HACID - Deliverable

Aggregation methods for collective solutions

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101070588. UK Research and Innovation (UKRI) funds the Nesta and Met Office contributions to the HACID project.

Deliverable number:	D4.1
Due date:	31.01.2025
Nature[1]:	R
Dissemination Level[2]:	PU
Work Package:	WP4
Lead Beneficiary:	MPG
Contributing Beneficiaries:	CNR

[1] The following codes are admitted:
- R: Document, report (excluding the periodic and final reports)
- DEM: Demonstrator, pilot, prototype, plan designs
- DEC: Websites, patents filing, press & media actions, videos, etc.
- DATA: Data sets, microdata, etc.
- DMP: Data management plan
- ETHICS: Deliverables related to ethics issues.
- SECURITY: Deliverables related to security issues
- OTHER: Software, technical diagram, algorithms, models, etc.

[2] The following codes are admitted:
- PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)
- SEN – Sensitive, limited under the conditions of the Grant Agreement
- Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444
- Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444
- Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Document History

Version	Date	Description	Author	Partner
V1.0	17/02/25	First draft	N. Zöller	MPG
V1.1	18/02/25	Revision	S. M. Herzog	MPG
V1.2	03/03/25	Revision	N. Zöller	MPG
V1.3	03/04/25	Revision	C. D'Onofrio	CNR
V1.4	03/24/25	Revision	V. Trianni	CNR
V1.5	08/05/25	Revision	C. D'Onofrio	CNR
V1.6	30/05/25	Final Revision	V. Trianni	CNR

Table of content

1. Introduction

The ability to aggregate individual expert solutions into collective decisions is critical in fields where decision-making is complex and uncertain. Work Package 4 (WP4) focuses on developing and evaluating methods for aggregating expert solutions for open-ended problems, with a focus on the integration of user metadata and domain knowledge in order to harness collective intelligence in situations where simple methods for closed problems (e.g., plurality voting) is not applicable. This report documents the aggregation methods explored to improve collective accuracy in open-ended problems.

The results presented in this document are based on the medical diagnostics use case (WP6) because here we had data available to develop and test methods. For the climate services use case, data collection is still in the planning phase but a discussion on whether and how the findings can be transferred to the climate services use case is given in the outlook.

In order to avoid duplication of efforts some of the following descriptions are in part reproduced from a preprint we published as part of the HACID project.[3] In the following we will proceed with a brief description of the data that the analyses in this report are based on.

1.1 Data

The empirical basis for this work is data from the Human Diagnosis Project (Human Dx), an online collaborative platform for medical professionals and trainees. Users from around the world can register on the platform, submit cases, review case details, and provide diagnoses. The cases submitted are published only if approved by an editorial board of licensed medical professionals. Each case is presented as a vignette mimicking information that physicians encounter in real-world practice and containing patient information such as symptoms, medical records, and clinical test results (see Figure 1A). When responding to a case, users can provide either a single diagnosis or a ranked list, commonly known as *differential diagnosis*, either as free text or by selecting from a medical taxonomy with an auto-complete feature that activates as they type. We refer to this response as a differential diagnosis, whether it contains one or multiple diagnoses. Once the users have submitted their differential diagnoses, they are shown the solution as provided by cases' authors and vetted by an expert panel, which may consist of one or several diagnoses. For our main analyses, we used a set of 2,133 medical cases and 40,762 differential diagnoses from 3,669 qualified physicians with different levels of professional experience. An additional 11,772 diagnoses were contributed by 1,037 medical students. For the analysis of diagnostic confidence, we are in the process of conducting a separate experiment for which we have currently collected 828 responses for 30 case vignettes.

[3] Zöller et al., Human-AI collectives produce the most accurate differential diagnoses, (2024), https://doi.org/10.48550/arXiv.2406.14981

1.2 Mapping open-ended diagnoses to SNOMED CT

A key challenge in aggregating open-ended medical diagnoses is recognizing when different textual diagnoses refer to the same medical concept. Differential diagnoses by medical experts often vary due to typos, synonyms, or spelling differences. To address this, we developed a processing pipeline that maps raw text responses to unique identifiers in the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT).[4] SNOMED CT is a comprehensive clinical terminology and coding system designed to standardize the representation of medical concepts and support the accurate communication of clinical information in healthcare. Our pipeline consists of the following steps:

1. **String Normalization**: Diagnoses—including correct ones from case authors—are standardized using the Norm 1 pipeline from the National Library of Medicine.[5] This process removes stop words, converts British to US English, singularizes plurals, and resolves acronyms.
2. **Concept Mapping**: Normalized diagnosis strings are compared to SNOMED CT entries (including synonyms), assigning a SNOMED CT ID (SCTID) only when a Jaccard similarity of 1 is achieved. If multiple SCTIDs match, the preferred one is selected based on predefined semantic tag priorities.

This approach matched 90% of correct case diagnoses, and 84% of diagnoses by medical experts. Unmatched cases were resolved using a sentence-transformer model[6] (based on PubMedBERT), which embeds SNOMED CT concepts and diagnosis strings into a 768-dimensional vector space. The unmatched diagnosis is then assigned the SNOMED CT concept with the highest cosine similarity. This method successfully mapped all remaining diagnoses. A sanity check confirmed that for 99.4% of previously matched diagnoses (see steps 1 and 2 above), both methods arrived at the same SCTID. For more details on the matching process see the articles by Kurvers et al.[7] and Zoeller et al.[3] that were published within this research project. The matching process is depicted for an example case in Figure 1B-D. Once all string diagnoses have been mapped to SCTIDs, they serve as a basis for the aggregation process of individual differential diagnoses into a collective one (Figure 1E).

---

[4] Donnelly, K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, *121*, 279–290. (2006).

[5] https://lhncbc.nlm.nih.gov/LSG/Projects/lvg/current/docs/userDoc/tools/norm.html

[6] Deka, P., Jurek-Loughrey, A., Deepak, P.: Improved methods to aid unsupervised evidence-based fact checking for online health news. Journal of Data Intelligence 3(4), 474–504 (2022) https://doi.org/10.26421/JDI3.4-5

[7] R.H.J.M. Kurvers, A.G. Nuzzolese, A. Russo, G. Barabucci, S.M. Herzog, & V. Trianni, Automating hybrid collective intelligence in open-ended medical diagnostics, Proc. Natl. Acad. Sci. U.S.A. 120 (34) e2221473120, https://doi.org/10.1073/pnas.2221473120 (2023).

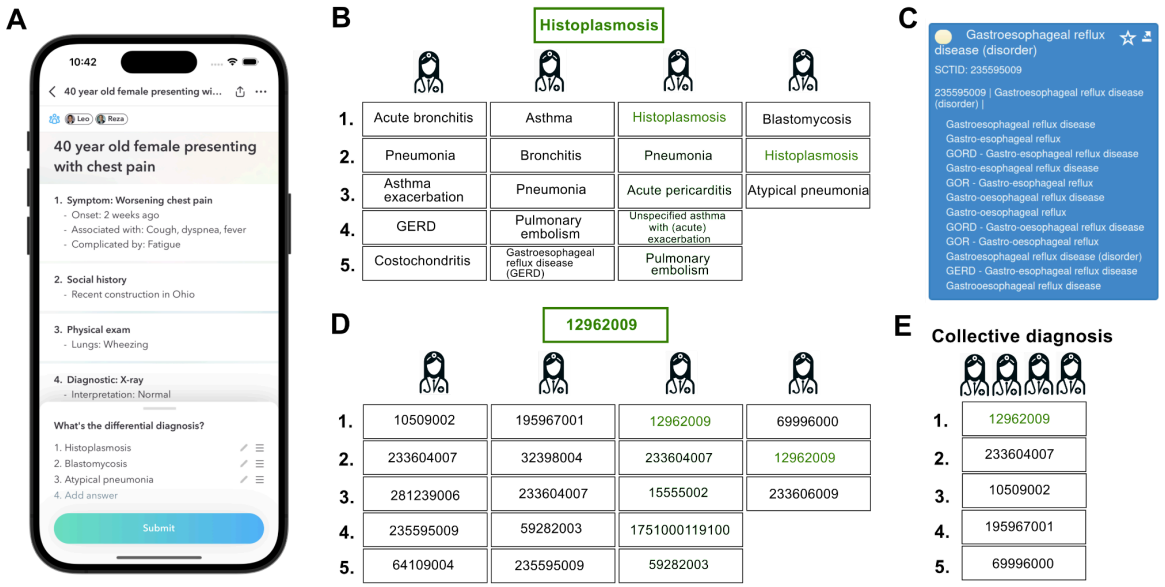

Figure 1. Illustration of the collective intelligence process, which aggregates individual diagnoses into a collective differential diagnosis. **A**: Screenshot of the Human Dx interface. **B**: Example of open-ended diagnoses from human experts. **C**: Example of a SNOMED CT entry, illustrating how synonyms map to a single SCTID. **D:** Diagnoses after matching to SCTIDs. **E**: The final collective diagnosis.

2. Aggregation of individual differential diagnoses into a collective solution

For open-ended questions the most simple and intuitive way to combine individual answers into a single collective response is to pick the most frequently proposed solution—a procedure commonly referred to as a “plurality voting”. However, this approach can be suboptimal (e.g. because of high variance in group member’s expertise). In this report, we explore two primary avenues for improving collective outcomes beyond simple vote aggregation:

1. Selecting the most competent contributors (or otherwise filtering participants) before aggregating their responses.
2. Weighting proposed solutions based on various factors, such as a participant’s competence (or proxies thereof), the position of each solution in a participant’s list, or its similarity to other proposed solutions.

