Active Literature Discovery for Scoping Evidence Reviews

How Many Needles are There?*

Byron C. Wallace†, Issa J. Dahabreh†, Kelly H. Moran‡△
Carla E. Brodley‡, Thomas A. Trikalinos†
†Health Services Policy & Practice, Brown University, Providence, RI
‡Computer Science, Tufts University, Medford, MA
△MIT Lincoln Laboratory, Lexington, MA
byron_wallace@brown.edu, issa_dahabreh@brown.edu, kmoran@ll.mit.edu
brodley@cs.tufts.edu, thomas_trikalinos@brown.edu

ABSTRACT

Scoping reviews of the biomedical literature are commonly used in health technology assessments to inform the planning of more detailed and resource-intensive evaluations. A typical task is to 'map' the literature addressing a specific clinical question, i.e., (i) identify as many relevant articles of interest as feasible under a constrained budget, and (ii) estimate how many such articles likely exist. These are competing objectives. Using active retrieval strategies (e.g., active learning) to realize the former aim immediately hinders our ability to achieve the latter: 'naive' estimates of the number of relevant articles taken over an enriched sampled acquired through selective sampling will be inflated. We propose a novel method for correcting such estimates. We demonstrate the efficacy of our approach on three systematic review datasets, showing that we can achieve both aims: rapid evidence discovery and acceptably accurate estimation of the number of relevant articles.

Categories and Subject Descriptors

I.2.l [**Computing Methodologies**]: Artificial Intelligence–*Applications and Expert Systems*

General Terms

Algorithms, Human Factors

Keywords

active learning, medical, information retrieval, applications, text classification

*This work sponsored in part by the United States Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

KDD-DMH'13, August 11-14, 2013. Chicago, USA.

1. INTRODUCTION AND MOTIVATION

Health technology assessments are increasingly used to inform decision making at all levels of health care [7]. A key tool of such assessments is the *systematic review*, a protocol-driven and resource-intensive evaluation of all available evidence addressing a clinical question [9, 18, 19]. Large systematic reviews can take several months to complete and are costly, because they are undertaken by researchers with substantial medical and methodology training. Therefore, substantial forethought and planning goes into prioritizing the questions that should be addressed with systematic reviews. Clinical or policy topics attain higher priority if a new systematic review is expected to provide information that will change current practice; reduce costs, resource use, needless practice variation or health inequalities; affect the well-being of an important population subgroup; and for which substantial empirical evidence has been generated.

Typically these prioritization exercises incorporate input from various stakeholding parties. Irrespective of the exact prioritization methodology being used, however, a common task during the planning of a systematic review is the construction of *evidence maps* or *horizon scans* of large domains of the biomedical literature [15, 6, 12]. These rapid-turnaround scoping exercises describe the volume and type of available evidence on a topic in broad strokes. For example, if the scoping review suggests that the available evidence on a candidate question is sparse, a full-blown systematic review may be premature to undertake. Conversely, if the estimated number of relevant articles is large, appropriate resources must be budgeted and/or the scope must be contained.

Among other goals, these scoping exercises invariably include two aims that must be achieved rapidly and within a tight budget. The first is to *maximize the number of relevant articles identified*, and the second is to *estimate the total number of eligible articles in the literature*. Identifying relevant articles in the planning stage provides direct content-relevant information to the stakeholders and the prioritization team. The importance of accurately estimating the scope of a full-blown systematic review is obvious.

