When is Machine Learning Data Good?: Valuing in Public Health Datafication

Divy Thakkar*
dthakkar@google.com
Google Research
Bangalore, India

Azra Ismail*
Georgia Institute of Technology
Atlanta, USA
azraismail@gatech.edu

Pratyush Kumar
pratyush@cse.iitm.ac.in
IIT Madras
Chennai, India

Alex Hanna
alexhanna@google.com
Google Research
San Francisco, USA

Nithya Sambasivan
nithyasamba@google.com
Google Research
Seattle, USA

Neha Kumar
neha.kumar@gatech.edu
Georgia Institute of Technology
Atlanta, USA

ABSTRACT

Data-driven approaches that form the foundation of advancements in machine learning (ML) are powered in large part by human infrastructures that enable the collection of large datasets. We study the movement of data through multiple stages of data processing in the context of public health in India, examining the data work performed by frontline health workers, data stewards, and ML developers. We conducted interviews with these stakeholders to understand their varied perspectives on valuing data across stages, working with data to attain this value, and challenges arising throughout. We discuss the tensions in valuing and how they might be addressed, as we emphasize the need for improved transparency and accountability when data are transformed from one stage of processing to the next.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Data work, India, public health, valuation

ACM Reference Format:
Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is Machine Learning Data Good?: Valuing in Public Health Datafication. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 30–May 06, 2022, New Orleans, LA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3501868

1 INTRODUCTION

Working with data, or *data work*, is of emergent interest to the field of Human-Computer Interaction (HCI), where recent research has actively investigated the challenges that arise around data procurement, organisation, management, visualisation, and more across a range of domains [61, 62, 74, 82]. Many of these studies focus on the public sector, such as healthcare and public welfare, to foreground the design challenges that result from data work [47, 54, 82]. Møller et al. draw attention to the role of the human(s) behind the data who perform the work of data collection and processing [61]. The workflows and collaborative practices of data scientists, machine learning (ML) developers, and data annotators are also topics of growing interest in HCI [62, 63, 103]. We expand on this scholarship, investigating what happens when there are multiple humans working on (versions of) the same data, as they go from field to function, *i.e.* from their collection to their use in ML models.

Our research examines datafication efforts in public health, augmenting a body of work within HCI that has progressively been engaging on topics around frontline health (e.g., [7, 21, 43]). We draw on scholarship that discusses how frontline health workers (FHWs) procure data within the communities where they work and hand them over to their supervisors. This prior work reports on the disconnects that exist between the locally relevant insights of frontline health workers (FHWs) and the information sought by state and healthcare authorities, leading the data collected by the FHWs to be viewed by other stakeholders as inaccurate, incomplete, or simply unreliable [7, 42]. Our research provides a deepened understanding of this disconnect in *valuing* by investigating the perspectives of multiple stakeholders or *data workers*, as data are collected by FHWs, passed on to *data stewards*, who prepare them for the *ML developers* putting them to use in ML models.

We present findings from interviews held with the range of data workers employed in or contributing to the public health domain in India. We analysed the data collected from these interviews to arrive at an understanding of the data *supply chain* in public health, or how data changes hands from stakeholder to stakeholder. We draw from the field of valuation studies, particularly Heuts and Mol's discussion of valuing in a supply chain of tomatoes, to analyze how data are valued by stakeholders differently at various stages of their collection, processing, and use [36]. Our participants shared how they know whether data are good, and the work they must do to make the data good. We reflect on (a) the interdependencies of data work through the supply chain, (b) the values and priorities of different data workers participating in this supply chain, and (c) the need to make the labor involved in data work more visible.

Our paper is laid out as follows. We begin by situating our work in the context of prior research on data quality, frontline health and datasets, and valuing of data and data work. Our findings present

*Both authors contributed equally to this research.

CHI '22, May 8–13, New Orleans, LA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3501868

the data supply chain in our focus area of public health ecologies. Starting with the ML development stage, we describe how data change hands from one stage to the next, the ways of valuing data across the supply chain, and the work undertaken to attain that value. Drawing on these findings, we then discuss where tensions in valuing arise and are visible, and implications for the generation and curation of ML datasets. We argue for greater alignment in valuing through transparent and accountable processes and structures that empower data stewards and data collectors who are part of the supply chain. Our research insights seek to advance scholarship towards understanding valuing behaviors and practices in data work, paving the way for greater transparency and accountability overall in ML development.

2 RELATED WORK

Data-driven approaches are becoming increasingly central to HCI and HCI-adjacent fields. In this paper, we connect with three bodies of work that are centered around data. First, we engage on topics around data quality in ML, extending this research with our empirical study of a multi-stakeholder context where questions around quality routinely surface. Second, we augment the body of work on data and global health, given that data-driven approaches are increasingly being adopted towards healthcare solutions. Finally, we build on a growing area of interest within HCI—that of data work and the care that data routinely entails.

2.1 Data Quality in ML

The quality of training data has a significant impact on the quality of ML algorithms developed [33]. The aphorism 'garbage in garbage out' has frequently been used in relation to ML [31]. Recent research by Sambasivan et al. on data work among ML practitioners reports that data work is highly under-valued compared to model development, and as many as 92% of respondents reported experiencing data cascades, where issues with data collection led to even greater challenges on the development end [82]. Our research augments this work by examining the supply chain of data for ML in the context of public health.

With respect to data quality, ML efforts have typically been concerned with missing data, duplicate data, inaccuracies, highly correlated variables, too many variables, unreliability of labels among other data-related challenges [33, 46, 75, 97]. Recent research on algorithmic bias has uncovered other data quality dimensions—such as unbalanced datasets, finding that existing datasets tend to be heavily skewed towards white, male, Western, urban, and English-speaking settings [92, 104]. Zou and Schiebinger have noted that over 45% of Imagenet data, which computer vision research draws from significantly, comes from the United States, where only 4% of the world's population resides [104]. India and China together contribute only 3% of Imagenet data, while representing 36% of the world's population, resulting in cultural biases. For example, a photograph of a traditional US bride is annotated accurately while that of a North Indian bride is recognised as 'performance art' and 'costume' [87]. Data gathered may also reveal systemic racist and sexist biases, as reflected in algorithms used for policing or determining social welfare benefits [26, 69]. Data fields themselves can also be problematic, as Keyes points out, such as binary conceptualisations of gender in gender recognition algorithms [49]. Such challenges are further compounded when people share data across contexts, but taken-for-granted norms and standards of data from these contexts are not included [65]. These broader questions around data gathering have long been the focus of the field of critical data studies, which has closely examined how the context around data—the people, institutions, instruments, policies, finances, and more—impact data collected and their use [20, 44, 48, 50].

Data collection and curation practices can have significant effects on the performance of ML algorithms. Buolamwini and Gebru's pivotal study uncovered misclassification by facial analysis algorithms based on gender and race as a result of datasets being overwhelmingly comprised of white and male subjects [13]. Ntoutsi et al. have described how bias in algorithms is introduced with decisions made by humans on data gathering and processing, such as relying on easily available data and over- or under-representation of certain groups [67]. In their survey of bias and fairness in ML, Mehrabi et al. created a taxonomy of 22 forms of bias that ML systems can exhibit, partially because of data practices [56]. Miceli et al. have brought focus to the power structures that influence data quality in computer vision [59].

Researchers have recently called for paying more attention to data gathering and processing, such as by documenting how data were labeled [5, 31, 39, 63]. Denton et al. have outlined a research agenda for documenting the genealogy of ML data, investigating the origin of and values engaged in the collection of benchmark datasets such as ImageNet [20]. Prior research has also examined the perceptions and needs of various communities such as AI developers, clinicians, designers to mitigate bias and mitigate varying expectations of data [28, 37, 65]. Similarly, ethnographic studies and survey reviews have explored narratives that shape datafication, mostly in settings in the Global North [24, 81, 91]. Our research builds on such prior work by investigating the origin and flow of data through its lifecycle comprising of several stakeholders. We extend this body of literature by looking at the broader data pipeline as a supply chain, documenting how datasets are constructed and how decisions made by people at each stage impact data quality for ML.

2.2 Data and Global Health

Data collection practices have long been central to the field of Global Health to track health outcomes within a population. For decades, the public health systems of many countries in the Global South have been engaged in paper-based data collection for reporting purposes [19]. Prior work has engaged with field practices of frontline workers and the challenges in overcoming data quality issues. Pervaiz et al. have examined the challenges in using data for developmental goals in Pakistan and propose a taxonomy of data cleaning challenges [76]. Batool et al. have described the strategies and management styles adapted by supervisors to detect data falsification by frontline workers in Pakistan [7]. As the costs of mobile phones and the internet have fallen dramatically, data collection in Global Health is increasing being digitised [19, 43, 70]. Recent work has begun to explore how AI/ML interventions can leverage these data streams.

Proposed AI/ML applications in Global Health include chatbots to support breastfeeding practices [98], ML to forecast the spread of tropical and infectious diseases like dengue and Ebola [14, 34, 72], and deep learning to support screening and adherence to medicines for diseases like tuberculosis [2, 55, 58]. AI systems are helping with early detection and diagnosis [1, 9, 101], drug discovery [16], as well as outcome prediction and prognosis evaluation [38, 94]. AI is also being used to support maternal and child health, which are of particular importance in many parts of the Global South. Here, interventions have been developed to support early screening of low birth weight and preterm infants [80], and maternal health programs [66, 71]. Increasingly, data sources outside the healthcare system are also making an appearance in public health. For instance, researchers have studied the use of social media posts and search queries to monitor and predict the spread of diseases [95], though these assume widespread technology penetration. Prior work has highlighted limitations around transparency and replicability in the use of large datasets for scientific analysis, even though the output instruments were not designed for such analysis [52].

