Boosting collective intelligence in medical diagnostics: Leveraging decision similarity as a predictor of accuracy when answers are open-ended rankings

Nikolas Zöller
Max-Planck Institute for Human Development
Berlin, Germany
zoeller@mpib-berlin.mpg.de

Stefan M. Herzog
Max-Planck Institute for Human Development
Berlin, Germany
herzog@mpib-berlin.mpg.de

Ralf H.J.M. Kurvers
Max-Planck Institute for Human Development
Berlin, Germany
kurvers@mpib-berlin.mpg.de

ABSTRACT

Leveraging collective intelligence and combining several decisions into a single one can outperform individual judgments in many domains. Additionally, distinguishing between high-performing and low-performing individuals can further boost accuracy and is especially important in high-stake contexts. In binary decision problems it has been shown that decision similarity to others is a predictor of accuracy and can be used to identify high performers even if no actual track record of performance is available. Here we apply and generalize this approach to open-ended medical diagnostics where diagnoses are given in the form of free-text, and incomplete rankings of varying length. We show that selecting decision makers based on prior decision similarity to others increases the average accuracy of both individual and collective diagnoses.

CCS CONCEPTS

• **Information systems → Similarity measures**; **Rank aggregation**; • **Applied computing → Health informatics**; • **Human-centered computing → Collaborative and social computing**.

KEYWORDS

Collective intelligence, medical diagnostics, rank aggregation, decision similarity

1 INTRODUCTION

Across a range of domains it has been shown that aggregating the independent judgments of multiple decision makers can substantially boost decision accuracy. In medical decision making this has been found both in binary decision making and open-ended diagnostic tasks [4, 6]. Another way to increase the quality of a diagnosis is to select only the most competent physicians to be part of the decision-making process. One straight forward way for doing this is on the basis of past performance [2, 7]. However, this information might not always be available, for example because the accuracy of a given diagnosis is only revealed after a longer time period or because past performance of medical professionals might not be recorded or available due to privacy regulation. Therefore, an alternative is to rely on proxies of accuracy instead to identify high-performing individuals.

It has recently been shown in the context of binary decision making that decision similarity can act as a predictor of accuracy, provided that average individual accuracy surpasses chance [5]. Here we investigate the potential of leveraging decision similarity in the domain of open-ended medical diagnostics where the structure of a given diagnosis is significantly more complicated than in binary choice problems. In the following we first describe the dataset that provides the empirical basis for our study. We then define the metrics for performance and similarity between given diagnoses. Next, we test whether similarity predicts performance and how it can be leveraged to increase diagnostic accuracy. Our results show that decision similarity can be used to identify high-performing individuals and groups; leading to an increase in performance of both individuals and groups.

2 DATA & METHOD

The empirical basis for this work is a large dataset from the Human Diagnosis Project (HDX), previously analyzed in [6]. HDX is an online collaborative platform for medical professionals which allows medical experts to submit and solve patient cases. These cases are reviewed by an expert panel and published only if they meet certain quality criteria. They consist of patient information including symptoms, medical records and clinical findings. Users from around the world can register on the platform, review case details and provide diagnoses. Users can enter diagnoses either as free text or select from a medical term catalog, and can enter one or several ordered diagnoses.

For our analyses, we use 1,333 medical cases each of which was solved by 10 users. There was substantial variation in the number of solves per user, ranging from users solving several hundred cases while others solved only a single case. In order to make the (open-ended) diagnoses of users comparable and identifiable we follow the preprocessing NLP steps described in [6]. This allows us to map free-text diagnoses to concepts (and their unique ids) in the *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT) [3]. SNOMED CT is a comprehensive clinical terminology and coding system to standardize the representation of medical concepts and support accurate communication of clinical information in healthcare.

To quantify the individual accuracy of users, we use the mean reciprocal rank [8] which is a well-established performance metric in the field of information retrieval

$$\mathrm{MRR} = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{1}{\mathrm{r}_i}, \tag{1}$$

where C corresponds to the set of cases the metric is evaluated on and r_i is the rank of the correct answer in the given diagnosis for case i.

To calculate the similarity between two rankings L with length l and S with length s, $l \geq s$, we use the extrapolated rank bias overlap [9]:

$$RBO_{ext}(L,S,p) = \frac{1-p}{p}\left(\sum_{d=1}^{l}\frac{X_d}{d}p^d + \sum_{d=s+1}^{l}\frac{X_s(d-s)}{sd}\right) + \left(\frac{X_l - X_s}{l} + \frac{X_s}{s}\right)p^l \quad (2)$$

where $X_d = |L_{:d} \cap S_{:d}|$ is the size of the overlap of the rankings L and S up to depths d, and p is a hyper-parameter that determines to which extend higher ranked items have more weight in the similarity score. We set $p = 0.5$ (main results were similar for the range $0.1 < p < 0.8$). RBO_{ext} has several properties that make it suitable for our problem: it is top-weighted, it handles non-conjoint and incomplete rankings, it can be applied to rankings of different lengths and it is bounded to values between 0 and 1.