Here we frame the problem following the typical workflow in a scoping review: standard Boolean queries are issued to a biomedical database (e.g., MEDLINE, via PubMed), returning a relatively large number of citations (often several

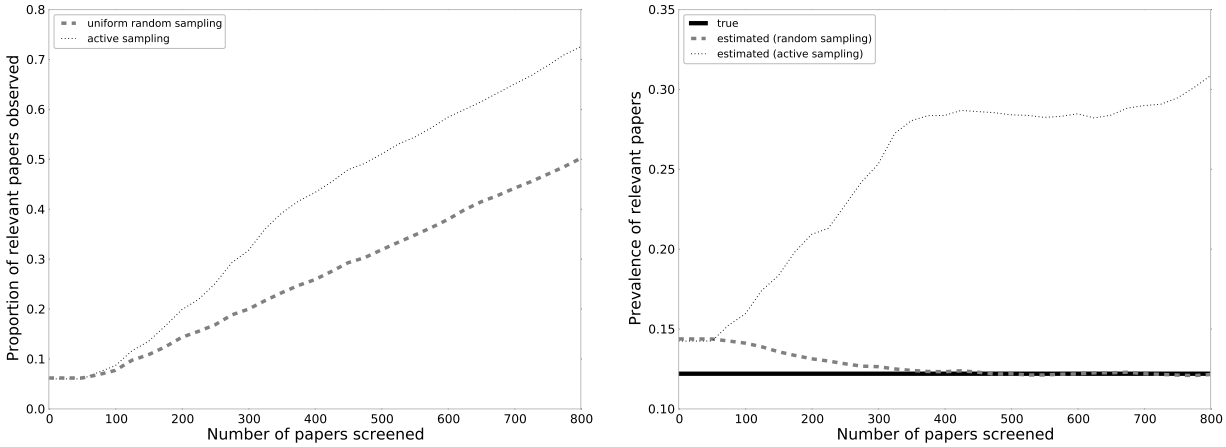

Figure 1: Selective compared to uniform-random sampling versus the number of labels provided (screening decisions). The left-hand plot shows the proportion of relevant articles identified; the right-hand plot shows the estimated prevalence of relevant citations. On the latter, the thick black line demarcates the true prevalence in the corpus (which is constant). Lines correspond to means over ten independent runs. Selective sampling drastically increases the amount of retrieved relevant literature, but (unsurprisingly) results in a poor estimate of the prevalence.

thousands). Such searches are sensitive by design to cast a net wide enough to capture most (and preferably, all) relevant articles. Assuming that budget constraints do not allow screening (labeling) all the citations in the retrieved corpus as relevant or irrelevant to the topic at hand, researchers typically examine a (uniform) random sample. The proportion of relevant articles in the sample is an unbiased estimate of the prevalence π of relevant articles in the pool.

However, because the search is not specific, π is typically small (often as low as $1-5\%$), and only a small number of relevant articles are found in the drawn sample. Thus while the second aim is achieved, the first is not. On the other hand, selective sampling, e.g., using ranking technologies [17] or *active learning* [24, 25] facilitates rapid discovery of relevant articles. However, when using such strategies the observed proportion of relevant articles in the (selectively) drawn sample will be an upwards biased estimate of π, because relevant articles have been oversampled by design.

This is illustrated in Figure 1, which shows the results of retrospectively simulating interactive information retrieval using a previously assembled dataset.[1] Selectively sampling articles preferentially with respect to their predicted probability of being relevant (left-hand side) results in rapid discovery of relevant articles, but induces a severely inflated estimate of the prevalence. Whereas sampling documents uniformly at random is inefficient in terms of identifying relevant articles, but the observed sample prevalence of relevant articles provides an accurate estimate of π. However *we would like to achieve both of these aims*, rather than one or the other.

To our knowledge, this trade-off between article identification and estimation of the proportion of relevant articles in a corpus has not previously been addressed. But this poses an important practical problem to evidence-gathering activities that could otherwise benefit from active ranking/retrieval strategies. To address this issue, *we present a novel method to leverage interactive information retrieval (IR) methods to rapidly find relevant articles while obtaining a reliable estimate of the proportion of relevant papers early on in the process.*

Our approach relies on two prevalence estimates. One is based on stochastically sampling articles with known probabilities (we set the probability of sampling each article to a value proportional to its predicted probability of being relevant), and then using inverse-weighting to correct for the induced sampling bias. The other relies on probability estimates to calculate the expected number of relevant citations remaining in the unlabeled set at each step (and combines this expectation with the number of relevant articles observed thus far). This approach can be used in conjunction with any learning algorithm that provides an estimate of the probability that individual citations are relevant. We demonstrate that this method allows for rapid evidence discovery while still providing a reasonable estimate of prevalence on three datasets from previously conducted systematic reviews.