Within HCI, emerging research has begun to explore the human-centered design of AI/ML systems in healthcare. For example, Beede et al. undertook an ethnographic study of a deep learning system for diabetic retinopathy in hospital settings [8]. Okolo et al. have highlighted the need for explainability in AI systems to help communicate model outcomes to people with low literacy [68]. Others have studied how AI could support the routine work of clinicians [100], and help with collaborative decision making across medical experts [15]. Sendak et al. have previously deployed an AI model in a clinical setting as a socio-technical system and argue for moving beyond model interpretability towards identifying stakeholder relationships, creating ongoing feedback loops and respecting professional discretion [86]. Gu et al. have also studied how pathologists might use imperfect AI in their work [32]. Despite the reliance of such interventions on large datasets, little research describes how public health data are collected, digitised, aggregated, and then processed into machine learning datasets. Prior work on the data collection practices of FHWs demonstrates how data quality is impacted by the context around data collection [42, 70]. We build on this thread by examining the data pipeline from data collection to the development of ML models.

2.3 Valuing in and of Data Work

The field of valuation studies focuses on understanding the tensions, determinants, contexts, and effects of valuation practices, or how assets are assigned value. We particularly draw on Heuts and Mol's discussion of valuing in a supply chain of tomatoes [36], to analyze how data are valued by stakeholders differently in a data supply chain. Examining valuation at various stages of data collection, processing, and use gives us an opportunity to identify the underlying practices and tensions that constantly shape and transform data to make them 'good'. In their study of the valuing of tomatoes across diverse stakeholders in the supply chain, Heuts and Mol closely examine this constant tinkering and negotiation of 'good' drawing on an ethics of care [36]. They identify five registers of valuing tomatoes—money, handling, historical time, what it is to be natural, and sensual appeal [36]. Value registers point towards shared relevance of valuing an entity, where the value of a 'good' entity differs and is likely to be in tension in different situations. What makes a tomato 'good to eat' involves the investment of care at each stage, which may not necessarily succeed [36].

Different values and considerations may drive practices for each stakeholder, but these practices collectively determine whether a tomato is 'good' and consumable when it reaches a customer. Mol further emphasises that the 'good' is not something to pass a judgment on, but something to *do*, in practice, as care goes on [60]. In our paper, we bring the analogy of tomatoes to data. In particular, we look at the case of datafication of public health to reflect on the *data supply chain*, or how data changes hands from stakeholder to stakeholder. In the process, we examine the valuing behaviors of our participants, or how they assign and add value to data, based on their partial perspectives. Our approach is closely related to emergent work that looks at data settings with a feminist *ethics of care*, which draws attention to emotional and material entanglements in the data lifecycle [18, 23, 53]. Prior work by Chen et al. frame the ethics of ML in healthcare through the lens of social justice [17]. Recent work has explored a logic of care for use of data by grassroots organisations to advocate for their communities, for example, for permanently affordable housing and surfacing resident concerns [4, 23, 57, 102].

Prior research in HCI has also examined value and valuation in design [11, 29, 30, 40, 89]. Speed and Maxwell discuss the role of value constellations and the mediation of value across value constellations for designers [89] . In their ethnographic work in charity shops, Elsden et al. call to recognise valuation as an ongoing process and design data work to support the performance of stakeholder's value [25]. Feinberg finds that data are infused with design process including data infrastructure, data collection and data aggregation [27]. Iansiti examines the incremental value of data to benefit consumers and technology companies. They value data across four factors: data quality, scale and scope of data, and data uniqueness [41]. Recent work by Scheuerman et al. brings focus to the disciplinary values in computer vision dataset development and argue for a closer look at the trade-offs such as efficiency at the expense of care, universality at the expense of contextuality and impartiality at the expense of positionality [84]. Studies have also explored the politics and power structures within data work [4, 59, 77]. Singh calls to situate data infrastructures through existing practices around dataset development and understand relationships that hold them together [88]. Our work builds on these studies and provides insight into the role of valuation and the data transformations caused due to hidden tensions across stages.

HCI researchers have examined the role of collaboration between data science workers for data work [6, 51, 62, 64, 73, 74, 78, 103]. Discussions on collaborative data work have focused on collaborations among data science workers employing collaborative practices across data science workflows [103]. Data science workers develop an intuitive sense of their data and employ strategies such as rationalization and decomposition to reduce tensions in the data workflow [62, 74]. Studies have also uncovered how data science workers need to adapt different coordination and communication strategies while working on multidisciplinary teams with domain experts and could benefit from building a process common ground [51, 54, 64]. Data science workers also face challenges working with data and

spend significant time data cleaning, data wrangling and working with 'dirty' data [35, 90]. We build on this work to discuss the collaboration practices within each stage of the data workflow and analyze the role of collaboration across the data workflow through the lens of valuation to make data 'good'.

3 METHODS

The goal of our research was to attain a deeper understanding of the origin and evolution of data in the public health system in India, before it comes to be used for ML. Our study draws from 46 semi-structured interviews with individuals involved in different parts of the data supply chain. Interviews were conducted from May 2020 to August 2020. Participants include data collectors, data stewards, and developers linked to ML applications being developed for public health such as healthcare resource allocation, improving public health program adherence, health outcome predictions, community health surveillance and healthcare worker evaluation. The study was approved by the author's organisation's ethics committee before the work was commenced; the committee was comprised of health ethics experts and approved the research design without raising concerns. Below we offer more context on the stakeholders we interviewed, our recruitment process, data collection, and data analysis.

3.1 Participant Information

Data collectors in all the projects were frontline health workers (FHWs) including government-funded health workers Accredited Social Health Activists (ASHAs), senior government-employed frontline health workers called Auxiliary Nurses and Midwives (ANMs) who were responsible for supervising ASHAs, and outreach workers for HIV and sexual health awareness and support. FHWs were responsible for providing access to healthcare services in their communities and performed data collection along with their care responsibilities [22]. FHWs in our study were involved in project domains including maternal and infant healthcare, sexual health, disease monitoring in communities and vaccination coverage. They collected demographic and health data from beneficiaries using paper forms and, occasionally, tablets.

Data collectors, FHWs, passed on data to data stewards at a Primary Healthcare Centre (PHC). Data stewards included data entry operators, data controllers, and project officers. Data stewards were tasked with cleaning, processing, analysing and validating the data from data collectors to upload it onto an information management system. Data stewards worked on similar project domains as data collectors and maintained data records in PHCs. Focus areas for data collectors and data stewards was also dependent on state public health priorities since each state dealt with different public health challenges and severity of diseases.

The final set of stakeholders are the developers, including ML researchers, data scientists, and engineers. ML developers perform advanced analysis and modelling on public health data, which is acquired often in conjunction with an NGO or a public health agency. ML developers were involved in projects such as efficient resource utilization, predicting community health outcomes, improving public health program adherence and understanding community social networks for improving program coverage.

3.2 Participant Recruitment

To recruit participants who were involved at different stages of data work, we recruited participants through professional networks, personal contacts, snowball and purposive sampling and stopped when our data had reached saturation [10, 85]. We acknowledge the tremendous support from three grassroots non-governmental organisations (NGOs) in recruiting frontline health workers for our study. Our dataset is intentionally geographically balanced to understand data practices across different public health programs in different states. To obtain a geographically balanced sample, we recruited FHW participants across sixteen villages and one city in India across five states including Rajasthan, Uttar Pradesh, Madhya Pradesh, Karnataka and Bihar. We recruited thirteen data stewards, working in jobs of data collection and analysis in public health organisations. Additionally, we recruited ten ML developer participants from three countries; ML developers were recruited by seeking references from author's professional networks, and personal contacts. All developers who participated in our study were actively working on problems at the intersection of ML and public health with the goal to deliver the software to public health agencies, NGOs or a partner organization. ML developers we interviewed came from academia, start-ups, and industry research labs and were located in cities in India, the United States, and Singapore. Some developers in our study were situated in countries outside of India because many academic and industry research projects at the intersection of ML and public health for India are globally distributed.

3.3 Data Collection

Due to travel constraints as a result of the COVID-19 pandemic, Divy conducted interviews via phone calls and video calls on Google Meet with study participants. He conducted interviews in Hindi, Kannada, and English depending on the participants' choice of language. As he was not fluent in Kannada, he was accompanied by a translator for interviews. In ten interviews, Divy was joined by an NGO representative who was familiar with FHWs and helped build rapport and trust with the FHW. After initial exchanges, the NGO representative did not interject during interviews and were not provided with the list of interview questions in advance. In other interviews, the NGO representative helped build rapport prior during participant recruitment and did not accompany during participant interviews. Interview questions with FHWs were focused on understanding their backgrounds in day-to-day workflows, public health data collection and analysis, incentives, system design, communication workflows, community interactions and their interactions with technology for their work. Interviews with data stewards and ML developers were additionally focused on capturing their experiences and challenges on working with data entry, processing, analysis, and modelling. Interviews lasted 45 - 60 minutes with all participants.