3 RESULTS

We start by testing the correlation between decision similarity and accuracy. We only included users who solved more than 5 cases in order to have a robust estimate of a user's decision similarity and performance in terms of diagnostic accuracy. First, we calculated the similarity for each given diagnosis by a user (i.e., a single diagnosis or ranked list of diagnoses), by comparing it to all other diagnoses given to the same case using equation 2. To get an overall estimate of a user's decision similarity we averaged over all cases a user had solved. Next, we calculated each user's MRR. Figure 1 shows that an individual's decision similarity strongly correlated with their accuracy (Pearson correlation $r_p = 0.68$, p-value< 0.001). This suggests that we can identify high-performing individuals via their record of past decision similarity, even if no information about the correct diagnosis or their actual diagnostic accuracy is known.

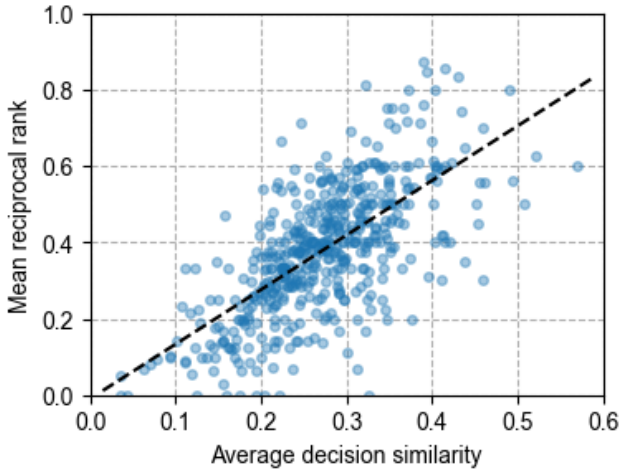

Figure 1: An individual's decision similarity to others strongly correlates with their accuracy (i.e., MRR; Pearson correlation $r_p = 0.68$, p-value< 0.001) .

Next, we studied if we can use this insight to produce more accuracy crowds. It has been shown in [6] that aggregating diagnoses of several users into a collective diagnosis improves accuracy and this effect increases with increasing group size. Here we test whether we can leverage decision similarity to select high-performing individuals and sub-groups to increase diagnostic accuracy even further; or since the time of medical professionals is valuable and cost-intensive whether we could reach a certain level of accuracy with fewer diagnosticians in the group. To this end, we employed a leave-one-out analysis: for each case we calculated the average decision similarity of a solver on all other cases that solver had solved (i.e., not including the focal case). In a similar way, to mimic past performance, we calculated the prior MRR of a solver for a particular case on all other cases that user had solved. For users that had solved fewer than 5 other cases, we assumed the median of all other solvers instead, since these were too few data points for a reliable estimate.

To simulate groups and determine their collective diagnosis we followed the procedures described in [6]. To form a collective diagnosis we first scored diseases according to their ranks in the individual diagnoses using a $1/rank$ scoring rule. We then aggregated the scores across all individuals in the group and sorted according to the score, resulting in a collective ranking on the group level. We compared three ways of selecting crowd members: 1) by random selection as done in [6], 2) by prior performance, i.e. selecting the users with the largest MRR, and 3) by selecting users with the highest decision similarity to others. Figure 2 shows how accuracy increased with group size for each of the three selection procedures. Selection by prior performance via MRR increased the diagnostic accuracy significantly both for individuals and subgroups. Interestingly, selection by prior decision similarity to others also increased the diagnostic accuracy significantly and almost to the same degree as selection via prior performance.

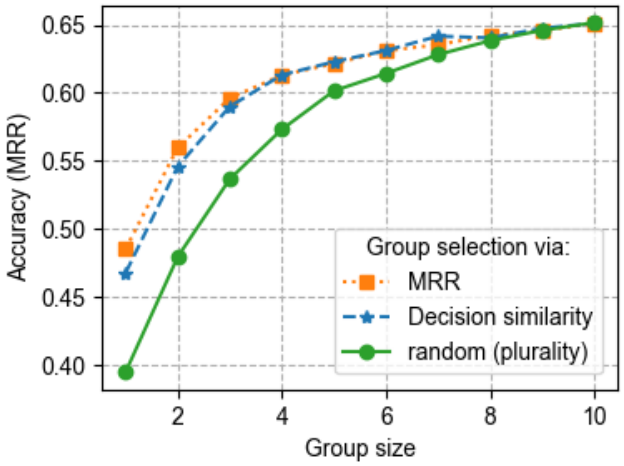

Figure 2: Accuracy improves as the size of the group contributing to the collective diagnosis increases. Accuracy also increases significantly if group members are selected based on prior performance or prior decision similarity.

