Inconsistent Ranking Assumptions in Medical Search and Their Downstream Consequences

Daniel Cohen
daniel_cohen@brown.edu
Brown University
Providence, RI, USA

Kevin Du
kevin_du@alumni.brown.edu
Brown University
Providence, RI, USA

Bhaskar Mitra
bmitra@microsoft.com
Microsoft
Montreal, Canada

Laura Mercurio
laura_mercurio@brown.edu
Alpert Medical School
Brown University
Providence, RI, USA

Navid Rekabsaz
navid.rekabsaz@jku.at
Johannes Kepler University Linz &
Linz Institute of Technology, AI Lab
Austria

Carsten Eickhoff
carsten@brown.edu
Brown University
Providence, RI, USA

ABSTRACT

Given a query, neural retrieval models predict point estimates of relevance for each document; however, a significant drawback of relying solely on point estimates is that they contain no indication of the model's confidence in its predictions. Despite this lack of information, downstream methods such as reranking, cutoff prediction, and none-of-the-above classification are still able to learn effective functions to accomplish their respective tasks. Unfortunately, these downstream methods can suffer poor performance when the initial ranking model loses confidence in its score predictions. This becomes increasingly important in high-stakes settings, such as medical searches that can influence health decision making.

Recent work has resolved this lack of information by introducing Bayesian uncertainty to capture the possible distribution of a document score. This paper presents the use of this uncertainty information as an indicator of how well downstream methods will function over a ranklist. We highlight a significant bias against certain disease-related queries within the posterior distribution of a neural model, and show that this bias in a model's predictive distribution propagates to downstream methods. Finally, we introduce a multi-distribution uncertainty metric, confidence decay, as a valid way of partially identifying these failure cases in an offline setting without the need of any user feedback.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; • **Social and professional topics** → *Computing / technology policy*; • **Computing methodologies** → *Neural networks.*

KEYWORDS

information retrieval, uncertainty, bias, fairness, medical search, bayesian

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531898

ACM Reference Format:
Daniel Cohen, Kevin Du, Bhaskar Mitra, Laura Mercurio, Navid Rekabsaz, and Carsten Eickhoff. 2022. Inconsistent Ranking Assumptions in Medical Search and Their Downstream Consequences. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531898

1 INTRODUCTION

In a conventional search system, a ranking model estimates the relevance of a set of documents to the user's information need. The score assigned to each document then represents the relative utility to the user, according to the probability ranking principle (PRP) [16]. This PRP is foundational to many information retrieval (IR) tasks, and is essential when the ranked documents are part of a later downstream task in order to reason over their relative importance [1–3, 11, 17]. For instance, in the case of the downstream task of cutoff prediction where the goal is to find an optimal truncation point of a ranked list to avoid wasted resources, the cutoff model is able to determine when the utility of a document has dropped to the point where the cost of showing the user a document is greater than the utility of the user seeing it. For the majority of recent work, these relevance, or utility, scores produced from a model are deterministic, so that a single query-document pair produces the same score regardless how many times it is passed through the model.

However, the PRP, and therefore subsequent downstream algorithms, assume that these scores are well calibrated with each other and that the model is fully confident in the scores assigned to query-document pairs. Although these assumptions present a solid theoretical foundation, it is unrealistic to assume that any modern neural retrieval model is fully confident or calibrated with respect to its outputs. Recent work supports such a situation by showing that document scores produced from neural retrieval models possess substantial uncertainty, which has historically been obfuscated by the common use of deterministic ranking models [4, 14]. Furthermore, the degree of this uncertainty heavily varies across documents such that not all documents are treated equally. Despite this uncertainty, downstream methods are still able to effectively reason over the utility of documents. The effectiveness of these methods suggests

that they are still able to learn and adapt as long as there exists a consistent violation of the PRP's theoretical assumptions.

Continuing with the cutoff task example, if the cutoff model learns that the query-document scores produced from the neural ranking model below x are unreliable when making comparisons between documents, the cutoff model would still be able to take advantage of the PRP as long as this behavior is consistent. Unfortunately, scores can depart from this pattern if the neural ranking model observes a rare query, and the cutoff model would not be able to effectively reason over the candidate document scores.

The consequences of this performance degradation would be even more severe when the above situation is systemically violated for search intents that correspond to specific groups, as a failure in those downstream functions may specifically marginalize certain populations. Figure 1 highlights this danger, where monitoring for a significant change in the model's predictive distribution for marginalized groups can preemptively identify these issues without the need for relevance judgments or other user feedback.