We obtained informed consent from all our participants, taking particular care to communicate the purpose of the study clearly to FHW participants who were unfamiliar with ML and were sharing sensitive information. The NGO representative and Divy assured FHWs multiple times that their responses were anonymous and their responses were not linked to their performance in any way.

Communicating this information was critical as surprise visits from higher ranking officers was regular practice and could induce bias into the data collection process. Sixteen FHW interviews were not audio recorded as per the participants' privacy preferences, since they worked directly with government institutions and did not feel comfortable being recorded. All participants consented to detailed note-taking during the study. We paid careful attention to the responses of our participants throughout, and did not sense fears or anxieties on their part. Divy left them with their contact information, inviting them to get in touch if they had any questions. This is all in addition to following standard procedures of informed consent (in participant's local language) and reinforcing that participants could end the interview if desired, at any time.

3.4 Data Analysis

All transcripts were translated to English by Divy. Participants' names were anonymised and replaced with codes in the analysis. NGO names have also been anonymised for this paper to protect the privacy and interests of frontline workers. We carried out inductive qualitative data analysis to summarise and interpret the interview data. The process of analysis was iterative: we began by identifying themes at sentence-level and identified larger themes emerging from these iterations. All authors read transcripts multiple times, developed affinity clusters, and derived key themes, which we iteratively refined [93]. Emergent themes were guided through our analysis and were organised across stakeholders by tracing the use of data from ML developers to the process of data aggreation and eventually collection. Within each stakeholder, we refined our themes across data operations, data contextualization and organizational structures that shape the perception of 'good' data and the practices to create, analyze or maintain it.

4 FINDINGS

In their study on the supply chain of tomatoes, Heuts and Mol describe the work performed on the tomato and the valuing by stakeholders at each stage to make the tomato good [36]. Taking inspiration from their analysis, we view data as part of a supply chain, where data are handed over from one set of stakeholders to the next. We first describe the work conducted by ML developers to develop ML models from the noisy datasets they receive, and the valuing driving this work. We then detail the data stewards' data entry and processing practices. Finally, we describe the work and valuing by data collectors. Across the supply chain, we highlight differences that arise in the valuation of data.

4.1 ML developers: Making Data Fit for Purpose

Our interviews with ML developers revealed that they spent a substantial amount of time making data good for developing ML models. Their perceptions of what entails a good dataset for ML resulted in several operations on the data to make them fit for purpose. In prior work, Sambasivan et al. have described how data work has a significant role to play in ML work, despite frequently being undervalued [82]. Our participants also pointed out the need to pay attention to how their data work could impact their models, and sought to contextualize the dataset to make appropriate data operations. They recognized their limited understanding of the context in which data were collected, and struggled to contextualize their models for solving relevant problems using ML. We describe their work and the valuing involved below.

4.1.1 Operations to make data good. Our ML developer participants reported facing several data quality issues for model training, and shared their concerns about lack of visibility into the data collection process. We found that they relied on partner organisations to procure data for them for specialised use cases and were not involved in the design of the data collection process. ML developers reported that the key constituents of good data included structured and standardised feature-rich data with validated ground-truth labels that would help them build robust ML models. However, developers frequently faced challenges in the process of working with data and encountered noisy data:

> *"The data would frequently have large number of missing rows and unfilled columns. In some cases key parameters like blood and urine reports are missing in maternal health data which is a critical field for health outcome prediction. It is difficult to trust this data to validate ground truth. The data quality is so bad that predictions of high risk pregnancies are totally off" — P37, ML developer*

Missing data fields were a commonly reported issue among our participants. Other issues included large clusters of near-perfect values in a dataset and swapped entities (for instance, height with weight). Developers reported receiving crude datasets that had not been collected for ML applications, typically from non-profit organisations or public health agencies. ML developers noted the process of data transformation towards achieving better modelling by performing various data operations towards improving data the quality and completeness of the dataset, and preparing them for the ML data pipeline:

> *"My goal is to try and reach the highest accuracy and noisy data makes that really difficult. I spend a large portion of my time in preparing the dataset even before getting to the modelling. If it were up to me, I would not do this task but it is a critical part given the quality of data."—P41, ML developer*

Noisy data introduced a number of additional, laborious tasks for the developer and consumed significant time to prepare the dataset for model training. There were several strategies that our participants employed to address such issues. ML developers reported hand curating datasets and applying less data-hungry techniques when developing ML models, especially when data quality was poor. ML developers shared that they identified data clusters and key trends through statistical analysis to understand and scope the problem further. These operations also helped identify noisy parts and gaps in the dataset. Developers reported following an iterative method in the data transformation process, first by retaining noisy data rows and later by retaining high-quality data and training their models. Models with higher accuracy and less noisy data were preferred. These strategies were implemented with the limited context that ML developers had on how and why the data

No.	Roles	Locations	Domains
P1-P23 (22F, 1M)	Data collectors (Frontline health workers)	Rajasthan, Uttar Pradesh, Madhya Pradesh, Karnataka, Bihar	Maternal Health (9), Sexual Health (6), Other (8)
P24-P36 (4F, 9M)	Data stewards	Karnataka, Maharashtra, Delhi	Maternal Health (3), Sexual Health (4), Other (6)
P37-P46 (2F, 8M)	ML developers	Karnataka, Singapore, United States	Maternal Health (3) , Sexual Health (1), Other (6)

Table 1: Demographics of research participants. "Other" refers to a broad umbrella of projects in domains such as tuberculosis, non-communicable diseases, preventive care, health outreach etc.

were collected a certain way. In making these data operations, they could lose valuable context captured in the data.

ML developers were thus engaged in making data good by addressing the noise in the data, and introducing some structure. Our paper offers insight into valuing by data collectors in later sections, how this might have introduced noise in the data, and implications for data operations. We next describe how ML developers tried to contextualize these data operations in the absence of insight into the broader context of data collection.

4.1.2 Contextualizing data operations. ML developers we interacted with frequently collaborated with partner organizations providing the data, to make better decisions on what operations to conduct to make data good. There were cases, however, where there was no suitable resolution that our participants had arrived at with their current level of understanding of the context. We describe these contextual challenges and workarounds below.

We found that developers struggled to find reliable documentation on how data had been formatted. In such cases, they relied on partner organisations to provide more information about the data format and organisation of the data, and to comprehend data fields that required domain expertise, such as those on medical conditions or terms that the ML developers were unfamiliar with. Developers also shared the challenge of connecting multiple datasets, even if they were related to the same set of beneficiaries, which such lack of documentation would have further made difficult. This made it challenging to track the long-term impact of their ML algorithms and to fully understand the improvements of their model on the beneficiaries' health outcomes:

> *"You can see that two datasets have intersecting people but there is no easy way to connect these datasets. It becomes very difficult to say with confidence that your algorithm has a long-term impact on someone's health if you cannot measure the outcomes over a long period."—P39, ML developer*

More context on how the data were collected and more standardized structures for collecting data could have helped in this case. Another challenge was that of multiple languages in the same dataset, because this introduced more steps in the preparation of the ML data pipeline. Public health data, especially textual data, was collected in regional languages and hence developers were forced to expend significant time and effort to process such data fields. The different data and written literacies of stakeholders collecting and aggregating data may have played a role here.

Labelling datasets was reported to be an important yet one of the most challenging tasks during data processing and ML model development. Developers frequently relied on their data provider to help them identify a set of rules or heuristics to produce the correct labels, especially in domains where specialised expertise was needed to understand the data and required output. In other cases, developers tried to decipher patterns in the data and label the data by themselves, by hiring gig workers or through crowdsourcing.

> *"If I am trying to assess which person is at more risk, I need to know which of their health parameters in data are important to determine the risk, right? It becomes really difficult to identify the right labels without experts and sometimes even when you are able to figure out the labels, the underlying data cannot be trusted or the labels are subjective"—P37, ML developer*

We also found that validating these labels through ground truth was considered to be a very important part of the labelling process. In the absence of partnerships with field experts, it was difficult for developers to know the accuracy and efficacy of existing labelling operations on the dataset, in turn affecting the model performance and result validation. Data labelling was noted to be difficult and time-consuming. ML developers reported challenges in understanding subjectivity and biases of the people who labelled the data when they relied on other experts. Labelling without an understanding of the context in which data were collected was hence noted to be particularly difficult; developers were unsure whether *"labelling was in accordance to the real needs of the problem" P40, ML developer*

Our participants also noted that good data protected the privacy of those from whom data were being collected. ML developers reported that frequently, NGOs and public health agencies did not have the means or technical know-how to completely anonymize datasets. As a result, these tasks were routinely performed by ML developers. ML developers thus did the work to create datasets without private details.

> *"I understand that they (NGO) do not have the technical means to make data completely anonymized so we have do it because we have to make sure we are following protocols submitted via IRB and that there are no questions when we go for publication"—P42, ML developer*

We see that there were several data operations that ML developers worked on that required contextualization such as connecting multiple datasets, handling data collected in multiple languages, labelling data and validating these labels, and anonymizing data. They relied on partner organizations to provide the domain expertise to be able to contextualize appropriately, but even with their support, these were hard challenges that were had not been fully resolved.

4.1.3 Contextualizing models. ML developers noted that there were a large number of interesting and exciting problems in their local communities but they were extremely selective in picking the problem because of the lack of availability of high-quality data and the difficulty in implementing the ML algorithms without a trusted on-ground partner. ML developers frequently assessed the feasibility of working on a specific problem based on the availability of high-quality data. In other cases, ML developers preferred to work in collaborative teams that could potentially work with partners such as NGOs and public health agencies with access to high-quality data.