2.1 Definition

To put this on more formal footing, let $U = \{u_1, u_2, ..., u_m\}$ be the set of users in a group of size m and let $C = \{c_1, c_2, ..., c_n\}$ be the set of nominated concepts by the users of that group. Approach (1) corresponds to a selection procedure $S: ▯ \rightarrow U$, which determines the composition of the group from a larger candidate pool ▯. With regard to approach (2), we

define a general equation for the score s_j of each nominated concept c_j to facilitate the ranking of the collective solution:

$$s_j = s(c_j) = \sum_{i=1}^{n} \sum_{u=1}^{m} \sum_{r=1}^{R} w_u \cdot rankscore(r) \cdot 1(\rho^{(u)}(c_i) = r) \cdot sim(c_i, c_j) \quad (1)$$

where

- $\rho^{(u)}(c_i) = r$ indicates that user u ranked concept c_i at position r,
- $1(\rho^{(u)}(c_i) = r)$ is an indicator function that equals 1 if true and 0 otherwise,
- w_u is the weight assigned to user u,
- $rankscore(r)$ is the weight function for the rank,
- $sim(c_i, c_j)$ is a similarity function between different nominated concepts,
- R is the maximum rank of the proposed differentials.

A collective solution in form of a ranking can then be constructed through $R_U = argsort_{c \in C} s(c)$.

As an alternative to directly using these similarities to adjust the relevance scores s_j, they can be used to build a graph where each nominated concept is represented as a node. To identify the most important nodes in this graph, we leverage the PageRank[8] algorithm, originally developed to rank web pages but now widely applied in various network contexts. We specifically use a modified version,[9] commonly known as *personalized* PageRank, which includes a personalization vector to favor selected nodes. In our case, we initialize this vector with the baseline scores $s_j = \sum_{u=1}^{m} \sum_{r=1}^{R} 1/r \cdot 1(\rho^{(u)}(c_j) = r)$, then allow the relevance scores to propagate through the similarity network. The damping parameter α, which controls the strength of this propagation, is set to $\alpha = 0.4$. Note that $\alpha = 0$ leads to the baseline result with no diffusion along the similarity network, while $\alpha = 1$ corresponds to eigenvector centrality of the similarity matrix with no influence of the frequency with which concepts were nominated. For implementation, we use the NetworkX[10] Python library.

In the remainder of this report we analyse for approach (1) different selection procedures $S: ? \rightarrow U$ based on past performance, decision similarity to others and metadata such as response times and diagnostic confidence, and compare the results to the baseline of random users selection. For approach (2), we compare different $rankscore$ scoring functions, test weights on the user level (e.g. based on confidence and past performance) and leverage

[8] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford infolab.

[9] Langville, A. N., & Meyer, C. D. (2005). A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Review*, *47*(1), 135–161. https://doi.org/10.1137/S0036144503424786.

[10] https://networkx.org, see also Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

relations in the domain knowledge graph to build similarities $sim(c_i, c_j)$ between nominated concepts.

2.2 Evaluation

The choice of performance metrics for an aggregated collective diagnosis depends on its intended use case. For instance, if the differential diagnosis generated by a collective is meant to serve as a consideration set to support a physician's decision-making, it may be sufficient for the correct diagnosis to be included in the list rather than necessarily ranked first. To account for this, we report multiple accuracy metrics including *top-5*, *top-3*, and *top-1* accuracy. A differential diagnosis is considered correct under these metrics if the correct diagnosis appears within the top five, top three, or top one ranked diagnoses, respectively. Accuracy is then defined as the proportion of cases in which this condition holds. Additionally, we report the *Mean Reciprocal Rank (MRR)*, a well-established performance metric in information retrieval, which captures how highly ranked the correct diagnosis tends to be on average. It is defined as $MRR = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{r_i}$, where C corresponds to the number of cases on which the metric is evaluated and r_i is the rank of the correct answer in the final list for case i (with $r_i = \infty$ if the correct answer is not present in the list).

2.3 Baseline aggregation and the effect of rank penalty

To compare different scoring functions based on the rank of items we set the similarity weight in equation (1) to the Kronecker delta: $sim(c_i, c_j) = \delta_{i,j}$. This means that a concept's score increases only when it is explicitly nominated by a user, rather than when a similar concept is nominated (for a discussion of how similarity between concepts can be quantified and leveraged see Section 4). Additionally, we set $w_u = 1$ giving all users equal weight. Equation (1) then simplifies to

$$s_j = \sum_{u=1}^{m}\sum_{r=1}^{R} rankscore(r) \cdot 1(\rho^{(u)}(c_j) = r).$$

Setting $rankscore(r) = 1$ corresponds to simple plurality voting where all nominated concepts contribute equally to the collective solution regardless of their rank in the differential diagnosis. We also test rank-biased scoring rules of $rankscore(r) = 1/r$ and $rankscore(r) = 1/r^2$ which introduce varying degrees of rank penalty.

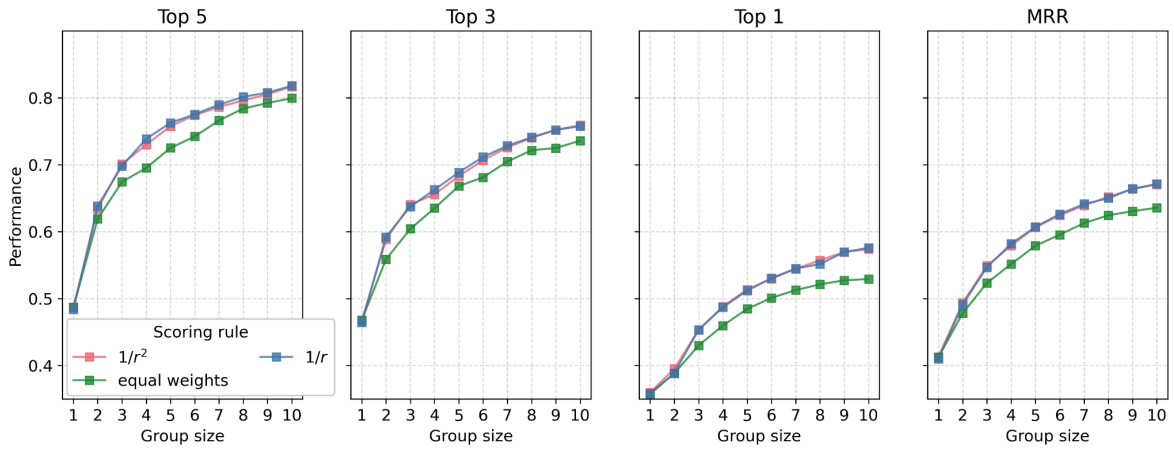

Figure 2. Performance of different scoring rules.

As shown in Figure 2, the collective intelligence effect is clearly visible for all scoring rules where a larger group size leads to higher diagnostic performance compared to individuals and smaller groups. Applying a rank-biased scoring rule of $rankscore(r) = 1/r$ significantly improves diagnostic performance compared to equal weighting. This aligns with the intuition that physicians rank diagnoses in their differential list based on estimated diagnostic probability. However, increasing the rank penalty with $rankscore(r) = 1/r^2$ does not further improve accuracy, which is in line with previously published research on collective intelligence in medical diagnostics.[11] In the remainder of this report we will therefore be working with a scoring rule of $rankscore(r) = 1/r$.

3. Integration of user metadata

In this section, we examine whether and how incorporating user metadata into aggregation algorithms can enhance performance. Previous research has demonstrated that selecting[12] individuals for a group or assigning weights[13] based on indicators of their individual ability can improve collective performance.

3.1 Expertise

Seniority. We begin our analysis by selecting users for the collective based on their level of professional experience. Specifically, we distinguish between physicians—including attending and resident physicians as well as fellows—and medical students. As in previous analyses, we apply a scoring function of $rankscore(r) = 1/r$ and randomly select members into the group. However, in this case, we conduct two separate sampling procedures: one in

[11] Barnett, M.L. et al. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Network Open, 2019, doi:10.1001/jamanetworkopen.2019.0096.
[12] Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677
[13] Budescu, D. V., & Chen, E. (2014). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*. https://doi.org/10.1287/mnsc.2014.1909

which members are drawn from the pool of physicians and another in which they are drawn from the pool of medical students who diagnosed the respective cases.

Figure 3 compares the diagnostic performance of collectives composed of physicians and those composed of medical students. The results indicate that, on average, physician groups consistently outperform groups of medical students.

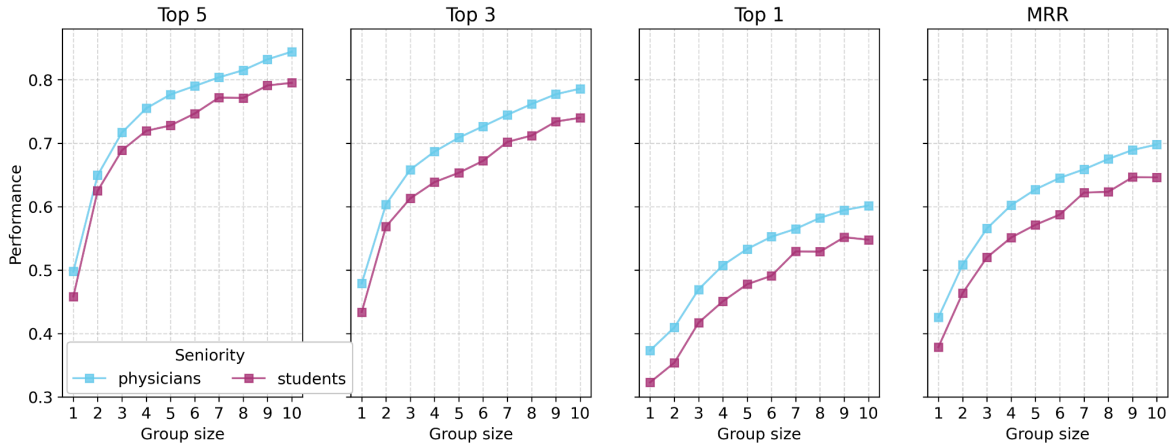

Figure 3. Performance difference between groups of physicians and groups of medical students.