We therefore propose leveraging this assumption and posit whether the trend of uncertainty within a rank list is indicative of its downstream performance. The hypothesis is that a substantial change in the retrieval model's confidence for a query would suggest that the downstream model's utility estimates will no longer be accurate due to the change in document score consistency with respect to the PRP.

To test this hypothesis, we leverage recent work in uncertainty modeling for IR on neural models trained with real-world logs from a medical search engine [15] to examine whether it is suitable to assume that uncertainty is an actionable information source when attempting to model the utility of documents for the downstream cutoff prediction task [1, 4, 14]. In collaboration with medical experts, we identify queries specific to seven diseases, and examine their performance on this downstream task to quantify whether the degradation is a random process over queries or whether it is specific to sub-populations within a collection.

To our knowledge, this paper is the first to consider the impact of disparate uncertainty in IR and the associated assumptions over model outputs. Its core contributions include:

- A demonstrated disparity in the uncertainty of relevance score predictions between documents ranked for queries between sub-populations, even after accounting for infrequent queries.
- The feasibility of using model uncertainty over sub-populations to monitor for negative downstream impacts without the need for relevance judgements or user feedback.

2 BACKGROUND

2.1 Downstream Reliance on Relevance Scores

Under the PRP, relevance scores describe the relative utility of documents toward a query's information needs. Methods in downstream tasks, such as fairness correction and cutoff prediction, rely on the consistency and calibration of relevance scores.

Within the realm of fairness correction, Mehrotra et al. [12] develop a framework for jointly optimizing supplier fairness and relevance of recommendations to consumers, measured through retrieval model scores as a surrogate for item utility. Similarly, Biega et al. [3] view fairness through user attention, aiming to rank documents such that a document's received attention is proportional to its relevance. Both of the above studies assume the relevance scores for all documents to be equally reliable. Both Singh and Joachims [18], who developed an algorithm for fair learning-to-rank maximizing the utility of the ranked list while satisfying fair exposure constraints, and Diaz et al. [6], who introduced stochasticity in ranking via Plackett-Luce sampling, implicitly assume consistent calibration of relevance scores when making downstream decisions.

Cutoff prediction is another task which relies on model-reported relevance scores to determine when the model is no longer effective. Although the objective is different from the aforementioned work on fairness, these cutoff algorithms make the same assumptions that the reported relevance scores across documents are reliable. Culpepper et al. [5] depends on a deterministic re-ranking model as a gold standard with which to compare the quality of rankings with different cutoffs. Lien et al. [11] uses relevance scores from a ranked list as input to an LSTM model that predicts the cutoff point. As mentioned, both of these methods assume high confidence in the relevance scores produced during re-ranking. With inconsistent behavior for relevance score predictions, it is unsafe to assume the reliability of the results produced by these downstream.

2.2 Uncertainty

Prior works have measured model uncertainty in IR systems. Both Penha and Hauff [14] and Cohen et al. [4] modify BERT-based rankers with a Bayesian approximation method of stochastic dropout sampling to capture predictive relevance distributions that can be used to measure model uncertainty and subsequent ranking uncertainty. While these past works examine uncertainty over all the queries, our contribution highlights the importance of understanding how the model's posterior, and therefore its predictive distribution, changes on rare or infrequent queries.

3 QUERY PREDICTION UNCERTAINTY FOR DISEASE GROUPS

3.1 Defining Disease Groups

Informally, *group fairness* methods compare some performance metric of a model between two groups [7]. In the context of medical search, this paper considers a group as a collection of users who issue queries for a common disease, such as all users who search for information related to *Kawasaki Disease*. Unfortunately, without fine-grained data on queries searched by each population provided by the search engine's click logs, in this study we rely on examining specific diseases that are relatively common to rare conditions impacting specific genetic profiles (e.g., Ashkenazi Jewish). We note that usage of the word "groups" in this paper refers to the disease with which a query or document is categorized and any populations relevant to that disease rather than a pre-defined demographic.