Finally, we investigated how case difficulty affected potential accuracy improvements when aggregating diagnoses via the different selection mechanisms. To determine the difficulty of a case, we calculated the average individual accuracy for that case (i.e., based on the reciprocal rank of the correct diagnosis of 10 individual solves). Figure 3 shows how accuracy improvements (i.e., collective minus average individual accuracy) changed with case difficulty for a simulated group size of 3. The greatest accuracy improvements are observed for moderately difficult cases. Very hard cases gain little from aggregation. Likewise, very easy cases gain little, which is most likely a ceiling effect because of the very high average individual accuracy for these cases. Comparing the different selection rules, we observe that selections based on prior performance and decision similarity amplify the overall tendencies of the plurality rule. Interestingly, there is no region of case difficulty for which the aggregation is expected to (on average) lead to lower collective performance, as is typically found in binary decision making where aggregation leads to poorer performance when average individual accuracy drops below 0.5 (so-called wicked environments).

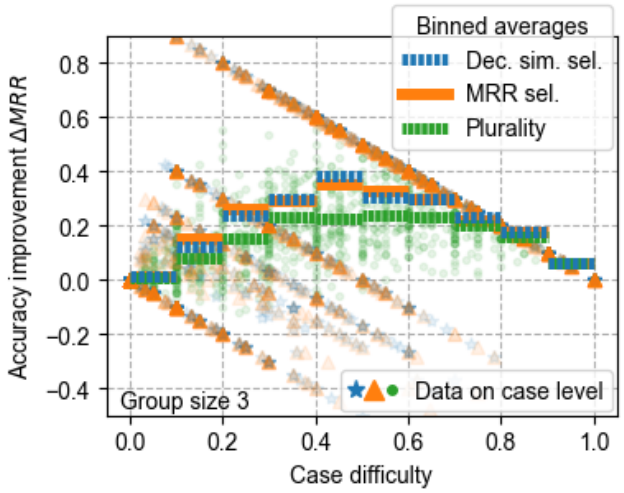

Figure 3: The relationship between case difficulty (i.e., average individual accuracy of a case) and the improvement of a crowd of size three as compared to individual accuracy, for different selection rules.

4 DISCUSSION

In this study we found a strong correlation between decision similarity to others and diagnostic accuracy in the domain of medical decision making where diagnoses where given in the form of open-ended rankings. Building on this, we utilized decision similarity as a tool to identify and select high-performing individuals and groups. This approach significantly improved individual and collective diagnostic accuracy compared to randomly-selected individuals and groups. Accuracy improvements were most pronounced for cases of intermediate difficulty. Importantly, there was no region of case difficulty for which the collective aggregation led on average to lower diagnostic accuracy, which is a known pattern in difficult binary problems (also called wicked cases) where the majority is wrong.

In real-world settings increasing diagnostic accuracy can save lives. Both, a collective intelligence approach, but also the selection of physicians based on past decision similarity show great potential in this regard. The time of medical professionals is, however, valuable, scarce and costly. While collective intelligence approaches tie up the resources of several professionals, selecting practitioners based on prior decision similarity can lead to a a more optimal allocation of professionals to patient cases. This seems particularly important for the increasingly relevant practice of online medical consultation via digital services and apps, a development that was accelerated through the challenge of social distancing which emerged during the Covid-19 pandemic [1].

One limitation of this study is the fact that cases were curated by HDX and as such might not be representative of real-world scenarios. Furthermore, the number of cases solved by individual users varied significantly, so that the number of data points underlying estimates for decision similarity and accuracy on the user level also varied. However, this variation in the data can also be regarded as a feature of the dataset and it is a promising result that the selection based on users' decision similarity yielded significant accuracy increases regardless. In future work we plan to employ a methodological approach based on hierarchical Bayesian models to more accurately account for the variation in the number of data points on the user level.

REFERENCES

[1] Bokolo Anthony Jnr. 2021. Implications of telehealth and digital care solutions during COVID-19 pandemic: a qualitative literature review. *Informatics for Health and Social Care* 46, 1 (2021), 68–83.
[2] Mark A Burgman. 2016. *Trusting judgements: how to get the best out of experts.* Cambridge University Press.
[3] Kevin Donnelly et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.
[4] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Patricia A Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. 2016. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences* 113, 31 (2016), 8777–8782.
[5] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Mehdi Moussaid, Giuseppe Argenziano, Iris Zalaudek, Patty A Carney, and Max Wolf. 2019. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science advances* 5, 11 (2019), eaaw9011.
[6] Ralf H. J. M. Kurvers, Andrea Giovanni Nuzzolese, Alessandro Russo, Gioele Barabucci, Stefan M. Herzog, and Vito Trianni. 2023. Automating hybrid collective intelligence in open-ended medical diagnostics. *Proceedings of the National Academy of Sciences* 120, 34 (2023), e2221473120. https://doi.org/10.1073/pnas.2221473120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2221473120
[7] Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science* 25, 5 (2014), 1106–1115.
[8] Ellen M Voorhees et al. 1999. The trec-8 question answering track report.. In *Trec*, Vol. 99. 77–82.
[9] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.