Track record. Next, we investigate whether selecting users for the collective based on their track record can enhance diagnostic performance. To assess this, we employ a leave-one-out strategy: for each case, we compute a user's performance (in MRR) based on all other cases they had diagnosed, excluding the focal case. For users who had diagnosed fewer than five other cases, we substitute their track record with the median MRR of all other solvers, as the available data points were insufficient for a reliable estimate.

Given that the dataset is highly imbalanced—some cases have more than 100 differential diagnoses, while others have as few as 10—we first sample 10 differential diagnoses per case to standardize the selection pool. To form groups of varying sizes, we then either randomly sample users from this set of 10 or select those with the highest leave-one-out track record.

As shown in Figure 4, selecting users based on a stronger track record significantly improves collective diagnostic performance compared to random selection.

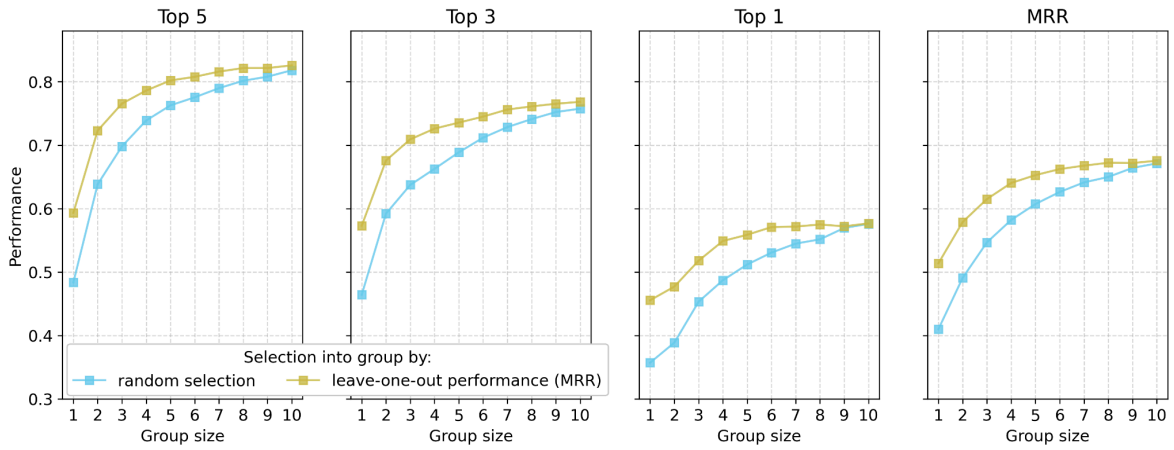

Figure 4. Selecting users by track record improves collective diagnostic performance over random selection.

3.2 Decision similarity

While relying on past performance is intuitive and straightforward, 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 track record instead of identifying high-performing individuals. One possible proxy is the track record in another domain, which has been used to improve collective accuracy in generating forecasts.[14] However, in the context of medical diagnostics, this strategy does not seem either promising or justifiable.

Another approach to pooling individual knowledge or beliefs is Cultural Consensus Theory,[15] originally developed in cognitive anthropology and later extended to diverse research settings. It employs formal cognitive models to estimate both a group's consensus truth and individual parameters such as knowledge level or response bias.

In the context of binary decision making it has been shown that decision similarity can act as a predictor of performance,[16] provided that average individual accuracy surpasses chance.

However, leveraging decision similarity in the domain of open-ended medical diagnostics where differential diagnoses have the form of rankings of varying length is significantly more complicated than in binary choice problems. To define the similarity between two rankings L

[14] Howe, P. D. L., Martinie, M., & Wilkening, T. (2024). Using cross-domain expertise to aggregate forecasts when within-domain expertise is unknown.Decision, 11(1), 35–59. https://doi.org/10.1037/dec0000212

[15] Batchelder, W. H., Anders, R., & Oravecz, Z. (2018, March). Cultural consensus theory. In J. T. Wixted (Ed.), Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (pp. 1–64). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn506

[16] Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., Zalaudek, I., Carney, P. A., & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11), eaaw9011. https://doi.org/10.1126/sciadv.aaw9011

with length l and S with length s, $l \geq s$, we use the extrapolated rank biased overlap,[17] a metric developed in the field of information retrieval to compare search engine results:

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

As a first step, we assess the correlation between decision similarity and diagnostic performance. To ensure robust estimates, we include only users who have solved more than five cases. In the reported results further down we set the rank bias strength $p = 0.5$ but observed similar effects over a wide range of values.

First, we quantify the decision similarity for each differential diagnosis provided by a user—whether a single diagnosis or a ranked list—by comparing it to all other diagnoses given for the same case using RBO_{ext}. To obtain an overall estimate of a user's decision similarity, we compute the average similarity across all cases they had solved.

Next, we evaluate each user's diagnostic performance using top-5, top-3, and top-1 accuracy, as well as MRR. As shown in Figure 5, we observe a strong correlation between an individual's decision similarity and their diagnostic performance. This finding suggests that decision similarity serves as a reliable proxy for diagnostic competence, allowing for the identification of skilled individuals without directly relying on gold-standard solutions which might not be available.

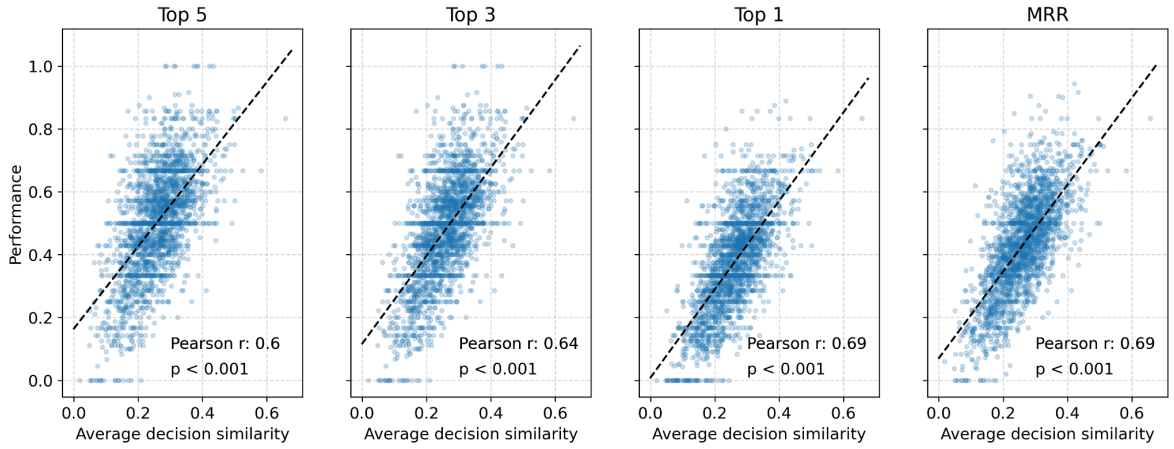

Figure 5. Decision similarity correlates strongly with diagnostic performance, indicating its potential as a proxy for competence. The dashed line corresponds to a simple linear regression.

Building on the previous findings, we apply a leave-one-out strategy similar to that used for track record assessment (see section 3.1). However, instead of evaluating past diagnostic performance, we compute each user's average decision similarity to all other diagnosticians across the cases they had solved, excluding the focal case. This average decision similarity is then used as a criterion for selecting individuals into a group.

[17] Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, *28*(4), 20:1-20:38. https://doi.org/10.1145/1852102.1852106

As shown in Figure 6, this selection procedure significantly improves diagnostic performance compared to a random selection baseline. Moreover, the improvement is comparable to that achieved by selecting users based on their leave-one-out track-record performance.

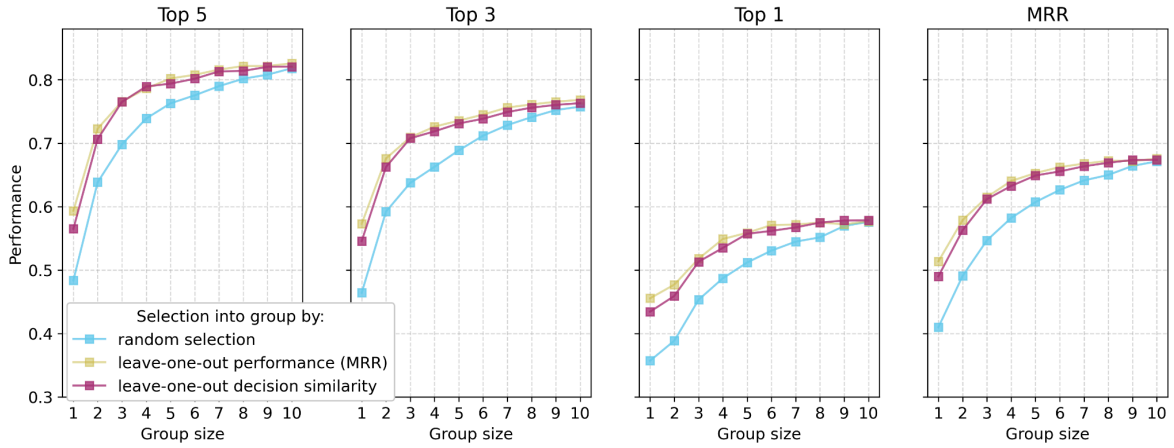

Figure 6. Selecting users by decision similarity improves diagnostic performance, almost matching the gains from performance-based selection.

3.3 Confidence

It seems intuitive to favor the diagnosis of a highly confident physician compared to one who expresses lower confidence. Meyen et al.[18] found that confidence-weighted plurality voting not only provided a closer match to actual decisions in real interacting groups, but also tended to yield higher accuracy than unweighted plurality voting. However, Silver et al.[19] demonstrated that incorporating confidence into decision-making is only beneficial if participants' confidence is well-calibrated to their accuracy.

