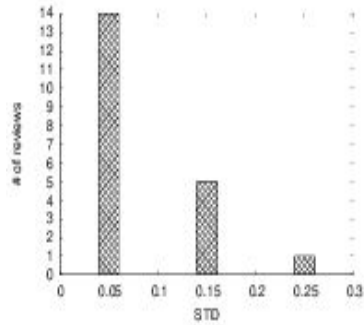
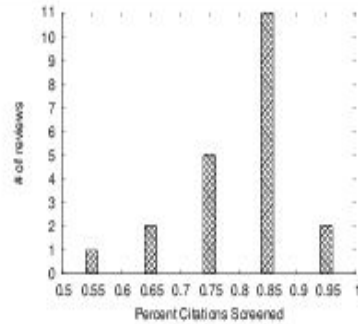
Prevalence [0.22% – 5.92%]



Prevalence [6.79% – 13.07%]



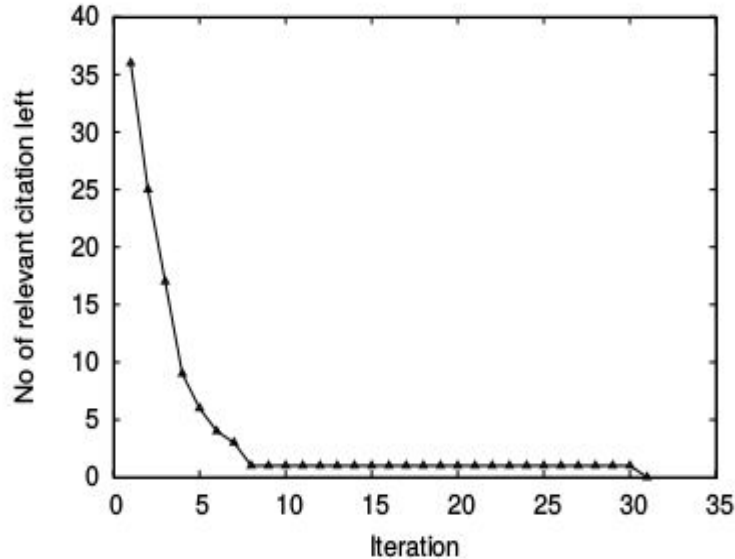
Prevalence [13.45% – 40.08%]



Observations:

- For low prevalence group, out of 20 reviews, 7 reviews need 40% of the total citations
- For the mid and high prevalence groups, 9 out of 20 and 11 out of 21 reviews need around 80% to 90% citations to be screened to get all the relevant citations.

An Interesting Case !!!!!



Observations:

- The figure represents the inclusion behavior of a particular random run of review 1.
- It gets almost all but the final relevant one after screening only 400 out of 2544 which is around 15%
- The final one cost around 1100 citations to screen more. Is the final one a outlier ??

Our Proposal

- Can we design an algorithm which takes best algorithms in various metrics and combine them?
- Method 21 outperforms the other methods in AUC and Recall and it has also lowest standard deviation in AUC
- Method 25 produces the highest F1 Measure
- Method 7 has the highest Precision
- We combine these three methods to give **RelRank**

Our Algorithm

Algorithm 1: RelRank: A Five Star rating algorithm using ensemble of max-margin based methods

Input : \mathcal{L} , Labeled dataset; \mathcal{U} , Unlabeled dataset

Output: Score, $\mathcal{S}_{1 \leq i \leq |\mathcal{U}|}$

```
1  $\mathcal{F}_{\mathcal{L}}, \mathcal{F}_{\mathcal{U}} \leftarrow \text{GenerateFeature}(\mathcal{L}, \mathcal{U}, \text{feature} = \text{W})$ 
2  $h1 \leftarrow \text{Train}(SVM^{perf}, \mathcal{F}_{\mathcal{L}})$ 
3  $h2 \leftarrow \text{Train}(SVM^{cost}, \mathcal{F}_{\mathcal{L}})$ 
4  $S_{h1} \leftarrow \text{Predict}(h1, \mathcal{F}_{\mathcal{U}})$ 
5  $S_{h2} \leftarrow \text{Predict}(h2, \mathcal{F}_{\mathcal{U}})$ 
6  $\mathcal{F}_{\mathcal{L}}, \mathcal{F}_{\mathcal{U}} \leftarrow \text{GenerateFeature}(\mathcal{L}, \mathcal{U}, \text{feature} = \text{U})$ 
7  $h3 \leftarrow \text{Train}(SVM^{cost}, \mathcal{F}_{\mathcal{L}})$ 
8  $S_{h3} \leftarrow \text{Predict}(h3, \mathcal{F}_{\mathcal{U}})$ 
9  $\mathcal{S} \leftarrow \text{GenerateCombinedScore}(\mathcal{U}, S_{h1}, S_{h2}, S_{h3})$ 
10 return  $\mathcal{S}$ 
```

Observations:

- RelRank can capture the precision of Method 7 at 5-star rating
- RelRank at 3-star has the top recall whereas RelRank at 4-star and 5-star ranks in the top rank groups
- In conclusion, it can capture the goodness in the three methods combined.

Conclusion

- Automating the production of systematic reviews is crucial in delivering the promises of evidence-based medicine
- We studied the most popular methods employed in the very first step in appraisal (citation screening)
- Various methods perform well on different prevalence groups and for different metrics
- Active Learning methods may consider filtering out outliers for better performance

Thank You