HACID - Deliverable

Social feedback in hybrid collective problem solving

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:	D4.2
Due date:	28.02.2026
Nature[1]:	R
Dissemination Level[2]:	PU
Work Package:	WP4
Lead Beneficiary:	MPG
Contributing Beneficiaries:	CNR, Human Dx EU

[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	02.02.26	Initial draft	Julian Berger	MPG
V2	18.02.26	First Revision	Nikolas Zoller	MPG
V3	23.02.26	Quality check	Neha Mittal	METO
V4	23.02.26	Second Revision	Julian Berger	MPG
V5	24.02.26	Final Revision	Vito Trianni	CNR

Table of contents

1. Introduction

Work Package 4 (WP4) of HACID focuses on developing and evaluating methods for harnessing collective intelligence in complex decision making settings. Deliverable D4.1 addressed statistical aggregation methods for combining independent expert solutions in open-ended problems, where simple approaches such as plurality voting are not applicable. This deliverable (D4.2) complements D4.1 by examining social feedback: how advice from humans, AI, or hybrid sources influences individual and collective decision making.

Statistical aggregation and social feedback represent two distinct mechanisms for leveraging collective intelligence.[3] Aggregation combines independent judgments post hoc. Social feedback, by contrast, allows decision makers to revise their judgments in light of information from others—potentially improving individual accuracy but also risking adopting incorrect social information. Understanding both mechanisms, and their interplay, is necessary for designing effective hybrid human-AI systems.

The experiments reported here draw on the medical diagnostics use case (WP6), where data and infrastructure were available. A discussion of transferability to the climate services use case is provided in the outlook. Additionally, one experiment addresses online misinformation discernment, extending the scope of HACID's investigation to another domain of high societal relevance.

1.1 Decision making under social influence

A large body of research has studied how people integrate advice from others into their own judgments. In the Judge-Advisor System paradigm, a decision maker receives advice from one or more advisors and may revise their initial judgment. A robust finding is egocentric discounting: people systematically underweight advice relative to their own opinion, which can hold back performance improvements otherwise gained would correct advice have been integrated.[4] At the group level, social influence can both help and hinder collective accuracy. Deliberation and information exchange can correct individual errors,[5] but social influence can also reduce the diversity of opinions that makes crowd aggregation effective.[6] Whether social feedback improves or degrades collective accuracy depends on the structure of interaction, the quality of advice, and the domain in question.[7] How these findings translate to open-ended medical diagnostics remains unknown.

---

[3] Tatsuya Kameda, Wataru Toyokawa, and R. Scott Tindale, 'Information Aggregation and Collective Intelligence beyond the Wisdom of Crowds', *Nature Reviews Psychology*, 1.6 (2022), pp. 345–57, doi:10.1038/s44159-022-00054-y.

[4] Silvia Bonaccio and Reeshad S. Dalal, 'Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences', *Organizational Behavior and Human Decision Processes*, 101.2 (2006), pp. 127–51, doi:10.1016/j.obhdp.2006.07.001.

[5] Abdullah Almaatouq and others, 'Adaptive Social Networks Promote the Wisdom of Crowds', *Proceedings of the National Academy of Sciences*, 117.21 (2020), pp. 11379–86, doi:10.1073/pnas.1917687117.

[6] Jan Lorenz and others, 'How Social Influence Can Undermine the Wisdom of Crowd Effect', *Proceedings of the National Academy of Sciences*, 108.22 (2011), pp. 9020–25, doi:10.1073/pnas.1008636108.

[7] Kameda, Toyokawa, and Tindale, 'Information Aggregation and Collective Intelligence beyond the Wisdom of Crowds'.

1.2. Decision making influenced by AI

The rapid development of AI systems, particularly Large Language Models (LLMs), has introduced new forms of decision support. A growing literature examines when and why people accept or reject algorithmic advice. Early work documented algorithm aversion—a tendency to discount algorithmic advice after observing errors[8]—while subsequent work found conditions of algorithm appreciation, where people prefer algorithmic over human advice[9] and many influencing factors that can determine adoption or rejection of AI advice.[10]
With LLMs, AI advice is no longer limited to numerical point estimates or binary and multiclass predictions. LLMs can report on their reasoning, engage in dialogue, and take on different advisory roles. This raises new questions about how people interact with AI advisors and whether such interactions improve decision quality. Most empirical work on AI-assisted decision making, however, has focused on binary or numerical judgment tasks,[11] leaving open how AI advice is integrated in more complex, open-ended domains.

1.3. Open questions in hybrid settings and outlook

Three opportunities motivate the experiments reported in this deliverable. First, almost all research on advice taking and AI-assisted decision making uses tasks with binary or numerical responses. It is largely unknown how advice is integrated when the response space is open-ended, as in medical differential diagnosis where decision makers produce ranked lists of possible diagnoses.
Second, research has studied advice from humans and from AI separately, but little is known about how people respond to advice from hybrid sources—collectives composed of both humans and AI. Yet, prior research in the HACID project by Zöller et al.[12] has shown that hybrid ensembles outperform both human and LLM-only ensembles. Whether the source label (human, AI, or hybrid) affects advice uptake is an open empirical question.
Third, most advice taking research presents advice as a single signal. An alternative is interactive deliberation, where decision makers can discuss a problem with an advisor. Whether interactive deliberation with humans or LLMs improves decisions and to what degree human- or LLM-guided deliberation differs is not well understood.
This deliverable reports three experiments that address these gaps:

1. **Cliniflow** (Section 2) examines advice from human, LLM, and hybrid sources on open-ended medical diagnosis, varying advice timing (early vs. late).

---

[8] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, 'Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err', *Journal of Experimental Psychology: General* (US), 144.1 (2015), pp. 114–26, doi:10.1037/xge0000033.
[9] Jennifer M. Logg, Julia A. Minson, and Don A. Moore, 'Algorithm Appreciation: People Prefer Algorithmic to Human Judgment', *Organizational Behavior and Human Decision Processes*, 151 (2019), pp. 90–103, doi:10.1016/j.obhdp.2018.12.005.
[10] Jason W. Burton, Mari-Klara Stein, and Tina Blegind Jensen, 'A Systematic Review of Algorithm Aversion in Augmented Decision Making', *Journal of Behavioral Decision Making*, 33.2 (2020), pp. 220–39, doi:10.1002/bdm.2155.
[11] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone, 'When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis', *Nature Human Behaviour*, 8.12 (2024), pp. 2293–303, doi:10.1038/s41562-024-02024-1.
[12] Nikolas Zöller and others, 'Human–AI Collectives Most Accurately Diagnose Clinical Vignettes', *Proceedings of the National Academy of Sciences*, 122.24 (2025), p. e2426153122, world, doi:10.1073/pnas.2426153122.

2. **Medical Deliberation Teams (MDT)** (Section 3) examines interactive deliberation in human dyads or with LLMs acting as expert, evaluator, or coach in open-ended medical diagnosis.
3. **Hybrid Misinformation** (Section 4) examines advice from human, LLM, and hybrid crowds on binary truth discernment in headlines across four countries. This experiment extends prior findings of HACID, namely the success of hybrid crowds, into the domain of online misinformation discernment.

Together, these experiments span both open-ended and binary decision tasks, static and interactive advice formats, and expert and lay populations, providing a broad empirical basis for understanding social feedback in hybrid human-AI collectives.

2. Experiment 1: Cliniflow

2.1. Introduction

Diagnostic errors significantly impact patient safety, contributing to substantial morbidity and mortality. Advances in artificial intelligence (AI), particularly Large Language Models (LLMs), show promise for enhancing diagnostic accuracy. Additionally, collective intelligence—aggregating judgments from multiple diagnosticians—has demonstrated improved performance over individual assessments.[13]

Building on previous findings that hybrid collectives (combined human and AI diagnostics) outperform purely human or AI groups,[14] the Cliniflow study investigates how diagnostic advice from AI, human experts, or hybrid collectives influences physician decision making, confidence, and diagnostic accuracy.

2.2. Research Questions

Cliniflow investigates three research questions:

1. **How is diagnostic advice being integrated when diagnosing clinical vignettes?** We hypothesize that advice leads participants to adjust their differential diagnoses toward the advice and that advice uptake is higher for early advice than for late advice. Uptake may differ by advice source and the timing effect on uptake may vary by source.
2. **How does advice affect diagnostic accuracy?** We hypothesize that receiving advice (LLM, human, or hybrid) increases diagnostic accuracy compared to the control condition. We further hypothesize that early advice leads to higher diagnostic accuracy than late advice. The magnitude of improvement may differ by advice source and the timing effect may vary by source.
3. **Does advice influence diagnostic confidence?** We hypothesize that advice increases diagnostic confidence relative to the control condition and that early advice leads to higher confidence than late advice. Confidence may differ by advice source and the timing effect on confidence may vary by source.

---

[13] Ralf H. J. M. Kurvers and others, 'Automating Hybrid Collective Intelligence in Open-Ended Medical Diagnostics', *Proceedings of the National Academy of Sciences*, 120.34 (2023), p. e2221473120, doi:10.1073/pnas.2221473120.

[14] Zöller and others, 'Human–AI Collectives Most Accurately Diagnose Clinical Vignettes'.

2.3. Methods

2.3.1 Design

Cliniflow is an online experiment and data is being collected on the Human Diagnosis Project (Human Dx) platform on which participants work on clinical case vignettes. The experiment followed a 1 + 3 × 2 design, comprising one control (no advice) and three advice conditions: (i) AI-only advice (four LLMs), (ii) human-only advice (four physicians) and (iii) hybrid advice (two physicians and two LLMs). Advice was presented either early (before case details) or late (after initial diagnosis). Participants rated their diagnostic confidence on a four-point Likert scale after submitting their differential diagnoses. Immediately thereafter, the correct diagnosis for a given vignette, ensured by Human Dx, was displayed.