In the context of medical diagnostics, findings on the relationship between confidence and accuracy have been mixed. Friedman et al.[20] reported a significant alignment between confidence and diagnostic accuracy, whereas Berner and Graber[21] identified overconfidence as a common source of diagnostic errors in medicine. Similarly, Meyen et al.[22] found that

[18] Meyen, S., Sigg, D. M. B., Luxburg, U. von, & Franz, V. H. (2021). Group decisions based on confidence weighted majority voting. *Cognitive Research: Principles and Implications*, *6*(1), 18. https://doi.org/10.1186/s41235-021-00279-0

[19] Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, *96*, 104157. https://doi.org/10.1016/j.jesp.2021.104157

[20] Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? *Journal of General Internal Medicine*, *20*(4), 334–339. https://doi.org/10.1111/j.1525-1497.2005.30145.x

[21] Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5, Supplement), S2–S23. https://doi.org/10.1016/j.amjmed.2008.01.001

[22] Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine*, *173*(21), 1952–1958. https://doi.org/10.1001/jamainternmed.2013.10081

physicians' confidence levels may be relatively insensitive to both diagnostic accuracy and case difficulty, further complicating the role of confidence in collective diagnostic decision-making.

To investigate the relationship between confidence and diagnostic accuracy, we conducted an experiment on the Human Dx platform, where users were asked to rate their diagnostic confidence in terms of a subjective probability. Following the completion of a case, a survey (shown in Figure 7) prompted users to provide their confidence ratings.

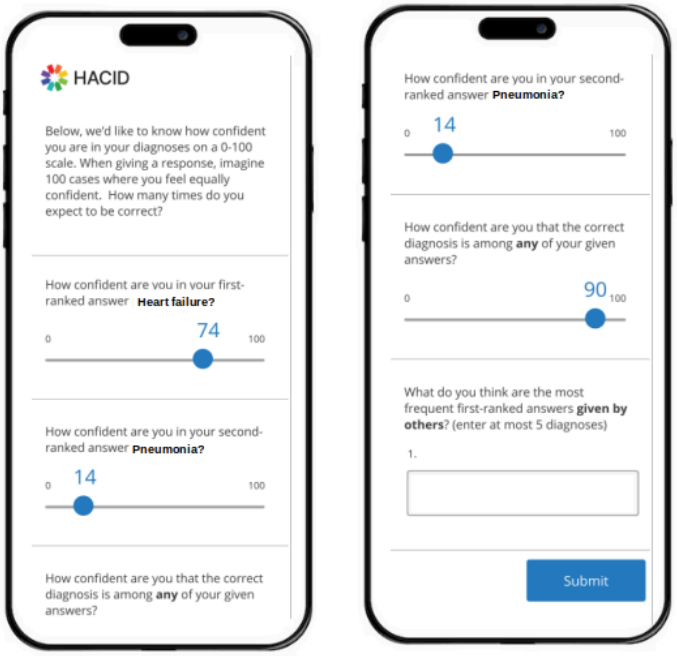

Figure 7. User interface facing for confidence elicitation.

In the following, we present preliminary results based on 30 medical case vignettes. A hierarchical Bayesian logistic regression model assessing the relationship between confidence in the primary diagnosis and diagnostic correctness revealed a clear association between confidence and accuracy (see Figure 8). The corresponding probability estimates fitted to the data at the case level, are presented in Figure 9.

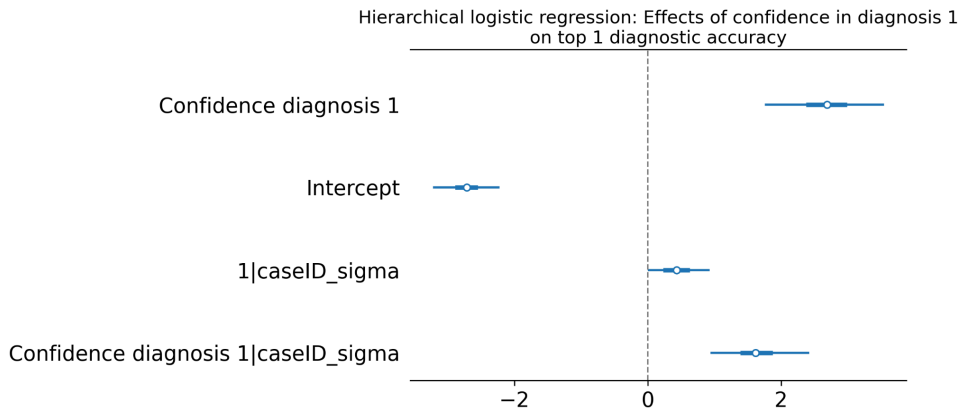

Figure 8. Higher confidence in the primary diagnosis is significantly associated with greater top-1 diagnostic accuracy. The intercept (Intercept) reflects the baseline log-odds of a correct diagnosis at zero confidence, while the slope (Confidence diagnosis 1) captures the effect of confidence on diagnostic accuracy. Random intercepts (1|caseID_sigma) and slopes (Confidence diagnosis 1|caseID_sigma) by *caseID* account for case-specific variability in both baseline accuracy and the confidence-accuracy relationship. Point estimates represent mean effect sizes, with the inner band showing the 50% HDI and the outer band the 95% HDI.

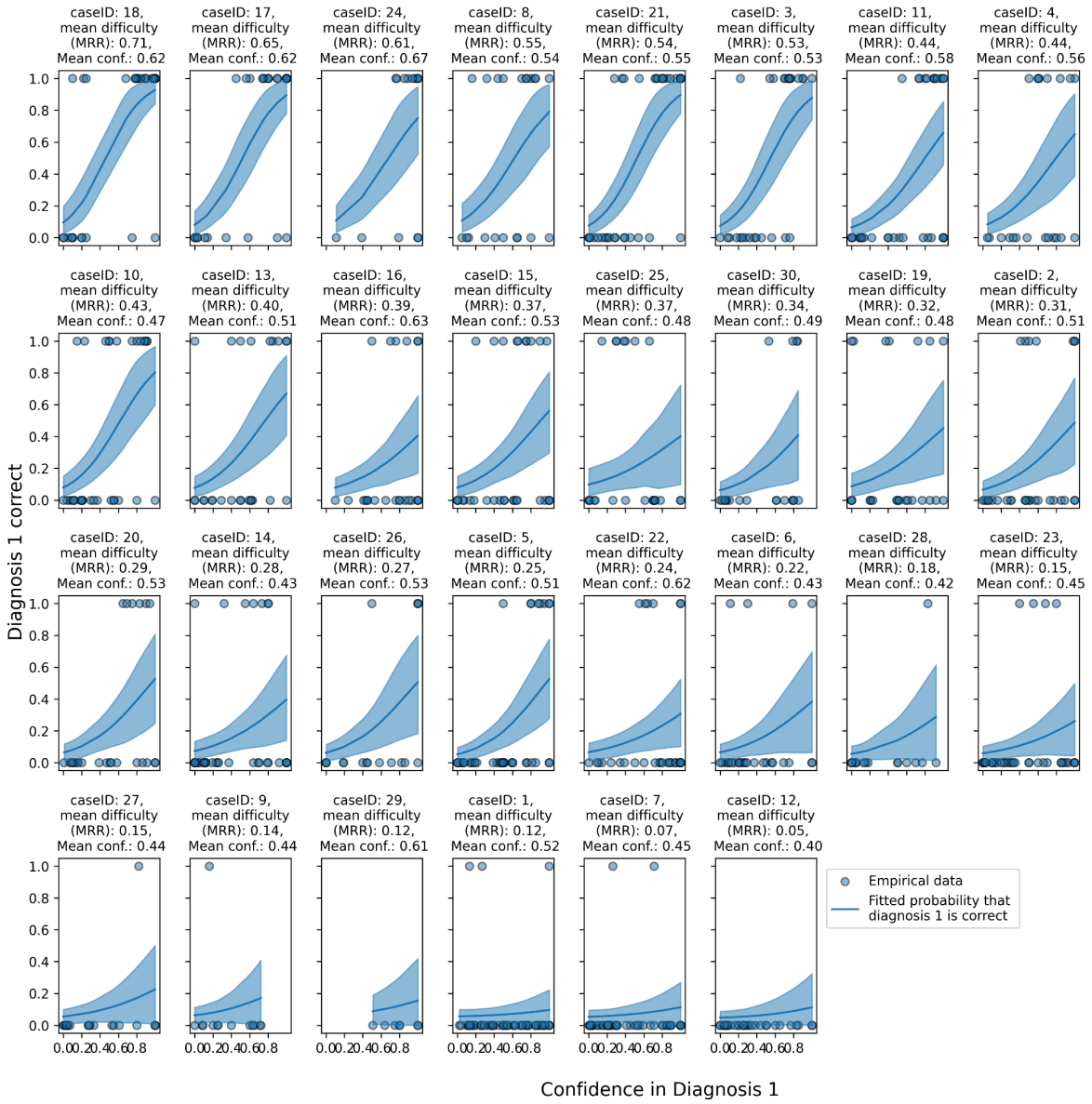

Figure 9. Case-level probability estimates illustrate the relationship between confidence and diagnostic correctness. The band indicates 95% HDI for the estimated probability.

Building on these findings, we examine the impact of selecting group members based on their confidence in their primary diagnosis and compare the resulting collective accuracy to a random selection baseline. Figure 8 illustrates that selecting only the most confident individuals significantly improves performance across all evaluated metrics. Additionally, Figure 9 presents results for a confidence-weighted aggregation approach, where individual responses are linearly weighted[23] by confidence in the primary diagnosis. This corresponds to a score of $s(c_j) = \sum_{i=1}^{n} \sum_{u=1}^{m} \sum_{r=1}^{R} conf_{1u} \cdot 1/r \cdot 1(\rho^{(u)}(c_i) = r)$ where $conf_{1u}$ is user u's confidence in their primary diagnosis that determines the weight on the user level w_u.