The remainder of this paper is structured as follows. We briefly review related work in the following section to provide a context for our contribution. We present our proposed method in Section 3. In Section 4 we describe our experimental setup and we report results in Section 5. We conclude in Section 6 by discussing future directions.

[1]The Chronic Obstructive Pulmonary Disease (*COPD*) systematic review [3]; see Section 4.

2. RELATED WORK

There is a wealth of literature on biomedical information retrieval (IR); for a comprehensive survey, see [16]. Most relevant to our work, several existing systems allow the user to provide relevance feedback to rerank or refine the results [27, 26, 21]. Leveraging classification technology to rank articles with respect to their predicted probability of relevance (as we do here) has been previously proposed [21, 10]. But most of these approaches have considered the problem from a general perspective of improving PubMed search, whereas we are specifically interested in developing methods that facilitate retrieval of relevant literature for EBM tasks such as systematic reviews, horizon scans and evidence scoping.

An exception to this is the recent work by Karimi et al. [14] in which they experimented with ranking articles for the literature retrieval step in systematic reviews, as opposed to retrieving them with Boolean queries as is usually done. They argued that in the case of systematic reviews, wherein comprehensiveness (sensitivity) takes precedence over efficiency (precision), Boolean queries offer advantages over ranking-based approaches. A similar argument might be applied to scoping and related evidence synthesis tasks, although in such cases sensitivity is generally not as important as it is for comprehensive systematic reviews. Karimi et al. highlighted the potential of combining Boolean and ranking queries by starting with the former and then using the latter to rank within the retrieved set. This is the strategy we pursue here; we assume that we start with a set of articles retrieved via a query, and we rank within this set, interactively requesting labels from the user.

However, in this paper our focus is not on the particular ranking function or strategy used. Instead, we look to develop a general method that allows for selective sampling (efficient discovery of relevant evidence) *and* accurate estimates of the volume of existing relevant literature. There has been a fair amount of work on using text classification to semi-automate citation screening for systematic reviews [5, 4], including approaches that use active learning [24, 25]. But this work has not addressed the issue of simultaneously estimating the volume of published relevant literature while rapidly identifying relevant studies.

Aside from work specific to biomedical text, there has recently been interest in inducing unbiased estimators of the expected loss of models during active learning [2, 11]. Most relevant to the present work, Ganti and Gray recently proposed an inverse-weighting based method for accomplishing this in the context of pool-based active learning, which they called *Unbiased Pool-based Active Learning*, or UPAL [11]. Our strategy is similar, in that we leverage inverse-weighting to mitigate bias, but we focus on estimating the *prevalence* rather than the expected loss. Moreover, we combine this with a second estimator based on predicted probabilities (see Section 3), which improve over time (i.e., as we acquire additional training data).

3. ACTIVE RANKING & PREVALENCE ESTIMATION

We now present our method, which we envision proceeding in rounds. At each round k, we would draw a set of articles $\mathcal{S}_k$ of cardinality B_k, following the sampling procedure described below. We will denote the set of articles the expert has labeled thus far (up to k) by $\mathcal{L}_{k-1}$, the set of as-yet unlabeled articles by $\mathcal{U}_{k-1}$ and the set of all articles (the entire corpus) by $\mathcal{D}$. The expert would be asked to label those articles in $\mathcal{S}_k$ not already in $\mathcal{L}_{k-1}$, i.e., $\mathcal{S}_k \setminus \mathcal{L}_{k-1}$. These articles would be removed from $\mathcal{U}_{k-1}$ and added to the labeled set to form $\mathcal{L}_k$. We then train a model over $\mathcal{L}_k$ and obtain predicted probabilities of relevance over all $x_i \in \mathcal{D}$, which we denote by $\mathbf{P}_k = (p_{k,1}, \ldots p_{k,|\mathcal{D}|})$.