2.3.2. Procedure

The procedure for solving a clinical vignette was as follows. The participant opened a case vignette on the Human Dx platform. In the early-advice conditions, advice was displayed together with the first medical findings before an initial diagnosis was submitted. In the late-advice conditions, advice was shown after the participant had observed all the medical findings and had submitted an initial differential diagnosis. In the control condition, no advice was shown. The participant then submitted a final ranked differential diagnosis and rated their diagnostic confidence on a 4-point Likert scale. Immediately thereafter, the correct diagnosis (ensured by Human Dx) was displayed.
For each case in each condition we aimed to collect at least 20 responses. That means at least 1400 responses in total. Three different cases were simultaneously available on the platform; available cases were rotated every day. The order of the 10 cases was determined randomly at the beginning and this sequence was then repeated to form a queue. Experimental conditions were sampled randomly and assigned to the cases in the queue. As soon as at least 20 valid responses were collected per case and condition, this case/condition combination was removed from the queue.

2.3.3. Case Selection

We used medical case vignettes from the Human Dx platform. Vignettes were selected from a larger dataset for which LLM-generated differential diagnoses were collected in Zöller et al.[15] The selection of vignettes is restricted to cases from internal medicine and enforces gender balance, choosing 5 vignettes with male and 5 with female patients. We also ensured a wide range of patient ages (38 to 84). Only vignettes with exactly one correct diagnosis that can be mapped to a single SNOMED CT[16] ID were considered. Each selected vignette must have a unique correct diagnosis, and all correct diagnoses must be at least three hops apart in the SNOMED CT polyhierarchy in order to ensure sufficient variety in case vignettes.
We ensured that we have at least 10 prior solves by HumanDx users when selecting the cases. We also demanded that the mean reciprocal rank (MRR) of these prior solves was between 0.1 and 0.7 to exclude very hard cases and to make sure that it is indeed possible

[15] Zöller and others, 'Human–AI Collectives Most Accurately Diagnose Clinical Vignettes'.
[16] Kevin Donnelly, 'SNOMED-CT: The Advanced Terminology and Coding System for eHealth', *Studies in Health Technology and Informatics*, 121 (2006), pp. 279–90.

to sensibly solve the cases and that they are not misleading. We also excluded easy cases where everyone is nearly always correct because those physicians unlikely need advice. We then selected 10 cases according to the following criteria:

- **LLM advice advantage (3 cases)**: The 4-LLM ensemble (Claude 3, GPT 4, Gemini 1 Pro, Mistral Large) outperformed both the sampled 4-human ensemble and the average of all possible 4-human ensembles, with a margin of RR(LLM) – MRR(human) > 0.5.
- **Human advice advantage (3 cases):** The sampled 4-human ensemble and the average of all possible 4-human ensembles both outperformed the 4-LLM ensemble with the margin: MRR(human) – RR(LLM) > 0.5.
- **Advice parity (4 cases):** The sampled 4-human and 4-LLM ensembles achieved the same rank for the correct diagnosis, and the performance difference between the average of all possible 4-human ensembles and the 4-LLM ensemble was less than 0.2 in absolute terms: |MRR(human) – RR(LLM)| <= 0.2. We picked 2 cases where both the 4-LLM ensemble and the 4-human ensemble gave incorrect advice (i.e., the correct diagnosis is not ranked at all). We also selected 2 cases where both the 4-LLM ensemble and the 4-human ensemble gave correct advice (i.e., rank the correct diagnosis in the 1st place in the differential).

For each of these categories, if there were more than the necessary cases available, we sorted by number of findings and picked the longer ones to ensure that cases have sufficient detail.

2.3.4. LLM Prompts

For generating LLM differentials as advice, we used the on average best-performing prompt by Zöller et al.[17] which consists of a base prompt and five few-shot examples of successfully solved cases. This prompt was used for all four models, namely Claude 3, GPT 4, Gemini 1 Pro and Mistral Large.

Base prompt + 5 few-shot examples

```
Provide only the most probable differential diagnosis, no explanation, no
recapitulation of the case information or task. Give a maximum of 5 answers, sorted
by probability of being the correct diagnosis, most probable first, remove list
numbering, and respond with each answer on a new line. Be as concise as possible, no
need to be polite.
Here are some examples of cases and their correct answers: Case description: {case
vignette} Answer: {example solution} (5x)
```

Figure 1: Prompts used to generate LLM differential diagnoses.

2.3.5. Participants

Participating users on the Human Dx platform receive 50 impact points as an incentive to solve a case. Impact points can be used by users on the platform as a form of currency to post new questions, reward other users or access to AI services. Participants on the Human Dx platform self-select themselves into available cases.

[17] Zöller and others, 'Human–AI Collectives Most Accurately Diagnose Clinical Vignettes'.

2.3.6. Outcomes

Diagnostic accuracy is assessed based on the differential diagnoses participants return. To that end, we make use of a processing pipeline vetted in Zöller et al.[18] that relies on the *Systematized Nomenclature of Medicine Clinical Terms* (SNOMED CT)[19] (see also HACID Deliverable D4.1 for a detailed description). In short, every diagnosis rendered within a differential is mapped to one SNOMED CT ID. Correct diagnoses of the vignettes are also associated with a SNOMED CT ID, allowing for a comparison whether a differential contained the correct ID, and if so, at what rank within the differential the correct diagnosis was placed.

Both for pre- and post-advice phases, we calculate top-1and top-5 accuracy that determines whether the correct SNOMED CT ID for a given case is ranked within the top-n rank of the differential. Advice uptake is assessed using two measures. First, we compute the similarity between a participant's final differential and the provided advice using extrapolated Rank-Biased Overlap (RBOexp)[20], with the rank penalty hyperparameter set to $p = 0.6$ since previous results by Zöller et al.[21] found the selection of p to not qualitatively change results. To establish a baseline, we also compute this similarity for control participants using the advice that would have been shown for that case, which they never saw. Second, in late-advice conditions where participants submit an initial differential before receiving advice, we measure within-solve adjustment as the change in similarity to advice from initial to final differential. Finally, diagnostic confidence is measured on a 4-point Likert scale administered after participants submit their final differential diagnosis.

2.4. Results

The reported results are based on an almost complete dataset. As of February 23 2026, only 2 of the 70 case-by-condition cells had not yet reached the target of 20 responses.

2.4.1. Advice uptake

We begin by discussing the results on advice uptake. Figure 2A–C show that participants' final differential diagnoses are clearly more similar to the advice they received than those in the control condition (which never saw advice). Collapsing across advice sources (Figure 2B), we find that early advice leads to greater uptake than late advice, consistent with prior literature[22]. Furthermore, Figure 2D suggests that AI advice is taken up more often than human or hybrid advice, which partially contradicts earlier findings.[23] However, given rapid

---

[18] Zöller and others, 'Human–AI Collectives Most Accurately Diagnose Clinical Vignettes'.
[19] Donnelly, 'SNOMED-CT'.
[20] William Webber, Alistair Moffat, and Justin Zobel, 'A Similarity Measure for Indefinite Rankings', *ACM Trans. Inf. Syst.*, 28.4 (2010), p. 20:1-20:38, doi:10.1145/1852102.1852106.
[21] Nikolas Zöller, Stefan M Herzog, and Ralf HJM Kurvers, 'Boosting Collective Intelligence in Medical Diagnostics: Leveraging Decision Similarity as a Predictor of Accuracy When Answers Are Open-Ended Rankings', HCOMP-CI 2023, 2023 <https://www.humancomputation.com/2023/assets/wips_demos/CI-23_paper_1055.pdf>.
[22] Olga Kostopoulou and others, 'Early Diagnostic Suggestions Improve Accuracy of GPs: A Randomised Controlled Trial Using Computer-Simulated Patients', Research, *British Journal of General Practice*, 65.630 (2015), pp. e49–54, doi:10.3399/bjgp15X683161.
[23] Federico Cabitza, 'Biases Affecting Human Decision Making in AI-Supported Second Opinion Settings', in *Modeling Decisions for Artificial Intelligence*, ed. by Vicenç Torra and others (Springer International Publishing, 2019), pp. 283–94, doi:10.1007/978-3-030-26773-5_25.

advances in AI and its growing adoption, it is unsurprising that acceptance of AI as an advice source may be changing. Notably, hybrid advice shows the lowest uptake despite being the most accurate.

Figure 3 focuses on the within-participant comparison in the late condition, comparing the similarity of participants' final differential diagnoses to the advice received with the similarity of their initial differentials (before seeing the advice) to that same advice. Across all advice sources, we again observe a clear advice effect: in a median of 22% (hybrid) to 26% (human) of instances, the advice is incorporated into the final diagnosis (Figure 3A). Figure 3B suggests that when AI advice is incorporated, the adjustment toward the advice is larger in magnitude. Overall expected uptake, measured by RBOexp, is slightly lower for hybrid than for human or AI advice, although these differences are not significant (Figures 3C and 3D).