[23] We also tested logodds and softmax weights but they performed significantly worse.

This approach is compared to a baseline where all group members contribute equally. The results indicate that top-1 accuracy (and consequently MRR) improves with confidence weighting, whereas top-3 and top-5 accuracy show little or no improvement. It is important to note that confidence weighting and confidence-based selection are distinct mechanisms that, in principle, can be applied in combination. However, our findings reveal that selecting group members based on confidence leads to larger performance gains than confidence weighting alone.

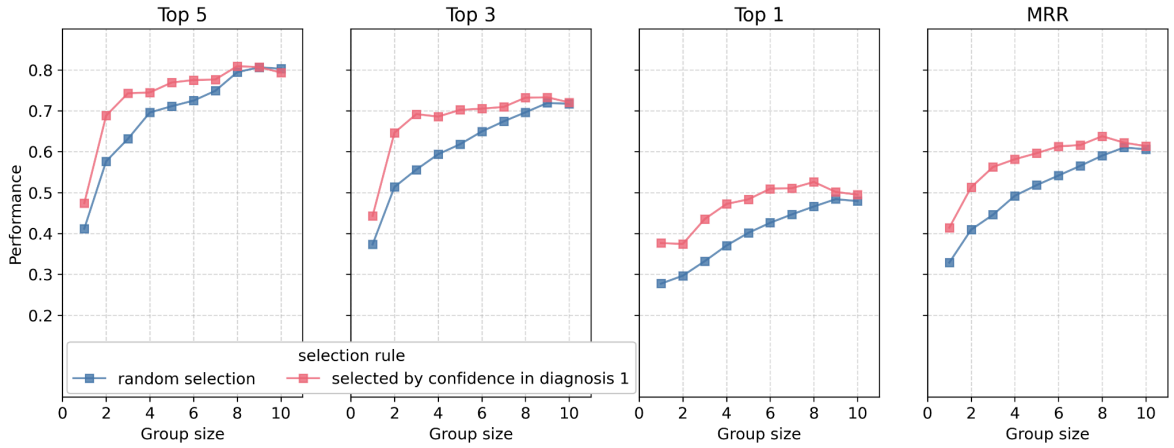

Figure 8. Selecting the most confident individuals into the group enhances collective diagnostic performance across all metrics.

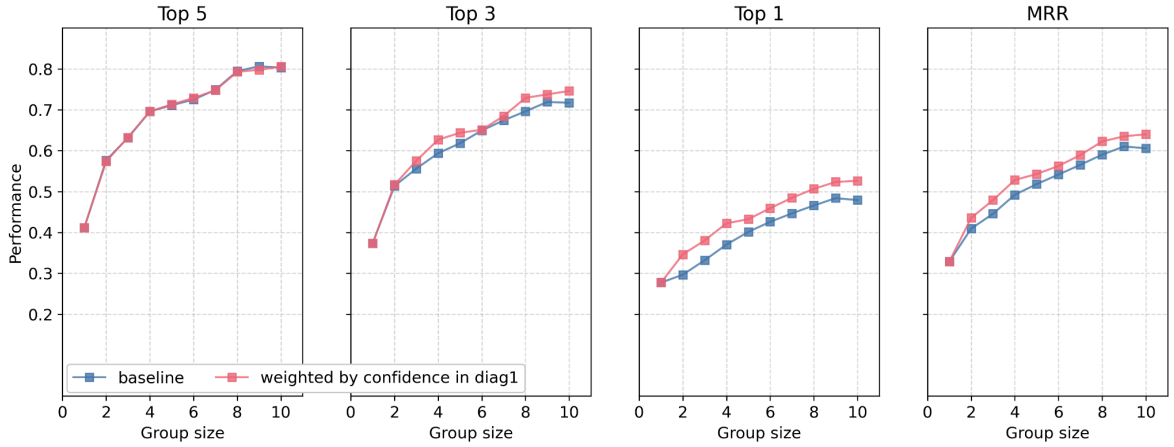

Figure 9. Confidence-weighted aggregation improves top-1 accuracy and MRR but has limited impact on top-3 and top-5 accuracy.

3.4 Response times

Finally, we examine the extent to which response times can be leveraged to identify competent diagnosticians. A recent study by van de Calseyde et al.[24] found that selecting the fastest diagnosis from a pool of diagnoses was, on average, more accurate than random selection and even outperformed the best individual diagnostician on average. However, their analysis focused solely on individual performance rather than collective decision-making.

Figure 10 presents the distribution of response times for all diagnoses in our main dataset, with a median response time of 112 seconds. To explore the relationship between diagnostic performance and response time, we compute the binned averages of top-1 accuracy across response time bins of 5 seconds (see Figure 11). For response time ranges where sufficient data are available, we observe a clear monotonic decrease in accuracy as response time increases. However, for very short (t<10 seconds) and very long (t>500 seconds) response times, accuracy estimates fluctuate considerably due to the limited number of observations. Here, we report results for top-1 accuracy, but findings for top-3, top-5, and MRR exhibit similar patterns.

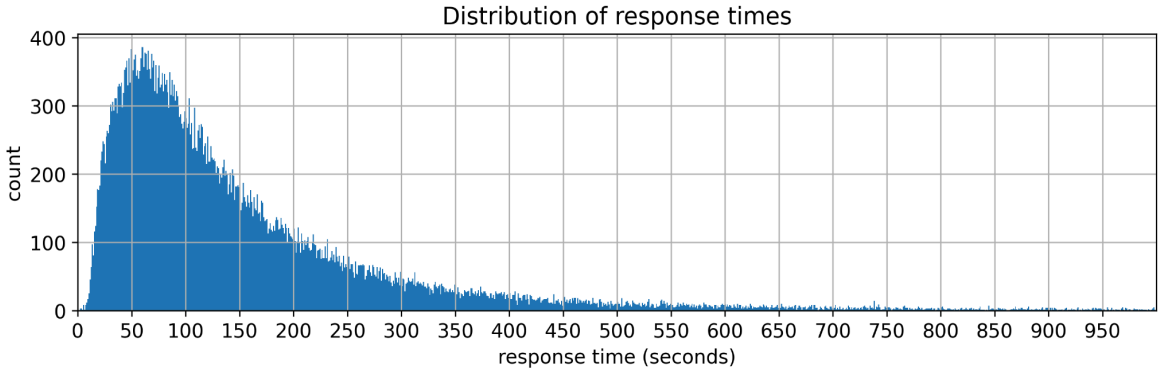

Figure 10. Distribution of response times for diagnoses.

[24] van de Calseyde, P., Efendic, E., van Dolder, D., van den Assem, M. J., Evans, A., Staal, J., Zwaan, L., Sherbino, J., & Norman, G. (2024). *Follow the fast: A simple algorithm for extracting wisdom from crowds* (SSRN Scholarly Paper No. 4959736). Social Science Research Network. https://doi.org/10.2139/ssrn.4959736

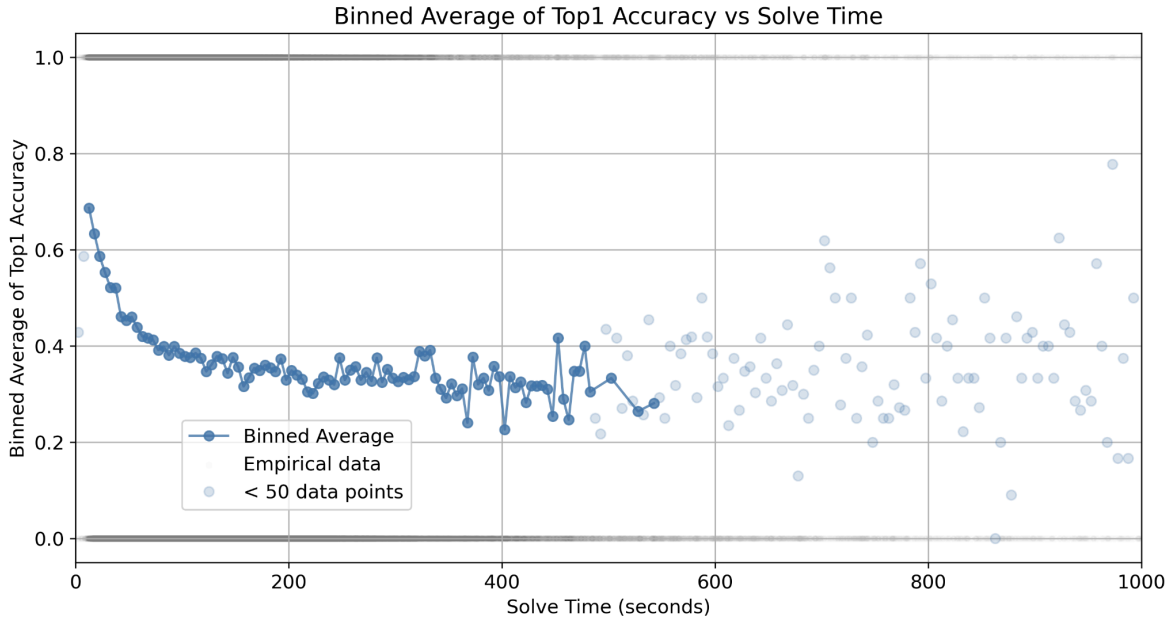

Figure 11. Diagnostic accuracy decreases monotonically with longer response times, except for very short and very long durations, where estimates fluctuate due to limited data.

To leverage the relationship between diagnostic performance and response time, we examine whether selecting the fastest diagnosticians into the group improves collective performance. Since cases vary in the number of responses, we first standardize the selection process by randomly sampling 10 diagnoses per case. From this subset, we then either select the fastest diagnosticians to form groups of varying sizes or apply random selection as a baseline. Figure 12 demonstrates that selecting the fastest diagnoses as the basis for aggregation leads to an improvement in diagnostic performance compared to random selection.