3.1.1 Mapping Queries to Groups. In order to map queries to a specific disease, for each disease, we collect a list of related queries from the collection of queries by keyword matching common synonyms and terms of the disease. Given the lack of data to create such a mapping from existing data, medical doctors identified seven diseases that are loosely correlated with specific demographics and

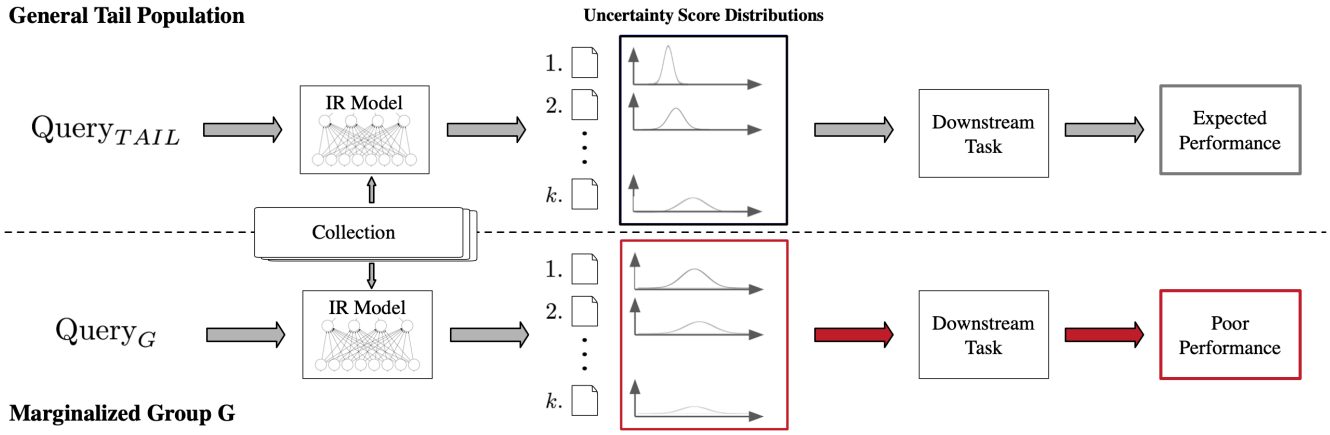

Figure 1: A visual representation of the current disparity between general population tail queries and marginalized query groups. The difference in score confidence can result in the poisoning of downstream methods that rely on certain score assumptions of the general population. By monitoring the distribution of uncertainty for certain populations, we can identify when disparate model confidence will lead to unfair downstream treatment.

provided keywords used to identify information related to these diseases. The queries were then selected from the TAIL dataset via keyword matching To compute the model's uncertainty levels for a disease, we then aggregate the top-k documents ranked by our model for each query. In our experiments, we use $k = 20$ to focus our findings on the impact uncertainty might have on the most relevant candidate documents.

3.2 Measuring Model Prediction Uncertainty

For a given query, a standard neural retrieval model produces a ranked list of documents with the highest predicted relevance scores. To capture the uncertainty within the retrieval model, we modify the model to reflect the predictive distribution rather than the conventional deterministic score to produce a relevance score distribution for each document. This distribution over the score for each query-document pair is created by considering different parameter configurations of the retrieval model weighed by how likely data supports such a parameterization. We use the recent framework of [4] that incorporates concrete Monte Carlo Dropout (MC-dropout) [8, 9], a simple method to make a network Bayesian-like, for efficient ranking.

3.2.1 Efficient MC-Dropout. In order to capture the predictive distribution, $P(y|q,d)$ for a q query and d document, we model the parameter distribution of the model, referred to as the posterior distribution:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta)P(\theta)}{\int_\theta P(\mathcal{D}|\theta)P(\theta)d\theta}, \tag{1}$$

where θ are the parameters of the retrieval model, and $\mathcal{D}$ is the provided training data. However, the *evidence*, $P(\mathcal{D})$, is intractable in all but the most simplest models. Therefore, we use an approximation of the true posterior by creating a variational distribution $q(\theta) \approx P(\theta|\mathcal{D})$ to reason about the uncertainty of the relevance scores. While there are numerous approximation methods, we utilize an efficient concrete Monte Carlo dropout [4, 9]. The method allows for standard stochastic gradient descent training with a learned dropout rate over the penultimate layers of a neural model. Once the model is trained, $q(\theta)$ is created via Monte Carlo sampling via dropout which allows for a Gaussian score distribution. This distribution of scores for a single document then captures the uncertainty of the ranking model for a given query and document pair.

3.3 Estimating and Comparing Group Uncertainty via Confidence Decay

While the above method produces a range of scores for each document, we require a measure to capture the uncertainty of sets of queries in order to quantify the model's uncertainty for a disease (hereby referred to as a "group", to maintain generality). We therefore extend the uncertainty defined in Section 3.2.1 to satisfy the group criteria by considering the set of top-k documents ranked for each query for all queries within the set defined by the group.