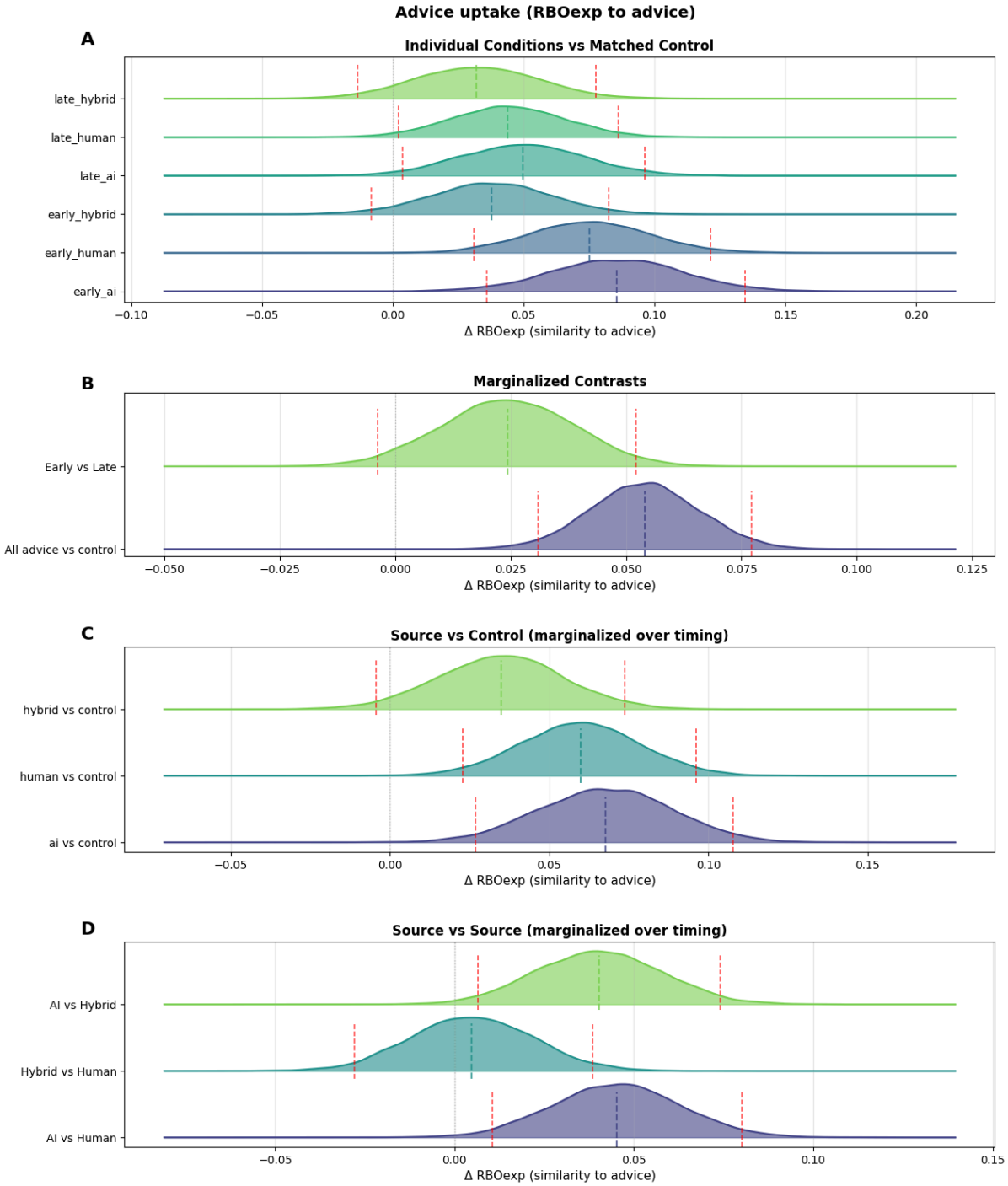

Figure 2. Advice uptake measured by similarity to advice (RBOexp). Posterior distributions of contrasts in expected RBOexp similarity between participants' final differential diagnoses and the advice provided, estimated from a hierarchical Bayesian Beta regression with varying intercepts by case. **(A)** Each advice condition compared to the matched control baseline (which never saw the advice). **(B)** Marginalized contrasts: overall advice effect (pooled across sources and timings) and early versus late timing effect (pooled across sources). **(C)** Source-specific effects versus control, marginalized over timing. **(D)** Pairwise source contrasts, marginalized over timing. Dashed lines indicate posterior medians; red dashed lines mark the boundaries of the 95% credible interval (CrI). The vertical grey dotted line marks zero (no difference).

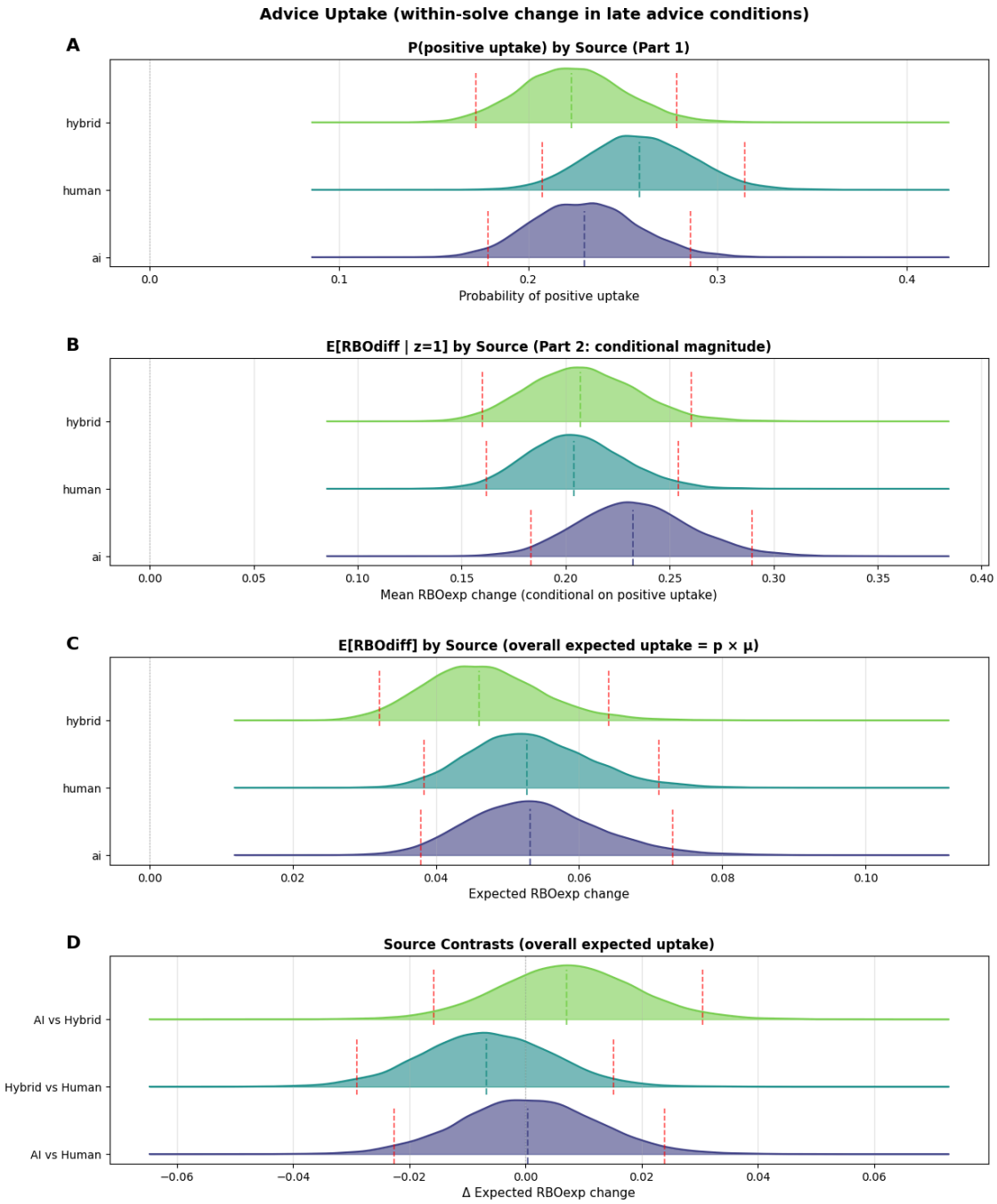

Figure 3. Within-solve advice uptake (late-advice conditions only). Results from a two-part hurdle model comparing participants' final differential diagnoses (after seeing advice) to their initial differentials (before seeing advice), using the change in RBOexp similarity to advice. **(A)** Posterior distributions of the probability of positive uptake (i.e., final differential more similar to advice than initial differential) by advice source. **(B)** Posterior distributions of the conditional magnitude of uptake (expected RBOexp change given positive uptake) by source. **(C)** Combined overall expected uptake (probability x magnitude) by source. **(D)** Pairwise source contrasts on combined expected uptake. Dashed lines indicate posterior medians; red dashed lines mark the 95% CrI boundaries.

2.4.2 Accuracy

Figure 4 shows that, for top-5 accuracy, advice improves performance: participants who received any form of advice were, on average, a median 5.8 percentage points more likely to include the correct diagnosis in their top five, with the 95% credible interval (CrI) just barely excluding zero and a 97.8% posterior probability of a positive effect.

For top-1 accuracy (Figure 5), however, the effect is small (~1.8 percentage points), uncertain, and not credibly different from zero (P(>0) = 76%). Hybrid human–AI advice yields the largest positive effect on accuracy, which is notable given that advice uptake was slightly lower for hybrid advice. This may be explained by the higher accuracy of hybrid advice, but it could also indicate a more selective and efficient uptake of hybrid recommendations. With respect to timing, we expected early advice to lead to higher accuracy. This expectation is not supported by any metric. For top-1 accuracy, there is a slight positive trend, but with substantial uncertainty (Figure 5B). Surprisingly, for top-5 accuracy (Figure 4B) the pattern reverses: late advice appears slightly better (P(late > early) ≈ 76%). This is noteworthy because prior studies reporting benefits of early advice on accuracy[24] focused on a single diagnosis rather than the more common format of a differential diagnosis. One possible explanation is that premature closure[25] primarily affects the main working diagnosis (rank 1), which is captured by top-1 accuracy. For broader consideration sets (i.e., more than one diagnosis), late advice may actually be beneficial.

2.4.3 Diagnostic confidence

Receiving advice increases expected diagnostic confidence by approximately 0.12 points on a 4-point scale (from ~2.56 to ~2.68 on average). The 95% CrI just barely includes zero, but the directional probability is 97.1% (Figure 6A–C). Human advice shows the largest estimated effect on confidence compared to control and is the only source whose 95% CrI excludes zero, followed by AI and then hybrid advice (Figure 6C); however, differences between sources (Figure 6D) are uncertain and not statistically credible.

Timing has essentially no effect on confidence: there is no evidence that early advice leads to higher confidence than late advice.