To attain aim 1 (to rapidly identify relevant articles), we then preferentially sample those articles with high probability of being relevant (as predicted by the model). Specifically, we use probability sampling with replacement to select B_k papers to be labeled at round k. (The sampling is stochastic, but we set the probabilities such that the expected number of articles sampled at each round is B, i.e., $\mathbb{E}(B_k) = B$.) Call the current vector of sampling probabilities $\mathbf{Z}_k = (z_{k,1}, \ldots z_{k,|\mathcal{D}|})$. By design (to attain aim 1), $\mathbf{Z_k}$ is a function of $\mathbf{P_k}$. In the simplest case $\mathbf{Z_k} = \mathbf{P_k}$, in which case the expected number of articles sampled at each round would be equal to the expected number of relevant articles in the corpus.

To achieve aim 2 (i.e., to estimate prevalence π in the original corpus), we use two estimators. The first estimator $\hat{\pi}^*_{HT,k}$ is based on the Horvitz–Thompson estimator [13],[2] in which the contribution of each observed label y_i in $\mathcal{S}_k$ is weighted by its inverse sampling probability ($z_{k,i}$).

$$\hat{\pi}_{HT,k} = \frac{\sum_{x_i \in \mathcal{S}_k} y_i / z_{k,i}}{\sum_{x_i \in \mathcal{S}_k} 1 / z_{k,i}} \tag{1}$$

Under typical large sample assumptions, $\hat{\pi}_{HT,k}$ is an unbiased estimator of π provided that all $z_{k,i}$ are positive. Because we sample with replacement in each round, the k independent estimates (one per iteration) can be regarded as *bootstrap* estimates, and we can average them to reduce variance [8]. Our first estimator is thus the mean:

$$\hat{\pi}^*_{HT,k} = \frac{1}{k} \sum_{t=1}^{k} \hat{\pi}_{HT,t} \tag{2}$$

The second estimator we use, which we denote by $\hat{\pi}_{M,k}$, is based on the predicted probabilities of relevance $\mathbf{P}_k$ and the observed labels (i.e., the labels in $\mathcal{L}$).

$$\hat{\pi}_{M,k} = \frac{\sum_{x_i \in \mathcal{L}_k} y_i + \sum_{x_i \in \mathcal{U}_k} p_{k,i}}{|\mathcal{D}|} \tag{3}$$

This estimator can be biased; it will depend on the probability estimator used. Here we use Support Vector Machines (SVMs) [1] as our base classifier and a variant of Platt scaling [20] to estimate class probabilities. In particular, we construct an ensemble of regressors, each induced on a random,

[2]The * here is to highlight that this is an average (Eq. 2).

balanced bootstrap sample of the training data. We have previously demonstrated that this method provides more reliable probability estimates for minority instances in imbalanced scenarios than standard Platt scaling (see [23]). Because it emphasizes providing reliable estimates for minority instances, we can expect that estimates from this model will be upwardly biased. In any case, in principle any probability estimator could be used in the proposed framework.

Note that a 'naive' estimator that ignores the sampling scheme (as in Eq. 4) is biased because the sampling weights are not uniform due to the preferential sampling scheme.

$$\hat{\pi}_{naive,k} = \frac{1}{|\mathcal{L}_k|} \sum_{x_i \in \mathcal{L}_k} y_i \tag{4}$$

We use two estimators because we expect each to perform well at different points during the retrieval process. Specifically, the Horwitz–Thompson estimator (1) is unbiased in expectation, but $B << |\mathcal{D}|$ and thus, empirically, the sampling probabilities z_i may be practically 0. We may therefore expect an upward bias in the estimate of π (if it is extremely unlikely that we draw irrelevant articles). Recall that to achieve aim 1 (to identify as many positives as possible) we set the sampling probabilities proportional to the predicted probability of relevance. Thus we observe the 'paradox' that as the model improves (as we achieve aim 1), the elements of $\mathbf{P}_k$ are expected to tend towards 0 or 1, resulting in potentially biased estimates of π. By contrast, in earlier rounds, the probability predictions would be more uncertain, and Eq 2 is more likely to be unbiased.