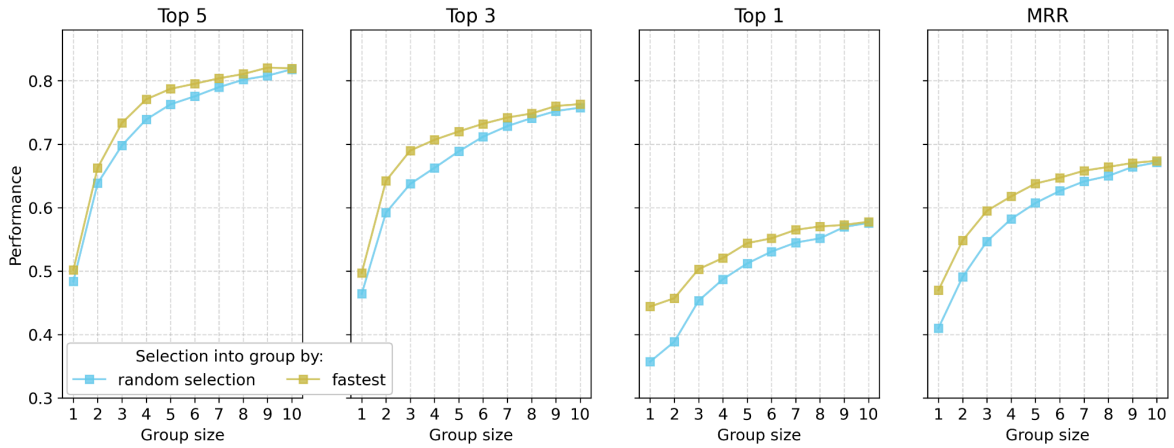

Figure 12. Selecting the fastest group members into the collective enhances collective diagnostic performance across all metrics.

4. Aggregation methods exploiting domain knowledge

So far, we have leveraged domain knowledge captured in the HACID Domain Knowledge Graph (DKG)—which integrates SNOMED CT as a core component—primarily during the matching step, where it supports the standardization of diagnostician responses that refer to the same concept but are expressed using different textual representations (see Section 1.2). This standardization is essential for enabling the automated aggregation of responses into coherent collective solutions. However, this use of the DKG has so far focused solely on concept equivalence and has not yet taken advantage of the richer semantic relationships embedded within the graph structure—such as hierarchical links or contextual similarities—which could further enhance reasoning and aggregation.

In this section, we extend our approach by incorporating these conceptual relationships from the domain knowledge graph into the scoring mechanism used to rank collective solutions, as introduced in Section 2.1. Specifically, we integrate the semantic similarity $sim(c_i, c_j)$ between concepts into the scoring function defined in equation (1). In this formulation, we set $w_u = 1$, meaning all diagnosticians are weighted equally, and assume a $1/r$ rank penalty, which was previously found to perform best (see Section 2.3). The similarity measure contributes to the overall score assigned to each nominated concept by factoring in its semantic proximity to other candidate concepts.

4.1 Hop distance on the polyhierarchy

The HACID domain knowledge graph encodes a wide range of semantic relationships between medical concepts, including those derived from SNOMED CT. One of the most frequent and significant is the *is-a* relationship, which is represented by the *broader* property in our ontology model and thus in the HACID knowledge graph. This relationship forms a polyhierarchy in the graph that organizes clinical concepts by specificity, linking more granular diagnoses to broader categories through subject-relation-object triples—for example, *Asthma - broader - Disorder of the respiratory system*, and *Acute Bronchitis - broader - Bronchitis*.

The rationale for leveraging the broader polyhierarchy in the aggregation process is best illustrated with an example. Consider a scenario in which three physicians diagnose a patient differently: one selects *Bronchitis*, another chooses *Acute Bronchitis*, and the third diagnoses *Pneumonia*. Since each diagnosis has a distinct SNOMED CT identifier, a simple plurality voting following automated matching would not favor any particular diagnosis. However, an improved aggregation approach should account for the semantic similarity between *Acute Bronchitis* and *Bronchitis*, ranking them as more probable compared to *Pneumonia*.

A straightforward example of a distance metric for measuring similarity within this polyhierarchy is the number of hops $h_{i,j}$ required to traverse the shortest path between two concept nodes. Formally, the distance can be defined as $d(c_i, c_j) = h_{i,j}$. A corresponding similarity metric can be defined accordingly as $sim(c_i, c_j) = 1/(1 + h_{i,j})$. For this purpose,

we treat the broader relationship as an undirected edge within the concept graph. Figure 13 illustrates the graph formed by the shortest paths on the polyhierarchy between diagnoses included in the differentials of three physicians.

Figure 14 compares the diagnostic performance between the baseline aggregation, the additional weighting via similarities between nominated concepts and the relevance diffusion of the baseline scores via personalized PageRank along the similarity network. We see that the additional weighting via similarities leads to a small increase in performance at least for top-5 and top-3 accuracies. The propagation of relevance scores via personalized PageRank also leads to very small improvements in accuracies and for larger group sizes seems to perform best. However the differences between the baseline aggregation and aggregation techniques that leverage the similarities between concepts are very small.

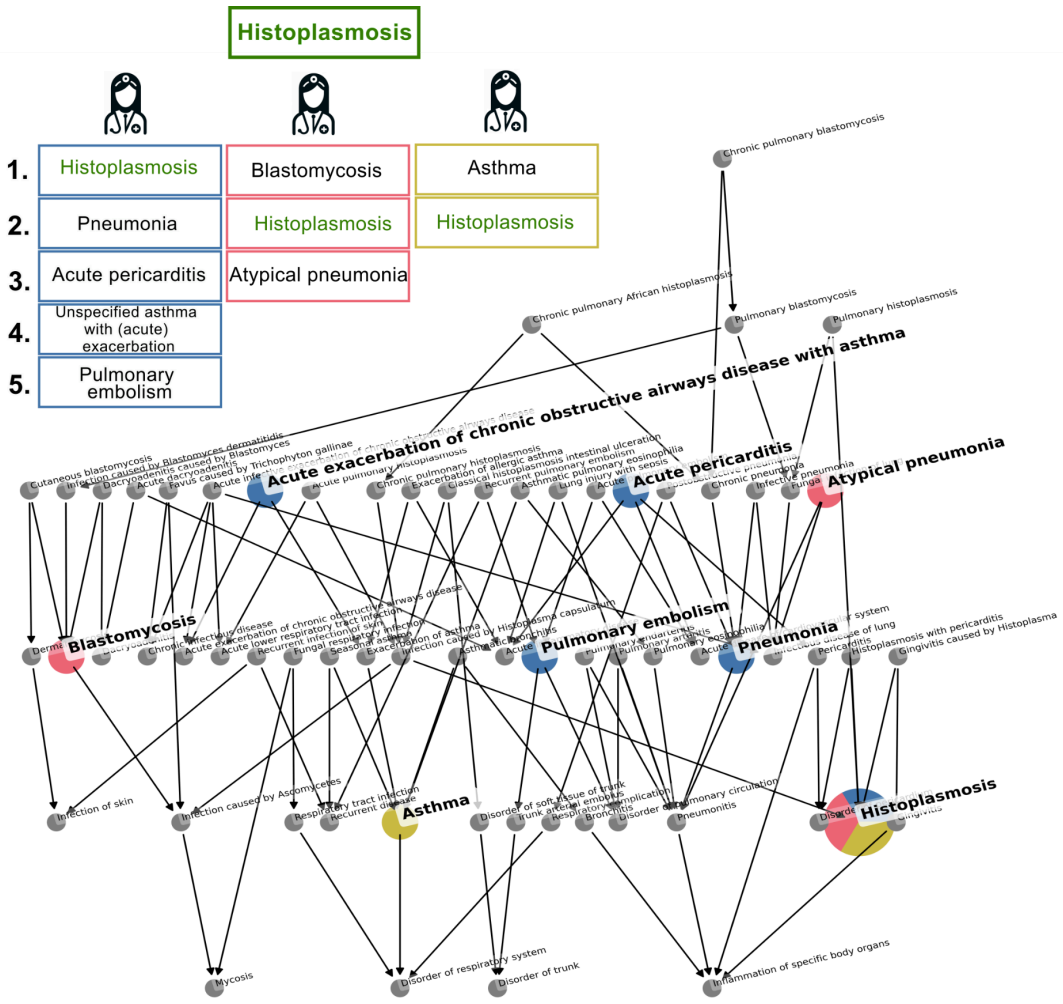

Figure 13. Example of three differential diagnoses and graph of shortest paths they span on the SNOMED CT polyhierarchy that is constructed from the *broader* relation.

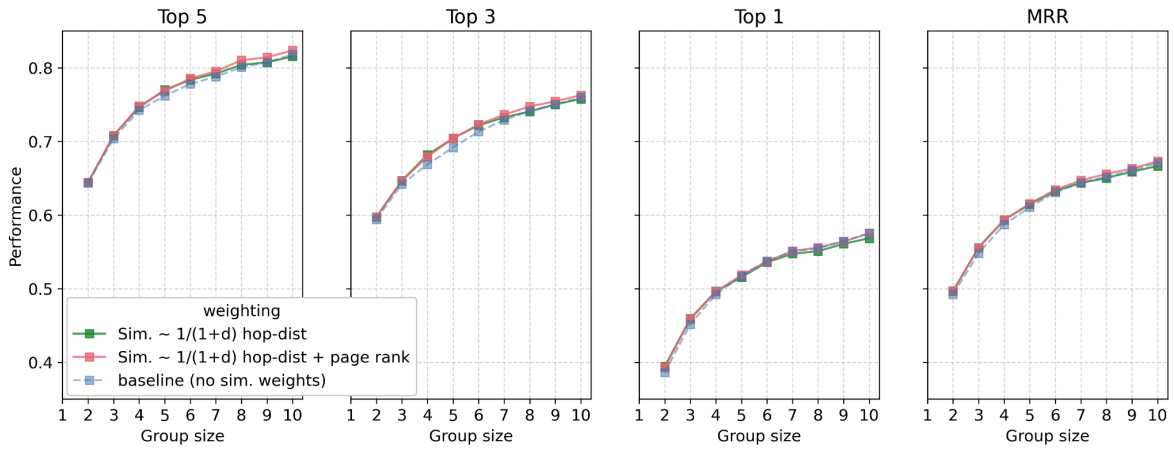

Figure 14. Performance comparison between baseline aggregation, weighting nominated concepts based on similarity in the SNOMED CT polyhierarchy and applying personalized PageRank to the similarity network of nominated concepts with the baseline scores as the personalization vector.