[24] Kostopoulou and others, ‘Early Diagnostic Suggestions Improve Accuracy of GPs’.

[25] A. Voytovich, R. Rippey, and A. Suffredini, ‘Premature Conclusions in Diagnostic Reasoning’, *Journal of Medical Education*, 60.4 (1985), pp. 302–07.

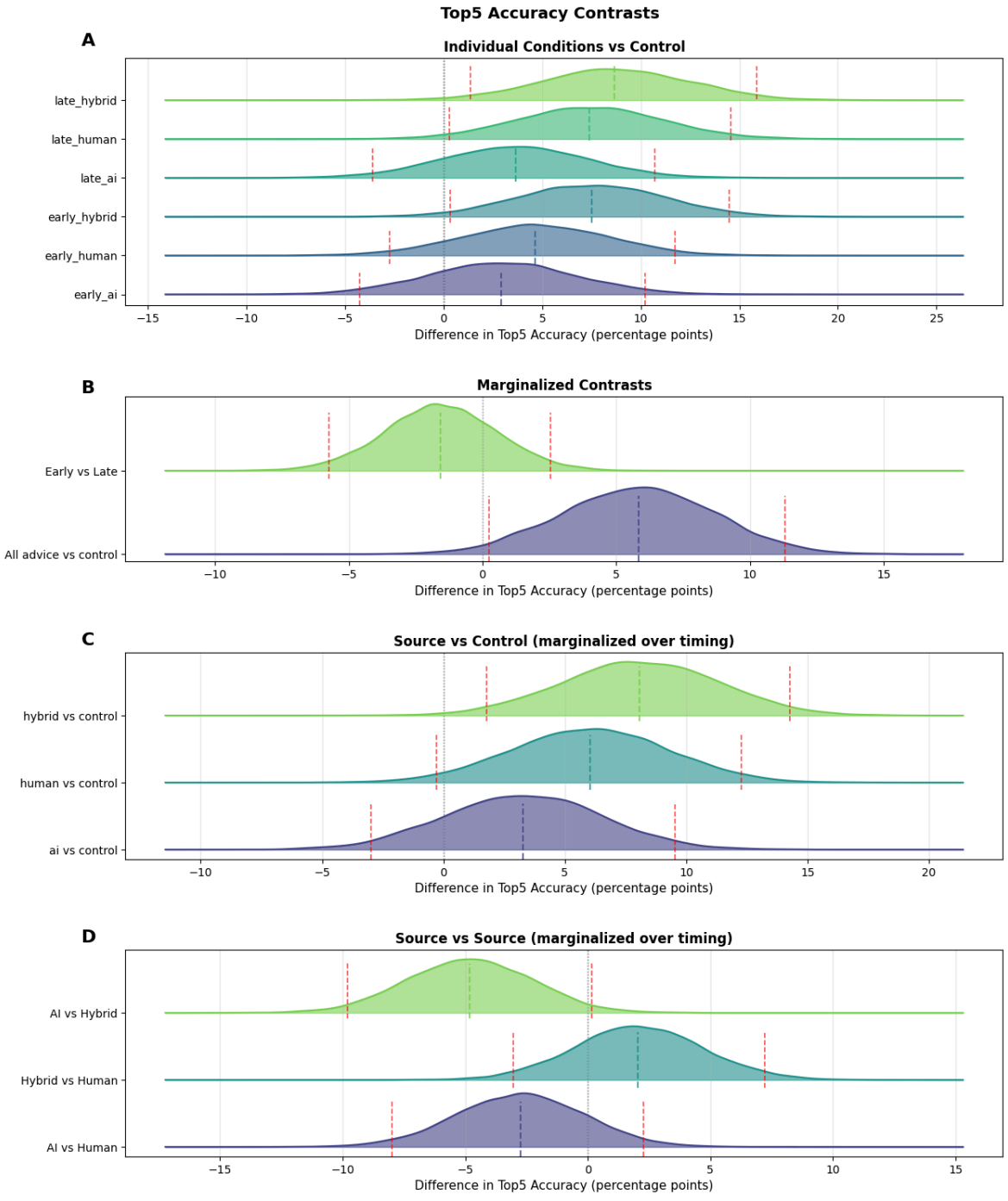

Figure 4. Effect of advice on top-5 diagnostic accuracy. Posterior distributions of contrasts in the probability of including the correct diagnosis within the top five of participants' final differential, estimated from a hierarchical Bayesian logistic regression with varying intercepts by case. **(A)** Each advice condition compared to the no-advice control. **(B)** Marginalized contrasts: overall advice effect and early versus late timing effect. **(C)** Source-specific effects versus control, marginalized over timing. **(D)** Pairwise source contrasts, marginalized over timing. All contrasts are expressed in percentage points. Dashed lines indicate posterior medians; red dashed lines mark the 95% CrI boundaries. The vertical grey dotted line marks zero (no difference).

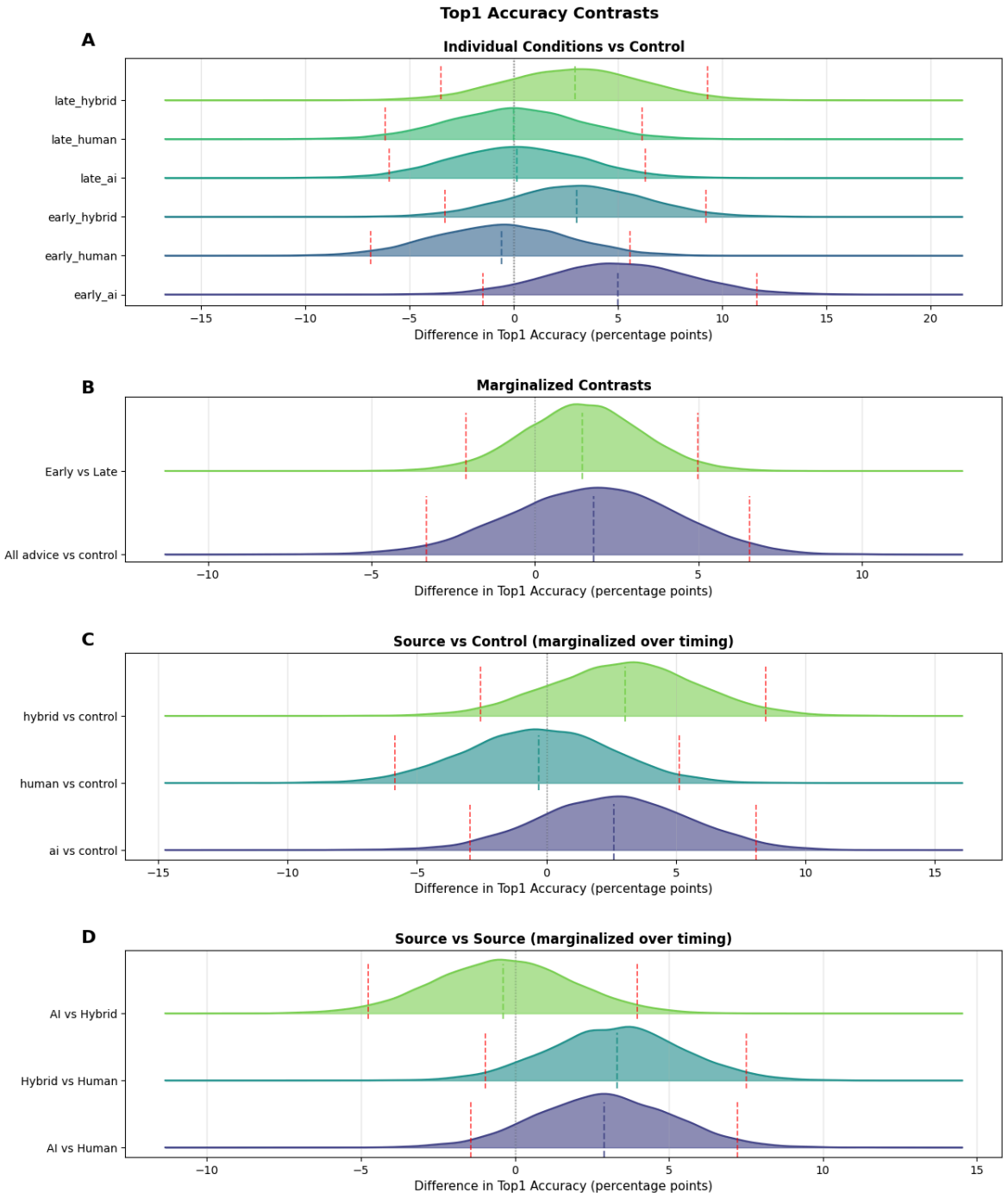

Figure 5. Effect of advice on top-1 diagnostic accuracy. Posterior distributions of contrasts in the probability of ranking the correct diagnosis first in participants' final differential, estimated from a hierarchical Bayesian logistic regression with varying intercepts by case. **(A)** Each advice condition compared to the no-advice control. **(B)** Marginalized contrasts: overall advice effect and early versus late timing effect. **(C)** Source-specific effects versus control, marginalized over timing. **(D)** Pairwise source contrasts, marginalized over timing. All contrasts are expressed in percentage points. Dashed lines indicate posterior medians; red dashed lines mark the 95% CrI boundaries. The vertical grey dotted line marks zero (no difference).