The estimator in (3), meanwhile, may be inaccurate in earlier iterations because the model is ill-informed. However, in later rounds the model probability estimates will improve, and we thus expect this estimate to approach the true prevalence π. Hence we expect the two estimators to be more accurate at different points in the interactive retrieval process, and we expect both to tend to over-estimate the true prevalence when they are off. We therefore take the pragmatic approach of taking the minimum of the two estimates as our estimate of π at any given iteration.

Algorithm 1 provides pseudo-code for the approach we have just described (in the pseudo-code we have elided the k subscripts, which denote the iteration; these are implicit).

Algorithm 1 Adjusted active retrieval

Input: Initial labeled articles $\mathcal{L}$, Unlabeled pool of articles $\mathcal{U}$, Desired batch-size B

$\mathbf{HT} \leftarrow \{\}$
while Not *STOPPING CRITERION* **do**
  $\mathcal{D} \leftarrow \mathcal{U} + \mathcal{L}$
  $f \leftarrow$ probability estimator induced over $\mathcal{L}$
  $\lambda \leftarrow \frac{\sum_{x_i \in D} f(x_i)}{B}$
  $\mathcal{S} \leftarrow \{\}$
  for all $x_i \in \mathcal{D}$ **do**
    $z_i \leftarrow \frac{f(x)}{\lambda}$
    add x_i to $\mathcal{S}$ with probability z_i
    $w_i \leftarrow \frac{1}{z_i}$
  end for
  request labels $y_i \in \{0, 1\} \forall x_i \in \mathcal{S} \setminus \mathcal{L}$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{S}$
  $\hat{\pi}_{HT} \leftarrow \frac{\sum_{x_i \in \mathcal{S}} w_i \cdot y_i}{\sum_{x_i} w_i}$
  add $\hat{\pi}_{HT}$ to $\mathbf{HT}$
  $\hat{\pi}^*_{HT} \leftarrow$ average($\mathbf{HT}$)
  $\hat{\pi}_M \leftarrow \frac{\sum_{i \in 0,1,\ldots \mathcal{L}} y_i + \sum_{x_i \in \mathcal{U}} f(x_i)}{|\mathcal{D}|}$
  $\hat{\pi} \leftarrow \min(\hat{\pi}^*_{HT}, \hat{\pi}_M)$
end while

4. DATASETS AND EXPERIMENTAL SETUP

We experimented with three systematic review datasets. The *proton beam* dataset is from a systematic review of comparative studies on charged particle radiotherapy versus alternate interventions for cancers [22]. It consists of 4,751 citations retrieved via a broad search, 243 of which are "relevant" (about 5%). The *COPD* dataset is from a systematic review and meta-analysis of all genetic association studies in chronic obstructive pulmonary disease [3] comprising 1,601 citations, 196 of which are "relevant" (about 12%). The *Sedatives* dataset from the set of systematic reviews of drug classes made available by Cohen [4].[3] This dataset comprises 1,655 articles, of which about 8% (132) are relevant.

[3]We selected this specific drugs dataset at random.

We encoded the articles comprising these documents using standard binary bag-of-words (BoW) representation and removing words that were either on the PubMed stopword list[4] or that did not occur at least three times. We collapsed titles, abstracts and MeSH terms (when available) into a single document before performing BoW encoding.

We used a 'batch' size of $B = 50$, i.e., we scaled the sampling probabilities $\mathbf{P}_k$ such that the expected number of articles drawn at each step k was 50. This is a relatively large batch-size by active learning standards (though not necessarily for information retrieval tasks). Slightly larger batches allow for more reliable estimates of the prevalence at each iteration. Given that we are dealing with low prevalences (often $< .1$), using small batch sizes would likely result in poor estimates. Note also that because we are sampling with replacement, the expert will only be asked to label the as-yet unlabeled articles drawn at each iteration; this number shrinks over time, as an increasing number of articles have already been labeled. This effectively shrinks the batch size as the process proceeds. That is, even though we calculate prevalence estimates over draws comprising 50 articles, in practice the expert would likely only review a fraction of these at each iteration, since some of them will have already been labeled.

We seeded each strategy with 4 articles (2 relevant and 2 irrelevant, randomly drawn from each class) and simulated interaction until labels were acquired for half of the dataset. We performed 10 independent runs of this experiment and report averages over these.

5. EXPERIMENTAL RESULTS

Results over the three systematic review datasets, *COPD*, *proton beam* and *Sedatives*, are shown in Figures 2, 3 and

[4]`http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/`

4, respectively. Consider first the left-hand sub-plots, which show the proportion of relevant evidence literature discovered (y-axis). This corresponds to the efficiency of the corresponding method in terms of facilitating rapid identification of relevant biomedical literature. In all cases, stochastically sampling with probabilities proportional to the predicted probability of relevance discovers a far greater amount of relevant literature than does sampling uniformly at random.[5]

The right-hand sub-plots show the estimated prevalences. The true prevalence in each dataset is demarcated by the thick black lines (this is constant, and only available once the entire corpus has been screened). The adjusted estimates are drastic improvements over the naive, unadjusted estimate (i.e., the observed proportions) in all cases. *Our method thus allows one to simultaneously rapidly identify relevant biomedical literature and estimate how much of it exists.*

6. CONCLUSIONS

We have addressed an important practical problem in many biomedical literature retrieval tasks: simultaneously attaining reliable estimates of the number of published relevant evidence while using selective sampling (ranking) to rapidly identify studies of interest. We demonstrated that this method facilitates rapid identification of relevant studies (compared to perusing them in an arbitrary order) while providing reasonable estimates of the overall prevalence of pertinent literature early on in the process of literature screening.

This method may be especially useful for horizon scan and evidence-mapping activities, in which one aims to rapidly characterize the body of existing literature on a topic [15, 6, 12]. Such activities are best informed both by finding relevant studies and by estimating how many might remain. Moving forward, we hope to develop a more principled means of combining the two prevalence estimates we have proposed. We also plan on conducting further empirical evaluations, and conducting experiments with live (rather than simulated) experts.

7. REFERENCES

[1] K. P. Bennett and C. Campbell. Support vector machines: hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, 2(2):1–13, 2000.

[2] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th annual International Conference on Machine Learning (ICML)*, pages 49–56. ACM, 2009.

[3] P. Castaldi, M. Cho, M. Cohn, F. Langerman, S. Moran, N. Tarragona, H. Moukhachen, R. Venugopal, D. Hasimja, E. Kao, et al. The COPD genetic association compendium: A comprehensive online database of COPD genetic associations. *Human Molecular Genetics*, 19(3):526–534, 2010.

[4] A. Cohen, W. Hersh, K. Peterson, and P.-Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *JAMIA*, 13:206–219, 2006.

[5] I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. *Knowledge Discovery in Databases (KDD)*, pages 125–136, 2004.

[6] K. Davis, N. Drey, and D. Gould. What are scoping studies? A review of the nursing literature. *International Journal of Nursing Studies*, 46(10):1386–1400, 2009.

[5]Sampling the top ranking document deterministically increases this proportion even more, but without sampling probabilities there would be no way to bias-correct via weighting.

[7] J. Eden, L. Levit, A. Berg, S. Morton, et al. *Finding what works in health care: Standards for systematic reviews*. National Academies Press, 2011.

[8] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26, 1979.

[9] M. Egger, G. D. Smith, and D. Altman. *Systematic reviews in health care: Meta-analysis in context*. BMJ books, 2008.

[10] J.-F. Fontaine, A. Barbosa-Silva, M. Schaefer, M. R. Huska, E. M. Muro, and M. A. Andrade-Navarro.

[11] R. Ganti and A. Gray. UPAL: Unbiased pool based active learning. *arXiv preprint arXiv:1111.1784*, 2011.

[12] M. Gwinn, D. Grossniklaus, W. Yu, S. Melillo, A. Wulf, J. Flome, W. Dotson, and M. Khoury. Horizon scanning for new genomic tests. *Genetics in Medicine*, 13(2):161–165, 2011.

[13] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[14] S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making*, 10(1):58, 2010.

[15] D. Levac, H. Colquhoun, K. O'Brien, et al. Scoping studies: Advancing the methodology. *Implementation Science*, 5(1):69, 2010.

[16] Z. Lu. Pubmed and beyond: A survey of web tools for searching biomedical literature. *Database: the journal of biological databases and curation*, 2011, 2011.

[17] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[18] C. D. Mulrow. The medical review article: State of the science. *Annals of Internal Medicine*, 106(3):485–488, 1987.

[19] C. D. Mulrow. *Systematic reviews: Synthesis of best evidence for health care decisions*. American College of Physicians, 1998.

[20] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[21] G. Poulter, D. Rubin, R. Altman, and C. Seoighe. MScanner: A classifier for retrieving medline citations. *BMC Bioinformatics*, 9(1):108, 2008.

[22] T. Terasawa, T. Dvorak, S. Ip, G. Raman, J. Lau, and T. A. Trikalinos. Charged Particle Radiation Therapy for Cancer: A Systematic Review. *Annals of Internal Medicine*, 2009.

[23] B. C. Wallace and I. J. Dahabreh. Class probability estimates are unreliable for imbalanced data (and how to fix them). In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 695–704. IEEE, 2012.

[24] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 173–182. ACM, 2010.

[25] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11(1):55, 2010.

[26] H. Yu, T. Kim, J. Oh, I. Ko, and S. Kim. Refmed: relevance feedback retrieval system for pubmed. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 2099–2100, New York, NY, USA, 2009. ACM.

[27] H. Yu, T. Kim, J. Oh, I. Ko, S. Kim, and W.-S. Han. Enabling multi-level relevance feedback on pubmed by integrating rank learning into DBMS. *BMC Bioinformatics*, 11:S6, 2010.