Formally, consider the set of all distinct queries in a dataset $Q = q_1, ..., q_n$, and the subset of queries $Q_A = \{q_{a_1}, ..., q_{a_m}\}$ that are categorized with group A (via the method described in Section 3.1.1), where $m << n$. Let $r(q,k)$ denote the set of top-k ranked documents for a query q, $D = \bigcup_{q \in Q} r(q,k)$ the set of all documents ranked in the top-k for any query, and $D_A = \bigcup_{q \in Q_A} r(q,k)$ the set of all documents ranked in the top-k for a query in group A.

In recent work, Hullermeier et al. [10] argue that directly measuring the variance of the predictive distribution is a valid measure of uncertainty for a given input. Unfortunately, this approach fails in the ranking situation as it is not a classification problem. Evaluating the predictive distribution for individual q_i, d_i pairs independent of other $d_j \in D$ can lead to incorrect conclusions. We therefore introduce an alternative metric to capture the uncertainty of D, the *decay* of a model's confidence across D_A. Specifically, we treat the task as a regression problem modeled by an exponential function that tries to predict the probability mass placed on the mean of the

distribution. In this regression, we consider the independent variable to be the mean relevance score of the predictive distribution of a document, x_μ, and the dependent variable to be the likelihood of the distribution at that relevance score, $p(x_\mu|d,q)$. The fitted curve will then follow the form $p(x_\mu|d,q) = ae^{bx_\mu}$, where a, b are fitted parameters. The shape of the curve captures the degree to which the variance, or uncertainty, across documents increases. A steep curve would indicate a rapid increase in uncertainty, whereas a flatter curve would suggest a gradual, predictable change in uncertainty as one goes down the ranklist.

4 EXPERIMENTS

4.1 Datasets

We evaluate using the TripClick dataset [15], an IR benchmark collected through click logs from a health web search engine. TripClick contains 700,000 unique free-text queries and 1.3 million pairs of query-document relevance signals.

The queries in the TripClick dataset are partitioned into HEAD, TORSO, and TAIL splits according to their search frequency. Queries issued less than 6 times are in the TAIL group, queries issued between 6 and 44 times are in the TORSO group, and queries issued more than 44 times are in the HEAD group. In this dataset, queries for our diseases of interest overwhelmingly fall in the TAIL split because many of our diseases of interest are relatively rare and appear primarily in their corresponding marginalized population group (e.g., Niemann Pick Disease or sarcoidosis). As comparing model confidence between head and tail queries would be trivial, we only consider the TAIL query set within TripClick for our experiments to better capture whether there exist disparities across already rare queries.

4.2 Measuring Downstream Task Sensitivity

Given the lack of publicly available clinical decision support datasets that include IR, we determine the impact of the above uncertainty quantification by examining it as a performance indicator of cutoff prediction [1] as a surrogate. While this setting has arguably negligible repercussions for users in current IR environments, it represents the susceptibility of neural models to changes in the underlying predictive distribution.

Using the reported relevance judgments from the TripClick dataset [15], we train the Choppy transformer architecture [1] on the general TAIL population using an 80-10-10 train, validation, and test split to best perform on infrequent queries. We then evaluate the performance of the trained cutoff model on individual disease groups in addition to the general TAIL population to determine whether the observed difference in uncertainty significantly impacts these downstream tasks and to examine how well our proposed curvature metric is indicative of downstream performance.

4.3 Characterizing Uncertainty Differences

To establish a significant difference in uncertainty between a disease and the general TAIL population as measured by the curvature metric described in Section 3.3, we compare whether the parameters of the exponential curve fit for a disease differ significantly from that of the general TAIL population. We add an interaction term i to our data and regression, $y = (a + s_a \cdot i)e^{(b+s_b \cdot i)x}$ where i is a binary indicator of whether a data point belongs to a disease or the general TAIL population, and a, b, s_a, s_b are the fitted parameters. We can then test for whether s_a and s_b are 0. Should we have sufficient evidence that either s_a and s_b is not 0, then we can conclude that the fit of the exponential curve for the disease differs from that for the general population.

Table 1: Cutoff performance of Choppy [1] as a percentage of oracle performance using the F_1 metric * indicates significance with respect to the TAIL Total Population using a two tailed t-test with $p < 0.05$.