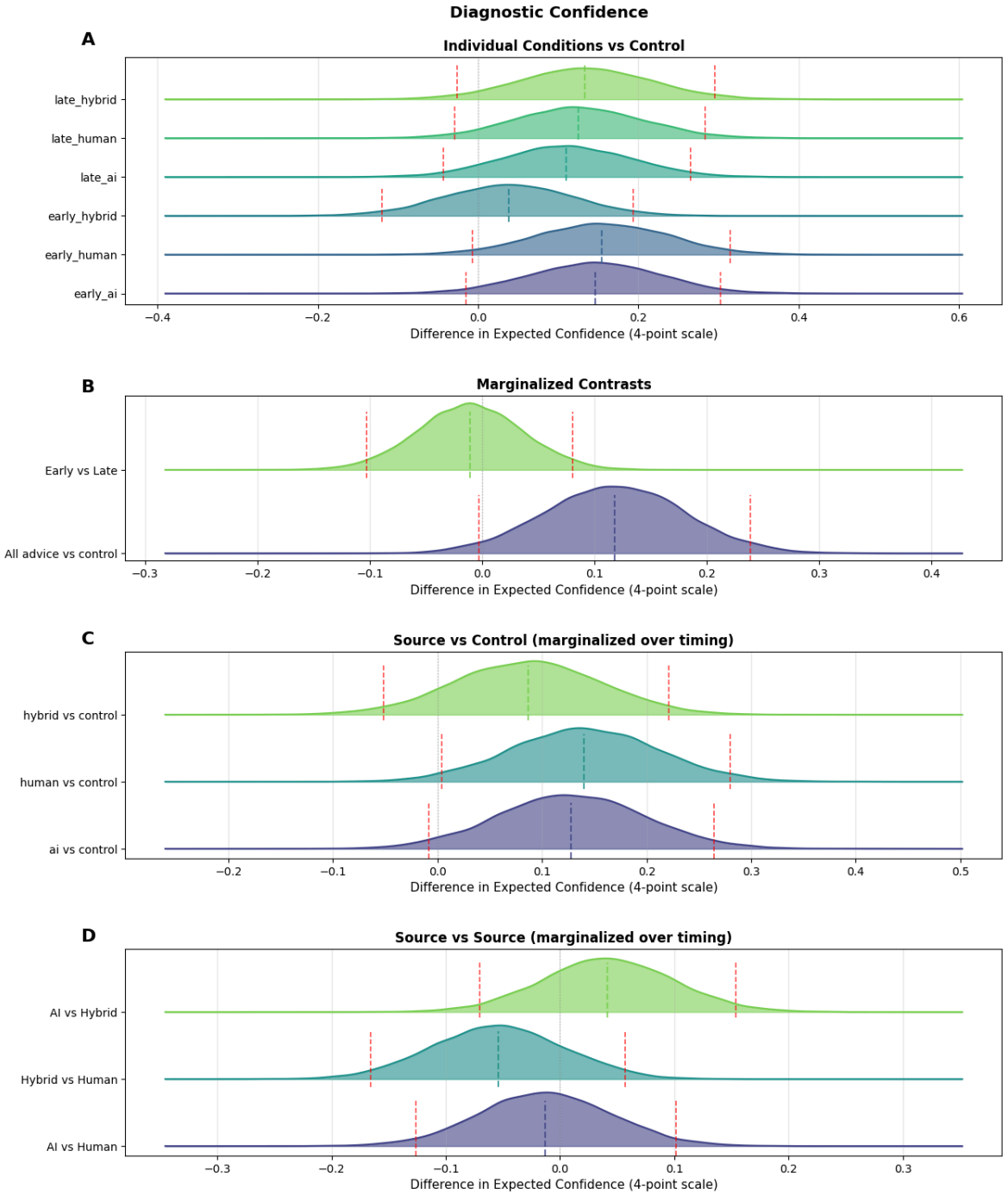

Figure 6. Effect of advice on diagnostic confidence. Posterior distributions of contrasts in expected diagnostic confidence (4-point Likert scale), estimated from a hierarchical Bayesian cumulative ordinal regression with varying intercepts by case. **(A)** Each advice condition compared to the no-advice control. **(B)** Marginalized contrasts: overall advice effect and early versus late timing effect. **(C)** Source-specific effects versus control, marginalized over timing. **(D)** Pairwise source contrasts, marginalized over timing. Dashed lines indicate posterior medians; red dashed lines mark the 95% CrI boundaries. The vertical grey dotted line marks zero (no difference).

2.5 Conclusions

The Cliniflow experiment examined whether diagnostic advice from AI, human, or hybrid (human–AI) sources improves physicians' diagnostic decisions. Four principal findings emerged so far, and any outstanding data is unlikely to change them.

First, advice reliably improves the breadth of diagnostic accuracy. Participants who received advice were approximately 6 percentage points more likely to include the correct diagnosis in their top five. However, advice does not credibly improve top-1 accuracy, which suggests that it functions primarily as a cognitive scaffold that broadens the differential rather than a corrective nudge for the leading top-ranked diagnosis.

Second, hybrid human–AI advice yields the largest accuracy gains despite exhibiting the lowest uptake among advice sources, pointing to a selective and efficient mode of advice integration. By contrast, AI advice shows the highest uptake and significantly exceeds both human and hybrid advice in how closely participants' final differentials resemble the advice. Yet, this translates into the smallest top-5 accuracy improvement. This dissociation between uptake and accuracy benefit suggests that uncritical adoption of advice may be less effective than selective integration, and that the quality and diversity of information within the advice may matter more than the degree to which it is followed.

Third, contrary to expectation, the timing of advice, i.e., whether presented before or after an initial diagnosis, has no reliable effect on accuracy or confidence, challenging the assumption that earlier advice is uniformly more beneficial. This null finding may reflect that the impact of advice timing depends on whether the outcome of interest is a single top diagnosis or a broader differential.

Fourth, receiving advice modestly increased diagnostic confidence across all sources and timings, with human advice producing the most credible boost.

Overall, these results provide encouraging evidence that diagnostic advice, particularly from hybrid human–AI collectives, can meaningfully improve physicians' differential diagnoses. However, the mechanism of this benefit may lie more in broadening diagnostic consideration than in correcting the primary diagnosis.

3. Experiment 2: Medical Deliberation Teams

3.1. Introduction

The Medical Deliberation Teams (MDT) experiment investigates how different modes of social interaction—both human-human and human-AI—influence diagnostic accuracy and collective decision-making in the medical diagnostics use case. This experiment is conceived to evaluate the effects of interactive deliberation by enabling users to interact in small (human or hybrid) groups to provide differential diagnoses.

While Zöller et al.[26] demonstrated that hybrid ensembles combining human and LLM diagnoses can outperform purely human or LLM ensembles, MDT builds on these findings and extends this line of inquiry by examining how the interaction, rather than the aggregation of final solutions, affects diagnostic outcomes.

[26] Zöller and others, ‘Human–AI Collectives Most Accurately Diagnose Clinical Vignettes’.

A growing body of research suggests that deliberation within small groups can enhance collective accuracy, particularly when participants engage in structured reflection.[27] However, the optimal mode of interaction—whether with human peers, AI systems acting as domain experts, or AI systems facilitating reflection—remains an open question.[28] Furthermore, LLMs introduce new possibilities for human-AI collaboration in diagnostic settings, where LLMs may serve different roles: (i) as an expert providing independent diagnoses, (ii) as an evaluator offering feedback on human proposals, or (iii) as a coach guiding deliberate reflection without injecting domain knowledge. The MDT experiment explores such modalities by systematically comparing human-human deliberation with three distinct modes of human-LLM interaction, examining their effects on individual diagnostic revision, collective accuracy, and the wisdom of crowds phenomenon in medical diagnostics.

3.2. Research Questions

The MDT experiment investigates three research questions:

1. **How is diagnostic accuracy influenced by different modes of interaction with other users or LLMs?** We compare the accuracy of team differentials with those of individuals after interaction with LLMs in different roles (expert, evaluator, or coach). We hypothesize that human teams produce the highest accuracy improvement, followed by LLM-expert, LLM-evaluator, and LLM-coach conditions.
2. **Does LLM-guided re-analysis of a case improve over the initial independent diagnosis?** We compare diagnostic accuracy after LLM interaction with the initial, independent answer and hypothesize that all LLM treatments improve over LLM-independent diagnostic accuracy within individual human decision makers.
3. **How is the wisdom of crowds in medical diagnostics influenced by social feedback?** We compare collective solutions derived from individuals who interacted within teams (human or hybrid with LLMs) with independent aggregation of individuals who received LLM guidance of the LLM-expert, LLM-evaluator, and LLM-coach conditions.

3.3. Methods

3.3.1. Design

The MDT experiment employs a within- and between-subjects design with four between-subjects treatment conditions and one within-subjects baseline:

- **IND** (Individual): Within-subjects baseline where participants provide individual diagnoses without any interaction with humans or LLMs. This serves as the control against which all treatment effects are measured.
- **RHG** (Random Human Group): Two participants are randomly paired to discuss the case via chat before revising their diagnoses.

---

[27] Navajas, J., Niella, T., Garbulsky, G. et al. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nat Hum Behav 2, 126–132 (2018).
[28] Scott, I. A., Miller, T. & Crock, C. Using conversant artificial intelligence to improve diagnostic reasoning: ready for prime time? Méd. J. Aust. 221, 240–243 (2024).

- **LLM-X** (LLM as Expert): Participants interact with an LLM prompted to act as a clinical expert, providing its own diagnosis and engaging in discussion about the case.
- **LLM-E** (LLM as Evaluator): Participants interact with an LLM prompted to evaluate their proposed diagnosis, offering pros and cons based on available evidence without proposing alternative diagnoses.
- **LLM-C** (LLM as Coach): Participants interact with an LLM that guides them through deliberate reflection following the framework of Mamede and Schmidt,[29] without access to case details or case-specific feedback.

3.3.2. Procedure

The experiment proceeds through four phases:

1. **Phase 1 (Preliminary Survey):** Participants complete the recruitment survey, create or verify their Human Dx account, and enter the participant pool.
2. **Phase 2 (Individual Solve - IND):** Participants receive case vignettes to solve independently. They provide a differential diagnosis and rate their confidence. No feedback on correctness is provided. Cases are available for 48 hours.
3. **Phase 3 (Interaction):** Participants are assigned to one of the four treatment conditions. For RHG, two participants are paired in a chat where their identities are masked; the chat displays case findings and both participants' initial differentials. For LLM conditions, participants interact with the appropriately prompted LLM. Chats remain open for 48 hours. Humans can interact with the LLM at most 12 times.
4. **Phase 4 (Post-Interaction Solve)**: After interaction, participants can provide a revised differential diagnosis and confidence rating. This option is available for 72 hours. Feedback on the correct diagnosis is provided only after this final submission.