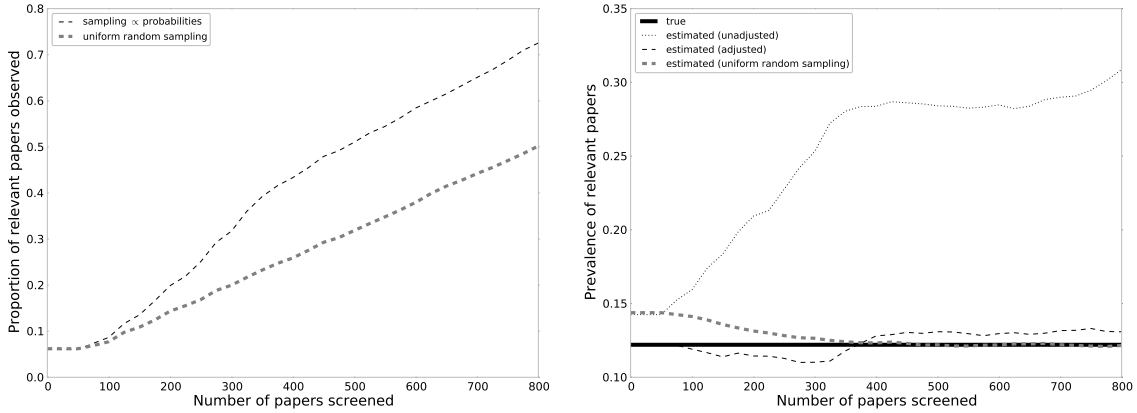

Figure 2: Results for the *COPD* [3] systematic review dataset. The x-axis in both sub-plots represents the number of citations labeled. The left-hand plot shows the proportion of relevant literature found, the right-hand plot shows the running estimate of the true prevalence of relevant articles. Results are averaged over 10 independent runs (each seeded with two randomly selected relevant and two randomly selected irrelevant articles).

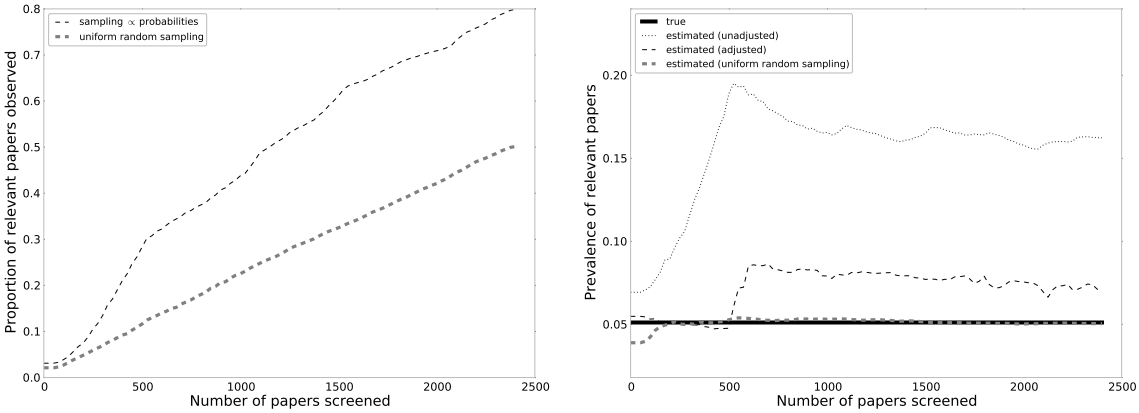

Figure 3: Results for the *proton beam* [22] systematic review dataset. Layout is similar to Figure 2.

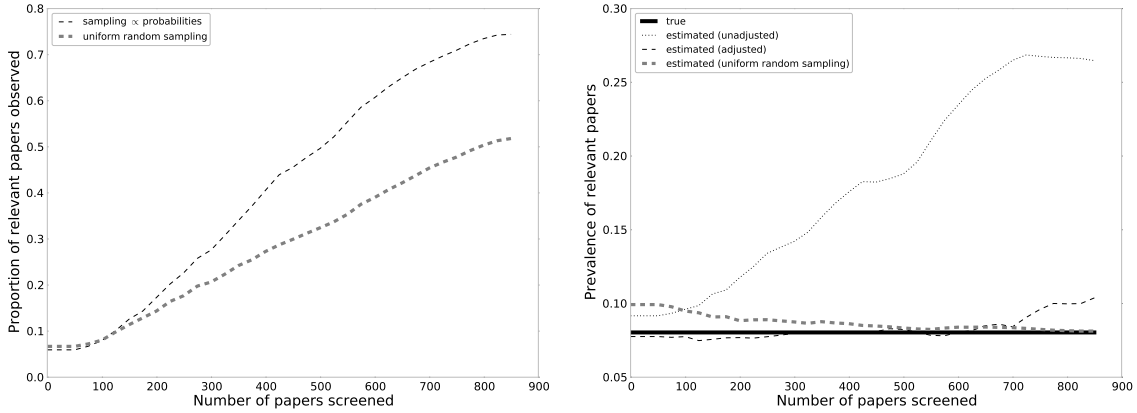

Figure 4: Results for the *Sedatives* [5] systematic review dataset. Layout is similar to Figure 2.