In cases where the data-providing agency and the developer were working closely, developers noted that it was easier to ask questions about the data to understand observed patterns from modelling. They reported working closely with partner organisations to validate key observations and insights, and to test for their robustness in practice:

> *"I can sometimes see correlation between two parameters but does that really lead to any actionable insight? I need to work with domain and field experts to understand that, getting access to such people is not easy."—P46, ML developer*

ML developers shared their concern for whether their models were solving the right problem by focusing and learning the right data features. They reported that the process of validating their ML solution into giving the most useful predictions was a task which needed field and domain expertise that they did not possess. For instance, these challenges included getting a grasp of the validity of causation and correlation of parameters on the output and building consensus and capacity among field workers to use their developed technology in the real world. We found that developers needed to find domain experts who could help identify and scope out the right problem and the features to focus on. However, easy access to such expertise was noted to be a limiting factor. We report that good data were accompanied with domain expertise and a clear understanding of the problem scope.

We also found that developers frequently performed analysis to understand the bias in the dataset and their models. They reported to being sensitive to not cause harms through their models, however we found that developers struggled to contextualise bias for different settings since they did not have the field expertise and relied on collaborations with field partners and subject-matter experts:

> *"Fairness is an issue. Even when we have many records only a few of them are more diverse than others so you cannot use all the records for your algorithm because of the lack of data balancing. It is also very difficult to know which data field can be a proxy for bias so it becomes difficult to balance outcomes as well"—P37, ML developer*

We note that developers had access to a complete or broader dataset compared to other stakeholders like FHWs and data stewards who deal with piecemeal data. This gave them an opportunity to provide insights on broader trends across regions or populations that may have otherwise not been visible. However, this required careful contextualization, including a consideration for bias in the data and models developed. Close collaboration with domain experts and project partners was critical in this regard.

4.1.4 Training and career motivations. All developers we interacted with shared that they had largely been taught to work with high-quality datasets during their universities and training. Developers shared the realities of working with public health data as starkly different from their training, and noted that they spent a significant time to ensure preparation of high-quality datasets from crude data rather than getting to directly work with high-quality data.

> *"It's mostly my students who work with the data, and I can tell you from the students' experience, there is a lot of time spent and they are not extremely fond of it. They want to work on cooler things like the algorithm and how to work with that. But data issues in themselves actually pose a technical challenge and we can have novel methods for such low-resource contexts."—P43, ML developer*

We found that our ML developer participants did not receive direct incentives for performing routine data work to develop data pipelines for ML models. However, application and invention of novel data mining techniques was valuable for their work although in most occasions existing techniques would help get their job done. Additionally, development of new algorithms to prove scientific robustness was noted to be an important consideration in some cases, especially for scientific publications. Though we found that ML developers were motivated by the potential for impact through their model development work, they were also being motivated to publish study results in reputed venues and showing promising results (such as accuracy, precision, recall, F1 scores) to further ML deployments and gain organisational recognition. Developers noted the pressure to work on a problem with technical depth which would be valued by the scientific community and their organization, also potentially aligned with their career aspirations, over problems with an easier technical solution but potentially higher field impact.

4.2 Data Stewards: Aggregating Data to Meet Reporting Requirements

We next share the perspective of the data stewards and how this contrasted with the work and valuing by ML developers. Before data reached the ML developer, they were aggregated and processed by data stewards. The data stewards we interviewed performed a range of operations on data collected by FHWs through paper forms, daily diaries and notebooks. These operations included data entry, processing, and analysis, as detailed below.

4.2.1 Making data legible and complete. Through our interviews, we found that the work of data stewards to be primarily motivated

by reporting requirements for aggregated data. They were tasked with digitising data records and generating reports to enable the generation of summaries of health performance indicator for government officials. They identified structure and completeness as key constituents of good data. A critical component of their role was translating paper forms into structured and complete data on health informational management systems, to support robust reporting on goals that were part of health policy by state and federal government:

> *"Getting data that is complete and readable is the basic need for us to provide reporting mechanisms for the state.—P28, data steward*

The format for digitising records in health information management systems was prescribed to our participants during the training they received at the start of their role. However, the data that our participants received was frequently not 'good' as desired. Our interviews revealed data stewards' concerns about the quality of the data provided by FHWs. They reported prevalent issues such as missing data fields (especially for medical history), data fudging, misrepresentative data entries, and data manipulation. They also struggled with deciphering handwriting in some cases:

> *"Data that comes to us has a lot of gaps... there are missing entries, sometimes you can take a quick look and see some glaring numbers which I know do not make sense."—P32, data steward*

Many of these challenges echoed those shared by ML developers. However, though data stewards tried to address these issues, the problems persisted beyond this stage. Data stewards also reported dealing with multiple forms and entries for the same individual (from a house visit or a health camp check-up), which made it a frustrating and lengthy process to digitise. This may have further contributed to the data quality issues experienced by ML developers, and may also explain the challenge that ML developers identified when connecting multiple datasets with the same individuals.

Our interviews also revealed that like ML developers, data stewards found 'good' data to be privacy-preserving. Several sensitive and private fields including socio-cultural factors such as caste, religion, and gender were captured during the data collection process. Data stewards performed several activities to preserve data privacy. Public health data were transferred from FHWs to Advanced FHWs in a Sub-Health Centre, to data controllers and stewards in a Public Health Centre (PHC). Data were aggregated and processed to remove personally identifiable information at the PHC, prior to which each stakeholder in the process had access to individual medical history and current reports of each beneficiary. Beyond the PHC, data were aggregated and trends were to be reported. Further, data stewards shared that part of their work involved ensuring that data entries had unique IDs and could be connected across different datasets, which was a difficult task due to different data collection structures that generated each dataset. Though both ML developers and data stewards cared about data privacy, the specific level of privacy and procedures followed likely differed. Even after the processing done by data stewards, ML developers found the need to perform additional privacy-preserving data operations.

4.2.2 Organizational structures for meeting data quality. We found that there were several organizational structures in place to check the quality of data collected by FHWs, though there were also significant gaps and misaligned incentives that impacted data quality. Data stewards were geographically located far away from FHWs (albeit in the same state as the FHW) and in most cases, they did not maintain any direct communication with FHWs. Data stewards were located in Public Health Centres (or regional NGO centres) where they followed defined protocol and performed work such as data entry, aggregation, and reporting. Data stewards noted that organisational structures were not set up so that data stewards could have direct oversight or influence on the data work of FHWs. Instead, data quality checks were combined with the supervision of health work performed by FHWs. Senior officers from the PHC visited FHWs for surprise visits and checked their registers to validate their work. However, the purpose of such checks was to ensure work was being done than to enforce data quality.

Data stewards highlighted a lack of access to tools and training for efficient ways to understand data quality. We found that they spent significant time validating data from FHWs through manual checks, going through the data provided by a single FHW to develop a broad impression of the quality of the data they were collecting. They reported any specific instances of continued data quality issues with an FHW to their supervisors. In some cases, especially at NGOs, data stewards directly communicated with FHWs to seek clarifications and report concerns.

> *"I will check the data in my own way to see if someone has left compulsory fields empty or if they have tried to enter perfect values but all this takes time and I cannot spend all day in doing checks..."—P30, data steward*

Asynchronous communication and data updation delays were also identified by our participants as a significant reason for the ineffectiveness of current data and perceived poor data quality. Data stewards often experienced a long delay between data collection to being ready for the data ingestion and reporting stage. Each time a political or bureaucratic leader was interested in the poor performance for any identified district, instead of accessing prior health records (which were perceived to be of poor data quality), a fresh round of data collection was ordered. This further burdened the public health system by adding additional data work for FHWs and data stewards.

Data stewards also shared their perception on financial incentives for FHWs. They reported that some FHWs optimised for activities that paid larger financial incentives which created focused reporting on some aspects of the public health scenario while other programs might have very little data and reporting since they had lower financial incentives. This may explain why ML developers found duplicate entries in their datasets, and also has implications for how data operations should be performed by developers. There was thus a gap between valuing of data by FHWs who prioritized incentives, and that by data stewards and ML developers who were aligned. Data stewards cared about breadth and completeness of reporting, and ML developers cared about feature-rich data.

4.2.3 Organizational structures for reporting. Data stewards performed practices to enable better health reporting mechanisms for state officials. However, they lacked incentives for good reporting,

did not have the autonomy to define reporting structures, and were overburdened and had to operate amidst resource constraints. They also had to work with limited visibiity into the impact of their data reporting work.

Our interviews revealed that data stewards were called upon on an *ad hoc* basis to provide specific data reports (such as immunisation rate in a particular district). Such requests were typically from senior officers and having organised data readily available was looked upon favourably. However, preparing these reports could involve combining data from different sources which was a difficult and time consuming task, causing frequent delays when specific data reports were requested. Data stewards reported that they did not receive any financial incentives based on their performance on data work, and did not have any control over the structure of data collection format or forms. However, we found that the release of funds by the state or federal government for specific programs was dependent on good performance on relevant health parameters.