4.2 Knowledge graph embeddings

In the previous section, we focused exclusively on the *broader* relation which forms the polyhierarchy mentioned above. However, our DKG also includes a variety of other relationship types that can further enrich our analyses. To capture these additional semantic links, we now turn to knowledge graph embeddings (KGEs).

Knowledge graphs represent information as sets of entities (e.g., medical concepts) and relationships (e.g., *finding site*) in subject–predicate–object triples. KGEs project these entities and relations into a continuous vector space, preserving structural and semantic relationships. By representing each entity (e.g., a specific disease or body part) and each relation (e.g., *broader* or *finding site*) as vectors, embedding algorithms can learn latent patterns such as transitivity or hierarchy. These learned representations can then be used for downstream tasks such as measuring semantic similarity, predicting missing links, or clustering related concepts, while retaining the original graph's informative structure. See the survey by Cao et al[25] for a systematic review of existing KGE techniques, particularly those based on representation space algorithms that exploit different types of relational symmetries.

Here, we train KGEs on all relation triples using the RotatE algorithm, implemented in PyKeen[26]. RotatE embeds entities and relations in a complex vector space, modeling different relation types as rotations. This makes it particularly suited to capturing the variety of relationship patterns that are found in knowledge graphs. We learn a complex vector representation with 512 dimensions, which is a typical size that strikes a good balance between performance and resource consumption. To transform each complex embedding into a real-valued vector, we concatenate the real and imaginary parts into a 1024-dimensional representation. We then define the similarity between two vectors x_i and x_j

[25] Cao, J., Fang, J., Meng, Z., & Liang, S. (2024). Knowledge Graph Embedding: A Survey from the Perspective of Representation Spaces. *ACM Comput. Surv.*, *56*(6), 159:1-159:42. https://doi.org/10.1145/3643806

[26] https://pykeen.readthedocs.io

using a Gaussian kernel: $sim(x_i, x_j) = exp(-\frac{(|x_i - x_j| - \alpha)^2}{2\sigma^2})$. To choose appropriate values for α and σ, we sampled 1,000 SNOMED CT concepts and identified each concept's closest neighbor. We set α to the minimum distance observed among these closest pairs, and defined a cutoff in the similarity function such that any pair of concepts with a distance strictly less than α receives a similarity score of 1. This ensures that the most tightly related concept pairs are treated as maximally similar. Next, we set σ to the mode of the distribution of those closest-neighbor distances. This choice ensures that concepts within typical neighborhood ranges are assigned high similarity scores, while allowing similarity to decay rapidly for more distant pairs.

Figure 15 presents a comparison between three approaches: the baseline aggregation, direct similarity weighting using KGEs, and a diffusion of baseline scores via personalized PageRank along the similarity network. Contrary to our expectations, incorporating KGEs does not enhance performance relative to the baseline, and direct similarity-based weighting even decreases accuracy. Additionally, letting baseline scores diffuse through the similarity network yields no significant improvement. These outcomes are surprising, given our initial assumption that leveraging the domain knowledge captured in KGEs would boost performance. However, we have yet to conduct extensive hyperparameter tuning or to explore alternative embedding algorithms beyond RotatE—avenues we intend to investigate in future work.

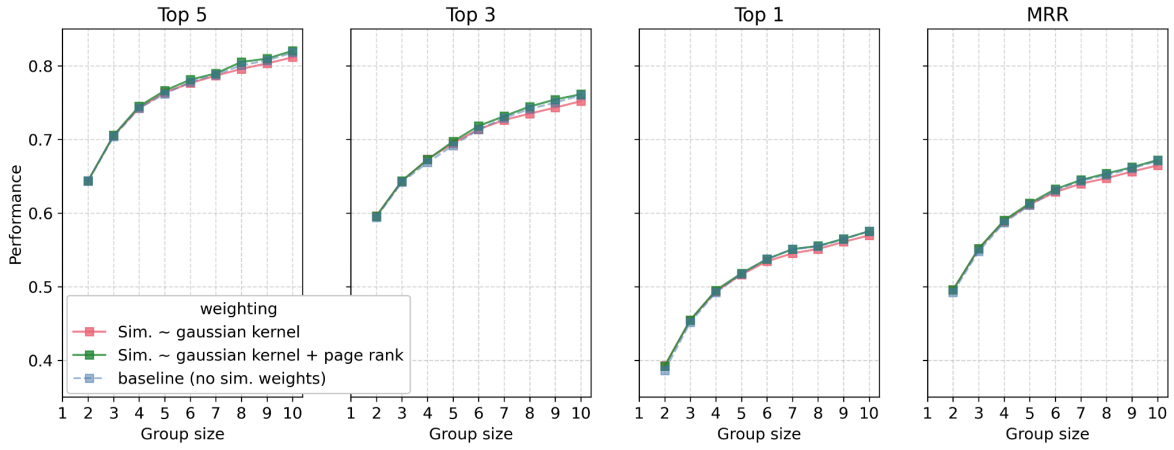

Figure 15. Performance comparison between the baseline aggregation, direct similarity-based weighting, and personalized PageRank on the nominated concepts' similarity network (using baseline scores as the personalization vector). Similarities are derived from RotatE knowledge graph embeddings trained on all relation triples, evaluated via a Gaussian kernel function.

4.3 Graph edit distance between subgraphs

We are currently developing a method to compare diagnoses using structured subgraphs derived from the HACID knowledge graph. The HACID knowledge graph integrates SNOMED CT along with other relevant domain knowledge and will be further enriched with data from Wikidata in the future. While our analysis is still in progress, we outline here our preliminary results. For each diagnosis, we extract a subgraph centered on the diagnosis

concept, including its immediate neighbors (1-hop) and their connections (2-hop). This is because our KG has been constructed exploiting the situation-description pattern, which exploits description nodes to formalise the co-occurence of multiple relations between a single subject and multiple objects (see Figure 16). Hence, the extracted subgraph captures key semantic properties such as:

- *type*, indicating semantic categories such as *disorder*, *clinical finding*, *organism*, or *morphologically abnormal structure*;
- *broader*, linking the concept to its parent nodes in the SNOMED CT polyhierarchy;
- *isDescribedBy*, linking the concept to its description nodes, and relationships as *hasInterpretation*, *findingSite*, *associatedMorphology*, *pathologicalProcess*, *causativeAgent*, etc., connecting description nodes to relevant concepts.

For instance, Figure 16 illustrates a subgraph centered on the concept *Histoplasmosis* where the relationships involved are *broader*, *type*, *pathologicalProcess*, and *causativeAgent*.

To quantify similarity between two diagnoses, we represent their respective subgraphs as *trees*—where the central concept is the root, 1-hop neighbors are children, and 2-hop distant concepts are leaves. We then compute the *tree edit distance*[27] *(TED)*, which measures the minimum number of edit operations (such as node substitutions, insertions, or deletions) needed to transform one tree into the other. Then, we define the similarity coefficient between two concepts c_i and c_j by the Gaussian kernel $sim(c_i, c_j) = exp\left(\frac{-(e_{i,j}-\alpha)^2}{2\sigma^2}\right)$, where $e_{i,j}$ represents the TED between the two root concepts, modulating the parameters α and σ as in Section 4.2. This coefficient informs the aggregation process by weighting diagnoses according to their structural similarity within the knowledge graph.

Figure 17 compares the diagnostic performance of the baseline aggregation with this extended approach, which incorporates additional weighting based on semantic similarities between nominated concepts. Contrary to our expectations, the inclusion of similarity-based weighting does not improve upon the baseline performance; in fact, it even results in a decrease in accuracy. However, these results are preliminary, as we have not yet conducted systematic tuning of the similarity parameters α and σ—a key direction for future work.

[27] Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical Computer Science*, *337*(1), 217–239. https://doi.org/10.1016/j.tcs.2004.12.030

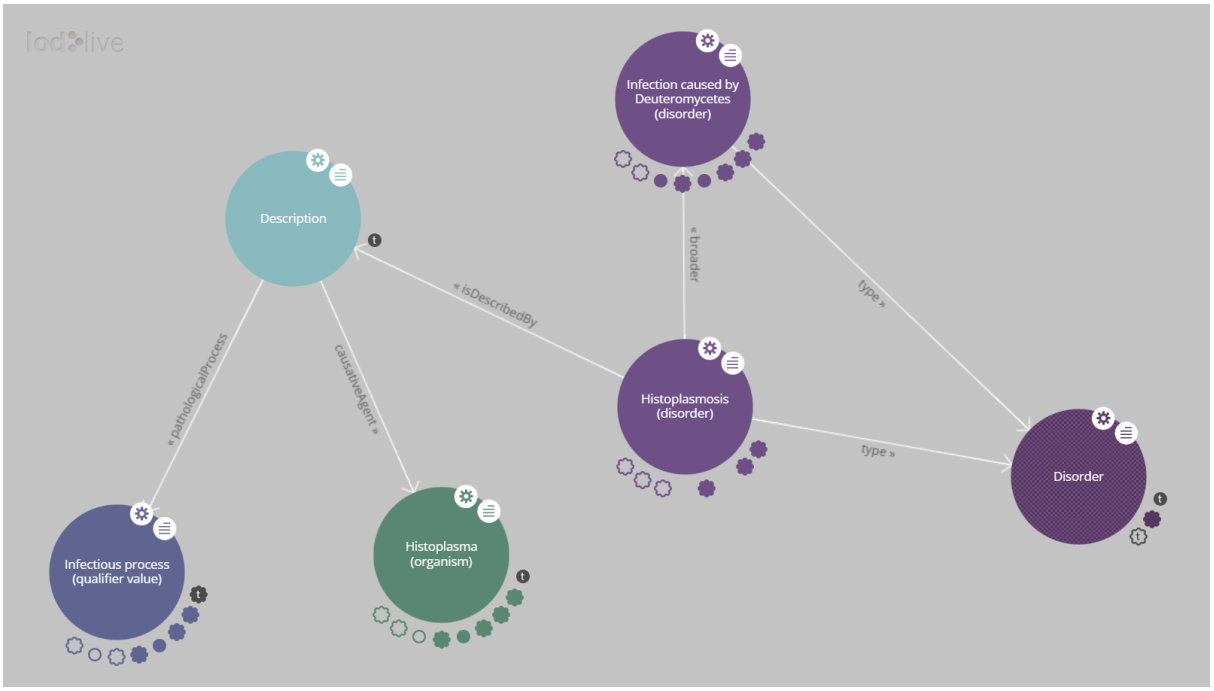

Figure 16. LodLive[28] view of a 2-hop subgraph extracted from the HACID knowledge graph illustrating the relationships around the concept ”Histoplasmosis”.