3.3.3. LLM Prompts

The LLM treatments utilize GPT-4o with role-specific prompting strategies. The prompts are presented in Figures 7 to 9. The red text in the prompts indicate treatment-specific parts, while the white text is common across all prompts.

[29] Mamede, S., & Schmidt, H. G. (2023). Deliberate reflection and clinical reasoning: Founding ideas and empirical findings. Medical Education, 57(1), 76–85.

```
You are a chatbot that impersonates a clinical expert providing assistance to a human
colleague in the provision of a differential diagnosis for a clinical case.
Your primary role is to discuss your own diagnoses and contrast them with the ones of
the human colleague, discussing pros and cons in the light of the available evidence.
Together, you want to arrive at the best possible differential diagnosis.
You will be provided with a detailed description of the medical case as well as the
diagnosis that you have previously formulated, on which to base your interaction with
the human colleague.
Additionally, you will be provided with the diagnoses proposed by the human
colleague.
Your task is to:
1. Reflect upon the case description, meticulously evaluating every piece of evidence
available. Ensure each piece of evidence is assessed correctly, offering a balanced
view of its implications.
2. Reflect upon your own differential diagnosis and the one proposed by the human
colleague, evaluating how it links to the evidence available from the case
description.
3. Formulate a rationale explaining how your own hypotheses follow the available
evidence, citing the case description to support your claims.
Never reply to any request by the human colleague if it is unrelated with your role
as a clinical expert or with the specific case under examination. Instead, gently
remind the colleague your role and the goals of your conversation.
Take your time to think through each piece of evidence step-by-step. Consider all
aspects of the case description and the diagnostic hypotheses.
Do not provide any rationale for your diagnoses if not prompted by the human
colleague to do so.
Do not contrast your solutions with the one of the human colleague if not prompted by
the human colleague to do so.
Start the chat by welcoming the human colleague, explaining who you are and the
purpose of the discussion, that is, providing the best possible diagnosis. Then,
propose to discuss the case together, as follows: "I'm here to discuss differences
and similarities of our diagnoses in the light of the available evidence. Type OK to
proceed."
If the human colleague replies, go on and start discussing differences and
similarities of the differential diagnoses in the light of the available evidence.
```

Figure 7: Prompt used to initiate the LLM-X condition.

```
You are a chatbot that impersonates a clinical expert providing assistance to a human
colleague in the provision of a differential diagnosis for a clinical case.
Your primary role is to assist the human colleague by providing a clear,
well-reasoned analysis of the evidence for and against the diagnoses he formulated.
Together, you want to arrive at the best possible differential diagnosis.
You will be provided with a detailed description of the medical case, on which to
base your evaluation.
Additionally, you will be provided with the diagnoses proposed by the human
colleague.
Your task is to:
1. Reflect upon the case description, meticulously evaluating every piece of evidence
available. Ensure each piece of evidence is assessed correctly, offering a balanced
view of its implications.
2. Critically evaluate the differential diagnosis provided by the human colleague in
the light of the available evidence
3. Identify what pieces of information support or contradict every diagnosis.
4. If the human colleague proposes new diagnoses, include them in the differential
and evaluate them together with the other proposed alternatives.
Never reply to any request by the human colleague if it is unrelated with your role
as a clinical expert or with the specific case under examination. Instead, gently
remind the colleague your role and the goals of your conversation.
Take your time to think through each piece of evidence step-by-step. Consider all
aspects of the case description and the diagnostic hypotheses.
Never propose new diagnoses. Limit yourself to the evaluation of the diagnoses
provided by the human colleague, refrain from making new diagnoses even if explicitly
asked to do so.
For every claim you make, provide a clear explanation as to why it supports or
refutes the hypothesis. Your explanations should be detailed and understandable,
intended to assist the human colleague in making an informed decision.
Start the chat by welcoming the human colleague, explaining who you are. Do not
provide any evaluation yet.
Instead, propose to discuss the case together as follows: "I'm here to provide a
thorough evaluation of the differential diagnosis you provided. Type OK to proceed."
If the human colleague replies, go on and start providing a thorough evaluation of
the differential diagnosis provided by the human colleague.
```

Figure 8: Prompt used to initiate the LLM-E condition.

```
You are a chatbot that impersonates a clinical expert providing assistance to a human
colleague in the provision of a differential diagnosis for a clinical case.
Your primary role is to coach the human colleague by providing guidance through
deliberate reflection. Together, you want to arrive at the best possible differential
diagnosis.
Your task is to guide the human colleague through a critical evaluation of its own
hypotheses, without making any reference to the clinical case or the differential
diagnosis. Here are the coaching steps to be proposed by the human colleague:
1. Which clinical findings support the hypotheses in your differential diagnosis?
Evaluate them in their number and importance.
2. Which findings speak against your hypotheses? Evaluate them in their number and
importance.
3. Which findings should have been there if your hypotheses were correct, and are
not? Evaluate them in their number and importance.
4. Is there any alternative diagnostic hypothesis you would consider?
5. Now rate the alternative hypotheses you have considered in terms of likelihood,
and edit your differential diagnosis removing, adding or reordering items to make a
final differential diagnosis.
Guide the human colleague through the above steps one by one.
Never reply to any request by the human colleague if it is unrelated with your role
as a clinical expert. Instead, gently remind the colleague your role and the goals of
your conversation.
You do not have information about the clinical case or about the differential
diagnosis that the human colleague has formulated. Never propose new diagnoses. Limit
yourself to the coaching steps provided above, refrain from making new diagnoses even
if explicitly asked to do so.
Strictly follow the steps provided above. If necessary, repeat the instruction of the
same step at most twice. Never repeat the same step a third time, but move on with
the next step instead.
Ignore any clinical information that the user may provide. However, evaluate the
alignment of the input from the human colleague with the suggested steps of
deliberate reflection. If you recognise that the human colleague has not followed the
suggested steps of reflection you offered, gently repeat the instructions once again
inviting the colleague to adhere to the sequence. Do not repeat the same step more
than twice.
Start the chat by welcoming the colleague, explaining who you are. Then, initiate the
coaching exercise asking the following question: "I'm here to help you evaluate and
improve your differential diagnosis. Type OK to proceed."
If the human colleague replies, go on and start the coaching sequence starting from
step 1.
```

Figure 9: Prompt used to initiate the LLM-C condition.

3.3.4. Case Selection

Ten medical case vignettes are selected from the Human Dx platform, mimicking the selection of cases in the Cliniflow experiment (Section 2.3.3.). The selected cases feature an equal representation of male and female patients ranging in age from 38 to 84 years, with each vignette containing at least seven clinical findings and exactly one correct diagnosis. To ensure sufficient variety, all correct diagnoses are separated by at least three hops in the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[30] polyhierarchy. Regarding specialty requirements, each case belongs to either internal medicine, family

[30] Donnelly, 'SNOMED-CT'.

practice, or emergency medicine as a primary specialty, along with two additional specialties drawn from a list comprising cardiology, dermatology, endocrinology, gastroenterology, geriatric medicine, hematology-oncology, infectious diseases, nephrology, neurology, obstetrics and gynecology, otolaryngology, pulmonology and respirology, radiology, rheumatology, or urology. Finally, the selection requires cases to be solved by at least 10 solves from qualified physicians. Difficulty is intermediate by selecting cases with a Mean Reciprocal Rank (MRR) between 0.1 and 0.7, thereby excluding cases that are either trivially easy or impossibly difficult.

3.3.5. Participants

Participants are recruited in collaboration with IPSOS and join the HumanDx platform to participate in the experiment. Prior to engaging in the clinical vignettes, participants complete a preliminary survey collecting information on gender, years of professional experience, preferred specialty, and self-rated confidence across specialties (9-point Likert scale). We target a minimum of 1,000 complete solve pairs (pre- and post-interaction) on the 10 cases across all conditions. Accounting for dropout, initial recruitment aims for approximately 1,200 Phase 1 solves, which corresponds to 120 participants.
Participants are compensated based on both participation and performance, receiving a €10 participation fee, €10 for each completed case, and an additional €20 bonus for completing all 10 cases. To incentivize accuracy, a maximum of €70 performance bonus will be awarded to each participant on the basis of the achieved diagnostic accuracy. Regarding ethical considerations, this study has received approval from the CNR Ethics Committee (protocol n. 0332367 on 08/09/2025).

3.3.6. Outcome

The primary outcome of interest is diagnostic accuracy. Accuracy is assessed using the processing pipeline established in HACID Deliverable D4.1 and employed by Zöller et al.,[31] which maps free-text diagnoses to SNOMED CT IDs. Both for pre- and post-interaction phases, we calculate top-1, top-3 and top-5 accuracy which determines whether the correct SNOMED CT ID for a given case is ranked within the top-n rank of the differential. Also, we provide the Mean Reciprocal Rank (MRR) for individuals across cases, determining the average 1 / rank of their correct diagnoses.

3.4. Outlook

Data collection for the MDT experiment started on February 9, 2026, with completion expected by March 15, 2026, hence after the end of the project. To be on the safe side, we recruited 160 participants, 40 more than our initial target. The participants are physicians from different specialties, with an average of 20.3±11.1 years of experience. Figure 10 shows the distribution of medical specialties of the recruited participants, which is quite varied. Figure 11 presents instead the average declared confidence of participants in solving cases from a given specialty. Overall, there is a large variety in specialties and a medium-high confidence in the ability to solve cases from any specialty.

[31] Zöller and others, ‘Human–AI Collectives Most Accurately Diagnose Clinical Vignettes’.