Despite their essential work in informing state policy, the data stewards we interviewed did not have information on whether their data were used and where they were reported: *"I type in the data, where all it goes after is unknown to me."—P26, data steward.* Our participants did not get any regular feedback on the efficacy of their data work, making it a frustrating experience because the value of their work was not immediately clear. Data stewards stated that in an ideal scenario, aggregated data would be used to observe trends to influence health policy but as far as they were aware, there was typically no action based on this data. The inaction was attributed to the lack of good quality data, which was further attributed to not having robust feedback mechanisms in the data collection process and the lack of oversight in the process:

> *"In most cases, the data is not acted upon after collection and ingestion. Data is mostly used for reporting, no further analysis is done on it."—P25, data steward*

We also found that data stewards navigated organisational limitations and resource constraints to perform their work. PHCs and NGOs had limited human resources to perform data entry, data cleaning, and analysis. Data stewards reported a large backlog of physical files that needed to be processed and digitised. The backlog caused a further delay in the availability of near real-time data for analysis which was sought from senior officials. Further, lack of automated techniques for cleaning and processing data were noted as other contributors to the long delays in data preparation. Data stewards noted their challenges in availability of credible statisticians in their departments to perform advanced statistical techniques for detailed analysis. This further burdened other departments for such analysis, increasing the feedback and response time for the data steward.

We see that reporting requirements were driven by bureaucratic needs that influenced valuing by data stewards. Data stewards were overburdened, lacked autonomy to define structures of reporting, and lacked visibility into how data were used. The complex organizational structures they were forced to work within along with limited training and resources impacted the level of data quality they were able to maintain and informed how the data were structured. This impacted the quality of the data made available to ML developers who faced similar challenges when conducting data operations for ML.

4.3 Data Collectors: Collecting Data as Part of Everyday Workflows

The valuing performed by ML developers and data stewards differed significantly from that of the FHWs collecting the data. FHWs performed data collection alongside their primary role of healthcare outreach. We found that organizational structures and the local context shaped data collection by FHWs and their valuing of the data. Below we highlight how data transformation took place, the factors that resulted in data quality issues identified by the other data stewards and ML developers, and implications for ML datasets.

4.3.1 Organizational structures driving data collection. We found that data collection was performed by FHWs for two primary uses—generating and maintaining public health records of care coverage including their own healthcare provision activities, and conducting specialised surveys on a need-basis. Both of these were closely linked to organizational structures for data collection, including financial incentives and supervision. Our FHW participants reported spending significant time on year-round data collection tasks. For instance, FHW P7 stated, *"We get a lot of notices to do a survey (for different programs such as spread of dengue) with the form to fill out. It is a very tiring process but we need to do it and a good part of my salary comes from doing surveys."* Data collection was perceived as a tedious and tiring process by our participants, spanning multiple forms, surveys, diaries. FHWs shared that the data collection process had several components which were opaque, repetitive, and redundant (such as multiple forms requiring similar data fields).

We found that FHWs associated performance on healthcare provision activities with their data work. They received financial incentives for each task completed, and data were used to demonstrate they work they had completed to receive payments. As described by P4, *"...register should be tip-top (polished in presentation) and complete."* Data were also viewed as a way to showcase their healthcare provision efforts to higher-ranking officers, especially during surprise checks. Additionally, good data and performance increased the visibility of the FHW within their PHC or NGO. 'Good' data for FHWs thus reflected their performance on healthcare tasks, and were well presented and complete. The linking of financial incentives to specific activities resulted in data stewards and ML developers receiving more data on some activities than others and missing fields, as identified earlier. FHWs also received financial incentives to procure new data for specialised program surveys. In such cases, FHWs reported their incentives to be lower than their expectations. For example, we found that FHWs were paid USD 14 per month to conduct COVID-19 surveys with a hundred houses every day. Low incentives may have impacted their motivation to collect high quality data in such cases.

The performance of Advanced FHWs was also tied to data, though their work was not based on financial incentives. They received fixed salaries and held the perceived prestigious status of a 'government employee' (government jobs are frequently sought after in India). Advanced FHWs oversaw a cluster of villages (typically five to six) and were responsible for the overall performance of their cluster. They were responsible for ensuring flourishing health

parameters such as low infant mortality rates. Lack of 'good' performance as reflected in data, led to warnings by Medical Officers.

We found that FHWs received training at the start of their job on maintaining diaries and structured notebooks for data work. Advanced FHWs received training to use tablets for data collection. Additionally, FHWs had to undergo two or three on-the-job training throughout the year, however these were focused on providing program information and medical training. FHWs did not receive any specific instructions or training for specialized surveys, even though each one was tied to a different program. Our participants used their prior experiences with data collection and supervision to deduce the expected outcomes of these surveys and proceed with data collection. The *ad hoc* nature of data collection may explain the challenges that data stewards and ML developers faced in connecting different datasets from the same beneficiaries. Beyond training, FHWs we interviewed used their personal smartphones to coordinate the process and workflow for data collection. Chat groups were used to trickle information top-down for new surveys along with information on training, reminders, and new directives from higher-ranking officials. The various channels for training and communication could be leveraged to better communicate data requirements to FHWs.

We see that the motivations driving data collection by FHWs were largely tied to their work responsibilities, and not the aggregate reporting that data stewards cared about or the generation of ML datasets. FHWs perceived data collection as tedious and redundant, and a significant time sink. We also found that these data were largely being collected in regional languages by FHWs, which posed a challenge for data stewards who were manually feeding data into information management systems.

4.3.2 Lack of transparency and feedback in data collection. FHWs noted the lack of transparency in the flow of the data after collection, despite the central role of data in their work. FHWs were required to submit their data records to a senior team member, who then passed them to data stewards. However, after submitting their data records, FHWs shared that they did not receive any regular feedback or communication on their data work which made it difficult for them to understand the use and monitoring of their data collection efforts. Despite not being clear of the immediate gains and value from all their data work and lack of feedback, FHWs continued to follow these process because of potential repercussions for non-compliance. Even as they waited on feedback, FHWs were asked by their supervisors to continue to submit forms and surveys for a large number of programs. The lack of feedback and transparency on the use of the data, especially survey data collected and not used for their immediate care work, may have contributed to data quality issues in the supply chain. These surveys may have been driven by the bureaucratic data needs identified by data stewards, burdening the healthcare system as a whole.

Despite the lack of feedback, FHWs shared that an immediate benefit of good data was informing their care work, beyond organizational requirements. The most cited example in our interviews was the generation of a *due list*, a list that helped classify and identify people who were due for routine visits, vaccination, medicine distribution, and health camp visits. Most FHWs generated this by manually tabulating entries in their registers. Even in cases where advanced FHWs generated the due list through their tablets, there was perceived fear of incorrect data or loss of data, compelling FHWs to check the compiled list against a manually generated list. Resource management was another important task where good data were seen as useful. Distribution of medicines, ORS packets, condoms, and other resources had to be recorded against the beneficiaries' names. Workers were instructed to maintain a register and were held accountable for each item; hence they spent significant time in tallying and recording entries. Aligning with these forms of valuing in the data supply chain more closely could help address the lack of feedback, while motivating collection of higher quality data.

4.3.3 Missing data when navigating diverse geographic and social topographies. FHWs were tasked with the goal of increasing access to health services in marginalised communities (e.g., caste minorities, religious minorities, and tribal communities). However, access to marginalised communities was restricted or limited due to social, cultural, and physical factors described below.

FHWs travelled to challenging locations in and around their assigned territories for data collection and health visits, frequently in challenging terrains and at the risk of physical safety, and they were required to manage their own transport. More experienced FHWs used public or shared transport, often at their own expense to travel to villages and the PHC. Within their villages, the location of houses was an important factor as FHWs preferred to visit certain locations that they perceived as 'unsafe' only during specific hours. FHWs built trust with indigenous (or *adivasi*) communities to provide health outreach and enable data collection; these locations were reported by FHWs to be more difficult to access than their own villages. These physical factors had a direct impact on sampling during data collection, which was dismissed by ML developers as simply missing or incomplete data fields. It is important to contextualise missing data in the context of where and how the data were collected.

FHWs also shared how societal norms around caste and religion played a significant role in their medical care coverage and data collection practices. Our participants related second-hand accounts from FHWs who had been chosen from marginalised communities (such as oppressed castes), who struggled with getting access to privileged households (such as upper-caste households) for data collection. Similarly, some FHWs reported challenges in gaining trust of marginalised communities. Such situations impacted data quality. ML developers need to consider the underlying societal structures that may be influencing data quality.

Additionally, FHWs reported that cultural beliefs frequently shaped the perception towards seeking and accepting medical care from FHWs (and relatedly, participating in data collection). Some families were reported to prefer home births and hold specific cultural beliefs around maternal care activities (such as nutrition, vaccination, post-delivery care, and more), which could conflict with the FHWs' responsibilities. FHWs shared that these practices were shaped by experience of family members in the household, and second-hand accounts. The lack of participation of such households could also manifest in the form of incomplete and missing data.