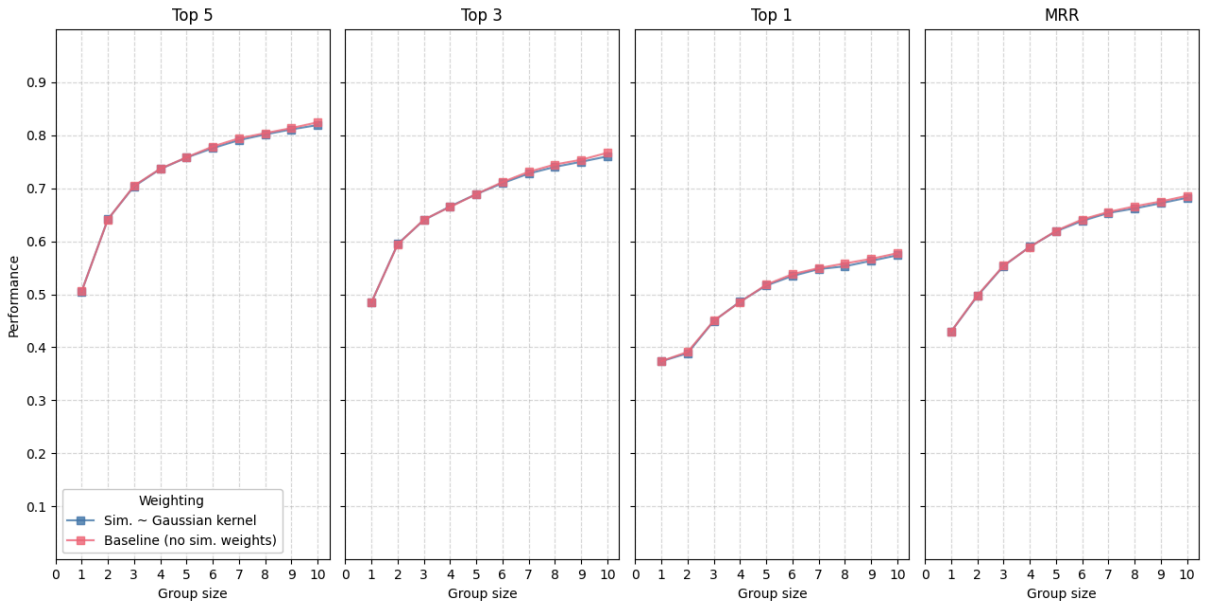

Figure 17. Performance comparison between baseline aggregation and the graph edit distance approach, which weights nominated concepts based on the similarity computed—via a Gaussian kernel function—between their corresponding subgraphs, retrieved from the HACID knowledge graph.

[28] https://github.com/LodLive/LodLive

4.4 Pretrained sentence transformer embeddings

In addition to knowledge graph-based approaches, pretrained sentence transformer embeddings are being considered to assess the semantic similarity between diagnostic terms and enhance the aggregation process.

Clinical term embeddings encode medical concepts as numerical vectors, capturing both semantic meaning and contextual relationships. In this vector space, semantically related terms—measured using a chosen similarity metric—are positioned closer together.

To integrate embeddings into the aggregation process, we:

- generate embeddings for all SNOMED CT concepts, incorporating *fully specified names*, *preferred terms* and *synonyms*;
- store the generated embeddings in a vector database, for efficient indexing and retrieval;
- for each pair of nominated concepts, compute the *Gaussian kernel similarity* between the embeddings of the corresponding fully specified names, to calculate vector-based similarities and assess semantic proximity;
- integrate similarity scores into the weighted ranking of collective solutions.

Specifically, in the latter, we define the similarity coefficient between two vectors x_i and x_j as $sim(x_i, x_j) = exp\left(\frac{-(|x_i - x_j| - \alpha)^2}{2\sigma^2}\right)$, with the parameters α and σ as defined above.

Figure 18 presents a comparison between the baseline aggregation and the direct similarity weighting using clinical term embeddings. We see that incorporating sentence transformer embeddings does not lead to a clear improvement over the baseline; however, the direct similarity-based weighting does yield a slight increase at least in top-5 and top-3 accuracy. As with the knowledge graph-based similarity, we anticipate more substantial gains following a thorough tuning of the parameters α and σ.

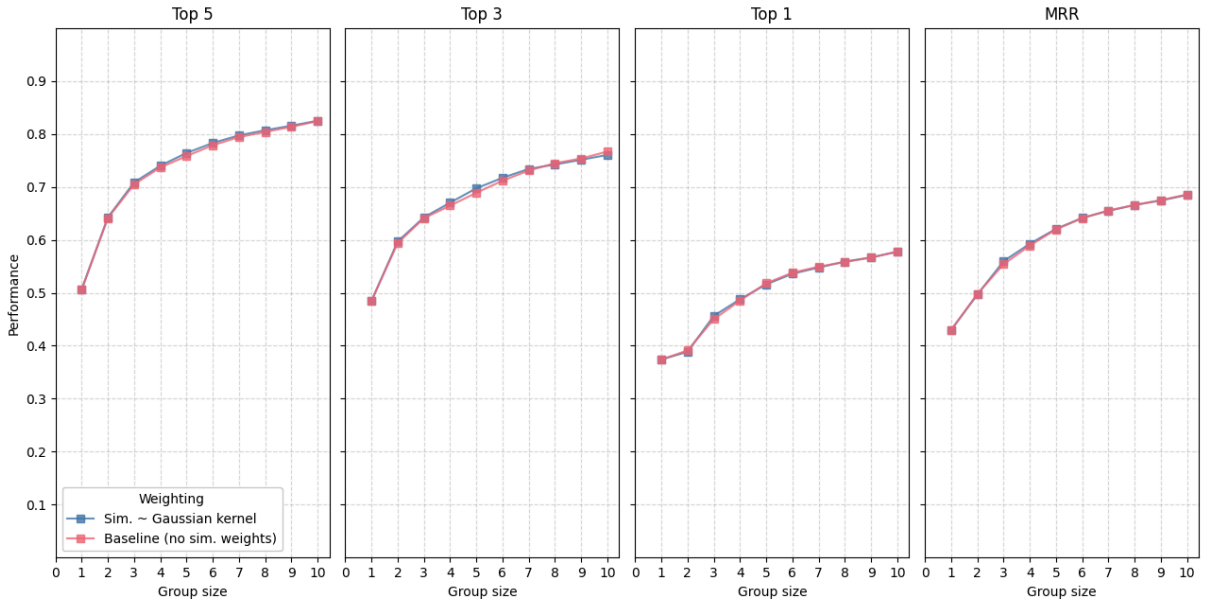

Figure 18. Performance comparison between baseline aggregation and direct similarity-based weighting of nominated concepts, where similarities are derived from clinical term embeddings and computed using a Gaussian kernel function.

5. Outlook and conclusion

In this deliverable, we explored multiple approaches to aggregating open-ended medical diagnoses into collective solutions. We demonstrated that simple plurality voting can be improved by:

1. Introducing rank penalties when solutions are given as rankings.
2. Incorporating user metadata such as track record, response time, confidence, and decision similarity.
3. Exploiting domain knowledge through the HACID DKG.

An important avenue for future research is to see whether combining these methods improves accuracy further. For example, selecting contributors based on response times while also weighting contributors based on their track record. We also plan to further develop and refine domain-knowledge-based aggregation approaches, since the methods tested so far have underperformed relative to initial expectations.

In this report, we focused on the medical diagnostics use case, which represents only one of the two application domains of HACID. We plan to adapt and apply the same techniques also to the climate services use case. Whereas the medical domain benefits from well-structured ontologies such as SNOMED CT and a clearly defined scope (i.e., identifying a correct diagnosis), the climate services domain lacks similarly standardized taxonomies, and its notion of solution is more abstract and context-dependent. Consequently, the HACID project is developing these ontologies from scratch (see WP2 and WP7).

A key challenge in the climate services use case is the absence of a clear gold-standard solution for each scenario. HACID conceptualizes solutions here as workflows—a series of tasks that must be identified and sequenced. The precise form of effective aggregation mechanisms can only be determined once additional expert data become available. Nevertheless, the core principles—such as user-based weighting, rank-informed scoring, and knowledge-graph-based similarity—remain broadly applicable to both climate services and other open-ended tasks with complex solution spaces. The methods presented here offer a robust starting point for harnessing collective intelligence not just in medical diagnostics but in diverse high-stakes environments.