Clinical Collabsheets: 53 Questions to Guide a Clinical Collaboration

Shems Saleh* SHEMS.SALEH@VECTORINSTITUTE.AI
Vector Institute
Toronto, ON, Canada

William Boag* WBOAG@MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA, USA

Lauren Erdman* LAUREN.ERDMAN@SICKKIDS.CA
University of Toronto, The Hospital for Sick Children, Vector Institute
Toronto, ON, Canada

Tristan Naumann TRISTAN@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Abstract

Clinical Machine Learning (ML) is a rapidly-growing field due to the digitization of hospital records, recent advances in ML techniques, and the ability to leverage increasing computational power for large and complex models. The high stakes and often unintuitive nature of clinical data make effective collaboration between clinicians and ML researchers one of the most important aspects of working in this interdisciplinary space. However, there are few resources codifying best practices for collaboration on Clinical ML projects. In this paper, we interviewed 18 experts in the Clinical ML field and distilled their advice and experiences into a list of questions (a *Helathcare Collabsheet*) ML scientists and clinicians can use to promote effective discussion when working on a new project. We intend this for a broad audience as checklist of discussion points to hit at a kickoff meeting, even for experienced researchers. This resource will enable more successful partnerships in Clinical ML with improved interdisciplinary communication and organization.

1. Introduction

Communication between interdisciplinary team members is critical and often an overlooked factor in the success of collaborations. This document helps facilitate collaborative Clinical Machine Learning (ML) work by providing many starting points of discussion through a series of questions and examples. Addressing the problem of miscommunication between clinical and ML collaborators is challenging because both the projects and individuals involved are diverse and deeply embedded in the context of the problem they are trying to solve. Adding to this challenge, is the fact that best practices and pitfalls of existing

approaches are constantly being identified in different subspecialites of ML (e.g. Fairness) and even ML collaborators may not be aware of how to practically integrate these findings into a clinical collaboration.

1.1. Motivation

Using tools that are safe and effective is critical in high-stakes domains such as healthcare. Dr. Robert Watcher's book, *The Digital Doctor* surveys many problems that have come about as the result of the imperfect deployment of Electronic Health Records (EHRs), including: the propagation of stale/incorrect patient information, harmful interruptions, alert fatigue, and even a specific case where a miscommunication at the human-computer interface resulted in a 39x overdose (Wachter, 2015). One major problem with using ML for patient care is that ML models can overlook the atypical, infrequent, and less supported cases in the data that might need the most attention. This is exacerbated by "hidden stratification" of clinical data, such as when a label like "pneumothorax" is an umbrella term that lumps together *treated* pneumothorax and *untreated* pneumothorax resulting in a model which performs worse on the rarer-but-more-severe untreated ailment (Oakden-Rayner* et al., 2019).

When working with a hospital, it is not obvious what questions are important to ask in order to help the team get on the same page and develop clear expectations. As observed in the book *The Checklist Manifesto*, good communication is one of the most important ingredients when handling complex situations, be it a surgery (Gawande, 2009), or in this case, a new collaboration.

1.2. Our contribution

The primary contribution of this work is to provide a checklist of questions for fostering collaborative conversations between data scientists and clinical domain experts. Specifically, providing a resource for researchers who are new to the field to use as a foundation for best practices when starting a collaboration and a reference for more experienced researchers. Some checklists (especially the in-the-moment ones used in hospitals and airplanes) are designed to be followed literally step-by-step while others (such as this one) tries to ensure nothing slips through the cracks. Crucially, however, we emphasize that checklists do best when they are taken seriously but not literally; a mindless "box ticking" approach will likely lead to zoning out and providing little value. This resource cannot drive the conversation start-to-finish, but it can help ensure the research team is on the same page. It acts as a basis for collaboration that avoids common pitfalls but by no means does the list of questions exhaustively capture all of the possible information needed for a collaboration.

As a secondary goal, we believe that these questions can serve as a helpful guide for clinicians to understand questions that data scientists want and need to know. There are many publicly available medical datasets, including MIMIC (Johnson et al., 2016), Philips eICU (Johnson et al., 2017), MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin

et al., 2019), and more. Much like Datasheets for Datasets has done for the general domain (Gebru et al., 2018), consideration of these questions can help dataset creators provide additional context about the creation and expected use of clinical data.

The rest of this paper is organized as follows. In Section 2, we discuss related work for the field Machine Learning for Health and clinical collaborations. In Sections 3, we describe how we synthesized the questions and categorized them (e.g. scoping, data, fairness, workflow, etc.). We also present the questions within these categories. We discuss limitations in Section 4 and conclusions in Section 5. Finally, we present the aggregated questions (annotated and summarized) in Appendices A and B, respectively.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work draws on insights from clinicians and ML scientists with diverse backgrounds, specialties, and institutions to provide a compendium of questions covering issues related to data bias, study feasibility, data storage and more. These questions are the distillation of many researchers' experience, we believe they represent items which most readers will identify with and provide a resource that is useful in the real world.

2. Related Work

Researchers have been using ML approaches to improve healthcare for decades, but due to constraints in computational power, algorithmic theory, and digitize data these innovations often had limited impact (Shortliffe and Buchanan, 1975; Miller et al., 1982). Recently, Clinical ML has seen large strides, especially in imaging tasks such as radiology (Lehman et al., 2018) and pathology (Wang et al., 2016), though in other areas as well (Topol, 2019; Baldi, 2018; Esteva et al., 2019; Beckmann and Lew, 2016; Wiens et al., 2019).

2.1. Why Collaborate?

Seemingly technical challenges such as missingness and dataset shift tend to be the result of workflow processes and policy decisions. For instance, there might be a doctor who only measures a patient's white blood cell count if they already suspect an infection and want to confirm. In that situation, a model may learn to associate the ordering of the test with the infection without actually providing meaningful knowledge that the care team didn't already know (Futoma et al., 2017; Agniel et al., 2018). Additionally, the interpersonal doctor-patient relationship is associated with differences in treatment patterns, which suggests that careful attention is needed when interpreting statistics about treatments and outcomes (Boag et al., 2018). Further, because EHRs are not static but rather are often updated by their vendors, the model may learn stale relationships from old data formats. This dataset shift is especially problematic because algorithmic approaches of correcting for this perform worse than domain-knowledge-informed approaches (Nestor et al., 2019). Of course the best practice would be to discuss with those who entered the data what the given variable measures actually measure and what assumptions are reasonable.

The framework of *Data Feminism* is especially relevant for these problems (D'Ignazio and Klein, 2020). This lens views data science as a form of power, and with that understanding, it challenges existing hierarchies, considers context, embraces pluralism, and makes labor visible. Of particular relevance to Clinical ML is the emphasis on ensuring that power is appropriately shared among the impacted stakeholders, whether that means patients who expect fairness from the medical system or nurses and doctors who need to be included in the design of tool workflows. In order to achieve goals of considering context and embracing pluralism, data scientists must collaborate with domain experts and system users.

2.2. How Collaborate?

There are numerous ways data scientists can engage with domain experts Mitchell et al. (2019). The most straightforward method is to include clinicians on the research team to help guide the direction of research and inform modeling choices. This approach was taken by the research team that developed Retina U-Net to align the clinical interest (e.g. determining cancer vs no cancer) with the granularity of the annotations (e.g. pixel level vs bounding boxes) (Jaeger1 et al., 2019). Similarly, a collaboration at Duke between the Department of Statistics, the Department of Medicine, and an Innovation Center housed in the hospital produced research to predict risk of sepsis in patients. They estimated the usefulness of this tool by performing experiments with case control matching and also a simulated real-time validation to measure the number of false positives that the tool would trigger (Futoma et al., 2017).

Just as how computer scientists have different perspectives on their field[1], so to do different clinicians have different opinions. Another approach could be to survey many clinicians to better understand the distribution of perspectives from domain experts. This strategy was employed by a research group at a Toronto hospital to explore how to usefully integrate explainable ML into clinical tools (e.g. explanations that depend on the time of the shift, per-instance explanations, etc.) (Tonekaboni et al., 2019). A similar example of this is crowdsourcing clinical judgments of medication risk assessment to pharmacists in order to better understand current expert opinion (Flynn et al., 2019).

These approaches can be combined, where domain experts both work on the research team and know how to incorporate clinical judgment into their models, which can be seen by a 2019 paper – coauthored not only by data scientists but also RNs and MDs – which explicitly encodes domain expertise into the model by adding "clincal concern" proxies (e.g. frequency of respiratory rate assessment and frequency of comments associated with blood saturation) into an early warning score (Rossetti et al., 2019).

Ultimately, the aim of clinical ML is to use pattern recognition to improve clinical care and patient outcomes. Much work is done on developing new technical advances in

1. For instance, the recent 2-hour debate about what Deep Learning philosophically means suggests that one one "ML expert" can necessarily speak for the community. https://www.youtube.com/watch?v=EeqwFjqFvJA

modelling, though as the field matures, more ML is being leveraged for clinical problems. In 2013, a team of PhDs and MDs collaborated to build and integrate a tool in the Beth Israel Emergency Department for automatically recommending a chief complaint based on the nurse's triage, which was able to standardize the data collection moving forward by replacing the previously free-text field (Jernite et al., 2013). Similarly, Michigan researchers partnered with care coordination organizations in an ongoing effort to predict ED utilization and use these predictions to work with high-risk patients for alternatives to the ED (Brannon et al., 2018). Interestingly, when an interdisciplinary team of researchers at the University of Pennsylvania integrated a sepsis early warning predictive model, they found no significant difference in mortality, discharge disposition, or transfer to the ICU, though there was a reduction in time-to-ICU transfer (Giannini et al., 2019). Additionally, Clinical ML systems have been integrated at Duke (sepsis), Johns Hopkins (sepsis), the University of Michigan (c. diff), Massachusetts General Hospital (mammogram), Stanford (end of life), and potentially others as well, but papers describing their methodology and impact on patient outcomes have not yet been published.