4.3.4 Collecting and protecting sensitive data. Even before gathering data from a specific household, FHWs shared that they spent considerable time building their relationships and trust in that household. This was critical because FHW worked and collected data on topics considered to be sensitive and personal. For example, in one NGO, older sex workers served as FHWs and helped identify and locate new sex workers. This was a task that could not be conducted by outsiders, given the discreet nature of sex work (because of its illegal status in India). Similarly, FHWs who worked on maternal health built trust with new brides in their villages with an intent to engage with them closely when they became expectant mothers. These new community members were identified through the routine surveys of FHWs and by leveraging trust with older community members. Trust between the FHW and their communities was a result of labour by the FHW and was critical to enabling the data collection process. However, FHWs also felt responsible for not abusing the carefully developed trust for repetitive data needs:

> *"Whenever I go to a house for a survey or to ask about health, the family will welcome me with respect... I also make sure to not keep going to the same house too many times to not overstep."—P11, FHW*

Being forced to collect repetitive data despite the inconvenience to households could have resulted in the false reporting noted by data stewards and data collectors. FHWs shared that they were frequently asked about the rationale for collecting sensitive information by community members. However, they could not clearly communicate the chain of access to sensitive data since they were not fully aware of the use of data after their collection. Sensitive information included images of legal identification, bank account numbers, sexual history, and more. FHWs noted to taking responsibility to alleviate their concerns and to ensure that the data collection workflows continued, despite the lack of transparency into data processes. We found that FHWs led negotiations to convince their communities about sharing sensitive information by providing details on the benefits that they would miss out if they did not provide specific data fields. Benefits typically included access to incentives in government-run health programs:

> *"People generally are afraid and ask why are we collecting bank account details, government ID proof, they think they will lose money if they give these details. I have to make them understand that they would not be robbed, if they do not give this information we cannot give them program benefits."—P22, FHW*

FHWs were tightly integrated into their communities and were trusted confidants for community members. They were privy to sensitive information regarding missed periods, domestic abuse, alcoholic partners, and sexual health, and provided advice and support to their communities for such issues beyond their medical responsibilities. The FHWs we interviewed did not report such information on structured forms, maintaining privacy by instead memorising the information. Beyond reporting 'good' data, FHWs were thus also making nuanced negotiations and decisions on *what data* to record, which would have resulted in missing data for data stewards and ML developers. P17's statement below emphasizes the importance of maintaining privacy when interacting with community members:

> *"They share information that they do not share with their own family, we are their trusted sahelis (friends)."—P17, FHW*

FHWs highlighted their practices to avoid accidental sharing of their work data through diaries, mobile phones, or tablets. Along with navigating concerns around accidental sharing of other people's private information, FHWs were also concerned about the risk of exposing the nature of their job to members of their own household. Families often knew that FHWs worked in the medical profession, but were at times unaware that they worked on topics considered taboo such as periods, condom usage, sex work, and sexually transmitted diseases. Specific details around taboo topics could have repercussions such as families urging FHWs to leave their jobs. These concerns were reported to be aggravated with shared access of their devices with elders and kids in their household. We found that community members had access to the FHW's cellphone number and could call at any time, resulting in privacy concerns. FHWs also dedicated evening hours when they had more private time to perform data work. 'Good' data practices thus involved maintaining privacy at all stages of the data collection process. However, the privacy of the data was beyond the control of FHWs once the data were submitted to the PHC, despite the reassurances they gave community members, and became the responsibility of data stewards and ML developers.

5 DISCUSSION

Our findings described the overlaps and conflicts in valuing by different stakeholders in the data supply chain. Below we discuss the tensions in valuing across stakeholders, and implications for work on ML datasets. We then consider the role of transparency and accountability in addressing the tensions that emerge in valuing in public health datafication.

5.1 Valuing Across the Data Supply Chain

In their study on the supply chain of tomatoes, Heuts and Mol highlight the constant tinkering by workers to make a tomato good at each stage, and the conflicts in valuing across stakeholders [36]. Along these lines, we uncovered the various practices that data collectors, data stewards, and ML developers were engaged in to make data good, and their efforts to continually try to do better. A lens of valuing allows us to move beyond arbitrary notions of data quality to focus on what are desirable and achievable data quality goals in a given context based on existing practices. We discuss tensions in valuing in three aspects that we identified to be relevant to all the stakeholders we studied—in data transformations performed to improve data quality, data contextualization, and as a result of organizational incentives.

5.1.1 Data Quality. ML developers and data stewards noted several challenges working with 'noisy' data from data collectors, including lack of structure, missing data, fudged data, and uneven data distribution. This impacted the eventual development of ML applications for developers and resulted in gaps on information management systems for the reporting needs of data stewards. Our analysis of the data supply chain suggests that it is imperative to not view data quality as separated from the context of the data collector, reinforcing prior work [82, 83]. Understanding valuation of 'good'

Stage	Valuation of 'good' data
ML Developers	- Structured and standardised data - Feature-rich data with validated labels - Improved model performance - Validation with field and domain experts - Contextualising bias and fairness
Data stewards	- Structure and completeness for complete entries on information management systems - Methods to decipher bad quality data - Ability to share feedback with data collectors - Synchronous data updation - Feedback on eventual data use
Frontline Workers (data collectors)	- Data work as a reflection of completion of tasks - Presentation of data registers - Training for new data workflows - Regular feedback on data work and use - Data to improve care work

Table 2: Summary

data across different stages is a step towards understanding these tensions and the surrounding ecologies.

We observe that data quality issues are not merely a result of "lazy" or unmotivated work from data collectors, as was frequently perceived by the ML developers and data stewards who participated in our study. Rather, data quality issues are embedded in the organizational, social, and cultural ecologies of the data collector through which they value data. For instance, data collectors faced challenges in accessing all communities equitably due to the safety risks, challenging geography, or social dynamics around caste or religion. These access challenges could be perceived as data distribution irregularities by developers. Similarly, data stewards frequently were mandated to work across an array of health programs, each requiring the use of a different information management system. They also faced challenges in accessing unique identifier fields, which were frequently missing. These challenges were experienced in ML development as data quality issues such as lack of structured data and issues in interoperability.

5.1.2 Data contextualisation. Building ML models far from the context where training data are collected and aggregated, as we found, can adversely impact ML development. We found that ML developers struggled to contextualise data due to lack of familiarity with the context around data collection, leading to challenges with labelling data and validating results, and unclear algorithmic outcomes. On the other hand, data collectors were respected health workers and seen as reliable confidants in their local communities. They were embedded within the communities that were providing the data and possessed deep knowledge of the topics on which data were being collected, which was gained by building relationships in their community. However, data collectors were not aware of the eventual use of their data. For them, 'good' data represented proof of work completion, impacting how they structured data. Improving transparency in data processes and leveraging similar goals across stakeholders could help better contextualize ML data operations. It is important to view the role of data collectors as community leaders with contextual and institutional knowledge. We recognise a timely opportunity to engage with their perspectives during the design of data collection workflows as well as deployment of ML models in the real world.

5.1.3 Organizational Structures. Data are valued differently based on the organizational structures and incentives for each stakeholder in the data supply chain. ML developers were motivated and incentivised to develop novel models, perceiving data work as mundane [82]. They cared about developing impactful ML applications, but lacked access to domain experts to identify such applications. Their training also did not prepare them to work with 'noisy' data leading to important context being lost through data operations they conducted, such as removing rows with incomplete fields. Data stewards were constrained by limited organizational resources, and unavailability of statistical experts and automated data cleaning workflows. Even earlier in the data supply chain, data collectors were provided with limited training on data workflows and lacked feedback on data use, resulting in confusion about data collection workflows and eventually leading to data quality issues. Financial incentives were also important to the data collectors who were frequently underpaid and worked long hours and impacted their choice of what data to collect, resulting in oversampling. Organizational structures could be designed to better support data work, such as through incentive structures that align with data needs, greater transparency around data workflows, and developing data literacy. Data collectors, in particular, could be empowered in their roles by communicating potential benefits of ML applications for them and their communities, and developing applications that benefit them more directly (*e.g.*, through automated due lists of households that they need to visit).

5.2 Transparency and Accountability in Data Transformation

Several of the tensions outlined above were related to the lack of transparency in data flows and organizational structures for accountability. Researchers within the FaccT (Fairness, Accountability, and Transparency) community have extensively studied such concerns in ML models. We view transparency and accountability from a relational perspective, where the level of transparency and accountability desired is determined through negotiation across multiple stakeholders, based on what is possible and appropriate in a given context.

5.2.1 Transparency. Prior work on transparency in ML datasets has largely focused on the development of artifacts and processes that document dataset development or summarize the dataset. For instance, Denton et al. have suggested mechanisms for recording the genealogy of data, to investigate the histories, values, and norms embedded in them [20]. ML developers in the public health context we studied had limited visibility into how the data were collected, the context of data collection, and data operations that had already been performed. More documentation could have informed subsequent data operations conducted by ML developers to improve data quality, and could also explain and help address biased ML predictions. However, generating documentation can be challenging in dynamic data settings. Though we viewed the development of datasets mostly as a linear process, there were several cycles involved at each stage. The data provided by organizations to developers had been put together over multiple data collection and aggregation cycles, making it difficult to document all the data operations. This was further complicated if ML developers requested organizations for more recent data or wanted access to real-time data, which is particularly useful in high-stakes settings like public health. In such cases, a focus on documenting the valuation of the stakeholders involved could be more practical, as these are less likely to change significantly over time.