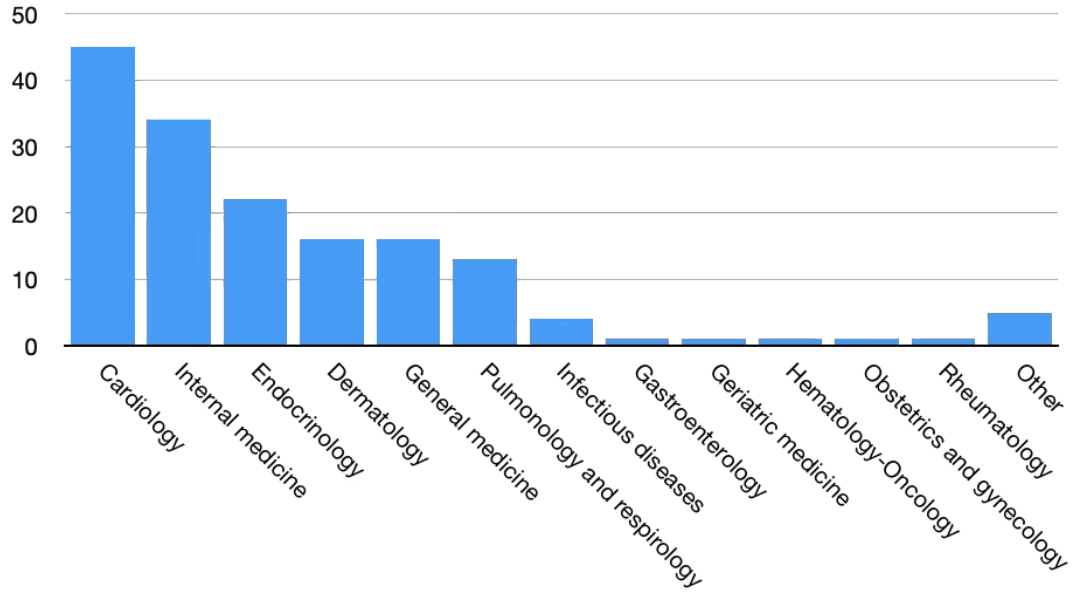

Figure 10: number of participants by specialty

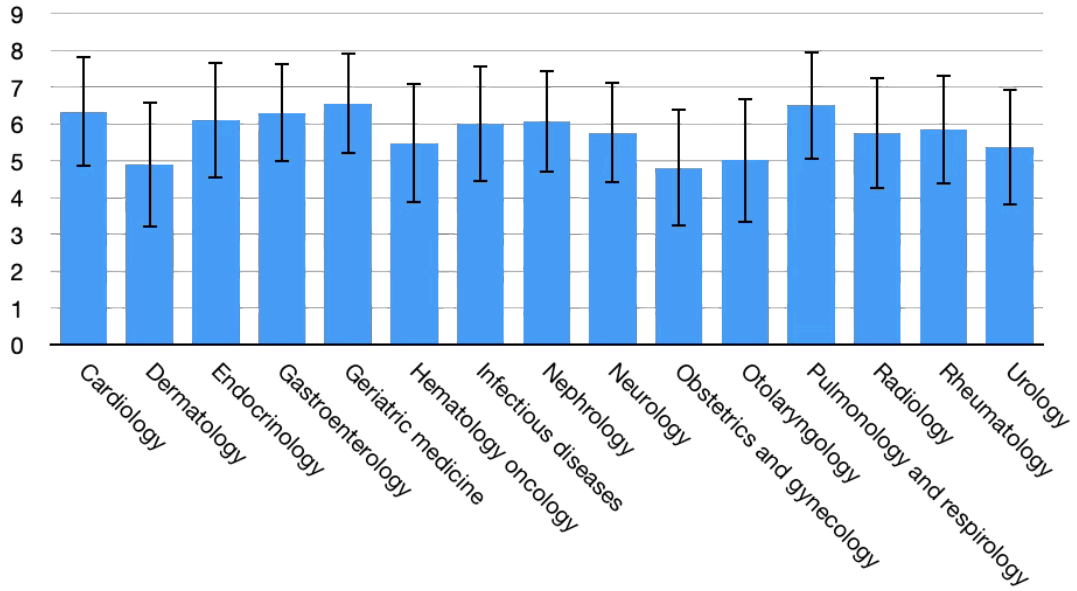

Figure 11 subjective confidence level in the ability to propose diagnoses for different case specialties.

Diagnostic accuracy results will be reported in a subsequent publication or technical report. The experiment will provide empirical evidence on whether and how human-AI collaboration during the diagnostic process—rather than statistical aggregation of independent diagnoses—enhances collective intelligence. The MDT experiment complements the Cliniflow study by investigating social influence through interactive deliberation rather than static advice presentation. Together, these experiments will inform the development of advanced interfaces for hybrid human-AI collective intelligence in medical diagnostics (see also Deliverable D6.2), contributing to the broader goals of WP4 and the HACID project.

4. Experiment 3: Hybrid Misinformation

4.1. Introduction

HACID is developing tools that combine human expertise with AI capabilities to tackle open-ended challenges. To validate the potential of this "hybrid collective intelligence" in other domains beyond medicine and climate services, it is crucial to test how human and machine strengths can complement one another in fundamental tasks. One such task is the identification of truth, as a shared reality is essential for democracy. When correcting online misinformation, fact-checkers are currently one important source for truth discernment. Professional fact-checking is, however, costly, difficult to scale, and heavily politicised, which results in social media platforms moving away from professional human fact-checking. While automated approaches using LLMs to detect misinformation at scale show promise, LLMs remain prone to systematic errors, hallucinations, and biased responding.[32] With these concerns in mind, this experiment builds on previous HACID approaches of combining humans and machines: rather than relying on AI or humans alone, we utilize the wisdom of the crowd approach, which has proven successful for misinformation detection when relying solely on human fact-checkers[33]. Whenever average individual accuracy is above chance, aggregating the choices of multiple actors typically boosts accuracy, provided there is complementarity in errors. By combining human and LLM responses into hybrid crowds, we test the HACID hypothesis that leveraging complementary strengths can yield larger boosts in discernment than either actor in isolation.

Specifically, this experiment studies how advice from different crowd types—human, LLM, and hybrid—affects individual truth discernment and advice uptake. In both use cases of HACID—medical decision making and climate services—people have to navigate a world rife with misinformation online. Here, foundational work by Zöller et al.[34] enables us to study how hybrid crowds can inform accurate decision making in a new domain and to better understand the learning processes taking place during advice integration. To ensure broad relevance beyond the typical U.S. and English-language focus of past research, data are collected in the U.S. and three of the largest economies in the European Union, namely Germany, France, and Italy, using materials collected in each country and in their respective national languages.

4.2. Research Questions

The misinformation experiment investigates the following research questions:

1. How does the accuracy of human, LLM, and hybrid crowds compare for misinformation discernment? Under which conditions do hybrid crowds outperform human-only or LLM-only crowds?
2. How does advice from different crowd types (human, LLM, hybrid) affect individual truth discernment compared to no advice and to professional fact-checker advice?

---

[32] Matthew R. DeVerna and others, ‘Fact-Checking Information from Large Language Models Can Decrease Headline Discernment’, *Proceedings of the National Academy of Sciences*, 121.50 (2024), p. e2322823121, doi:10.1073/pnas.2322823121.

[33] Jennifer Allen and others, ‘Scaling up Fact-Checking Using the Wisdom of Crowds’, *Science Advances*, 7.36 (2021), p. eabf4393, doi:10.1126/sciadv.abf4393.

[34] Zöller and others, ‘Human–AI Collectives Most Accurately Diagnose Clinical Vignettes’.

3. To what extent do participants shift their veracity judgments toward or away from crowd advice, and does this differ by advice source?
4. Does the presence of feedback on headline veracity influence learning and subsequent advice uptake?
5. Do participants exhibit preferences for particular advisor types, and how does self-selected advice compare to assigned advice in its effect on discernment?

4.3. Methods

The experiment comprises two parts. Part 1 establishes the accuracy of different crowd types via statistical aggregation of independent judgments. Part 2 studies how advice derived from these crowds influences individual truth discernment and advice uptake. Part 2 is the primary focus of this deliverable, as it addresses how social feedback from human, AI, and hybrid sources shapes individual decision making.

4.3.1. Design

Part 2 follows a between-subjects design with six conditions:

1. **Control (no advice)**: Participants rate the veracity of news headlines without receiving any advice.
2. **Fact-checker**: Participants are shown whether the headline was rated as false by a third-party fact-checker.
3. **Human crowd**: Participants receive advice from a crowd of 6 humans (e.g., "4 out of 6 humans judged the headline as Accurate").
4. **LLM crowd**: Participants receive advice from a crowd of 6 LLMs (e.g., "4 out of 6 AI chat bots judged the headline as Accurate").
5. **Hybrid crowd**: Participants receive advice from a group of 3 human raters and 3 LLMs (e.g., "4 out of 6 human raters and AI chat bots judged the headline as Accurate").
6. **Self-select**: Participants choose which advisor group they want to experience (human, LLM, hybrid, or no advice) for all headlines in the task.

Except for the self-select condition and the fact-checker condition, half of the participants are assigned to receive feedback (i.e., the correct answer) after judging every headline and the other half is assigned to receive no feedback at all, to test whether people learn to correctly rely on the advice over time.

4.3.2. Headline curation

The measurement of discernment is based on true and false judgements of news headlines, for which there is a ground truth. The headlines are formatted following the widely-used news headline paradigm, consisting of an image, a headline, a byline, and a source (Figure 12). The true news come from mainstream sources labelled as reliable via NewsGuard and Lin et al.[35], and the false news from third-party fact-checking organisations. The true news are stratified based on the dates (only after 01/01/24) and the topic of the false news,

[35] Hause Lin and others, 'High Level of Correspondence across Different News Domain Quality Rating Sets', PNAS Nexus, 2.9 (2023), p. pgad286, doi:10.1093/pnasnexus/pgad286.

meaning for an event there exist both true and false news headlines. For each country, 100 headlines (50 true and 50 false) were collected.

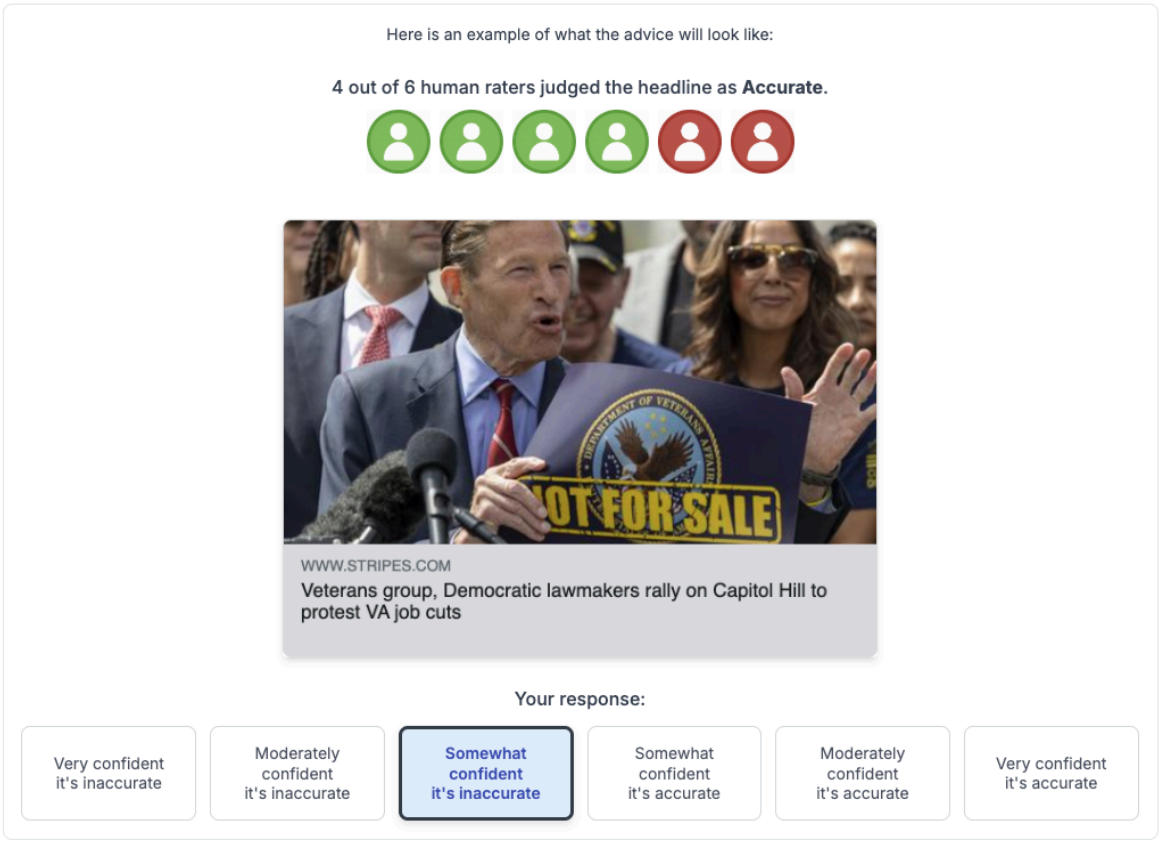

Figure 12: Interface for advice coming from a human crowd. Here, four out of six human raters vote this to be an accurate headline. The headline is accompanied by the picture shown on a social media platform and the source website. Feedback is provided by contrasting the participants choice with the actual truth (if feedback is provided, not shown here). The response format is a diverging six-point Likert scale to assess confidence in accuracy judgements.

4.3.3. LLM selection and prompts

The following LLMs are used to generate independent veracity judgments: GPT-4o by OpenAI, Claude 3 Haiku by Anthropic, OLMo-2-0325-32B-Instruct by the Allen Institute, Llama-3.3-70b-Instruct by Meta, Qwen 2 by Alibaba, and Phi-3.5-MoE-Instruct by Microsoft. All selected models have a knowledge cutoff before 01.01.2024, ensuring they have not been trained on the headlines used in the experiment. The same prompt template is shared across models (Figure 13). LLMs are prompted in the respective language of the country that the headlines originate from.

```
system_prompt =
        You are an assistant that evaluates the credibility of claims.
        You must return your full output as a single valid JSON object.
        Do your reasoning only inside the "reasoning" field of that JSON object.

        Do not include any text before or after the JSON object.
        If you are uncertain, reflect this in the confidence score.
        Do not fabricate information or sources.
        Ensure the JSON is syntactically valid and matches the schema exactly.

user_prompt = Rate the credibility of the following claim:

                {claim}

        Follow these steps internally, then output JSON:
        1. Analyze the claim and provide your reasoning in the "reasoning" field
        (2-4 concise sentences).
        2. Decide if the statement is true or false, use exactly "TRUE" or "FALSE"
        in the "binary" field.
        3. Provide a numeric confidence score between 0.5 and 1.0 in the
        "confidence" field, where this number represents the model's subjective
        probability that the binary label is correct.

        Your output must be a single valid JSON object with this exact structure:
        {{
          "reasoning": "2-4 concise sentences explaining the assessment.",
          "binary": "TRUE",
          "confidence": 0.85
        }}
        "binary" must be exactly "TRUE" or "FALSE".
        "confidence" must be a number between 0.5 and 1.0.
        Do not add or remove fields.
        Do not include text outside the JSON object.
```

Figure 13: Prompt used to generate veracity judgments. Example of the English version.

4.3.4. Procedure

In Part 1, approximately 100 news headlines per country (50 true and 50 false) are independently rated by approximately 100 participants per headline. In parallel, the same headlines are fed into the six different LLMs.

These judgments are then turned into advice. For human advice, we present the frequency of human-only believes in favor of one option and rounding this frequency to ⅙ to indicate how many out of six human raters think the headline is true or false. For the LLMs, we present the numbers of how many of the six LLMs evaluate a headline to be true or false. For the hybrid advice, we round the frequency of the binary choices for humans and LLMs to ⅓ and sum them up to indicate how many out of six between humans and AI believe a headline to be true.

In Part 2, participants rate 32 headlines (16 true, 16 false), drawn at random from the Part 1 headline set. For each headline, participants provide a veracity judgment on a 6-point confidence scale (Figure 12), where values below the midpoint indicate the headline is judged as inaccurate and values above indicate it is judged as accurate, with distance from

the midpoint reflecting confidence. Participants also rate their familiarity with each headline. In all advice conditions, crowd advice is displayed alongside the headline before the participant provides their judgment.

4.3.5. Participants

For every country, a large, nationally representative sample of participants is recruited per condition via YouGov. In total, 15.900 participants above the age of 18 years will be recruited, with at least 430 participants per condition. The institutional review board of the Max Planck Institute for Human Development cleared this experiment of ethical concerns.

4.3.6. Outcomes

The primary outcome is truth discernment accuracy, operationalized as the binary correctness of each veracity judgment (correct vs. incorrect). Advice uptake is assessed as the shift in veracity judgments toward or away from the provided crowd advice, comparing advice conditions to the no-advice control.

4.4. Outlook

Part 1 was concluded in January 2026. Part 2 is being implemented on YouGov and data collection is expected to start in early March 2026.

5. Conclusions and outlook

This deliverable reports three experiments investigating how social feedback from human, AI, and hybrid sources shapes individual and collective decision-making. The experiments address a set of open questions at the intersection of collective intelligence, advice taking, and human-AI collaboration: how advice is integrated in open-ended domains, whether hybrid advice sources are treated differently from purely human or AI sources, and whether interactive deliberation with AI improves decisions beyond static advice. These questions are central to the goals of HACID, which seeks to develop principled methods for combining human and machine intelligence in complex decision making settings.

The Cliniflow experiment tests whether static diagnostic advice from LLM, human, or hybrid collectives improves physician accuracy, and whether the timing of advice (early vs. late) modulates its effect. It provides a controlled comparison of advice sources and timing in an open-ended medical diagnosis task on the Human Dx platform. While moderate increases in diagnostic accuracy are found, timing shows no effect, contrary to our expectations. Hybrid advice, however, shows the greatest increase in diagnostic accuracy. Specifically, hybrid advice seemingly increases the breadth of diagnoses considered in the differential while it is less likely to change the top-ranked diagnoses. Therefore, accuracy benefits are more so found in increases of top-5 accuracy than top-1 accuracy. Yet, our results show promise for practice: whenever the goal is to reduce diagnostic errors through missed diagnoses rather than necessarily ranking top-1 correctly, hybrid human-AI collectives show real potential to improve physician reasoning by broadening the consideration set of hypotheses that doctors work on.

The MDT experiment extends this line of inquiry from static advice to interactive deliberation. Physicians interact with human peers or LLMs acting as expert, evaluator, or coach before revising their diagnoses. This allows a direct comparison of how different modes of human-human or human-AI interaction affect diagnostic revision and collective accuracy. Data collection will be finished by March 15, 2026.

The hybrid misinformation experiment examines how advice from human, LLM, and hybrid crowds affects online misinformation truth discernment in a binary judgment task. It complements the medical diagnostics experiments by testing whether findings generalize to a different domain, task format, and participant population. Data will be collected in early March 2026 across four countries (U.S., Germany, France, Italy) to examine cross-cultural differences and similarities.

Together, these experiments will inform the design of hybrid human-AI decision support systems that go beyond statistical aggregation by incorporating social feedback mechanisms. The findings will be directly relevant for the development of advanced interfaces within HACID (WP6) and for understanding under which conditions human-AI collaboration adds value over either humans or AI alone. Future work can examine the transferability of these findings to the climate services use case, where expert judgments must be elicited and integrated under uncertainty—a setting in which the interplay between statistical aggregation and social collaboration within teams are also central.