3. Questions

The goal and framing of this work focuses on avoiding common pitfalls in collaborations between healthcare professionals and ML researchers. We then distilled their advice for clinical collaborations in a short list of questions. We have divided the proposed questions into categorical contexts within which we believe they will be most useful. Categories have been inspired by fields such as HCI, Project Management, and Bioethics. The categories we have classified the questions into are:

1. Project Scoping
2. Workflow Considerations
3. Defining Roles and Responsibilities
4. Data Generating Process
5. Data Composition
6. Data Access and Privacy
7. Storage, System Integration, and Compute Infrastructure
8. Safety and Fairness

The Collabsheet questions can be found in their respective sections below, in addition to an aggregate in the appendices for ease of readability and usability. Appendix A contains the questions annotated with examples and more descriptions while Appendix B has a summarized, more portable, version of the questions.

To identify questions viewed as most important by the community, we surveyed 20 people in the clinical ML community. With data from the 14 respondents, we underlined

questions that passed the following threshold:
$2 * \#Favourite + \#Underrated - \#Unnecessary > 14$
The threshold was chosen based on a gap in the distribution of responses. The top three sections based on the number of highlighted questions in order are: Data Generating Process, Project Scoping, and Safety and Fairness. The top three questions are:

- **Problem Definition:** Forget ML for a second, what is the thing you want to improve? (e.g. triage patients by severity of illness) How is this task currently done?(e.g. nurse monitors situation and manually adjusts fluid)
- **Algorithmic Fairness:** What are the risks from from false positives, false negatives, tradeoffs between the two, mitigating strategies, etc.? How are these metrics distributed across sensitive groups?
- **Project Goals:** How will this analysis be used? (e.g. to learn about biology, test in clinic, put in back end of app, etc) What clinical impact will need to be measured using this tool? What actionable outputs would you like from this model? What is the metric for success at this type of data? At what level of granularity is the data available?

We compared the average number of clinical ML projects respondents had reported working on versus ratings of questions as unnecessary, underrated, and favourite (not mutually exclusive) using a Welch's two sample t-test. Below we note interesting findings in the total number of respondents who rated these questions as favorite, underrated, and unnecessary but also how trends in these ratings seem to associate with experience.

- **Label Validity:** How is the label defined? Could we validate a few of these labels by hand in a chart review?(more projects, favourite)
- **Noise / Error:** What are possible sources of noise in the data? How can I differentiate between an outlier and an error? Are the labels meaningful on their own or a proxy for something meaningful? Would two people with the same credentials give the same label? (more projects, favourite)
- **Integration Requirements:** System and performance requirements needed for a deployed environment? (fewer projects, unnecessary)
- **Data Availability:** What data would be needed to answer this problem? What data would be available to answer this problem? (fewer projects, unnecessary)

3.1. Project Scoping

A key to the success of a project is proper problem framing and scoping. In the context of an interdisciplinary collaboration between ML scientists and medical professionals, scoping translates to: clearly defining the problem, discussing its compatibility for an ML, defining goals for the project in terms of clinical impact/research, and validating the quality of a metric as an indicator of improvement/success.

Project Scoping

1. **Problem Definition:** Forget ML for a second, what is the thing you want to improve? (e.g. triage patients by severity of illness) How is this task currently done?(e.g. nurse monitors situation and manually adjusts fluid)
2. **First Sanity Check:** Is a human able do this task with enough time/effort? If not, is there a reason to believe ML would be able to do it? (e.g. have other people been able to use ML for this?)
3. **Data Sample:** Can we see a small sample of the data in a controlled environment? (e.g. perhaps radiology reports start with "as compared with study on DATE" requiring the previous radiograph as additional input)
4. **Benefits of AI:** How can the benefits of ML (e.g. pattern recognition, speed, consistency, scale) address this problem?(e.g. doing task quickly to allow continuous monitoring)
5. **Measurable Goal:** What is a specific, falsifiable outcome you hope to achieve? (e.g. decrease readmission rates by 2%)
6. **Project Goals:** How will this analysis be used? (e.g. to learn about biology, test in clinic, put in back end of app, etc) What clinical impact will need to be measured using this tool? What actionable outputs would you like from this model? What is the metric for success at this type of data? At what level of granularity is the data available?
7. **Participant Goals:** What does each side hope to get out of this collaboration? (e.g. researchers interested in conference paper or clinical journal, hospital operations team potentially not interested in academic metrics when deciding to allow deployment)
8. **Metric Proxy:** Is your metric (e.g. AUROC, precision, BLEU, etc) a good enough proxy for success? (e.g. nearest neighbor search for radiology report generation might score high on BLEU despite being factually incorrect)
9. **Methodological Baseline:** What is a reasonable, strong baseline for this task?(e.g. bag-of-words works well for many text classification tasks)

3.2. Workflow Considerations

As discussed above, a given tool's sophistication will not always determine its effectiveness. Sometimes a deep learning algorithm won't lead to a meaningful care improvement (Steiner et al., 2018) and sometimes a simple checklist can reduce infections by 40% (Gawande). The culture of an organization has an incredible amount of inertia, and it is important to understand how the tool would either fit into that workflow or otherwise how the care team would be convinced to incorporate the tool (e.g. seeing the Chief of Surgery use it, work out the kinks, and slowly encourage the rest of their team to follow suit). There are anecdotes of clinics that have gone digital (i.e. adopted an EHR) but which pass a piece of cardboard around with their patients in order to continue to track where the patient is and who is responsible for them (Wachter, 2015). That is an example of technology working against the workflow, not with it.

Technology does not exist in a vacuum. Intuitive and user-friendly interfaces are crucial to realizing the benefits of AI-based recommendations; a Google study found that providing state-of-the-art Deep Learning recommendations for breast cancer detection only improved performance for humans when the interface was thoughtfully designed and usable (Steiner et al., 2018). We can learn a lesson from one of the simplest tools, the humble checklist. A well-designed checklist is able to eliminate avoidable mistakes and to facilitate a culture of teamwork and communication so that care teams can handle complex situations when they arise, such as by "breaking the ice" for a new team of surgeons and nurses to go through the "pre-flight checklist" together before a surgery (Gawande, 2009).

Workflow Considerations

1. **Context:** Is the aim to augment the patient experience, the provider experience, or both? In what way(s)? (e.g. early warning/alarm, triage/ranking, etc) Can this be adopted easily? Will this change be tolerated by current users? (e.g. nurse might be able to interpret 1 new number, but not 5)
2. **Prediction Time Horizon:** What is the time horizon within which this prediction will be useful? Is there a way to feasibly do this based on when data is available and how we plan to deliver the information?
3. **Unintended Consequences:** If this tool were to be adopted at scale, what are potential unintended consequences? (e.g. burnout, doubling demand without addressing capacity, etc)
4. **Minimum Viable Product:** Can we build a Minimum Viable Product to test this?
5. **Model Retraining:** How often would the model parameters need to be updated with new data?

3.3. Defining Roles and Responsibilities

Collaborations with the end goal of supporting and improving healthcare through different venues require conversations with people who posses different expertise and responsibilities within the organization. In this section, we present questions that aim at identifying different roles that will be vital to the success of the project. In addition, we ask questions targeted at identifying every party's expectations and responsibilities, especially in relation to timelines and communication.

We have divided the main, active, roles into six categories below. We expect recurring engagement with people in active roles in different stages of the collaboration:

- **Clinical point of contact:** This individual is the project lead from the clinical side of the project. They are the first person to reach out to with non-technical questions and is able to answer/direct questions as needed.
- **Clinical champion:** This individual is a senior-level clinical contact who is able to champion this project to a healthcare organization's leadership.
- **Subject matter expert:** This will most likely be a scientist who has intimate expertise of the scientific background of the analysis.
- **Data analyst:** This individual has worked with this data for the purpose of analysis and are familiar with the data from an analysis perspective.
- **Data manager:** This person understand the processes and procedures involved in extracting, sharing, handling the data.
- **IT system/technical point of contact:** This individual is familiar with the technical backing of the system where the data was collected or is currently stored.

Defining Roles and Responsibilities

1. **Group roles:** Identify the following roles: clinical point of contact to direct questions to, clinical champion to move this forward at an institutional level, subject matter expert (e.g. scientist), data analyst(s), data manager(s), IT system/technical point of contact.
2. **Timelines:** What is the timeline for technical milestones? How will we know if we have fallen behind? Are there relevant conference or funding deadlines for this project?
3. **Communication Cadence:** How often are check-ins appropriate?
4. **Institutional Support:** Can we talk to the people who fill in the data?
5. **Code Maintenance:** Who would be responsible for writing and maintaining the integrated, live code?
6. **User Engagement:**How early can we engage users of the system?
7. **Information Tracking:** How will information be tracked? (e.g. one single running google doc summarizing all calls)

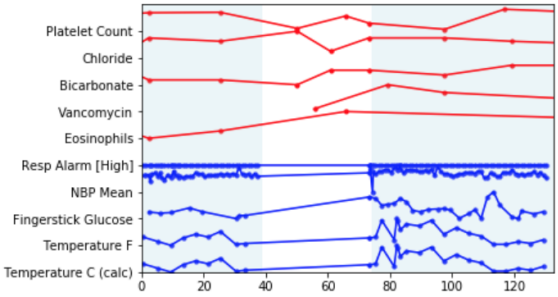
3.4. Data

It is critically important to understand the data generating process, because although ML provides a toolbox for pattern recognition in data, the analysis can only be as good as the data which is collected. There are so many nuances to the data that can have a huge impact on the accuracy and validity of an analysis.

Figure 1 shows that MIMIC happens to have lab events from the patient's entire admission but only has chart events when a patient is in the ICU. This trend may be very easy to miss by just looking at a dataframe of timestamped events.

Figure 2 demonstrates the concept of dataset shift: between 2006 and 2019, a pediatric hospital changed how often it uses different ultrasound machines during patient care. This can result in markedly different data collection practices, as demonstrated by Figure 3. This is a subtle example of "measurement bias" (Suresh and Guttag, 2019).

Figure 1: This patient's single hospital stay has multiple ICU stays (shaded light blue).



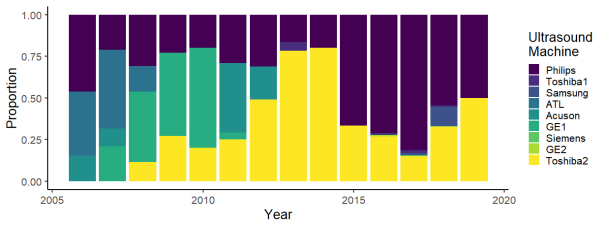


Figure 2: A shift in data generation demonstrated by a frequency plot of ultrasound machines used in a pediatric hospital for renal ultrasound from 2005-2019.

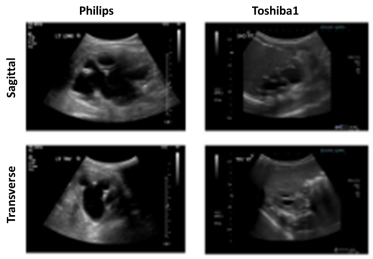


Figure 3: The qualitative difference produced by two different ultrasound machines with the same ultrasound views, sagittal and transverse kidney.

Data Generating Process

1. **Missing Data:** What ways can missing data arise? Will that impact model design? (e.g. model requires hourly buckets of heart rate but those values aren't recorded frequently enough) Is the missing data not measured or not recorded / recorded in another format?

2. **Data Input:** How is the data recorded? Who enters the data? Are any parts automatically entered or shortcuts used to enter the data? Does that have any accuracy implications?

3. **Context:** Are there any assumptions or constraints that the people who input the data are working under? (e.g. writing important parts to be understood by another nurse vs writing exhaustively for compliance with protocols)

4. **Label Validity:** How is the label defined? (e.g. consider True case if they have an ICD for inflammatory bowel disease 1 time? 3 times? If they have ICD code for endoscopy?) Could we validate a few of these labels by hand in a chart review?

5. **Domain Knowledge:** Was human judgment involved in recording this data? If so, how much? For clinical tests, why was this test ordered? What previous information / thresholds lead to this data collection decision?

6. **Gotchas:** Are there any unintuitive surprises that are not reflected in the final data snapshot? (e.g. timestamps reflect time of entry not necessarily time of event; low acuity (controls) patients all receiving imaging from the same machine ; survey-based questions delivered differently by site)

7. **Confounders:** Are there any anticipated confounders to the model we're trying to build? Are these measured explicitly or implicitly in our data?

8. **Dataset Shift:** Were there any changes that might affect data recording in the time period during which this data was collected? (e.g. switching EHR vendors, new generation of lab tests have increased sensitivity to trace amounts of a biomarker, combining data from multiple sites, shift in patient population based on global event, policy change, etc)

9. **Data Curation:** What pre-processing is done to curate the data before the algorithm sees it? (e.g. decision combine some similar methods for measuring blood pressure) Is pre-processing system-based and / or scientist-based? How complex is the curation process? (e.g. easily-retrievable subset of EHR columns, patching disparate unstructured systems into a single stream of data, etc)

Data Composition

1. **Data Availability:** What data would be needed to answer this problem? What data would be available to answer this problem?
2. **Data Provenance:** What was the original purpose behind collecting this data? Was the data collected retrospectively or prospectively? Are there limits to what questions can or cannot be answered by this data? (e.g. Were there any randomly-assigned interventions?)
3. **Multiple Datasets:** If there are multiple datasets being combined for this project, are they linked (e.g. by patient ID)? If not, what is the benefit of bringing multiple datasets together?
4. **Data Characteristics:** What are the dimensions of the data? What data modalities exist in the dataset? (e.g. images, text, billing codes, etc)
5. **Subjects:** How many subjects does the data represent? What is the unit of observation? (e.g. patient-level, admission-level, organ-level)
6. **Noise / Error:** What are possible sources of noise in the data? How can I differentiate between an outlier and an error? (e.g. how can I tell if a value is biologically or physiologically plausible?) Are the labels meaningful on their own or a proxy for something meaningful? Would two people with the same credentials give the same label?
7. **Cohort:** What is the inclusion/exclusion criteria for the cohort? (e.g. all patients from a given time frame, patients who died in the hospital or were discharged to a skilled nursing facility, etc)

3.5. Data Access and Privacy

It is important to set expectations during that process around the steps involved in accessing the data, data privacy, and post-analysis data retention.

1. **Data Access:** What is required for access to this data? Who is authorized to grant access to the data? Is there an established process that can be outlined? How can the data be accessed by new team members?
2. **Timeline Expectations:** How long does data access usually take?
3. **Handling Sensitive Data:** What is the procedure for handling sensitive data? Who is authorized to access the data in order to de-identify it?
4. **Benefit/Harm Tradeoffs:** What potentially-identifiable fields would still be useful in the analysis?
5. **Post-Analysis Data Retention:** Can this data be used for future analyses beyond the scope of the current project? Can the data be stored for replicability?

3.6. Storage, System Integration, and Compute Infrastructure

Storage and compute limitations can make certain projects a non-starter. It's important to know these details early on so that gaps in capacity can be filled and those who can help navigate these issues are identified.

For example, if the planned analysis involves deep learning, a graphical processing unit (GPU) will make the analysis much more efficient to conduct. Public GPUs exist but data security may prevent the use of these with sensitive data. If this is known early on, access to GPUs within a secure environment can be purchased or acquired while other aspects of the project (REB/IRB approval, data access, preprocessing, etc) are in progress.

1. **Data Storage:** Where will the data live? Does the IRB/REB allow moving the data? If that's not possible, what is the process to analyze it in place?
2. **Computing Resources:** Are the computing available resources sufficient for this analysis? (i.e. What's needed? What's available?)
3. **Integration Requirements:** What system/performance requirements will be needed for a deployed environment? (e.g. real-time access to data vs daily snapshots; queries to avoid overloading the database; etc)
4. **Deployment:** How will the code run on the system? (e.g. will the project run in python on the hospital IT infrastructure via docker? will the model parameters need to be hard-coded into the EHR with the MUMPS language? etc)
5. **Replication Considerations:** Once the project is over, how can we recreate the cohort to replicate the results as a starting point for future projects? (e.g. a stable query to reproduce the same cohort, if need be)

3.7. Safety and Fairness

Clinical ML is high-risk and has a long tail of dangerous outliers. This section is concerned with the possibility of algorithmic harm, especially at scale. Clinical ML Safety can apply to any group that appears to be at risk for inadvertent algorithmic harm (e.g. patients with a rare breast cancer pathology). Equity and fairness considerations arise when algorithmic tools run the risk of exacerbating historical disadvantages for groups. Suresh and Guttag (2019) outline a framework of ways ML bias can arise, including:

- representation bias (e.g. if a dataset was not trained on Hispanic populations)
- historical bias (e.g. social determinants of health make some populations less healthy)
- measurement bias (e.g. rural clinics have worse measurement tools than academic medical centers)
- aggregation bias (one-size-fits-all model on a heterogeneous population)
- evaluation bias (e.g. label is a bad proxy for success)
- deployment bias (e.g. the model's predictions are being systematically ignored)

The unintuitive nature of hidden stratifications makes it essential to talk with a domain expert about what the dangerous outliers are. Sensitive cases need to be identified so that care can be taken to test on data points in that category.[2] Responsible ML researchers understand the robustness of their model for validity and generalizability.[3]

Safety and Fairness

1. **Sensitive Variables:** Which sensitive variables are available? (gender, ethnicity, SES, etc.) What variables may be correlated with these sensitive variables?
2. **Algorithmic Fairness:** What are the risks and tradeoffs for from false positives vs false negatives? How are these risks distributed across sensitive groups?
3. **Data Generation Bias:** What are some ways biases could arise throughout the data generating process? (e.g. rural hospitals have worse measurement tools)
4. **Wrong Treatment Risk:** What is the harm of over- vs under-treatment?
5. **Hidden Stratification:** Are there rare, high-risk sub-populations in this dataset which need special care? (e.g. should labels distinguish *untreated pneumothorax* vs *pneumothorax*?)
6. **Generalization:** Is there a mechanism for reproducing the results in order to answer clarifications about the paper? Is there a defined procedure for validating replicability across sites?

2. For more reading on this topic, see Dr. Luke Oakden-Rayner's blog series on Medical ML Safety: https://lukeoakdenrayner.wordpress.com/2018/07/11/medical-ai-safety-we-have-a-problem.
3. McDermott et al. (2019) propose a framing of reproducibility in ML work into 3 types of replication: technical replication (under identical technical settings), statistical replication (replication using a different data sample/cohort), conceptual replication (replication of the conceptual idea, without access to the original data or code.

4. Limitations

When collecting information and writing this paper, the framing we had in mind focused on a collaboration between a healthcare professional working in a hospital and a ML scientist with the final goal of integration. Our research is primarily based in North America, therefore this work also focuses on the North American (Canada/US) context. Although this resource could be useful for other contexts, it is important to keep these limitations in mind.

In addition, we recognize that a lack of upfront discussion is not the only thing that limits collaborations. Academic silos and severe incentive misalignment also prevent Clinical ML research from flourishing. For example, hospitals may be reluctant to share data with collaborators for privacy / monetization purposes. Also, it tends to be the case that in the academic world a publication with a novel model is rewarded higher than a translational effort to test and integrate a model in a live system. It is worth emphathisizing that academics are one part of the group that would be required for a translational and integration effort. Those problems are indeed hard, and stem from policy and financial issues beyond what ML scientists can solve. Nonetheless, when right-place-right-time situations do occur, it is important to make the most of such collaborations. Sub-optimal communication can result in unclear expectations, and that is a problem that could potentially be avoided or improved upon.

5. Conclusions

The promise of Clinical ML is vast but rests on effective interdisciplinary work. In this paper, we seek to capture details that help in framing a new collaboration between healthcare professionals and ML scientists. We believe that this list of questions allows collaborators to spend time understanding each other's needs and limitations and take them into account to work together effectively

Acknowledgments

We sincerely thank all of the people who have taken the time to share their advice and experiences.

Thank you to Alistair Johnston,Amol Verma, Anand Avati, Andrea Smith, Bret Nestor, Corey Chivers, Craig Stewart, Di Jin, Emily Alesentzer, Erik Drysdale, Harini Suresh, Harry Hsu, Heather Berlin, Irene Chen, Joe Futoma, Jospeh Cohen, Joyce Ho, Karandeep Singh, Mark Sendak, Matthew McDermott, Michael Draugelis, Muhammad Mamdani, Robert Grant, Russ Greiner, Shalmali Joshi, Susan Regli, Tasmie Sarker, Taylor Killian, Thomas McCoy, Tzu-Ming, Wei-Hung Weng for discussions about this paper and often about their experience on clinical projects or deployments and/or contributing responses

and comments to the survey.

This work was made possible through input into how healthcare collaborations were done by researchers from Duke, Harvard, MILA, MIT, Massachusetts General Hospital, Microsoft Research, SickKids Hospital, St. Michael's Hospital, Stanford, University of Alberta, University of Michigan Ann Arbor, University of Montreal, University of Pennsylvania Medicine, Vector Institute, and University of Toronto.

References

Denis Agniel, Isaac Kohane, and Griffin Weber. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, 361:k1479, 04 2018. doi: 10.1136/bmj.k1479.

Pierre Baldi. Deep learning in biomedical data science. *Annual review of biomedical data science*, 1:181–205, 2018.

Jacques S Beckmann and Daniel Lew. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome medicine*, 8(1): 134, 2016.

Willie Boag, Harini Suresh, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Racial disparities and mistrust in end-of-life care. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 587–602, Palo Alto, California, 17–18 Aug 2018. PMLR. URL http://proceedings.mlr.press/v85/boag18a.html.

Elliott Brannon, Tianshi Wang, Jeremy Lapedis, Paul Valenstein, Michael Klinkman, Ellen Bunting, Alice Stanulis, and Karandeep Singh. Towards a learning health system to reduce emergency department visits at a population level. *AMIA Annual Symposium proceedings*, 2018:295–304, 12 2018.

Catherine D'Ignazio and Lauren Klein. *Data Feminism.* 2020.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Allen J. Flynn, Greg Farris, George Meng, Jack Allan, Sara Kurosu, Natalie Lampa, and Koki Sasagawa. Engaging pharmacists to crowdsource a fine-grained medication risk scale: An initial measurement study using paired comparisons of medications. *AMIA 2019 Annual Symposium*, 2019. URL https://symposium2019.zerista.com/event/member/602050.

Joseph Futoma, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O'Brien. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 243–254, Boston, Massachusetts, 18–19 Aug 2017. PMLR. URL http://proceedings.mlr.press/v68/futoma17a.html.

Atul Gawande. *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books, 2009.

Autil Gawande. Why doctors hate their computers). URL https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers?mc_cid=cbc74122a0&mc_eid=93cb8fa0f9&fbclid=IwAR3UArfvnQldK07FjIAuldDaekgxOHC2xPqTGVs4WEkLmRm-7bQwrcSUo2k.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets, 2018.

Heather Giannini, Jennifer Ginestra, Corey Chivers, Michael Draugelis, Asaf Hanish, William Schweickert, Barry Fuchs, Laurie Meadows, Michael Lynch, Patrick Donnelly, Kimberly Pavan, Neil Fishman, Clarence Hanson, and Craig Umscheid. A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice. *Critical Care Medicine*, 47:1, 08 2019. doi: 10.1097/CCM.0000000000003891.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031 [cs, eess]*, January 2019. URL http://arxiv.org/abs/1901.07031. arXiv: 1901.07031.

Paul F. Jaeger1, Simon A. A. Kohl1, Sebastian Bickelhaupt, Fabian Isensee1, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H. Maier-Hein. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In *Proceedings of Machine Learning Research*, Proceedings of Machine Learning Research, Vancouver, Canada, December 2019. PMLR. URL https://ml4health.github.io/2019/pdf/232_ml4h_preprint.pdf.

Yacine Jernite, Yoni Halpern, Steven Horng, and David Sontag. Predicting chief complaints at triage time in the emergency department. In *NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare*, 2013.

A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

Alistair E. W. Johnson, Tom J. Pollard, Seth Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]*, January 2019. URL http://arxiv.org/abs/1901.07042. arXiv: 1901.07042.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. eicu collaborative research database, 2017. URL http://eicu-crd.mit.edu/.

Constance D. Lehman, Adam Yala, Tal Schuster, Brian Dontchos, Manisha Bahl, Kyle Swanson, and Regina Barzilay. Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology*, 290(1), 2018. doi: https://doi.org/10.1148/radiol.2018180694.

Matthew McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Marzyeh Ghassemi, and Luca Foschini. Reproducibility in machine learning for health. *arXiv preprint arXiv:1907.01463*, 2019.

Randolph A. Miller, Melissa A. McNeil, Sue M. Challinor, Jr Fred E. Masarie, and MD Jack D. Myers. Internist-1: An experimental computer-based diagnostic consultant for general internal medicine. 307:468–476, 1982. URL https://static1.squarespace.com/static/59d5ac1780bd5ef9c396eda6/t/5d474315b0f5980001a2823d/1564951317375/Singh_7.pdf.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks, 2019.

Luke Oakden-Rayner*, Jared Dunnmon*, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of Machine Learning Research*, Proceedings of Machine Learning Research, Vancouver, Canada, December 2019. PMLR. URL https://arxiv.org/pdf/1909.12475.pdf.

Sarah Collins Rossetti, Chris Knaplund, Dave Albers, Abdul Tariq, Kui Tang, David Vawdrey, Natalie H. Yip, Patricia C. Dykes, Jeffrey G. Klann, Min Jeoung Kang, Jose

Garcia, Li-Heng Fu, Kumiko Schnock, and Kenrick Cato. Leveraging clinical expertise as a feature - not an outcome - of predictive models: Evaluation of an early warning system use case. *AMIA 2019 Annual Symposium*, 2019. URL https://symposium2019.zerista.com/event/member/602052.

Edward H. Shortliffe and Bruce G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3):351 – 379, 1975. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(75)90047-4. URL http://www.sciencedirect.com/science/article/pii/0025556475900474.

David Steiner, Bob MacDonald, Yun Liu, Peter Truszkowski, Jason Hipp, Christopher Lee Gammage, Florence Thng, Lily Peng, and Martin Stumpe. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *American Journal of Surgical Pathology*, 2018. URL https://journals.lww.com/ajsp/Fulltext/2018/12000/Impact_of_Deep_Learning_Assistance_on_the.7.aspx.

Harini Suresh and John Guttag. A framework for understanding unintended consequences of machine learning, 2019. URL https://arxiv.org/abs/1901.10002.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 359–380, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR. URL http://proceedings.mlr.press/v106/tonekaboni19a.html.

Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. 2019. ISBN 1541644638.

Robert Wachter. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*. 2015. ISBN 978-1260019608.

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. Deep learning for identifying metastatic breast cancer, 2016.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25 (9):1337–1340, 2019.

Appendices

Appendix A. Annotated Healthcare Collabsheets

A.1. Project Scoping

1. **Problem Definition:** Forget ML for a second, what is the thing you want to improve? (e.g. triage patients by severity of illness) How is this task currently done?(e.g. nurse monitors situation and manually adjusts fluid)
2. **First Sanity Check:** Is a human able do this task with enough time/effort? If not, is there a reason to believe ML would be able to do it? (e.g. have other people been able to use ML for this?)
3. **Data Sample:** Can we see a small sample of the data in a controlled environment? (e.g. perhaps radiology reports start with "as compared with study on DATE" requiring the previous radiograph as additional input)
4. **Benefits of AI:** How can the benefits of ML (e.g. pattern recognition, speed, consistency, scale) address this problem?(e.g. doing task quickly to allow continuous monitoring)
5. **Measurable Goal:** What is a specific, falsifiable outcome you hope to achieve? (e.g. decrease readmission rates by 2%)
6. **Project Goals:** How will this analysis be used? (e.g. to learn about biology, test in clinic, put in back end of app, etc) What clinical impact will need to be measured using this tool? What actionable outputs would you like from this model? What is the metric for success at this type of data? At what level of granularity is the data available?
7. **Participant Goals:** What does each side hope to get out of this collaboration? (e.g. researchers interested in conference paper or clinical journal, hospital operations team potentially not interested in academic metrics when deciding to allow deployment)
8. **Metric Proxy:** Is your metric (e.g. AUROC, precision, BLEU, etc) a good enough proxy for success? (e.g. nearest neighbor search for radiology report generation might score high on BLEU despite being factually incorrect)
9. **Methodological Baseline:** What is a reasonable, strong baseline for this task?(e.g. bag-of-words works well for many text classification tasks)

A.2. Workflow Considerations

1. **Context:** Is the aim to augment the patient experience, the provider experience, or both? In what way(s)? (e.g. early warning/alarm, triage/ranking, etc) Can this be adopted easily? Will this change be tolerated by current users? (e.g. nurse might be able to interpret 1 new number, but not 5)
2. **Prediction Time Horizon:** What is the time horizon within which this prediction will be useful? Is there a way to feasibly do this based on when data is available and how we plan to deliver the information?
3. **Unintended Consequences:** If this tool were to be adopted at scale, what are potential unintended consequences? (e.g. burnout, doubling demand without addressing capacity, etc)
4. **Minimum Viable Product:** Can we build a Minimum Viable Product to test this?
5. **Model Retraining:** How often would the model parameters need to be updated with new data?

A.3. Defining Roles and Responsibilities

1. **Group roles:** Identify the following roles: clinical point of contact to direct questions to, clinical champion to move this forward at an institutional level, subject matter expert (e.g. scientist), data analyst(s), data manager(s), IT system/technical point of contact.
2. **Timelines:** What is the timeline for technical milestones? How will we know if we have fallen behind? Are there relevant conference or funding deadlines for this project?
3. **Communication Cadence:** How often are check-ins appropriate?
4. **Institutional Support:** Can we talk to the people who fill in the data?
5. **Code Maintenance:** Who would be responsible for writing and maintaining the integrated, live code?
6. **User Engagement:**How early can we engage users of the system?
7. **Information Tracking:** How will information be tracked? (e.g. one single running google doc summarizing all calls)

A.4. Data Generation Process

1. **Missing Data:** What ways can missing data arise? Will that impact model design? (e.g. model requires hourly buckets of heart rate but those values aren't recorded frequently enough) Is the missing data not measured or not recorded / recorded in another format?
2. **Data Input:** How is the data recorded? Who enters the data? Are any parts automatically entered or shortcuts used to enter the data? Does that have any accuracy implications?
3. **Context:** Are there any assumptions or constraints that the people who input the data are working under? (e.g. writing important parts to be understood by another nurse vs writing exhaustively for compliance with protocols)
4. **Label Validity:** How is the label defined? (e.g. consider True case if they have an ICD for inflammatory bowel disease 1 time? 3 times? If they have ICD code for endoscopy?) Could we validate a few of these labels by hand in a chart review?
5. **Domain Knowledge:** Was human judgment involved in recording this data? If so, how much? For clinical tests, why was this test ordered? What previous information / thresholds lead to this data collection decision?
6. **Gotchas:** Are there any unintuitive surprises that are not reflected in the final data snapshot? (e.g. timestamps reflect time of entry not necessarily time of event; low acuity (controls) patients all receiving imaging from the same machine ; survey-based questions delivered differently by site)
7. **Confounders:** Are there any anticipated confounders to the model we're trying to build? Are these measured explicitly or implicitly in our data?
8. **Dataset Shift:** Were there any changes that might affect data recording in the time period during which this data was collected? (e.g. switching EHR vendors, new generation of lab tests have increased sensitivity to trace amounts of a biomarker, combining data from multiple sites, shift in patient population based on global event, policy change, etc)
9. **Data Curation:** What pre-processing is done to curate the data before the algorithm sees it? (e.g. decision combine some similar methods for measuring blood pressure) Is pre-processing system-based and / or scientist-based? How complex is the curation process? (e.g. easily-retrievable subset of EHR columns, patching disparate unstructured systems into a single stream of data, etc)

A.5. Data Composition

1. **Data Availability:** What data would be needed to answer this problem? What data would be available to answer this problem?
2. **Data Provenance:** What was the original purpose behind collecting this data? Was the data collected retrospectively or prospectively? Are there limits to what questions can or cannot be answered by this data? (e.g. Were there any randomly-assigned interventions?)
3. **Multiple Datasets:** If there are multiple datasets being combined for this project, are they linked (e.g. by patient ID)? If not, what is the benefit of bringing multiple datasets together?
4. **Data Characteristics:** What are the dimensions of the data? What data modalities exist in the dataset? (e.g. images, text, billing codes, etc)
5. **Subjects:** How many subjects does the data represent? What is the unit of observation? (e.g. patient-level, admission-level, organ-level)
6. **Noise / Error:** What are possible sources of noise in the data? How can I differentiate between an outlier and an error? (e.g. how can I tell if a value is biologically or physiologically plausible?) Are the labels meaningful on their own or a proxy for something meaningful? Would two people with the same credentials give the same label?
7. **Cohort:** What is the inclusion/exclusion criteria for the cohort? (e.g. all patients from a given time frame, patients who died in the hospital or were discharged to a skilled nursing facility, etc)

A.6. Data Access and Privacy

1. **Data Access:** What is required for access to this data? Who is authorized to grant access to the data? Is there an established process that can be outlined? How can the data be accessed by new team members?
2. **Timeline Expectations:** How long does data access usually take?
3. **Handling Sensitive Data:** What is the procedure for handling sensitive data? Who is authorized to access the data in order to de-identify it?
4. **Benefit/Harm Tradeoffs:** What potentially-identifiable fields would still be useful in the analysis?
5. **Post-Analysis Data Retention:** Can this data be used for future analyses beyond the scope of the current project? Are we able to store the data for replicability?

A.7. Storage, System Integration, and Compute Infrastructure

1. **Data Storage:** Where will the data live? Can we move the data? Is there an IRB/REB consideration to moving the data? If that's not possible, what is the process to analyze it in place?
2. **Computing Resources:** Are the computing resources available sufficient for this analysis? (i.e. What's needed? What's available?)
3. **Integration Requirements:** What system/performance requirements will be needed for a deployed environment? (e.g. real-time access to data vs daily snapshots, queries to avoid overloading the database)
4. **Deployment:** How will the code run on the system? (e.g. perhaps the project runs in python on the hospital IT infrastructure via docker, perhaps the model parameters need to be hard-coded into the EHR with the MUMPS language, etc)
5. **Replication Considerations:** Once the project is over, how can we recreate the cohort to replicate the results as a starting point for future projects? (e.g. a stable query to reproduce the same cohort, if need be)

A.8. Safety and Fairness

1. **Sensitive Variables:** Which sensitive variables are available? (gender, ethnicity, SES, etc.) What variables are likely to be highly correlated with these sensitive variables?
2. **Algorithmic Fairness:** What are the risks from from false positives, false negatives, tradeoffs between the two, mitigating strategies, etc.? How are these metrics distributed across sensitive groups?
3. **Wrong Treatment Risk:** What is the harm of over-treatment / under-treatment?
4. **Data Generation Bias:** What are some ways biases could arise throughout the data generating process? (e.g. your dataset wasn't trained on Hispanic populations, rural hospitals have worse measurement tools than academic medical centers, designing one-size-fits-all models for a heterogeneous population, etc)
5. **Hidden Stratification:** Are there rare, high-risk sub-populations in this dataset which need special care or are excluded entirely from the dataset? (e.g. perhaps labels should be more fine-grained so as to distinguish "untreated pneumothorax" vs "pneumothorax")
6. **Generalization:** Is there a mechanism for reproducing the results of *this* work in order to answer questions about the paper as an artifact? Is there a defined procedure for validating replicability across sites?

Appendix B. Healthcare Collabsheets Summarized

B.1. Project Scoping

1. **Problem Definition:** Forget ML for a second, what is the thing you want to improve? How is it currently done?
2. **First Sanity Check:** Is a human able do this task with enough effort? If not, why ML?
3. **Data Sample:** Can we see a small sample of the data in a controlled environment?
4. **Benefits of ML:** How can the benefits of ML address this problem?
5. **Measurable Goal** What is a specific, falsifiable outcome you hope to achieve?
6. **Project Goals:** How will this analysis be used? Clinical impact will need to be measured? Desired actionable outputs?
7. **Participant Goals:** What does each side hope to get out of this collaboration?
8. **Metric Proxy:** Is your metric (e.g. AUROC) a good enough proxy for success?
9. **Methodological Baseline:** What is a reasonable, strong baseline for this task?

B.2. Workflow Considerations

1. **Context:** Is the aim to augment the patient/provider experience? both? In what way(s)? Can this be adopted easily?
2. **Prediction Time Horizon:** Time horizon within which this prediction will be useful? how we plan to deliver the result?
3. **Unintended Consequences:** e.g. burnout, doubling demand without capacity, etc.
4. **Minimum Viable Product:** Can we build a Minimum Viable Product to test this?
5. **Model Retraining:** Frequency of model parameter updates with new data?

B.3. Defining Roles and Responsibilities

1. **Group roles:** Identify the following roles: clinical point of contact, clinical champion, subject matter expert (e.g. scientist), the data analyst(s), the data manager(s), IT system/technical point of contact.
2. **Timelines:** Timeline for technical milestones? How to know if we are behind?
3. **Communication Cadence:** How often should check-ins be?
4. **Institutional Support:** Can we talk to the people who fill in the data?
5. **Code Maintenance:** Who would be responsible for writing and maintaining the integrated, live code?
6. **User Engagement:**How early can we engage users of the system?
7. **Information Tracking:** How will project information be tracked?

B.4. Data Generation Process

1. **Missing Data:** How can missing data arise? Will that impact model design? Is the missing data not measured or not recorded / recorded in another format?
2. **Data Input:** How is the data recorded? Who enters it? Are any parts automatically entered or shortcuts used to enter the data? Does that have any accuracy implications?
3. **Context:** Are there any assumptions or constraints that the people who input the data are working under?
4. **Label Validity:** Label definition? Could we validate a few of the labels by hand?
5. **Domain Knowledge:** Was human judgment involved in recording this data? how much? For clinical tests, why was a test ordered? What previous information / thresholds lead to this data collection decision?
6. **Gotchas:** Unintuitive surprises reflected in the final data snapshot? (e.g. timestamps reflect time of entry instead of time of event; low acuity (controls) patients all receiving imaging from the same machine ; survey-based questions delivered differently by site)
7. **Confounders:** Anticipated confounders to the model we're considering? Measured explicitly or implicitly in our data?
8. **Dataset Shift:** Were there any changes that might affect data recording in the time period covered by this data?
9. **Data Curation:** What pre-processing is done to curate the data? Is pre-processing system-based and / or scientist-based? How complex is the curation process?

B.5. Data Composition

1. **Data Availability:** Needed? Available?
2. **Data Provenance:** Original purpose behind collecting this data? Collected retrospectively or prospectively? Limits to what questions can(not) be answered by this data?
3. **Multiple Datasets:** If there are multiple datasets being combined for this project, are they linked? what is the benefit of bringing multiple datasets together?
4. **Data Characteristics:** What are the dimensions of the data? What types of data exist in the dataset?
5. **Subjects:** How many subjects are represented? What is the unit of observation?
6. **Noise / Error:** Sources of noise in the data? How to differentiate between an outlier and an error? Are the labels meaningful or a proxy for something meaningful? Would two people with the same credentials give the same label?
7. **Cohort:** What is the inclusion/exclusion criteria for the cohort?

B.6. Data Access and Privacy

1. **Data Access:** Requirements? Who is authorized to grant access? Established process that can be outlined? How can the data be accessed by new team members?
2. **Timeline Expectations:** How long does data access usually take?
3. **Handling Sensitive Data:** Procedure for handling sensitive data? Who is authorized to access the data in order to de-identify it?
4. **Benefit/Harm Tradeoffs:** What potentially-identifiable fields would still be useful in the analysis?
5. **Post-Analysis Data Retention:** Can this data be used for future analyses beyond the scope of the current project? Are we able to store the data for replicability?

B.7. Storage, System Integration, and Compute Infrastructure

1. **Data Storage:** Where will the data live? Can it be moved? IRB/REB consideration to moving the data?
2. **Computing Resources:** Needed vs available.
3. **Integration Requirements:** System and performance requirements needed for a deployed environment?
4. **Deployment:** How will the code run on the system? (e.g. python on hospital IT infrastructure via docker, hard-coded model parameters into the EHR with MUMPS language)
5. **Replication Considerations:** Post-project, how to recreate cohort to replicate results?

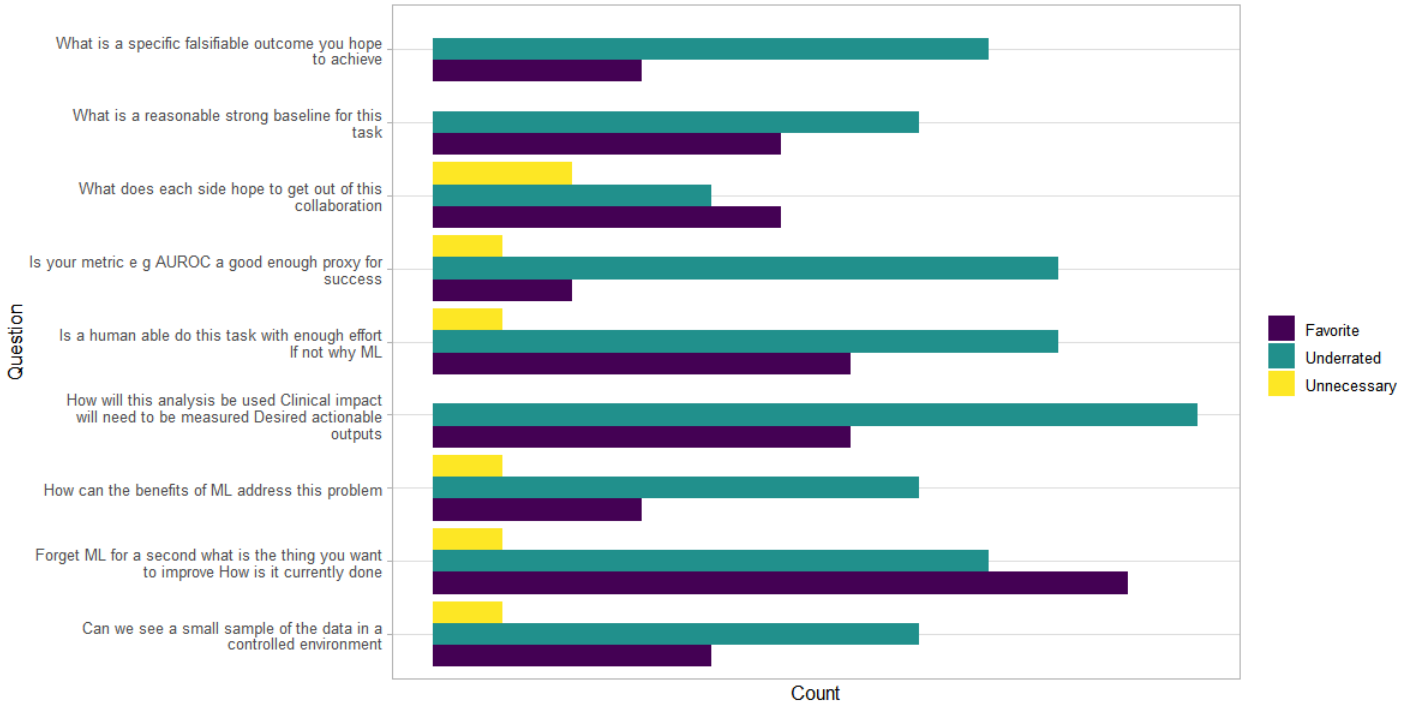
B.8. Safety and Fairness

1. **Sensitive Variables:** Which sensitive variables are available? (gender, ethnicity, SES, etc.) What variables are likely to be highly correlated with these sensitive variables?
2. **Algorithmic Fairness:** What are the risks from from false positives, false negatives, tradeoffs between the two, mitigating strategies, etc.? How are these metrics distributed across sensitive groups?
3. **Wrong Treatment Risk:** What is the harm of over-treatment / under-treatment?
4. **Data Generation Bias:** e.g. dataset not trained on Hispanic populations, rural hospitals have worse measurement tools, designing one-size-fits-all models for a heterogeneous population, etc
5. **Hidden Stratification:** Rare, high-risk, subpopulations in this dataset which need special care / are excluded from the dataset?
6. **Generalization:** Mechanism for reproducing the results of *this* work in order to answer questions about the paper as an artifact? Defined procedure for validating replicability across sites?

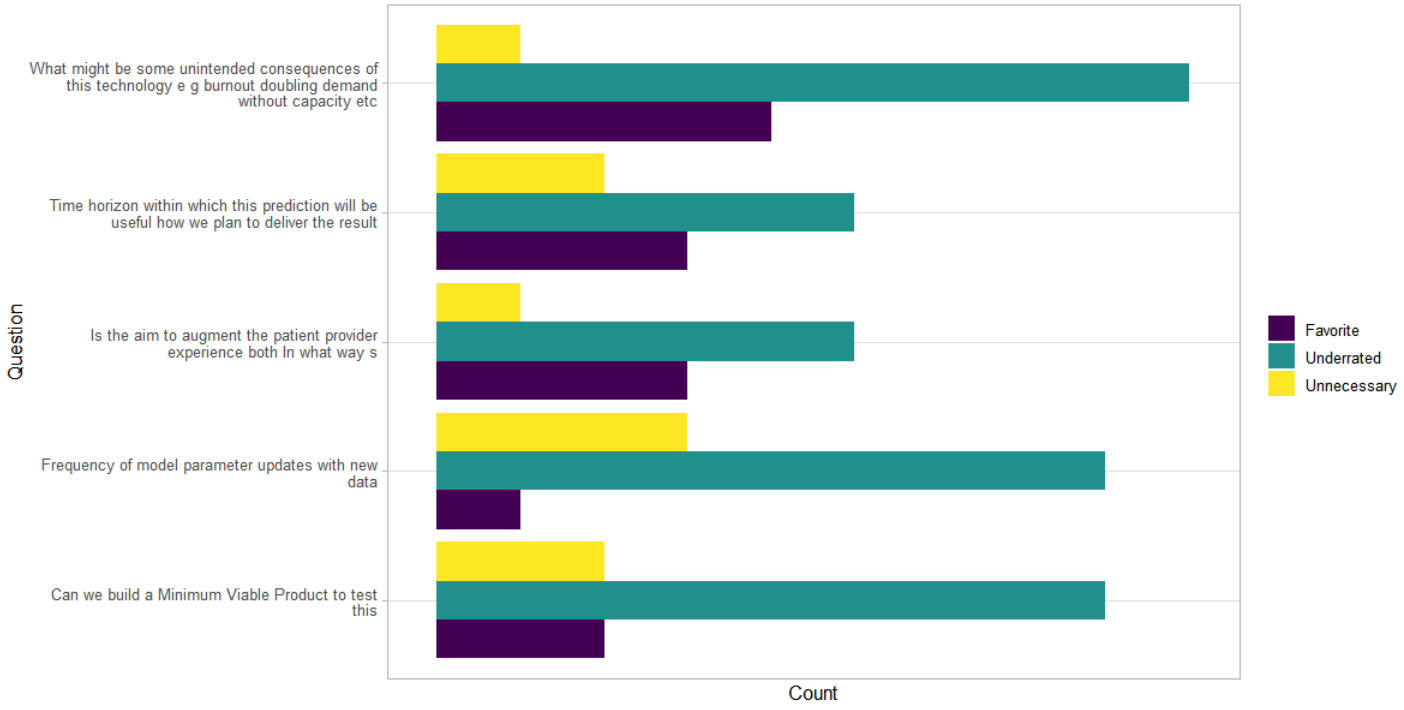
Appendix C. Survey Results

In the subsections below, we have included survey results per section. Each graph represents the number of times questions were marked "Favourite", "Underrated", or "Unnecessary". Analysis of the response differences based on the number of clinical collaborations can be found here: docs.google.com/spreadsheets/d/12sOU–wABaxDIzEWSTGtbNR15eMz5RYkOFXI80znUnA

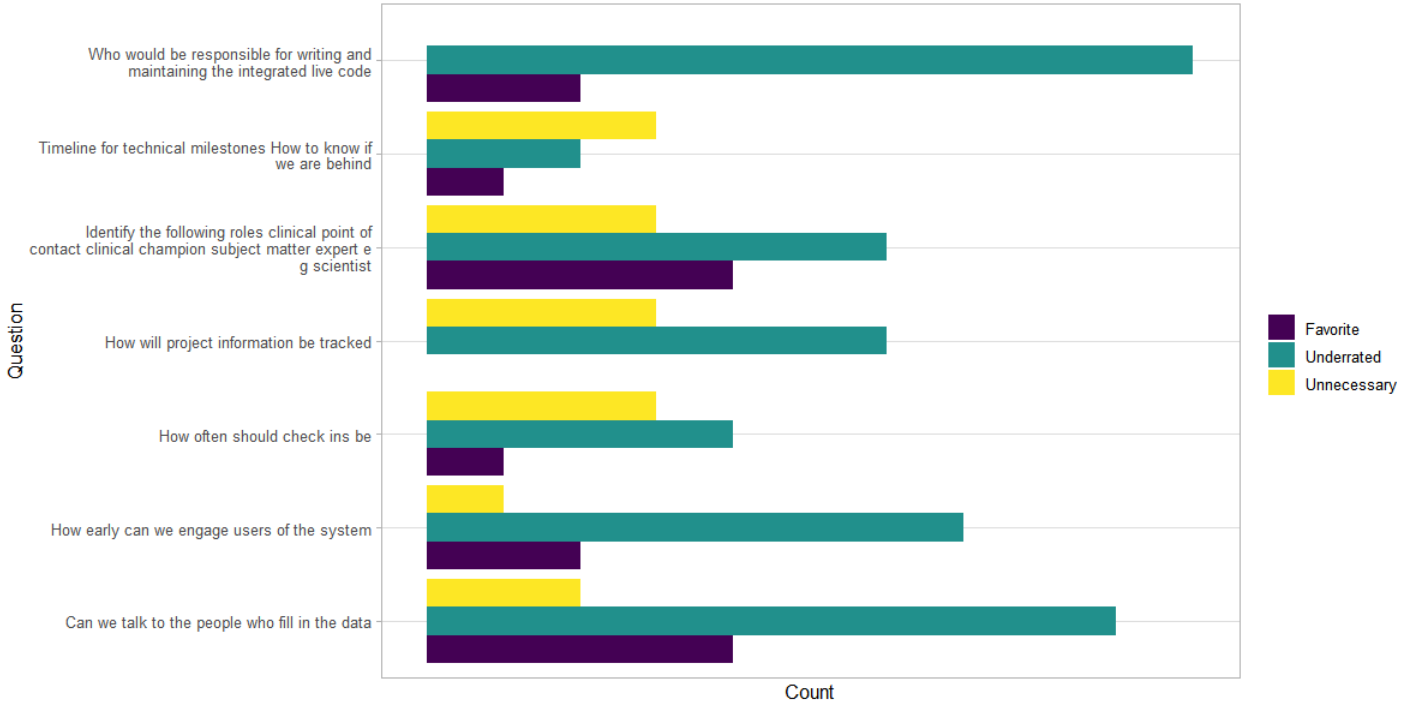
C.1. Project Scoping



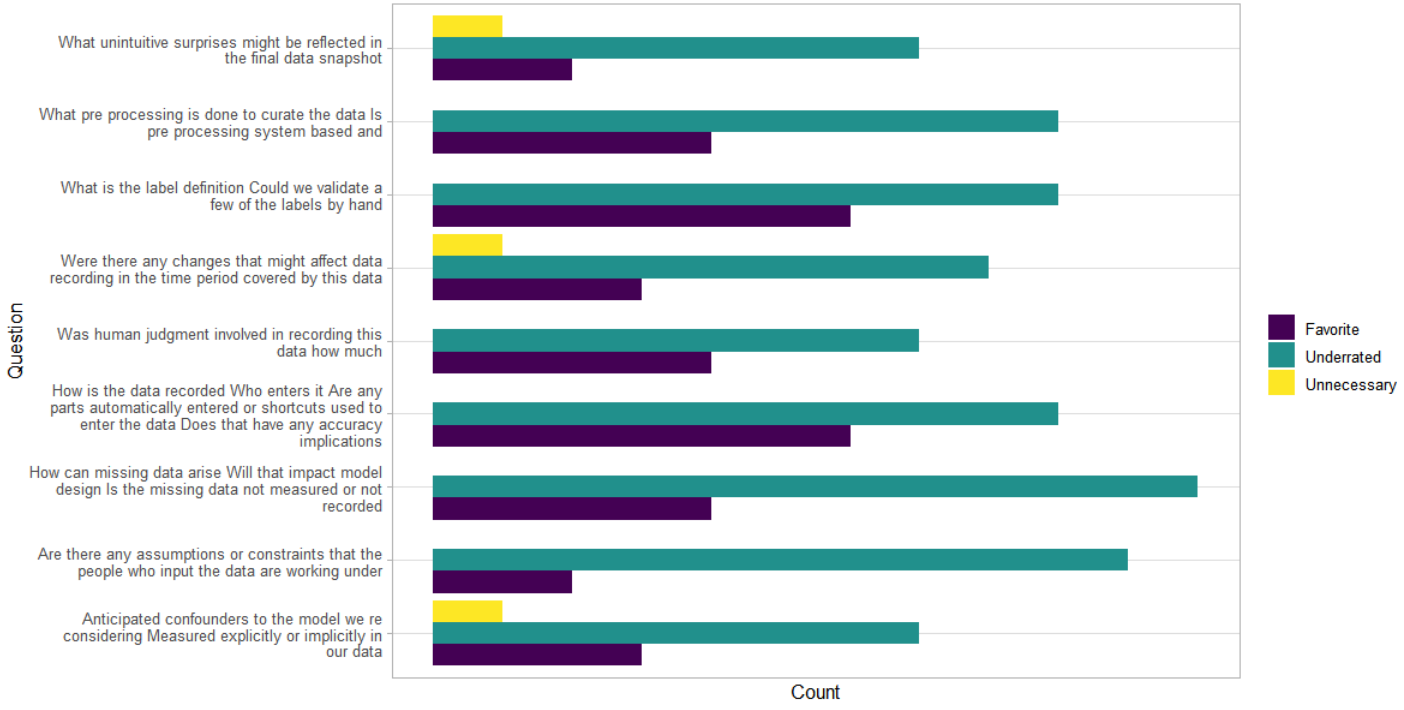
C.2. Workflow Considerations



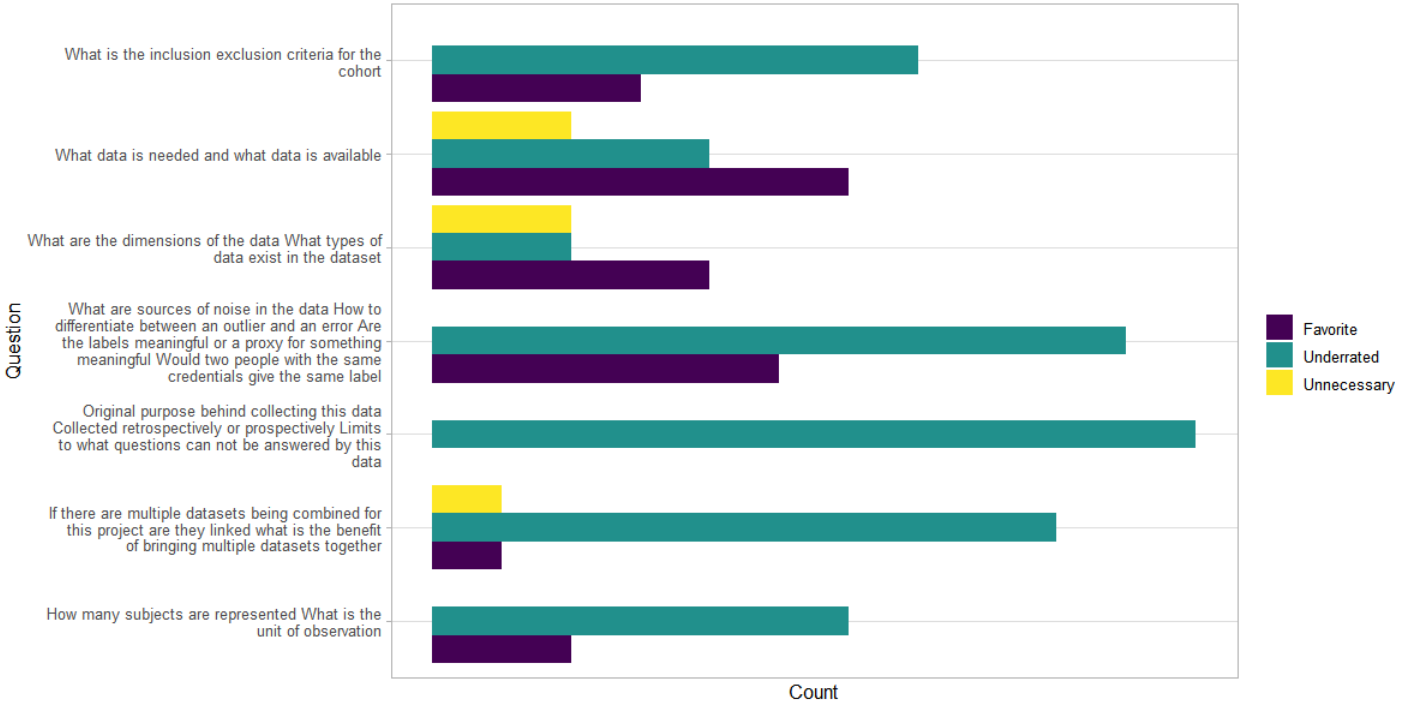
C.3. Defining Roles and Responsibilities



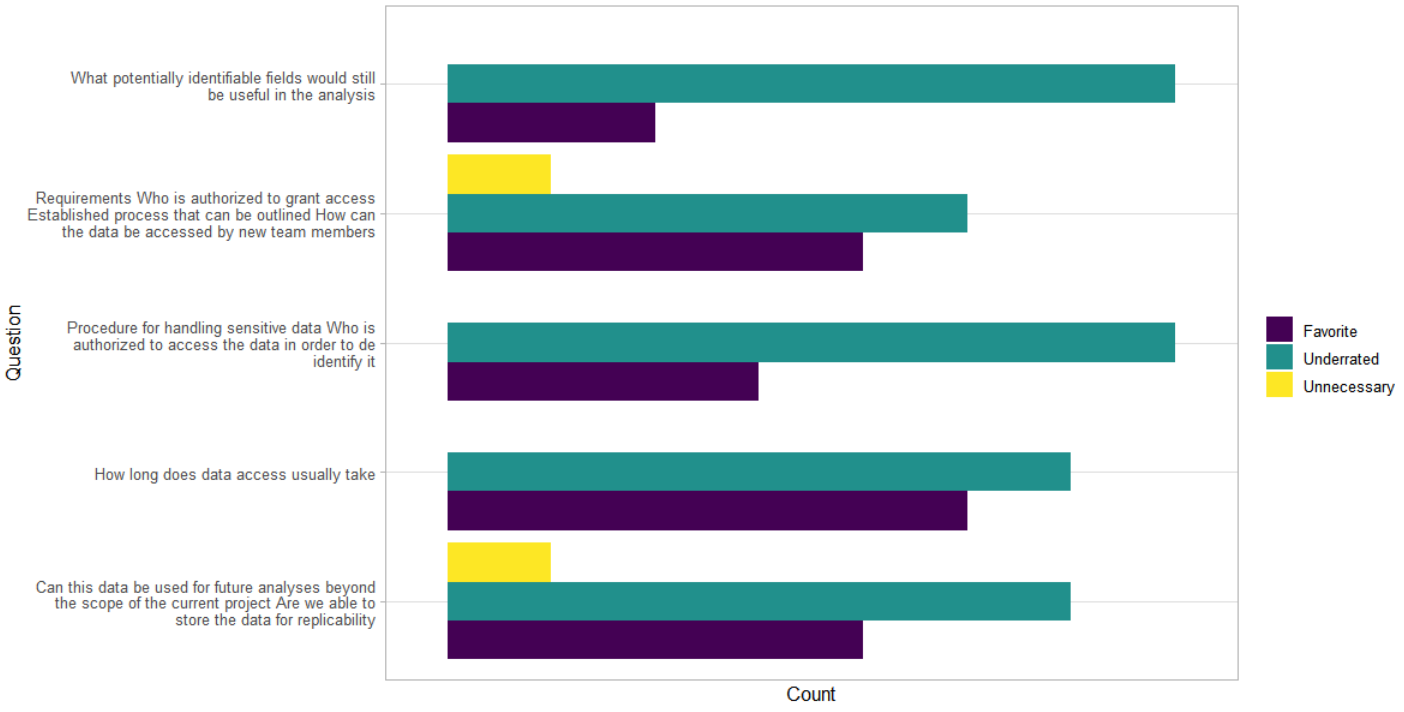
C.4. Data Generating Process



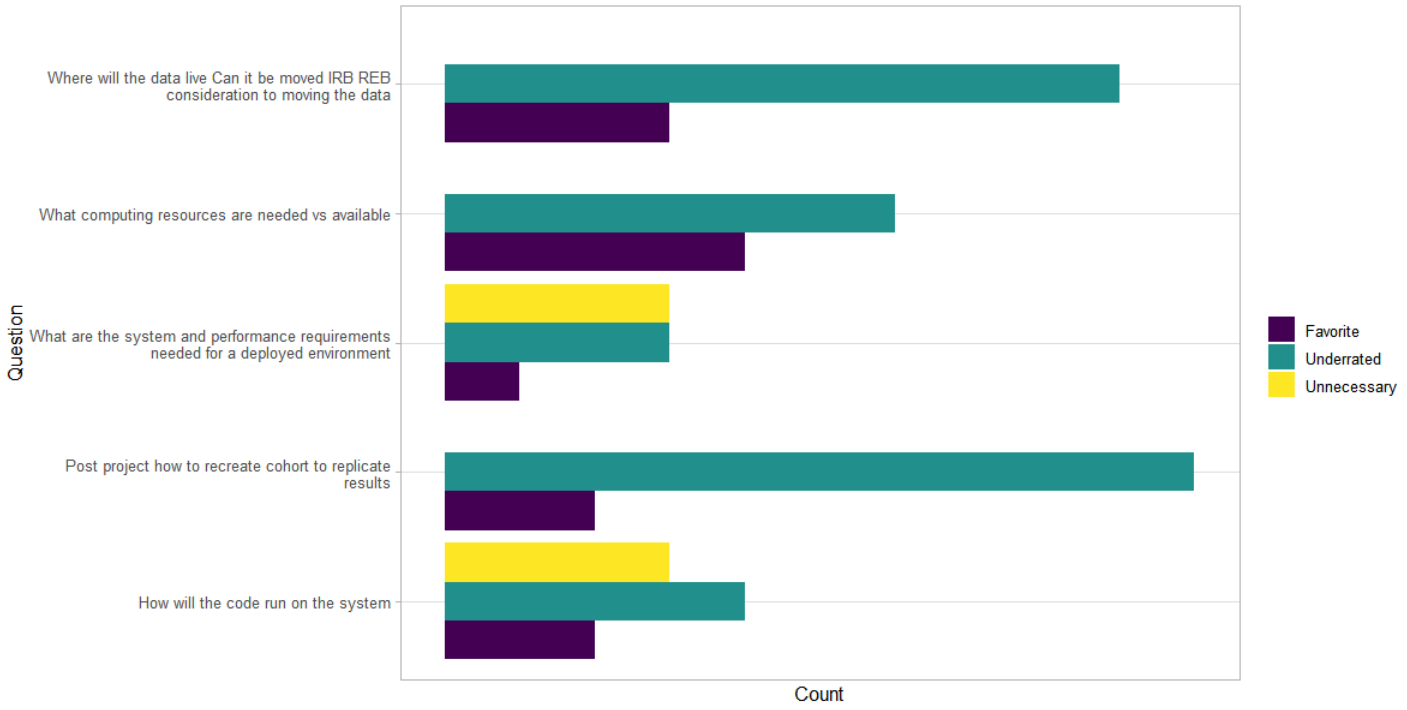
C.5. Data Composition



C.6. Data Access and Privacy



C.7. Storage, System Integration, and Compute Infrastructure



C.8. Safety and Fairness