

Using the Delphi Method to Strategize about Health AI

Whitney Welsh, Ph.D.^{1,2}, and Shelley Rusincovitch, MMCi, FAMIA^{1,3}

¹Duke AI Health, ²Duke Social Science Research Institute, ³Duke Clinical and Translational Science Institute

Introduction

The **Delphi method** is an iterative, group-based process for exploring whether a consensus exists on a given topic. A panel of subject matter experts completes multiple rounds of questionnaires, with the chance to change their responses in each round based on anonymized feedback about how their responses compare with the group's and the reasoning that other participants provide for their choices (Keeney, Hasson & McKenna 2011; Rowe & Wright 1999). The method is best suited to investigate "issues about which uncertain or incomplete data exists" (Neiderberger & Renn 2023).

Artificial intelligence in health is a rapidly evolving field, and expert opinion is an especially valuable resource while common knowledge and best practices are still uncodified. To probe for consensus on the barriers and facilitators of innovation in Health AI, we deployed the Delphi method during the Duke Summit on AI for Health Innovation. Held over three days in October 2024, the in-person summit brought together experts from the fields of engineering and health in order to foster a community of practice around health-oriented AI development. Representatives from industry, medicine, academia, and funders came together to discuss the current and future landscape of Health AI development and innovation, providing an ideal opportunity for a Delphi study on this topic.

Methods

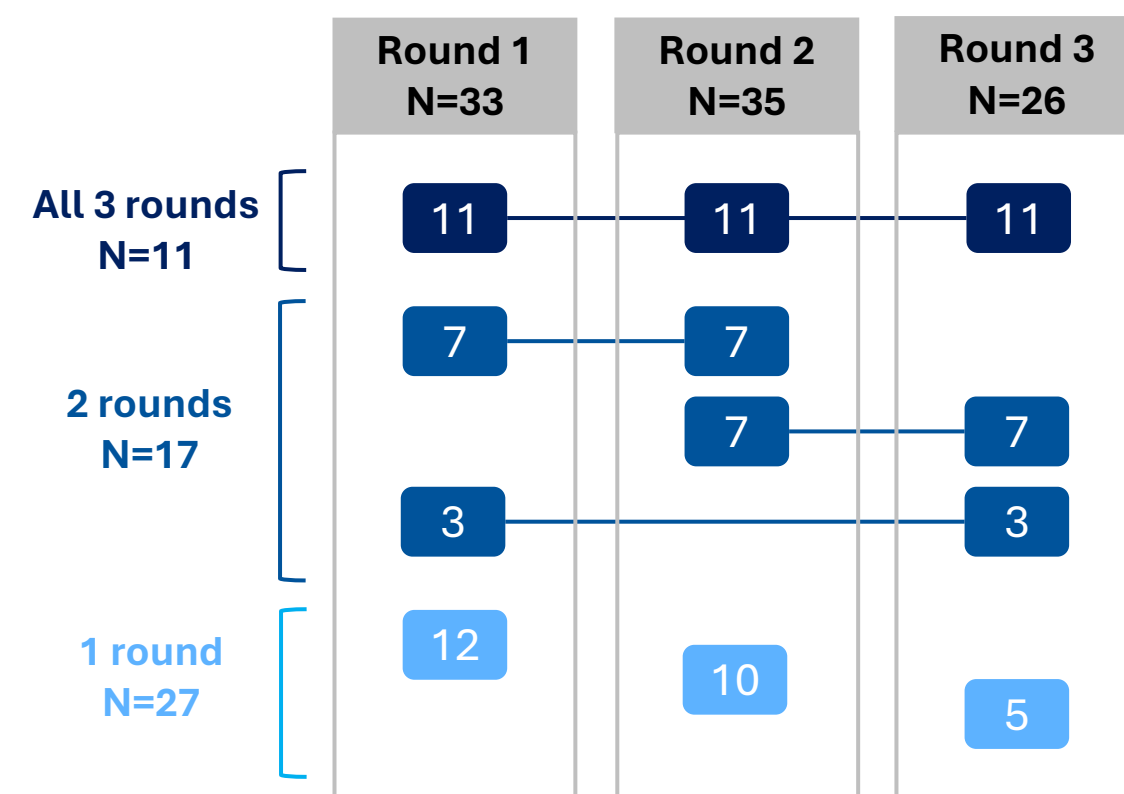
We employed a classical Delphi technique with three rounds of questionnaires. In the **round 1** survey, distributed the week before the event, summit registrants were asked to complete an open-ended survey to generate a list of statements for use in later rounds. We asked three questions:

1. What is the greatest barrier to innovation in Health AI?
2. Which is the most needed training or skillset that people in Health AI are not getting, or not getting enough of, currently?
3. Where would implementing AI result in the most significant impact on productivity?

For the **round 2** survey, completed on the first day of the summit, participants were asked to rate each of the responses generated in round 1 as very, somewhat, or not significant. They were also asked to indicate which of the responses was their top choice and provide their reasoning. For the **round 3** survey, completed on the second day of the summit, participants were given anonymized feedback about how the group rated the statements and the reasons for their top choices. Participants were then asked to re-rate the statements, and again select their top choice and provide their reasoning.

We invited all summit registrants on our list at the time of survey distribution (126 for round 1, 138 for rounds 2 and 3) to participate in each of the three rounds. 55 people participated in at least one round. **Figure 1** shows number of participants by which round(s) they participated in.

Figure 1: Participation across three rounds



Participants were assigned to engineering, health, or unknown categories based on their professional affiliation, degree type, and area of work. **Engineering** comprises primarily engineers, computer scientists, data scientists, and informaticists. **Health** comprises primarily clinicians, health researchers, and health policy specialists. About the same number of engineering and health participants took part in each round (**Table 1**).

Table 1: Participants by type and round

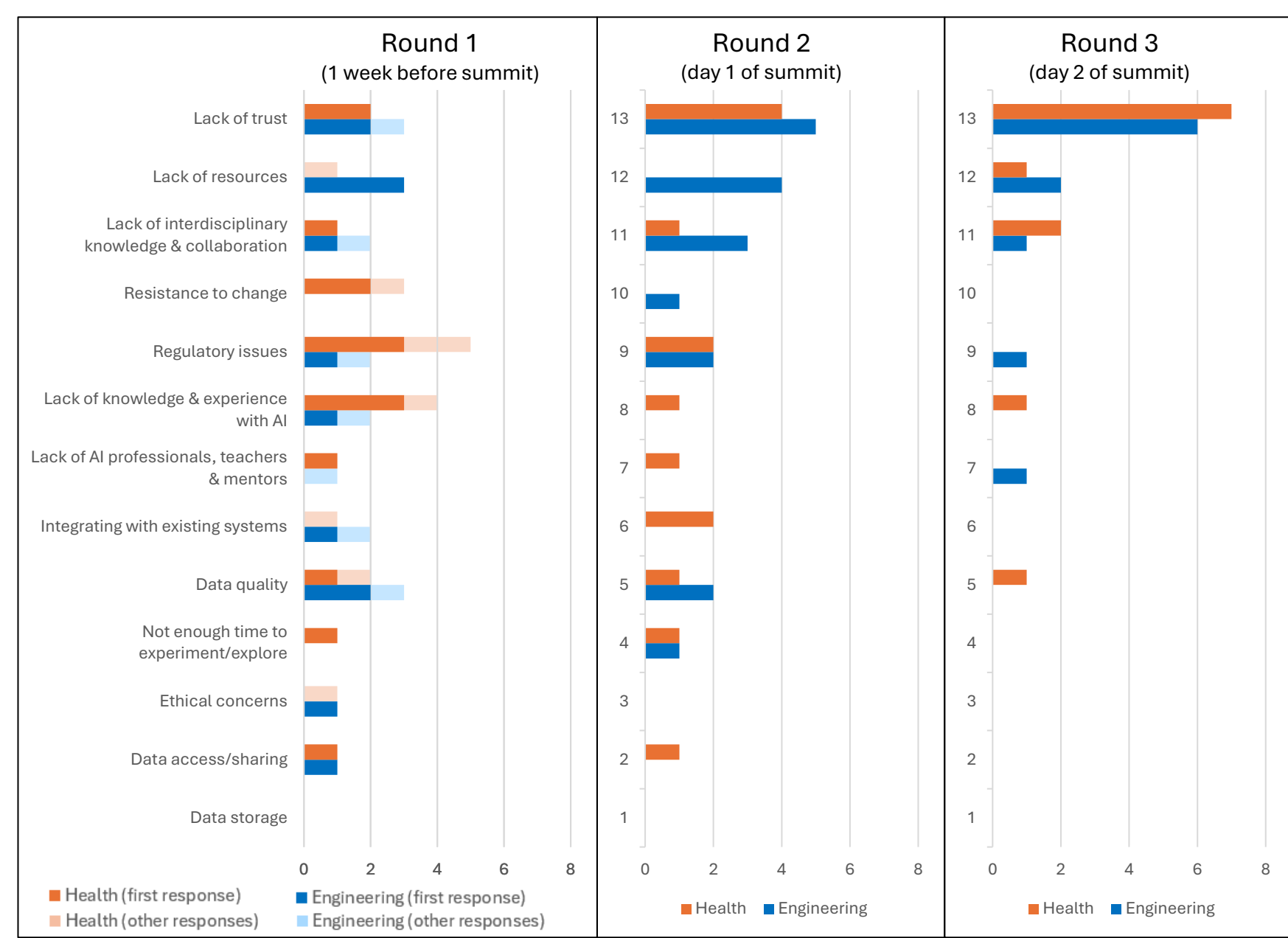
	Engineering	Health	Unknown	Total
Total	25	26	4	55
all 3 rounds	5	5	1	11
only 2 rounds	8	7	2	17
only 1 round	12	14	1	27
Round 1	13	17	3	33
Round 2	19	14	2	35
Round 3	11	12	3	26

References

Keeney, Sinead, Felicity Hasson, and Hugh McKenna. 2011. *The Delphi Technique in Nursing and Health Research*. Oxford: Wiley Blackwell.
 Neiderberger, Marlen, and Orwin Renn. 2023. *Delphi Methods in the Health and Social Sciences: Concepts, Applications, and Case Studies*. Springer.
 Rowe, Gene and George Wright. 1999. "The Delphi technique as a forecasting tool: issues and analysis." *International Journal of Forecasting*, 15:353-75.

Barriers to innovation in Health AI

Figure 2: Which is the most significant barrier to innovation in Health AI?



"As a clinician I need to trust health AI's accuracy and transparency to feel confident using it in patient care. Without that trust, adopting AI in critical decisions becomes a significant challenge."

Lack of trust: "If the public doesn't know what or why the AI is recommending as diagnosis or treatment they will reject the diagnosis and treatment and want a human to redo and deliver the message."

Lack of resources: "We have more AI-related ideas and need for assisted processes than individuals who can set aside time from their core job responsibilities to focus on them."

Lack of resources: "Not enough funding to consider moving things beyond proof of concepts into scale."

Interdisciplinary knowledge: "Collaboration is key to adoption."

For AI to be effectively and equitably integrated into health care, interdisciplinary expertise is essential to include diverse viewpoints and skills for development, validation, and implementation."

Figure 5: How significant of a barrier is each of the following to innovation in Health AI?

Barrier	Total (n=35) Round 2	Total (n=35) Round 3	Engineering (n=10) Round 2	Engineering (n=10) Round 3	Health (n=14) Round 2	Health (n=12) Round 3
2. Regulatory issues	74%	74%	70%	70%	79%	79%
7. Data access/sharing	71%	71%	60%	60%	83%	83%
6. Data quality	63%	63%	60%	60%	71%	71%
8. Lack of trust, confidence in results	60%	60%	60%	60%	64%	64%
9. Integrating with existing systems	57%	57%	50%	50%	64%	64%
6. Lack of interdisciplinary knowledge & collaboration	51%	51%	54%	54%	57%	57%
9. Ethical concerns	51%	51%	50%	50%	57%	57%
10. Resistance to change	49%	49%	50%	50%	47%	47%
5. Lack of knowledge & experience with AI	46%	46%	50%	50%	47%	47%
11. Lack of AI professionals, teachers & mentors	40%	40%	46%	46%	37%	37%
12. Not enough time to experiment/explore	34%	34%	27%	27%	26%	26%
13. Data storage	23%	23%	8%	8%	20%	20%

Results

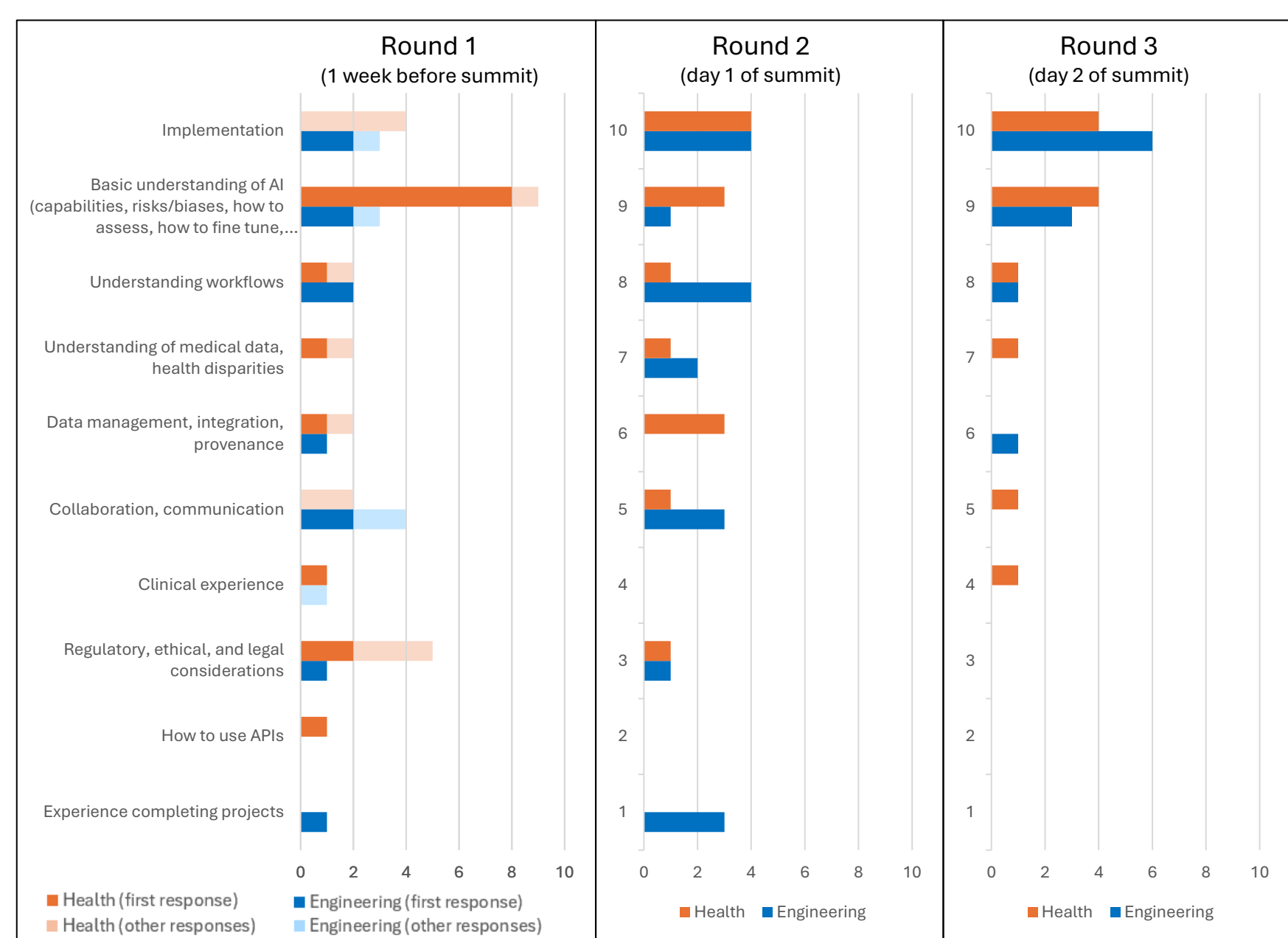
Top choice
Figures 2-4 show participants' choice of the single-most significant option for each of the three rounds. In round 1, participants were asked for their most significant response, but some listed multiple items. For the questions about barriers and impact, a clear consensus emerged around **trust** and **documentation**, respectively. For the question about needed skills, participants consolidated around two options, **implementation** and **basic understanding of AI**. Across all three questions, the engineering and health participants clustered around different options in rounds 1 and 2, but largely converged on the consensus choices in round 3.

Of the participants who took part in more than one round, 75% changed their answer to the barriers question, 79% to the impact questions, and 88% to the needed skills question. Even within the consensus categories, 38% of participants who chose lack of trust, 82% of those who chose implementation, and 53% of those who chose documentation in round 3 changed their responses at least once (56%, 90%, and 67%, respectively, of those who participated in more than one round). The participants who changed their response to the needed skills and impact question were fairly evenly split between engineering and health. However, only about half of the engineering participants changed their response to the barriers question, compared to 90% of the health participants.

Everyone who participated in multiple rounds changed their top choice on at least one of the three questions. Unfortunately, most of the participants who changed their response did not provide a reason for their new choice. There were no clear patterns in the way people changed their responses, beyond trending toward the consensus choice. Interestingly, the reasons given for both the barriers and needed skills responses (including those other than implementation) tended to relate to implementation.

Needed skills that people in Health AI are not getting, or not getting enough of, currently

Figure 3: Which is the most needed training or skillset that people in Health AI are not getting, or not getting enough of, currently?



"If a solution is not properly implemented or faces resistance, the solution, no matter how useful, will not be effective."

"Providers need not be expert AI scientists but they will need to be experts at delivering the message from the AI and implementing all the follow up steps."

Basic understanding: "AI is only as useful as the knowledge behind the use. The better informed, the more ingrained and adopted folks will become."

Understanding of how to use AI and adjust the use will help with the implementation. I see this basic understanding as going hand in hand with implementation."

Clinical experience: "Patients are all very different and have different family history, environmental risk factors, health literacy, compliance, etc. and its just not something we keep stored in EHRs that must be considered to make decisions."

Figure 6: How needed is each of the following training or skillsets for people in Health AI?

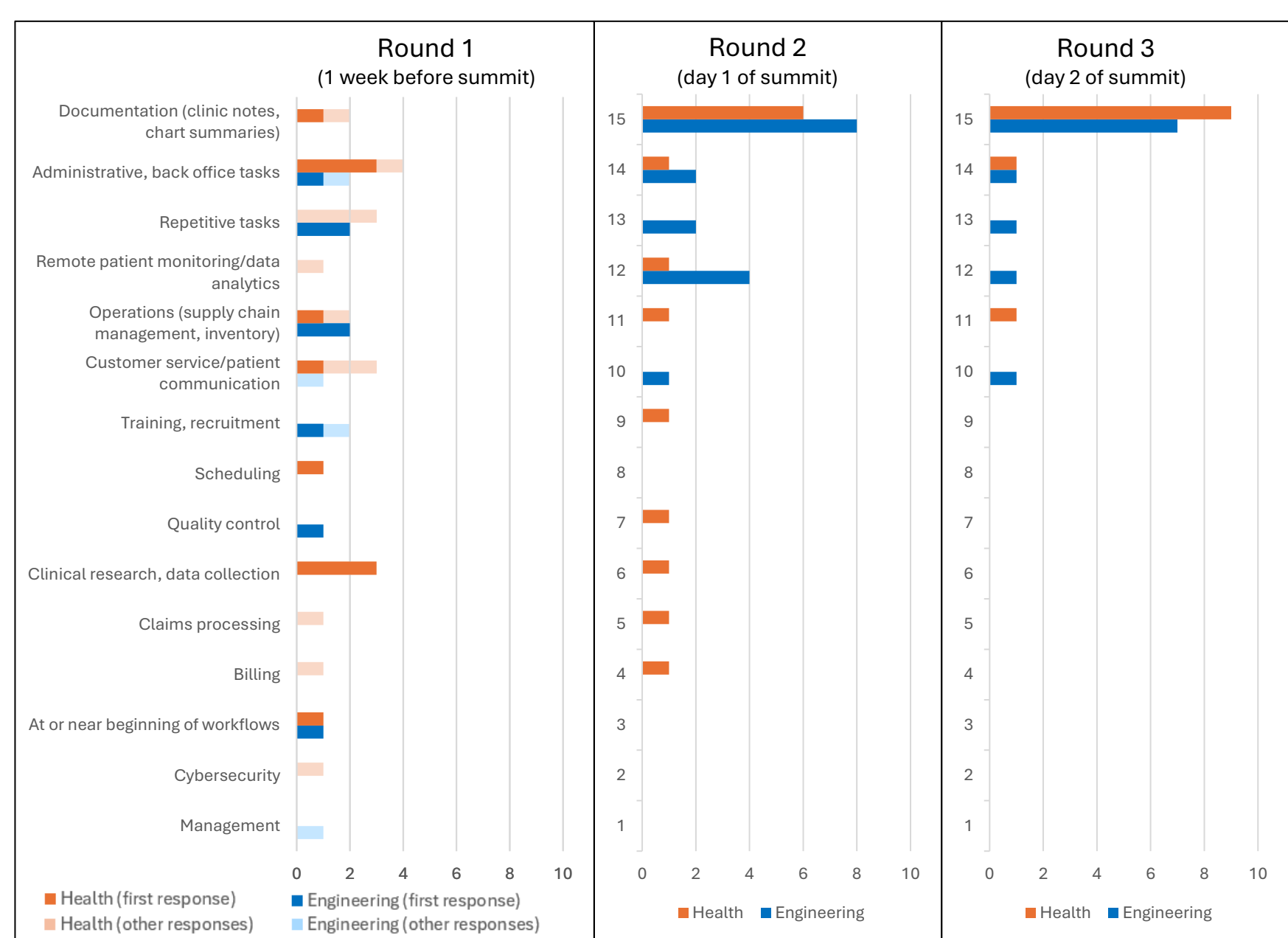
Skillset	Total (n=35) Round 2	Total (n=35) Round 3	Engineering (n=10) Round 2	Engineering (n=10) Round 3	Health (n=14) Round 2	Health (n=12) Round 3
2. Collaboration, communication	69%	69%	60%	60%	79%	79%
4. Basic understanding of AI	63%	63%	60%	60%	79%	79%
1. Implementation	60%	60%	60%	60%	79%	79%
3. Understanding workflows	57%	57%	60%	60%	71%	71%
5. Regulatory, ethical, and legal considerations	54%	54%	50%	50%	71%	71%
9. Experience completing projects	54%	54%	50%	50%	64%	64%
6. Data management, integration, provenance	51%	51%	54%	54%	64%	64%
8. Clinical experience, expertise	49%	49%	50%	50%	64%	64%
10. How to use APIs	23%	23%	12%	12%	36%	36%

Rating each response
Figures 5-7 show the ranked order of the responses for each question by participant type, based on the percent of participants who rated each response as very significant in rounds 2 and 3. The health and engineering rankings look more similar in round 3 than round 2 for the questions about needed skills and impact, suggesting convergence. However, for the barrier question, the round 2 rankings are more alike than the round 3 rankings.

In round 3, the order is not precisely the same for the engineering and health participants, but the top-tier, middle-tier, and bottom-tier are fairly consistent for the questions about needed skills and impact. For the question about barriers, however, there is quite a bit of variation. This is because, for the questions about gaps and impacts, most of the statements saw an increase in very significant ratings (compare the red-shaded columns in Figures 5-7) from round 2 to round 3 among both engineering and health participants. However, for the question about barriers, there was an increase in very significant ratings among the health participants, but a decrease on most statements among the engineering participants, suggesting that they became more optimistic about overcoming these barriers as the summit went on.

Where to implement AI for maximum impact on productivity

Figure 4: Where would implementing AI result in the most significant impact on productivity?



Documentation: "Biggest problem we hear about with regard to burnout"

"In the near term, automating administrative tasks like documentation will free up clinicians to focus on patient care."

Documentation: "Helps clinicians regain time and input less effort to focus more on patient care and clinical outcomes."

Clinical notes, documentation and other works of authorship will be low hanging fruit for AI to make the fastest and highest impact."

Back office tasks: "Healthcare is riddled with inefficiencies and redundancies. Thus, removing or streamlining simple tasks will allow people to focus on the empathic, human, and soft-skills necessary in healthcare."

Figure 7: How significant would the impact on productivity be if an organization implemented AI in each of the following places in their business models or workflows?

Location	Total (n=35) Round 2	Total (n=35) Round 3	Engineering (n=10) Round 2	Engineering (n=10) Round 3	Health (n=14) Round 2	Health (n=12) Round 3
1. Repetitive tasks	83%	83%	84%	84%	80%	80%
2. Documentation (clinic notes, chart summaries)	71%	71%	68%	68%	79%	79%
3. Administrative, back office tasks	66%	66%	77%	77%	71%	71%
4. Operations (supply chain management, inventory)	60%	60%	58%	58%	71%	71%
6. Clinical research, data collection	54%	54%	59%	59%	68%	68%
7. Remote patient monitoring/data analytics	54%	54%	59%	59%	57%	57%
8. Claims processing	54%	54%	59%	59%	68%	68%
9. Scheduling	51%	51%	65%	65%	57%	57%
5. Billing	49%	49%	58%	58%	70%	70%
11. Cybersecurity	37%	37%	54%	54%	42%	42%
10. Quality control	31%	31%	35%	35%	30%	30%
12. At or near beginning of workflows	31%	31%	35%	35%	29%	29%
13. Customer service/patient communication	26%	26%	23%	23%	20%	20%
15. Training, recruitment	20%	20%	19%	19%	14%	14%
14. Management	6%	6%	15%	15%	0%	0%

Discussion

Overall, the Delphi method was successful in ascertaining consensus on barriers and facilitators of innovation in Health AI, and moreover, consensus among two distinct spheres of expertise, engineering and health. We were especially interested to see the evolution of responses. The summit was organized into a single track, so we expected that participants would generally be exposed to similar experiences during the event.

One area where the two groups did diverge over time was on the relative importance of the barriers. Engineering participants perceived the barriers as less significant over the course of the conference, in contrast to the health participants. At the same time, health participants posted a large increase in the need for clinical experience from round 2 to round 3. Taken together, it raises the question of whether the engineers' lack of clinical experience may be leading them to underestimate the barriers, though more research is needed to understand the precise dynamics at play.

Interestingly, there were also some differences in the highest ranked response depending on whether participants were asked to select a top choice or rate each item. Notably, while implementation scored the highest among the ranked answers for needed skills, trust was the top choice for barriers among health participants but not engineering, and documentation was third and fourth choice, respectively, for impact. Rather, repetitive tasks, which barely registered when participants were asked to choose the most significant impact, was consistently at the top when they were asked to rate each response.

This study offers a number of avenues for future research. For instance, it would be interesting to see if the results can be replicated with other groups, or if this group was correct in the barriers, needed skills, and impacts it identified. It would also be informative to probe further into why participants changed their responses. Another area of future work is to further refine our implementation of Delphi method to further actively engage participants in reflecting upon their changes of perspectives and new learnings during the duration of an event, and to design our learning experiences with the benefit of these more iterative feedback loops.