Query Group	Cutoff Score
TAIL Total Population	50.2%
Asthma	46.9%
HIV	49.8%
HbSS	44.3 %*
Kawasaki	0%*
Niemann Pick Disease	0%*
Sarcoidosis	61.7%
Tay Sachs	0%*

4.4 Retrieval Models

We use a BERT-based retrieval architecture [13] and append two concrete MC-dropout layers to the end of the model in a fashion similar to Cohen et al. [4]. We run all experiments using tiny-BERT; we ensure tiny-BERT achieves comparable performance to benchmarks as it outperforms all non-transformer architectures in Rekabsaz et al. [15].

5 RESULTS AND DISCUSSION

In this section, we first examine whether there exists any real downstream discrepancy between disease groups as well as general infrequent queries within the TAIL dataset. After establishing these results, we then discuss whether the proposed confidence decay metric provides any grounds to indicate downstream performance impact.

5.1 Downstream Impacts

As we propose that confidence decay is indicative of downstream performance, we must first establish that there exists a substantial discrepancy in the cutoff prediction task both with respect to general infrequent tail queries and across disease groups, in accordance with their confidence decay measure. As indicated in Table 1, the cutoff model trained on the general tail population is able to achieve satisfactory performance. However, certain disease groups significantly suffer in performance, particularly Kawasaki disease, Niemann Pick disease, and Tay Sachs disease. While the other diseases achieve lower performance compared to the general TAIL population, it is not nearly to the degree to as the above mentioned groups. We note the performance of the sarcoidosis disease group in that it outperforms the TAIL baseline, indicating that the degradation in downstream performance is not solely due to the low number of queries within the dataset (n in Table 2). With this performance degradation noted, we now discuss our main hypothesis: whether this downstream performance is indicated by the confidence decay of the disease groups.

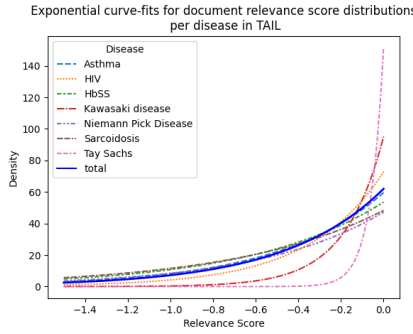

Figure 2: A visual comparison of uncertainty (as measured by curvature) for documents corresponding to a disease.

5.2 Group Confidence Decay

Table 2: Summary statistics of the parameters (a, b) of an exponentially fitted curve, $p(x) = ae^{bx}$, for queries in the TAIL dataset split. n indicates the number of documents ranked by queries categorized with a given disease. * indicates a significant difference in the fitted parameter value between a particular disease and the general TAIL total population, as determined by a two-tailed t-test with $p < 0.05$ (after applying the Bonferroni correction).

Disease	n	mean	std
Total population	88,110	(62.024, 2.154)	(0.1494, 0.009)
Tay Sachs	**60**	**(151.392*, 16.366*)**	**(13.177, 1.949)**
Niemann Pick	**100**	**(47.240*, 1.719*)**	**(3.922, 0.303)**
HbSS	**3,120**	**(53.601*, 1.607*)**	**(0.835, 0.042)**
Kawasaki	**60**	**(95.086*, 5.461*)**	**(5.763, 0.571)**
HIV	**16,520**	**(72.828*, 2.848*)**	**(0.461, 0.027)**
Sarcoidosis	**280**	**(48.306*, 1.434*)**	**(20.720, 0.134)**
Asthma	**28,840**	**(59.616*, 1.985*)**	**(0.230, 0.014)**

We report the results of comparing confidence decay metrics between different groups and the overall population in Table 2 to contextualize the cutoff performance results.

While the exponential parameters are significantly different, it is less evident in Figure 2 when visually comparing the "total" curve to the curves for different diseases that the curvature differs between the documents for the general TAIL distribution and documents for our diseases of interest. However, we note that the two disease groups, Tay Sachs and Kawasaki, with the largest curvature parameters (b in ae^{bx}) indicating the fastest confidence decay also achieve the worst downstream task performance of 0%. This suggests that the significance of the parameters alone in Table 2 is not enough to capture potential degradation in downstream performance; one can identify disenfranchised groups through a curvature parameter 2.5-8 times the magnitude of the mean TAIL query populations curvature parameter.