Our findings also demonstrated how data stewards and data collectors lacked visibility into data flows which resulted in conflicts in valuing. Data collectors frequently engaged in monotonous, repetitive, and redundant data collection routines provided to them by their supervisors without adequate training or context on the use of data. They also lacked feedback on data they submitted as part of organizational requirements to show performance. Similarly, data stewards also noted the lack of transparency in the flow of the data that they had aggregated. They assumed that the data were not being used. Such conflicts are likely in public health settings where data may have originally been generated for measuring program health rather than for ML. Workshops for data stewards and data collectors could be conducted to provide training and standard procedures on data practices, and to offer more context on data use. All stakeholders could also be engaged in developing a shared understanding of the taxonomy of 'good' data to help establish shared goals and motivations in the data supply chain. Visualizations of data transformations performed and what the data will be used towards could also function as a shared artifact for transparency across the supply chain. In addition to addressing data quality issues, transparency towards common goals could help mitigate concerns around protecting privacy of end-beneficiaries, and result in more balanced ML datasets [74, 99].

We also need to consider who is responsible for doing the work of improving transparency in dataset development. For there to be transparency in a relational sense, information sharing and understanding needs to be facilitated on both sides. In most cases, however, the group with more power defines the level of transparency desired and determines who will do the work to be transparent. For instance, in the case of data collectors, they do not have the power to ask for more transparency from officials but have a high reporting burden. Moreover, whether a process is seen as transparent is a subjective assessment. Organizational structures and processes can help address this power imbalance and ambiguity to some extent, as we discuss in the next section. Transparency efforts also frequently operate on the assumption is that more transparency is always desirable. High expectations of transparency, however, can increase the burden of making processes transparent and the cognitive load of making sense of information available. Instead, we suggest a focus on what kind of information each stakeholder cares about and who it needs to come from, which a focus on valuation can help with.

5.2.2 Accountability. Our findings also highlighted the need for better structures for accountability to improve data quality for ML. Prior work on accountability in ML has largely focused on how software developers can be held accountable for model outcomes (*e.g.* [39, 79]), and we extend this focus to the development of ML datasets. We employ Bovens' definition of accountability from the social sciences, which has become popular in computer science recently [45, 96]. Bovens refers to accountability as *"a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences"* [12]. To apply this definition effectively, we need to determine *who* is accountable *to whom*, *for what*, and *how*. We found several organizational structures and work practices in place for data collectors and data stewards that impacted valuation of data. At the individual level, we found that valuation of data work for all stakeholders was driven by monetary incentives, performativity, career aspirations, and even perceived duties towards end beneficiaries. Thus, accountability was closely tied to organizational structures, incentives, and personal motivations. We note that though each stakeholder's valuation was driven by their perceived job requirements, their work was also contributing to the formation of the ML datasets, though not always with their knowledge.

Though all stakeholders were engaged in the development of ML datasets, we suggest that accountability of ML outcomes be shared between ML developers and the organizations providing data, given the investment of these two groups in the outcome. Given the differences in valuation, it would be unfair for other stakeholders to be held accountable. ML developers could be held accountable by the intermediary organizations for the performance of their models, not just in the lab but also in the real world. This could motivate organizations to provide context on data collection and offer their domain expertise. Intermediary organizations providing data could be held accountable specifically for bias in ML datasets, and ML developers are well placed to highlight these biases. Our findings

uncovered how bias was frequently linked to poor workflow organization, excessive bureaucratic data collection requirements, and challenges that data collectors in reaching all communities. Accountability for bias could also compel institutions to reconsider whether their workflows are equitable to communities and workers, while improving data quality as well.

Within the organizations providing data, we observed that data stewards were held accountable for generating reports as the need arose. However, there is also a need for accountability by the organization and the officers that they reported to. Data stewards worked with limited resources, leading to a backlog in work. Organizations and governments should be accountable for providing them with more resources to complete, as well as training for statistical operations so that they feel equipped to do their work. Along these lines, data collectors should be able to hold superiors accountable for giving feedback. Given that both data stewards and data collectors have less power, enforcing such structures is a practical challenge. One possibility is to hold institutions and officers accountable for data quality, in addition to health outcomes. This could also motivate organizations to reduce redundant workflows and employ structured data collection mechanisms, instead of overburdening workers.

Eventually, all actors should be accountable to the communities who are the target beneficiaries and are providing data. Data collectors were accountable for community health based on incentive structures, but were not accountable for privacy negotiations through they engaged in privacy-preserving behaviors. Data stewards and ML developers were similarly not held accountable for the same, though there were some processes that ML developers followed based on internal or external ethics review boards that they had to comply with. Protocols for handling sensitive data along with redressal systems for community members in case of a privacy breach could help increase community trust. We note that privacy norms are culturally situated, and we need to be careful about whose notion of privacy is imposed [3]. Data collectors engaging in privacy negotiations with community members on a regular basis could help develop an understanding of privacy preferences and design appropriate protocols that are culturally situated. Protocols could then be further negotiated with ML developers, who may offer additional considerations based on the institutions they have to interface with. Finally, we suggest that a similar exercise could be undertaken to hold ML developers and institutions accountable for ML outcomes as well once models are deployed in the field, with redressal mechanisms for community members and mediating actors like FHWs or data collectors.

6 CONCLUSION

The growing prevalence of data-driven approaches makes it imperative for us to understand how data are produced in different application domains, and the valuation driving data work at various stages of the data supply chain. Data-driven approaches that form the foundation of advancements in ML are powered in large part by human infrastructures that enable large datasets collected across multiple stages. Taking the case of datafication of public health in India, we examined the movement of data through various stages, where the data workers included frontline health workers, data stewards, and ML developers. We presented our analysis of interviews conducted with these stakeholders to draw attention to how data are valued differently across stages of data work. We discussed how data are worked upon to attain this value, as well as the tensions in valuation that arise through the process. Finally, we offered recommendations for how data supply chains could be designed to bring transparency and accountability in the creation and use of data for ML development.

ACKNOWLEDGMENTS

We are grateful to our anonymous reviewers for their encouragement and feedback on this work, and to our study participants for sharing their valuable time and expertise. We would like to thank Ben Hutchinson, Lora Aroyo, Ashwani Sharma, Milind Tambe, Manish Gupta, Dong Whi Yoo, and Karthik S. Bhat for their feedback. This work is supported in part by the National Science Foundation under Grant No. 1745463. We are extremely thankful to ARMMAN, Khushibaby and Swasti, NGOs based in India for facilitating conversations with frontline health workers.

REFERENCES

[1] Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. 2018. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine* 1, 1 (2018), 1–8.

[2] Fábio S Aguiar, Rodrigo C Torres, João VF Pinto, Afrânio L Kritski, José M Seixas, and Fernanda CQ Mello. 2016. Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil. *Medical & biological engineering & computing* 54, 11 (2016), 1751–1759.

[3] Syed Ishtiaque Ahmed, Md Romael Haque, Shion Guha, Md Rashidujjaman Rifat, and Nicola Dell. 2017. Privacy, security, and surveillance in the Global South: A study of biometric mobile SIM registration in Bangladesh. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 906–918.

[4] Karen S Baker and Helena Karasti. 2018. Data care and its politics: Designing for local collective data management as a neglected thing. In *Proceedings of the 15th Participatory Design Conference: Full Papers-Volume 1*. 1–12.

[5] Agathe Balayn, Bogdan Kulynych, and Seda Guerses. 2021. Exploring Data Pipelines through the Process Lens: a Reference Model forComputer Vision. *arXiv preprint arXiv:2107.01824* (2021).

[6] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–52.

[7] Amna Batool, Kentaro Toyama, Tiffany Veinot, Beenish Fatima, and Mustafa Naseem. 2021. Detecting Data Falsification by Front-line Development Workers: A Case Study of Vaccination in Pakistan. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[8] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology* 16, 11 (2019), 703–715.

[10] Patrick Biernacki and Dan Waldorf. 1981. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research* 10, 2 (1981), 141–163.

[11] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1125–1134.

[12] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework. *European law journal* 13, 4 (2007), 447–468.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[14] Massimo Buscema, Masoud Asadi-Zeydabadi, Weldon Lodwick, Alphonse Nde Nembot, Alvin Bronstein, and Francis Newman. 2020. Analysis of the

Ebola Outbreak in 2014 and 2018 in West Africa and Congo by Using Artificial Adaptive Systems. *Applied Artificial Intelligence* 34, 8 (2020), 597–617.

[15] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[16] HC Stephen Chan, Hanbin Shan, Thamani Dahoun, Horst Vogel, and Shuguang Yuan. 2019. Advancing drug discovery via artificial intelligence. *Trends in pharmacological sciences* 40, 8 (2019), 592–604.

[17] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical Machine Learning in Health. *arXiv preprint arXiv:2009.10576* (2020).

[18] Marika Cifor, Patricia Garcia, TL Cowan, Jasmine Rault, Tonia Sutherland, Anita Say Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, and Lisa Nakamura. 2019. Feminist data manifest-no.

[19] Nicola Dell, Trevor Perrier, Neha Kumar, Mitchell Lee, Rachel Powers, and Gaetano Borriello. 2015. Digital Workflows in Global Development Organizations. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1659–1669.

[20] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *arXiv preprint arXiv:2007.07399* (2020).

[21] Brian DeRenzi, Nicola Dell, Jeremy Wacksman, Scott Lee, and Neal Lesh. 2017. Supporting Community Health Workers in India through Voice-and Web-Based Feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2770–2781.

[22] Bhavna Dhingra and Ashok Kumar Dutta. 2011. National rural health mission. *The Indian Journal of Pediatrics* 78, 12 (2011), 1520–1526.

[23] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism.* MIT Press.