Lastly, we address the issue of the Niemann Pick Disease group's performance. As its curve in Figure 2 closely resembles the curve of the general TAIL group, it suggests that there is additional information relevant to determining downstream performance that is not contained within the confidence decay curve.

6 CONCLUSION

In this paper, we introduce the concept of measuring *confidence decay* to capture how well certain query groups will perform on downstream tasks without the need for user feedback, making the approach of uncertainty modeling a viable option when direct feedback is infeasible or impossible. While this work focused entirely on disease populations, the prevalence of genetic depositions to certain diseases represents promising future work, such as identifying (1) at-risk groups (2) the cause of this disparity, and (3) how to mitigate unfair treatment of marginalized groups.

ACKNOWLEDGMENTS

This research is supported in part by the NSF (IIS-1956221), the State of Upper Austria and the Austria's Federal Ministry of Education, Science, and Research (LIT-2021-YOU-215). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, the U.S. Government, or the Austrian Government.

REFERENCES

[1] Dara Bahri, Yi Tay, Che Zheng, Donald Metzler, and Andrew Tomkins. 2020. *Choppy: Cut Transformer for Ranked List Truncation.* ACM, New York, NY, USA, 1513–1516. https://doi.org/10.1145/3397271.3401188

[2] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, Anchorage AK USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[3] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. *The 41st International ACM SIGIR* (June 2018), 405–414. https://doi.org/10.1145/3209978.3210063 arXiv: 1805.01788.

[4] Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. 2021. Not All Relevance Scores are Equal: Efficient Uncertainty and Calibration Modeling for Deep Retrieval Models. *arXiv:2105.04651 [cs]* (May 2021). https://doi.org/10.1145/3404835.3462951 arXiv: 2105.04651.

[5] J. Shane Culpepper, Charles L. A. Clarke, and Jimmy Lin. 2016. Dynamic Cutoff Prediction in Multi-Stage Retrieval Systems. In *Proceedings of the 21st ADCS*. ACM, Caulfield VIC Australia, 17–24. https://doi.org/10.1145/3015022.3015026

[6] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. *Proceedings of the 29th ACM International CIKM* (Oct. 2020), 275–284. https://doi.org/10.1145/3340531.3411962 arXiv: 2004.13157.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd ITCS* (Cambridge, Massachusetts) *(ITCS '12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[8] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142 [cs, stat]* (Oct. 2016). http://arxiv.org/abs/1506.02142 arXiv: 1506.02142.

[9] Yarin Gal, Jiri Hron, and Alex Kendall. 2017. Concrete Dropout. *arXiv:1705.07832 [stat]* (May 2017). http://arxiv.org/abs/1705.07832 arXiv: 1705.07832.

[10] Eyke Hüllermeier and Willem Waegeman. 2020. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *arXiv:1910.09457 [cs, stat]* (Sept. 2020). http://arxiv.org/abs/1910.09457 arXiv: 1910.09457.

[11] Yen-Chieh Lien, Daniel Cohen, and W. Bruce Croft. 2019. An Assumption-Free Approach to the Dynamic Truncation of Ranked Lists. In *Proceedings of the 2019 ACM SIGIR*. ACM, Santa Clara CA USA, 79–82. https://doi.org/10.1145/3341981.3344234

[12] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation

of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International CIKM*. ACM, Torino Italy, 2243–2251. https://doi.org/10.1145/3269206.3272027

[13] Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. *arXiv:1901.04085 [cs]* (April 2020). http://arxiv.org/abs/1901.04085 arXiv: 1901.04085.

[14] Gustavo Penha and Claudia Hauff. 2021. On the Calibration and Uncertainty of Neural Learning to Rank Models. *arXiv:2101.04356 [cs]* (Jan. 2021). http://arxiv.org/abs/2101.04356 arXiv: 2101.04356.

[15] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. TripClick: The Log Files of a Large Health Web Search Engine. In *In Proceedings of the 44th International ACM SIGIR (SIGIR'21), July 11–15, 2021, Virtual Event, Canada*. ACM. https://doi.org/10.1145/3404835.3463242

[16] S.E. Robertson. 1977. The Probability Ranking Principle in IR. *Journal of Documentation* 33, 4 (Jan. 1977), 294–304. https://doi.org/10.1108/eb026647 Publisher: MCB UP Ltd.

[17] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD*. ACM, London United Kingdom, 2219–2228. https://doi.org/10.1145/3219819.3220088

[18] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. arXiv:1902.04056 [cs.LG]