[24] Paul Dourish and Edgar Gómez Cruz. 2018. Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society* 5, 2 (2018), 2053951718784083.

[25] Chris Elsden, Kate Symons, Raluca Bunduchi, Chris Speed, and John Vines. 2019. Sorting out valuation in the charity shop: Designing for data-driven innovation through value translation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[26] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[27] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2952–2963.

[28] Brittany Fiore-Gartland and Gina Neff. 2015. Communication, mediation, and the expectations of data: Data valences across health and wellness communities. *International Journal of Communication* 9 (2015), 19.

[29] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of Washington technical report* 2-12 (2002).

[30] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*. Springer, 55–95.

[31] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.

[32] Hongyan Gu, Jingbin Huang, Lauren Hung, and Xiang'Anthony' Chen. 2021. Lessons learned from designing an AI-enabled diagnosis tool for pathologists. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[33] Venkat Gudivada, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10, 1 (2017), 1–20.

[34] Pi Guo, Tao Liu, Qin Zhang, Li Wang, Jianpeng Xiao, Qingying Zhang, Ganfeng Luo, Zhihao Li, Jianfeng He, Yonghui Zhang, and Wenjun Ma. 2017. Developing a dengue forecast model using machine learning: A case study in China. *PLoS neglected tropical diseases* 11, 10 (2017), e0005973.

[35] Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. 2011. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 65–74.

[36] Frank Heuts and Annemarie Mol. 2013. What is a good tomato? A case of valuing in practice. *Valuation Studies* 1, 2 (2013), 125–146.

[37] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[38] Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. 2020. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters* 471 (2020), 61–71.

[39] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[40] Eleanor Hutchinson, Susan Nayiga, Christine Nabirye, Lilian Taaka, and Sarah G Staedke. 2018. Data value and care value in the practice of health systems: A case study in Uganda. *Social science & medicine* 211 (2018), 123–130.

[41] Marco Iansiti. 2021. The Value of Data and Its Impact on Competition. *Available at SSRN* (2021).

[42] Azra Ismail and Neha Kumar. 2018. Engaging Solidarity in Data Collection Practices for Community Health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 76.

[43] Azra Ismail and Neha Kumar. 2019. Empowerment on the Margins: The Online Experiences of Community Health Workers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 99.

[44] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 306–316.

[45] Severin Kacianka and Alexander Pretschner. 2021. Designing Accountable Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 424–437.

[46] Ramesha Karunasena, Mohammad Sarparajul Ambiya, Arunesh Sinha, Ruchit Nagar, Saachi Dalal, Divy Thakkar, and Milind Tambe. 2020. Measuring Data Collection Quality for Community Healthcare. *arXiv preprint arXiv:2011.02962* (2020).

[47] Naveena Karusala, Jennifer Wilson, Phebe Vayanos, and Eric Rice. 2019. Street-Level Realities of Data Practices in Homeless Services Provision. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[48] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 45–55.

[49] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 88.

[50] Rob Kitchin and Tracey Lauriault. 2014. Towards critical data studies: Charting and unpacking data assemblages and their work. (2014).

[51] Meghana Kshirsagar, Caleb Robinson, Siyu Yang, Shahrzad Gholami, Ivan Klyuzhin, Sumit Mukherjee, Md Nasir, Anthony Ortiz, Felipe Oviedo, Darren Tanner, et al. 2021. Becoming Good at AI for Good. *arXiv preprint arXiv:2104.11757* (2021).

[52] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[53] Yanni Alexander Loukissas. 2019. *All data are local: Thinking critically in a data-driven society.* MIT Press.

[54] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[55] Aditya Mate, Jackson A Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Interventions. *arXiv preprint arXiv:2007.04432* (2020).

[56] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[57] Amanda Meng, Carl DiSalvo, and Ellen Zegura. 2019. Collaborative data work towards a caring democracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[58] Syeda Shaizadi Meraj, Razali Yaakob, Azreen Azman, Siti Nuralain Mohd Rum, Azree Shahrel, and Ahmad Nazri. 2019. Artificial Intelligence in Diagnosing Tuberculosis: A Review. *Int. Journal on Advanced Sci., Eng. and Inform. Technology* 9, 1 (2019), 81–91.

[59] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *arXiv preprint arXiv:2007.14886* (2020).

[60] Annemarie Mol, Ingunn Moser, and Jeannette Pols. 2015. *Care in practice: On tinkering in clinics, homes and farms*. Vol. 8. transcript Verlag.

[61] Naja Holten Møller, Claus Bossen, Kathleen H Pine, Trine Rask Nielsen, and Gina Neff. 2020. Who does the work of data? *Interactions* 27, 3 (2020), 52–55.

[62] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[63] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian

Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[64] Andrew B Neang, Will Sutherland, Michael W Beach, and Charlotte P Lee. 2021. Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[65] Gina Neff, Anissa Tanweer, Brittany Fiore-Gartland, and Laura Osburn. 2017. Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big data* 5, 2 (2017), 85–97.

[66] Siddharth Nishtala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindra, and Milind Tambe. 2020. Missed calls, Automated Calls and Health Support: Using AI to improve maternal health outcomes by increasing program engagement. *arXiv preprint arXiv:2006.07590* (2020).

[67] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[68] Chinasa T Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[69] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[70] Joyojeet Pal, Anjuli Dasika, Ahmad Hasan, Jackie Wolf, Nick Reid, Vaishnav Kameswaran, Purva Yardi, Allyson Mackay, Abram Wagner, Bhramar Mukherjee, et al. 2017. Changing data practices for community health workers: Introducing digital data collection in West Bengal, India. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. ACM, 17.

[71] Ankur Pandey, Inshita Mutreja, Saru Brar, and Pushpendra Singh. 2020. Exploring Automated Q&A Support System for Maternal and Child Health in Rural India. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 349–350.

[72] Elisavet Parselia, Charalampos Kontoes, Alexia Tsouni, Christos Hadjichristodoulou, Ioannis Kioutsioukis, Gkikas Magiorkinis, and Nikolaos I Stilianakis. 2019. Satellite Earth Observation Data in Epidemiological Modeling of Malaria, Dengue and West Nile Virus: A Scoping Review. *Remote Sensing* 11, 16 (2019), 1862.

[73] Samir Passi and Steven Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2436–2447.

[74] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. 2 (2018), 136: 1–136: 28. Issue CSCW.

[75] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345* (2020).

[76] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. 2019. Examining the challenges in development data pipeline. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. 13–21.

[77] Kathleen H Pine and Max Liboiron. 2015. The politics of measurement and action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3147–3156.

[78] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[79] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[80] Katelyn J. Rittenhouse, Bellington Vwalika, Alexander Keil, Jennifer Winston, Marie Stoner, Joan T. Price, Monica Kapasa, Mulaya Mubambe, Vanilla Banda, Whyson Muunga, and Jeffrey S. A. Stringer. 2019. Improving preterm newborn identification in low-resource settings with machine learning. *PLOS ONE* 14, 2 (Feb. 2019), e0198919. https://doi.org/10.1371/journal.pone.0198919

[81] Minna Ruckenstein and Natasha Dow Schüll. 2017. The datafication of health. *Annual Review of Anthropology* 46 (2017), 261–278.

[82] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[83] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. From Field Experts to Data Collectors: Deskilling of Domain Expertise in AI Development. In *CHI 2022*.

[84] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *arXiv preprint arXiv:2108.04308* (2021).

[85] I Seidman. 2006. A guide for researchers in education and the social sciences.

[86] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 99–109.

[87] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[88] Ranjit Singh. 2009. Study the Imbrication: A Methodological Maxim to follow the multiple lives of data. *Lives of Data* 56 (2009), 51.

[89] Chris Speed and Deborah Maxwell. 2015. Designing through value constellations. *interactions* 22, 5 (2015), 38–43.

[90] Charles Sutton, Timothy Hobson, James Geddes, and Rich Caruana. 2018. Data diff: Interpretable, executable summaries of changes in distributions for data wrangling. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2279–2288.

[91] Alex S Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-place: Thinking through the relations between data and community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2863–2872.

[92] Divy Thakkar, Neha Kumar, and Nithya Sambasivan. 2020. Towards an AI-powered future that works for vocational workers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[93] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[94] Kenneth Thomsen, Lars Iversen, Therese Louise Titlestad, and Ole Winther. 2020. Systematic review of machine learning for diagnosis and prognosis in dermatology. *Journal of Dermatological Treatment* 31, 5 (2020), 496–510.

[95] Edward Velasco, Tumacha Agheneza, Kerstin Denecke, Goeran Kirchner, and Tim Eckmanns. 2014. Social media and internet-based data in global systems for public health surveillance: a systematic review. *The Milbank Quarterly* 92, 1 (2014), 7–33.

[96] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 1–18.

[97] Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication Reliability–An Empirical Approach to Interpreting Inter-rater Reliability. *arXiv preprint arXiv:2106.07393* (2021).

[98] Deepika Yadav, Prerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 170.

[99] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[100] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[101] Lindsay E Young, Jerome Mayaud, Sze-Chuan Suen, Milind Tambe, and Eric Rice. 2020. Modeling the dynamism of HIV information diffusion in multiplex networks of homeless youth. *Social Networks* 63 (2020), 112–121.

[102] Ellen Zegura, Carl DiSalvo, and Amanda Meng. 2018. Care and the practice of data science for social good. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–9.

[103] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[104] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair.