Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3600211.3604766

RESEARCH-ARTICLE

Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI

HUBERT DARIUSZ ZAJAÇ, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

NATALIA ROZALIA AVLONA, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

FINN KENSING, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

TARIQ OSMAN ANDERSEN, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

IRINA SHKLOVSKI, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

Open Access Support provided by:

University of Copenhagen

PDF Download
3600211.3604766.pdf
26 January 2026
Total Citations: 15
Total Downloads: 1703

Published: 08 August 2023

Citation in BibTeX format

AIES '23: AAAI/ACM Conference on AI, Ethics, and Society
August 8 - 10, 2023
QC, Montréal, Canada

Conference Sponsors:
SIGAI

Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI

Hubert D. Zając*
hdz@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Natalia R. Avlona*
naav@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Tariq O. Andersen
tariq@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Finn Kensing
kensing@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Irina Shklovski
ias@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

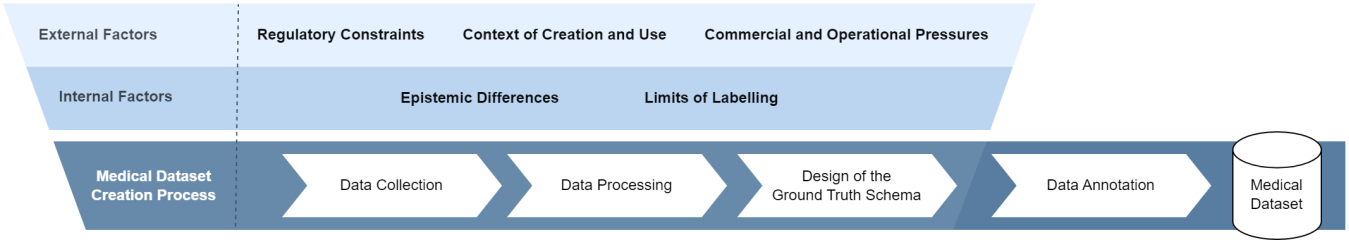

Figure 1: A simplified medical dataset creation process expanded with the design of ground truth schema and factors conditioning the pre-annotation stages.

ABSTRACT

One of the core goals of responsible AI development is ensuring high-quality training datasets. Many researchers have pointed to the importance of the annotation step in the creation of high-quality data, but less attention has been paid to the work that enables data annotation. We define this work as the design of ground truth schema and explore the challenges involved in the creation of datasets in the medical domain even before any annotations are made. Based on extensive work in three health-tech organisations, we describe five external and internal factors that condition medical dataset creation processes. Three external factors include regulatory constraints, the context of creation and use, and commercial and operational pressures. These factors condition medical data collection and shape the ground truth schema design. Two internal factors include epistemic differences and limits of labelling. These directly shape the design of the ground truth schema. Discussions of what constitutes high-quality data need to pay attention to the factors that shape and constrain what is possible to be created, to ensure responsible AI design.

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs International 4.0 License.

AIES '23, August 08–10, 2023, Montréal, QC, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604766

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

Medical Datasets, Data Creation, Responsible Artificial Intelligence and Machine Learning

ACM Reference Format:
Hubert D. Zając, Natalia R. Avlona, Tariq O. Andersen, Finn Kensing, and Irina Shklovski. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3600211.3604766

1 INTRODUCTION

Advances in applications of Artificial Intelligence (AI) in the medical domain promise to improve efficiency, promote accuracy and bring cost savings across many areas of medical subspecialty, yet there are also many concerns about ethics and responsibility in the deployment of these technologies [16]. The idea of responsible AI has been extensively discussed in the literature and received much attention from both commercial entities and regulatory bodies [15, 47]. There is considerable agreement that high-quality training data is key to the development of responsible AI systems [28]. Yet research shows that the creation of high-quality data also tends to be an undervalued step in the development of machine learning systems [57, 58].

The process of dataset creation is typically broken down into three steps - data collection, data pre-processing and cleaning, and finally, data annotation[1, 50]. This is especially so in the medical domain where high-quality training data is obtained through a

range of annotation practices such as data quality enhancement [10], generating labels using Natural Language Processing models [23], deriving image labels from medical documentation [34], and following labelling guidelines and principles focusing on fairness and inclusion [36, 59]. This paper investigates the factors that affect the creation of high-quality medical datasets demonstrating that the preparatory work involved in the design of ground truth schema used in data annotation is an important preceding step that tends to be overlooked in the literature. Following the work of Mueller and colleagues [49], we define the ground truth schema as a collection of relational labels and metrics, as well as their definitions and examples that are used during data labelling.

Recent research on the creation of training datasets [21] has discussed annotation activities as a matter of power relations in projects crowdsourced in the Global South [41, 42, 45], the social design of labelled data by domain experts [49], and annotation process recommendations [19]. While understanding data annotation is important, data design work begins before the first data points are labelled. Data is always designed and constructed through situated and emergent processes [18, 49] as domain experts, data scientists, other stakeholders, and diverse political interests imprint their values on the data. However, little is known about the preparatory work necessary to produce high-quality data [31]. Accounts of decisions that shaped the datasets are rarely documented and get dismissed as soon as the data creation work concludes [58], thus become impossible to inspect in the future [49, 51, 61].

In this article, we consider **what factors affect the design of medical datasets prior to data annotation.** We ground our findings in ethnographic research conducted across three organisations developing medical AI for (I) screening chest x-rays, (II) supporting the diagnosis of lung and pancreatic diseases (III) automating patients-to-clinical trials matchmaking. We explore the decisions made by medical professionals, data scientists, designers, and other relevant stakeholders in their quest to create medical AI datasets in highly constrained environments. Our data include approximately 50 hours of observations, interviews with 46 medical professionals, data scientists, and designers, as well as observation notes, email communication, reports, and artefacts. We followed a grounded theory approach [9]that led us to identify and define factors that influence the design of the ground truth schema that underpins the production of high-quality training data.

Our contribution is twofold. First, we identify five factors, three external and two internal, that influence medical dataset creation by affecting data collection, ground truth schema design, and data annotation stages (see Fig. 1). The external factors condition the medical dataset creation processes by determining the data collection and shaping the possibilities for the design of ground truth schemas:

- Regulatory Constraints
- Context of Creation and Use
- Commercial and Operational Pressures

The internal factors define the negotiations between the medical and technical domains:

- Epistemic Differences
- Limits of Labelling

Second, we show how these factors affect the final shape and quality of the resulting medical datasets. While we define each factor separately for analytical purposes, the factors are interrelated and affect each other, structuring the limits of responsible data creation approaches. We argue that these factors condition the stages that precede data labelling and mediate the design of what is aspired to be responsible AI.

2 RELATED WORK

While the idea of responsible AI has received much attention from both commercial entities and regulatory bodies, concerns about the quality of data and the challenges in the creation of quality data are increasingly in focus. The now-emerging guidelines list several data-related challenges as key obstacles that hinder the path towards responsible AI: skewed data (issues that originate during data collection), tainted data (issues that stem from labelling e.g. hidden stratification [52]), or limited features (an inadequate number of features represented in data) [4]. There is broad agreement that dataset creation processes deserve greater attention, despite scholars repeatedly pointing to a strong bias against data work [14, 57, 58].

2.1 How datasets are created and annotated

In computer science, dataset creation is often seen as an activity constituting a step in the larger development processes of ML-based systems [1, 11, 26, 50, 65]. However, scholars have also discussed the dataset creation process on its own merits. For example, Hutchinson drew parallels between software development and dataset creation practices by sharing conceptual stages like requirement analysis, design, implementation, testing, and maintenance[31]. Similarly, increased focus can be observed in the medical area, where researchers describe in greater detail the creation of publicly available medical datasets [8, 13, 33, 35, 40, 66]. Typically, dataset creation is described as a process that spans all activities related to work on medical data, collected under the umbrella of data collection, data cleaning and processing, and data annotation.

Data annotation is one of the most researched aspects of dataset creation. Data annotation or labelling usually happens as part of the curation or preparation step of larger data science projects, following data acquisition and cleaning, and preceding feature engineering [1]. These activities are usually iterative and highly collaborative. Linguistic scholars and Natural Language Processing researchers [19, 30, 63] offer guidance on how to carry out data labelling. They distinguish three focal points: the creation and improvement of an annotation guide [19], schema [63], or manual [30]; the labelling performed by trained annotators; and the adjudication of the annotated data.

In this paper, we use the terms data labelling and data annotation interchangeably and understand them as the action of assigning and adjudicating predefined labels to concrete data points. When considering this step alone, there is a multitude of decisions that need to be taken to complete it. Scholars have pointed to data annotation activities as a site of political struggle, challenges to the labour conditions, as well as the stage in dataset creation that can result in adverse downstream outcomes for trained models. For example, Schumann et al. [59] and Hanley et al. [24] demonstrate how the design of categories (or labels) can reinforce harmful stereotypes and exclude underrepresented groups of people. Badly annotated

data can reduce the performance of AI models [10, 23, 34, 47, 54] and perpetuate exclusion and inequality [36, 59].

In the medical domain, data annotation challenges can be compounded by the requirements for specialised knowledge and training. Despite initiatives like the Unified Medical System [38], the clinical meaning of labels can be unclear [51], and medical knowledge remains difficult to capture for computer use. Li and colleagues [37] explored the inter- and intra-rater agreement between six radiologists of different experience levels when labelling chest x-rays. In some cases, even the experienced radiologists reached only a moderate level of agreement with themselves [39]. This could occur due to not following the best medical practices when labelling data, due to resource constraints [57] or because of the disconnect between the practices of labelling and the actual usage of medical data in regular practice [51].

What much of this research points to is the fact that labelling and annotation as practices are heavily reliant on the creation of annotation guides and schemas [19]. Yet, despite the growing interest in the creation of datasets, current discussions tend to omit and overlook the pre-labelling activities and their potential impact on the quality of the resulting training data [67].

2.2 The design of the ground truth schema

Many scholars investigated the dynamic and situated work of domain experts, data scientists, designers, and other stakeholders engaged with data [25, 48, 57, 60]. For example, Muller and colleagues investigated how domain experts label data, highlighting that the ground truth contained in datasets is a human contribution resulting from improvised and iterative adjustments to principled design processes [49]. Discussing the design of ground truth schema implies that ground truth captured in medical AI datasets is not an objective representation of reality but is a result of a situated design process [6]. In other words, data is never raw [22], instead, all data is actively constructed [3, 43, 53]. Feinberg emphasises the importance of recognising the subjectivity involved in dataset creation and the need to consider the potential biases and limitations inherent in choices that stem from the social and organisational context in which data is produced [18].

Researchers who investigate AI datasets suggest that access to all of the "subtle design decisions" made during the dataset creation, is vital to ensuring a high-quality labelling process [17, 51] and thus high-quality datasets. However, documenting design decisions in data science work is not common [53, 56, 68]. To address this gap, researchers developed a range of documentation frameworks to support the accountability, use, and maintenance of complex datasets [2, 44]. These frameworks range from general purpose and qualitative - Datasheets for Datasets [20], NLP-focused - Data Statements [5], quantitative - Dataset Nutrition Label [27], to fairness focused - data briefs [17] and accountability [31]. Some of these tools [17, 20, 31] include a query for the origin of the labels, but most do not pay much attention to the pre-labelling activities involved in annotation schema creation.

While the existing scholarship has problematised the stage of the data labelling and the power relations and conditions affecting the data annotation work [49], little is known about the stages preceding the data labelling. Particularly, how these stages influence the final shape of medical datasets. We explore the collaborative and situated work of medical professionals, data scientists, and designers that takes place before the labelling stage, within the design stage proposed by Hutchinson et al. [31] or the preparatory work proposed by Fort [19].

3 METHODOLOGY

We investigated three organisations in the Global North developing medical AI-based systems that engaged in the medical dataset creation processes. We focused on the work conducted before the data annotation task by participants described in Table 3).

3.1 Research context and data collection

3.1.1 ORG I. was an interdisciplinary collaboration between academia, business, and the public healthcare sector, aiming to create AI-based chest x-ray prioritisation software for global use. The project's first step was designing the ground truth schema for labelling chest x-rays, which is the process investigated in this study.

Our engagement in ORG I spanned May 2021 to Feb 2023. During that time, we conducted participatory observations of the design process of the ground truth schema. The working group developing the system was based in a Northern European country (Table 3.1). Additionally, a feedback group comprising medical professionals from the Northern European country and an East African country provided feedback on the schema (Table 3.2). We participated in fifteen working group meetings ranging from 26 minutes to 2 hours and 12 minutes in length. Additionally, we conducted twelve interviews and observed external medical professionals evaluating and providing feedback on the intermediate results of the design work. Additional material included observation notes, meeting summaries from other participants, a work progress report, email communication, and produced artefacts - a labelling guide and the ground truth schema.

3.1.2 ORG II. was a large tech company in Western Europe with part of the business involved in the development of complex medical devices. We primarily engaged with sections of the company that focused on the development of AI-based diagnostic tools and systems for oncological radiology.

Our work with ORG II was split into a preliminary exploratory period online from February to May 2022 and in situ participant observations and semi-structured interviews conducted in June 2022 in a Western European country. Due to the size of the organisation, we employed snowball sampling. In ORG II, we conducted thirteen semi-structured interviews with experts (Table 3.3), with an average duration of 65 minutes.

3.1.3 ORG III. was a mid-size start-up in Western Europe that aimed at developing an AI-based platform for matching patients with advanced clinical trials for new drug and experimental procedure development. The company primarily dealt with two data sources. First, they collected data from medical practitioners and their patients. Second, they collected data from public registries in the EU and US and pharmaceutical companies about clinical trial requirements or experimental treatments. Their goal was to match the patients with unmet medical needs and their physicians with the requirements of BioPharma companies that need to enhance drug development and recruit participants for clinical trials.

ORG I	Position	Exp.
P1	Radiologist	Junior
P2	ML Engineer	Senior
P3	ML Engineer	Senior
P4	Computer Scientist	Senior
P5	Data Scientist	Senior
P6	Radiologist	Senior
P7	Radiologist	Senior
P8	HCI Researcher	Junior
P9	HCI Researcher	Senior

ORG I	Position	Exp.
P10	Radiologist	Senior
P11	Radiologist	Junior
P12	Radiologist	Junior
P13	Radiologist	Mid
P14	Radiologist	Senior
P15	Radiologist	Senior
P16	Radiologist	Senior
P17	Radiologist	Senior
P18	Radiologist	Senior
P19	Radiologist	Senior
P20	Radiologist	Senior
P21	Physician	Junior

ORG II	Position	Exp.
P21	Data scientist	Senior
P22	Product Owner	Mid
P23	Strategic Designer	Senior
P24	Data scientist	Mid
P25	Usability Designer	Senior
P26	Data scientist	Senior
P27	Data scientist	Senior
P28	Data Designer	Mid
P29	Interaction Designer	Senior
P30	Data scientist	Senior
P31	Data Designer	Senior
P32	HCI Researcher	Mid
P33	Data Designer	Senior

ORG III	Position	Exp.
P34	Product owner	Mid
P35	Software Engineer	Junior
P36	Software Engineer	Mid
P37	Software Engineer	Mid
P38	Data Scientist	Mid
P39	Data Scientist	Senior
P40	UX Designer	Senior
P41	Software Developer	Mid
P42	Medical Operations	Senior
P43	Quality Assurance	Senior
P44	UX Designer	Mid
P45	Neurobiologist	Senior
P46	Product Owner	Mid

Table 1: List of participants, their simplified positions, and experience levels. Respectively in ORG I (working group), ORG I (feedback group, participants 10-14 were located in the northern European country, and participants 15-21 were located in the East African country), ORG II, and ORG III.

Our engagement with ORG III spanned February to May 2022. The preliminary period involved online semi-formal meetings and interviews from February to April 2022. In situ ethnographic research was conducted during May and June 2022 at the headquarters of ORG III in Western Europe. We conducted participant observation by joining the daily stand-up sessions of the engineering department and shadowing the workflow of the AI team experts leading the data labelling process for the match-making platform. In total, we interviewed 13 participants (Table 3.4).

3.2 Data analysis

The main focus of our analysis was to identify factors affecting medical dataset creation. We analysed decisions made during the design work, tensions and misunderstandings that needed to be reconciled, looking both outside and within the organisations where the design work took place. We explicitly decided to explore the wider socioeconomic factors that condition the medical dataset creation and influence the final AI-based systems even before the first label is annotated.

Data analysis relied on techniques of grounded theory and situational analysis [9, 12]. First, we conducted line-to-line open coding, coming up with 850 initial codes. We then reflexively proceeded to thematic coding, in an iterative manner, discussing the themes and patterns emerging in our three sites of ethnographic inquiry. During this step, we designed visual maps to lay out the human, technological, and discursive dynamics of the organisations under study [12]. Second, we conducted axial coding to reflexively group the available themes into dimensions. Finally, we assessed these dimensions against the codes and situational maps, converging on the five final factors (regulatory constraints, context of creation and use, commercial and operational pressures, epistemic differences, and limits of labelling).

3.3 Positionality statement

Our qualitative data was obtained from three health-tech organisations in the Global North. The analysis was shaped by the following standpoints. First, we differentiated our roles in studying the three organisations. Researchers in ORG I had the dual position of the expert who on the one hand, designed the labelling software whilst they conducted participant observation and semi-constructed interviews in order to study the process of the ground truth schema design. Researchers working with ORG II and ORG III employed ethnographic methods as a research approach without having a prior engagement with the organisations. Second, we are researchers currently working for Northern European institutions. Third, we have mixed epistemic backgrounds in computer science and law and policy. Finally, we emphasise the situatedness of our research, which focuses on the development of medical AI at the specific loci of our studied organisations. We acknowledge that the factors we identify as defining the medical dataset creation bear the geographical and epistemic limitations of the Northern European context. On this note, we acknowledge that the divide between Global North and Global South we make below has been problematised by scholars in human geography and decolonial studies as a limiting one, reinforcing stereotypes and reducing the polyphony of southern standpoints [29, 64]. For this reason, we use this divide in this paper to (I) acknowledge the limitations of our standpoints in a northern institution and the privilege of our funded projects; (II) tackle assumptions about data universalism [46] by showing the particularities of the northern context in medical datasets creation and their effect on the intended use of such data in different contexts.

4 FINDINGS: FIVE FACTORS THAT INFLUENCE MEDICAL DATASET CREATION

The datasets used for medical AI benefit from the impression that they are a result of an age-old medical practice that is seamlessly transitioning to the digital age, unaffected by external influences, and focused on the pursuit of medical excellence. However, the reality is often different. Our ethnographic data suggest that even before medical professionals have had the chance to annotate or make their first label, many critical design decisions have been made, which frame the labelling space, thus limiting the extent to which medical professionals can use their expertise.

Our analysis challenged our initial understanding of the dataset creation process drawn from the literature. Our data made clear that the preparatory work should be conceptualised as a crucial stage in dataset creation taking place before data labelling because it defines what becomes captured as ground truth within a training dataset. This is the step where the ground truth schema is designed, which, when applied to an unlabelled dataset through expert annotation, embeds the intended ground truth within it.

Regulatory Constraints
• Extent of Collected Data
• Predetermination of Purpose
Context of Creation and Use
• Geographic context of use
• Demographic context of production
• Linguistic context
Commercial and Operational Pressures
• Business model and organisation scalability
• Competition and health tech market
• Intended future use within healthcare type

Table 2: External factors and their dimensions

We identified five factors that influenced the creation of medical datasets in the organisations we studied. Three of these factors were external to the activities directly involved in pre-labelling activities. External factors defined and delineated the limits and possibilities for labelling activities. Two internal factors on the other hand affected the negotiations around what needed to be labelled and how the labelling was to proceed through the design of the schema. Below we describe each factor and demonstrate how they affected the final shape of the medical datasets focusing on the data collection and ground truth schema design stages.

It is important to note that the organisations and processes examined in this paper were largely driven by data scientists as the owners of the dataset creation process, with representatives of other domains contributing to the dataset creation activities. As a result, data science as an epistemology dominated the design work by primarily embedding data scientists' perspectives, inadvertently compromising other domain-based practices and understandings. As datasets in our research were created for the purpose of AI development, the power distribution was uneven, leaving little room for misconceptions from data scientists to be challenged and addressed.

4.1 External factors: defining the ground truth schema design space

Despite the best intentions of the experts engaged in the medical dataset creation process, many of their decisions and actions were structured by different external factors. We identified three such factors - **Regulatory Constraints, Context of Creation and Use**, and **Commercial and Operational Pressures** - that shaped the space of medical dataset creation and thus influenced the final shape of the datasets themselves even before the labelling could begin (Table 2). Each factor consists of several distinct features. We describe these below in detail.

4.1.1 Regulatory constraints. The medical data space is highly controlled through a variety of local, national, and international regulatory constraints. This was particularly challenging for the data collection step of the process. We observed two areas where compliance with regulatory standards affected the creation of medical data: **the extent of the collected data** and **the predetermination of purpose**. Experts in all of the organisations we studied were concerned about compliance with diverse standards that intersected with their work on medical dataset creation. These standards originated from European binding legislative acts, international standard organisations, or industry standards. GDPR, the main legal standard for data protection in the European Union, was the most prominent example of a binding legislative act, regulating the conditions under which personal data is collected and processed. The industry and international organisations imposed, among others, ISO 2700013001, HIPAA, and Good Medical or Good Manufacturing Practices. In ORG III, a data scientist (P39) listed 21 unique regulations they felt they needed to consider. As a larger and more mature organisation, ORG II also had internal ethics boards, which at times imposed even stricter interpretations. However, these standards and limits legitimised the data collection and processing activities.

Constraints on data collection. While experts in all organisations were striving to create what they saw as high-quality data, complying with relevant regulatory standards required concessions from all participants. For data scientists, the regulatory constraints delimited what data was available for collection, at times inadvertently introducing bias in different ways. For example, P26, a data scientist from ORG II, explained: *"what is the data that we are allowed to use, especially if you look at ... bias ... people will want to look at bias and, and see if ... their product was fair to all, some demographics, and [we are] just not able to use the data because of privacy issues or GDPR"*. Similarly, in ORG II, the contractual agreement with a single local hospital, on the one hand, provided a controlled supply of high-quality data, on the other hand, reduced data representativeness: *"we have a strong relationship with them. How do you expect that the data is not going to be biased right?"* (P24). While ORG II was able to create highly detailed and structured training data for their models, this data was clearly not representative of populations that would eventually encounter the resulting technologies.

Limitations imposed on data collection could compromise the resulting datasets in ways that created challenges for subsequent data creation steps. For example, participants of ORG I could collect only chest x-rays and their linked radiological reports. Privacy concerns here also resulted in the loss of the chronological links between the images during data collection. This selection significantly diverged from the usual assortment of data available to radiologists in clinical practice, introducing challenges at the later stages of medical dataset creation, such as schema creation and annotation.

Regulatory standards and contractual agreements determined the purpose and context of use. Data protection regulations have recently focused intently on the purpose of use as one area of emphasis, tied to notions of data minimisation and data subject notification. Companies in our research had to negotiate the legal basis for their data collection with contracted data providers such as hospitals. For example, GDPR and contractual agreements with a local hospital bounded ORG II to use the collected data within the predefined purpose and context. Deviations from the initially stated purposes and context of use required new agreements that could be obtained only through significant time and resource investments. As a product owner (P22) explained the process of collecting data from the local hospital, *"maybe the new study that we want to do has a slightly different scope and it's not covered by the original contract, then we need to make a new contract"*. ORG I encountered a similar predicament where the data collection phase was negotiated based on what the data scientists believed to be a necessary and sufficient dataset given the available resources and legal constraints of local regulations. By the time domain experts explained that the dataset was lacking important data dimensions, it was too late.

4.1.2 Context of creation and use. The context of production and the context of use influenced the creation of medical datasets. In our studies, each medical dataset was created for a specific intended use that was embedded in the collected medical data, e.g., clinical trial repositories, hospitals, and patients. These sources cover specific geographical populations, which has consequences for the final medical dataset. We identified three dimensions where that influence was prevalent: **the geographic context of use, the demographic context of production**, and **the linguistic context** (Table 2).

The geographic context of use affected the selection of labels. While medicine strives to deliver replicable results that generalise across populations, the ground truth schemas are designed to serve specific needs in specific contexts. Some of them are defined by the intended use of the future AI-based systems in the geographic context, in which they are going to be used. In ORG I, the project group designed the first version of the ground truth schema based on local data from a Northern European country. As a result, the first version of the schema captured the locally prevalent conditions well but missed conditions relevant within the countries of intended use, which were almost never encountered locally. To account for that, direct and indirect input from medical professionals from the East African country was collected and incorporated into the schema during joint design work, as seen in this exchange between a radiologist and an ML engineer.
"So if you wanted that in the hierarchy, it could be there." (P1)
Is it aortic unfolding? Because I clearly remember this sentence from [the East African country] reports, "aortic unfolding due to chronic hypertension" (P2). Yet despite having a broader ground truth schema, the same project also struggled to ensure enough examples of common medical conditions across expected countries of use available for annotation, since the data was originally only collected from one country.

The demographic context affected representativeness concerns In both ORG I and ORG II, data in medical datasets were collected from a single country, which had several consequences. For example in ORG II, the data was predominantly collected from a single local hospital, where ORG II had a contractual agreement. Not only was this problematic due to a more homogeneous patient population, but the collected medical imaging data originated on machines from the same producer. This created many concerns since imaging machines from different manufacturers often produce slightly different artefacts in their output. Yet the information about which machines were used to produce the images was rarely included in the resulting dataset.

Similarly, due to the characteristics of the population embedded in medical datasets, experts worried about how portable the resulting AI models would be. As a usability designer (P25) from ORG II noted, *"you can have all sorts of differences in patient demographics ... and you cannot just apply a model that you train on population A to population B"*. However, despite the designers' and data scientists' awareness, a senior radiologist from the East African country emphasised that *"in the [developing world]*[1] *we are usually consumers, not producers of tech. We may find ourselves hitched to tech that doesn't serve our needs"* (P15). When evaluating the ground truth schema, the same medical professional elaborated, *"I've done this for 10 years since my graduation. I've never seen certain diseases like cystic fibrosis, but whenever I read the books, there's a lot of stuff about cystic fibrosis [prevalent in the Global North],"* which highlights the effect of local ground truth schemas on the transferability of the final AI-based systems.

Linguistic context and local understanding of medical terms challenged the application and transferability of the ground truth schemas. The design of ground truth schemas included naming the labels, defining and organising their relations, and providing examples. However, medical concepts are not always used in the same way across different countries. In ORG I when discussing the naming convention for a chest x-ray finding, one radiologist noted *"I know that it's not proper, but [in the Northern European country] they use 'infiltrat' as a synonym of consolidation ... I think the direct translation consolidation would be 'consolidering' but they don't use that, they use 'infiltrat'... I think maybe our infiltrate is broader"* (P1). As a result, a presentation of infiltration by an AI-based system could be understood differently by medical professionals from different countries. To account for that, data scientists and medical professionals evaluated the ground truth schema against English translations. In ORG III, which operates globally, the data scientists and designers recounted a similar challenge of re-translating medical terms during the data annotation process. The limitations of the locality of medical terms prohibited the aspiration of designing a ground truth schema that can operate universally. As a UX designer (P40) remarked: *"there are also challenges around that because different cultures will refer to different diseases in different ways. It's global and we re-translate some of our stuff into different pages. We also have to consider localisation, how you turn this medical term into a layman term, but that's also relevant in like different countries as well."*

4.1.3 Commercial and operational pressures. The three organisations each had a different business model and exhibited different relations to the market and the public sector. This often determined the availability of the resources (human and material) allocated for dataset creation and affected the organisations' ability to collect data and design the ground truth schema. We identified three dimensions of commercial and operational pressures (Table 2): **business model and scalability of the organisation, the competition in the health tech market**, and **intended future use within the healthcare type**.

The business model and scalability of the organisation determined the amount of collected and labelled data. Every investigated organisation represented a different business model. ORG I intersected with the public sector, whilst ORG II and III were situated entirely in the private sector. The business models of the organisation determined the way in which data was collected. The business model of ORG III relied on providing free use of the AI-based platform to patients but also providing paid services to BioPharma by enrolling patients into clinical trials. To do that, ORG III collected data from the public clinical trial registries in the EU and US, as well as patient medical information. Such data collection was heavily dependent on the organisation's scalability, as well as the "fine" balance between the data requested by their BioPharma clients and the data that could have been collected. As a data scientist (P38) explained: "sometimes it's difficult to decide what kind of data you collect, right? Or what patients. (...) there's a balance between what's actually feasible to collect and what will give us the highest chance of getting as much data as possible. So those

[1] edited to avoid pejorative language

I think are tricky decisions." These conditions affected how much data was finally collected, hence, the ideal of representativeness of the created dataset was compromised.

In ORG I, the budget allocation for the data annotation process played a vital role in the amount of data possible to be labelled by medical professionals. Due to the high cost of labelling by experienced medical professionals, ORG I had to cap the maximum number of labelled images. This cap limited the number of distinct labels that could be annotated in the created dataset and remained statistically significant. "We have a limited budget for the test data that we can collect because we need several radiologists board-certified possibly to look at images" (P3). The limited resources defined the amount of data that was possible to be annotated, putting ORG I at a competitive disadvantage: "What the [competitors] do (...) there is no way we can reach what they do. They have 127 findings and they use a hundred plus radiologists to annotate, and they annotated 800,000 images each image by three radiologists. So the scale is completely different" (P2).

Market standards and industry competition affect the design of the ground truth schemas. Since all organisations under study operated in the health tech sector, the experts engaged in the processes of designing ground truth schemas had to both consider existing state-of-the-art solutions and methods, as well as address market competition. In ORG I, the choice of a specific machine learning model architecture was dictated by the industry standard. However, this choice had consequences for the label needs during the design of the ground truth schemas. At the same time, addressing market competition influenced the work on the ground truth schema design, as seen here, "so this is [a competitor's system] and this is their output. they ... split consolidation and nodules, which at this stage of the hierarchy we are not doing. And so I was wondering why we're not doing it" (P2). In this organisation competition directly influenced the design work.

Due to the large size of ORG II, the matter of competition fed to internal business processes whose results other experts relied on during the dataset creation, as explained by a product owner (P22), "it's a combination of ... alignment with the business priorities and that is also strongly driven by customer requests and customer demands. So that is actually very important ... try to find the alignment". Finally, market competition created time pressures that could structure and limit how data creation had to be organised: "if you want to validate something properly, it costs time. If you want to validate across domains, it costs time. And we are often in very competitive domains where being fast to market or, or fast at the FDA is also important. So there are some time trade-offs, need to be made there." (P27).

The intended use and the type of healthcare system affected the content and the level of detail of the ground truth schemas. Visions of future intended use permeated the design work on the ground truth schemas. The imagined intended use of a future AI-based system factored into decisions about the validity of label choices. Imagined use did not fit in with current domain-specific practices and resulted in confusion and concerns during the design of the ground truth schema. Consider the following discussion between a medical professional and data scientists from ORG I about the implication of different intended uses of the future system for the selection of labels.

We have two priorities, one is decision support. So it might be easy for you to see the mass, so that won't help you. But there's also the pre-screening - prioritisation. So that might be relevant to detect mass prematurely, right? (P3)

So if you use it for like a warning, a prioritisation, it can be useful, but for detection... we can see a mass. It's not difficult to find (P1).

Epistemic Differences
• Miscommunication between domains
• Misapprehension of medical practice
• Misapprehension of medical knowledge
Limits of Labelling
• Domain expert buy-in
• Onboarding to the labelling task
• Labelling hardware and software
• Similarity to the clinical practice

Table 3: Internal factors and their dimensions

Medical AI-based systems in our organisations were designed to operate across the world within public or private healthcare systems. Yet medical systems in different countries operate differently based on public values, profit, incentives, and conventions. The design decisions during dataset creation are a product of all these components. The dependency on the healthcare type was well captured by a data scientist from ORG I when discussing the level of detail of the ground truth schema, *"if it was in the US where you actually pay, then from a business point of view, you really wanna find everything. First of all, you don't get sued, and secondly, you can make a lot of money by treating them. But here it's very different, right? Because it's a public system and you only treat things that are necessary, that need to be treated, right?"* (P4). These concerns manifested in debates about what could and needed to be annotated as expert annotators infused the values of their local system into data creation activities.

4.2 Internal factors: designing the ground truth schema

While external factors were key in shaping what data was collected and made available for annotation and highlighted the importance of local considerations and their implication for the resulting datasets, two internal factors drove debates, discussions, and disagreements that affected the ground truth schema and the resulting datasets. These were **Epistemic Differences** and **Limits of Labelling** (Table 3). The effort going into the creation of medical datasets as training data had two purposes that sometimes came into conflict. First, medical datasets were seen as a means of capturing the current state of medical knowledge and the tacit knowledge of medical professionals who focused on medical practice and clinical usefulness. Second, the same datasets served computer scientists as complex input data to solve problems through mathematical operations, where consistency and accuracy were in the spotlight. These two perspectives, while not opposing, often prioritised distinct qualities of the same datasets.

4.2.1 Epistemic differences. While in ORG II and ORG III, we engaged with relatively homogeneous teams within each company, in ORG I, our research process was focused on supporting the data creation process by working together with the data science and

radiologist teams. As such, in ORG I, we were able to observe first-hand how teams with domain expertise often disagreed on what constituted legitimate knowledge as they discussed what was worth annotating and how things ought to be annotated. We consider three sources of epistemic differences that affected the final design of the ground truth schemas (Table 3), communication challenges within the teams, misapprehension of medical practice, and misapprehension of medical knowledge. Within these dimensions, team members from different domains expressed diverging priorities, values, and understanding of concepts, which needed to be reassured and negotiated.

Communication challenges within teams. The three organisations involved stakeholders from different backgrounds, such as health, data science, and design. All of these brought their own traditions, meanings, and domain knowledge that needed to be shared, translated, and understood by other parties for worthwhile collaboration. It is no secret that interdisciplinary teams must spend time finding common ground before they can work together productively [7]. In our research, we observed how medical professionals, designers, and data scientists constantly translated and explained concepts from their respective domains to maintain a shared understanding. For example, at the beginning of the study in ORG I, medical professionals designed labels based on their, at times naive assumptions of machine learning capabilities, such as when they included two medical concepts under the same label, *"but couldn't that be, if you put nodule, mass in the same category, couldn't you just program it, later on, to say that if the thing that they have marked nodule/mass is over I think ... five millimetres or something, you call it a mass"* (P1), which was not possible given the collected data and was later clarified through a joint discussion. Similarly in ORG II medical professionals had to explain to data scientists that to detect some types of cancer it is necessary to look at more than just the organ in question, and that doctors need to use other information, such as the condition of bile ducts or the blood flow around the organ, affecting data collection and subsequent labelling set up.

Misapprehension of medical practice. Across the organisations the expectations for the quality of the datasets were closely aligned with concepts such as consistency or bias. This focus was clearly visible when discussing the goal of the labelling task in ORG 1. In the pursuit of consistent and unbiased data, data scientists initially framed labelling as a *"different task"* to clinical work: *"We need to know what's in the image and we need it without them being biased towards looking for only stasis"* (P6). As a result, the labelling task did not provide what was seen by the data scientists as "extraneous and potentially biasing" information, such as the background information of a patient. However, situating the labelling task further away from the medical practice affected the quality of the input medical professionals could provide, impairing the ability of medical professionals to use their knowledge. As one senior radiologist (P10) noted: *"Asking a radiologist to categorise something on a picture only without getting any information on the patient. Is like asking a surgeon to look at the scars on a patient and having him tell you what kind of surgery that patient had"*.

The pursuit of objective and unbiased labels isolated labelling from what data scientists saw as extraneous, potentially biasing information. Yet this transformed the work of the radiologists into a new task that was incompatible with medical practice. To deliver the expected results in this new unfamiliar process, radiologists attempted to reconstruct their medical practice by drawing from their tacit knowledge or, simply, guessing: *I have to create something about the patient myself, which is, [or] might not be true. And I then describe the picture from there...* (P10).

Misapprehension of medical knowledge. Specific data was needed to train AI models that provide clinically useful functionalities. However, due to the misapprehension of practice, the assumptions about what clinical knowledge was possible to extract from the clinical data provided were also, at times, flawed. As the schema went through iterative rounds of design, we observed how both sides struggled to understand why particular data was requested or why a particular request seemed to be difficult to fulfil. For example, in ORG I, radiologists were asked to assign one of three possible values as a patient's general state based solely on a single chest x-ray, so that relevant cases could be later prioritised using the resulting AI system. This task proved to be particularly problematic to radiologists who do not use such metrics in their daily practice, so they had to develop a range of new approaches to assign them, like *"I chose to interpret it from the view that it could be the worst situation"* (P12) or *"I think it was mostly a gut feeling"* (P11). In the end, the radiologists produced the kind of data that data scientists expected to see as labels. However, what these labels actually captured diverged from the original intention.

4.2.2 Limits of labelling. Finally, we turn to the mechanics of labelling itself that affected the final design of the ground truth schema. We observed schema design and testing in situ directly in ORG I, while in ORG II and ORG III, our data come from post-hoc interviews. We find that four features affected the final design of the ground truth schema (Table 3), domain expert buy-in, onboarding to the labelling task, clinical practice familiarity, and labelling hardware and software. These dimensions manifested when evaluating the labelling processes. Unlike the *Epistemic Differences*, where data science was the defining domain, the *Limits of Labelling* emerged as medical professionals confronted the intermediate results of the epistemic negotiations discussed above. These limits altered what kind of data was collected and affected the quality of the labelling.

Domain expert buy-in. Our data showed that domain expert buy-in was crucial and required concessions on the type and amount of collected data. Some ML models require specific types of annotated data, such as *"what we're asking them is for each patient to go through 500 images and for each image to annotate [...] at pixel level"* (P21). Not only are such tasks typically outside of the scope of clinical practice but are also mentally challenging. For example, when P1 was asked to oversee the labelling process performed by external radiologists, they recalled: *"I think that he [a senior radiologist] opened the program, saw how difficult it was, and just closed it and just never had the energy to start it again"* (P1). Monetary compensation turned out to be a necessary but not sufficient strategy in ORG I for recruiting medical professionals with high expertise to annotate data.

Once the experts agreed to annotate data, **limited training for the labelling task reduced the chance for a "shared mindset".** Additional metrics were a relevant part of the ground truth schemas. These metrics usually included concepts not used in daily clinical practice. In ORG I, the medical professionals were supplied with

written guidelines to boost common understanding and were briefly introduced to the labelling task. However, some annotators referred to the guidelines only when in doubt: *[the labelling software worked] right out of the box ... I didn't really read this part because it was not necessary* (P12). Not knowing the exact guidelines, medical professionals relied on an intuitive understanding of the metrics and labels, which often resulted in discrepancies between the annotators as they attributed different meanings.

Hardware configuration and user interface of the labelling software affected the quality of the annotations. These challenges were observed to a greater extent in ORG I, as to assess medical data like CT scans and x-rays, radiologists usually use diagnostic displays. Thus, when they annotate on a *"non-diagnostic screen, you miss details ... maybe small, smaller changes would be missed ... we don't annotate them because we cannot see them"* (P13). Similar comments were shared during the evaluation of the labelling software, medical professionals marked the location of findings using touchpads, which resulted in frustration and low precision.

Labelling software design could have influenced the final quality of the medical dataset to an even greater extent if not caught during the evaluation. Labelling medical data requires "[a] professional tool that could do the job in a very efficient way" (P21). However, the design of this software could have influenced radiologists in ORG I to overreport radiological findings per x-ray during an evaluation period, *"...maybe it's an interface. Maybe they forgot the normal button was there because they only saw the [labels]"* (P1).

The overreporting was not solely caused by the labelling software. **Expectations and habits influenced what medical professionals noticed in medical data.** For example, a radiologist who reported on an evaluation of the ground truth schema in ORG I reported, *"I told my participants that there would be some normal, but they have not marked any of them normal or I can't find them"* (P1). This phenomenon was later explained by a senior radiologist who pointed to the expectation of labelling a dataset with findings and the fact that when the ratio of abnormal to normal cases is skewed, radiologists tend to overreport to remain on the safe side, *"that's [why] they thought they saw something that was not there"* (P6).

5 DISCUSSION

In the creation of high-quality training data, our research shows that the design of ground-truth schema is a crucial but often overlooked step. We highlight five factors that represent external and internal constraints that directly affect the quality of the resulting medical datasets. The external constraints condition the data collection process, affecting this way the design of the ground truth schema, while the internal constraints strongly affect the resulting ground truth schema and can lead to disagreements and debates among domain experts, predominantly data scientists and medical professionals.

5.1 Conditioning the data collection

Our findings demonstrate that the regulatory constraints, along with the geographical, demographic, and linguistic context of creation and intended use, and the organisations' scalability crucially affect the amount and type of data that was possible to be collected by the organisations we studied. In this sense, specific data quality metrics were already compromised since the first stage of the medical datasets creation. For example, in ORG I and II the geographical and demographic distribution of the collected data reflected not only how much data was possible to be collected by the contractual agreements in place but also manifested a lack of representativeness, given the regional and local source of data collection.

In ORG III, the aspirations for creating datasets of global coverage stumbled upon the linguistic contextuality of medical terms, which proved to become an issue during the ground truth schema design for the match-making platform. Similarly, in ORG I, the geographical, demographic, and linguistic context of the medical data collection shaped the type of the collected data, such as that when the experts came to decide on how to design the ground truth schema, dilemmas did not only concern the different understanding of the same medical terms across countries and continents but also possible omissions of local lung diseases. In this sense, the aspiration of designing "transferable" ground truth schemas proved to be both dependent and limited by the standards that regulate the data collection and the context of its collection.

A further insight that emerged in our studies was that the business models and scalability of each organisation affected differently its capacity to collect data. For example, ORG I, being a small-size start-up, having however the public sector involved in its entity, had easier access to timely data (x-ray images of multiple years) from regional hospitals. Yet, the organisation's limited scalability defined the amount of data that was possible to be labelled by medical professionals. In Org III, a similarly small-size start-up, the data collection from both public registries and patients was shaped by the organisation's availability of resources. The constraints were imposed on the recruitment of data scientists designing the platform's ground truth schema and medical professionals who assisted the patients in submitting their medical information into an appropriate and structured format. On the other hand, in ORG II, due to its large size and scalability, the limitations of the data collection were shaped by market demands. This was reflected in the need to collect quality data, i.e., particularly structured, consistent, and contextual medical images from a controlled environment (the contracted local hospital). This push for one type of quality reduced another, in this case, the representativeness of the acquired data.

So far, scholarship has defined and treated data acquisition as a particular step in the data creation process, existing in a vacuum [1, 11, 26, 50, 65]. Very little is known about how this step influences the stages that precede the data labelling, eventually affecting the shape of the final medical dataset. Our studies show that regulatory constraints, the context of data creation and use, and the business models and scalability of the organisations, crucially affect the extent and the type of data that is possible to be collected and processed.

5.2 Conditioning the ground truth design

Within this context, we identified the design of the ground truth schema as a crucial stage of medical dataset creation. In our studies, the externally imposed constraints shaped the amount and type of data that reached the stage of designing ground truth schema. This has implications for scholarly discussions that focus on developing documentation frameworks that support the responsible and informed use of complex datasets [2, 5, 20, 31, 44]. We showed that

the decisions taken during the design of the ground truth schemas were foundational to the succeeding stages of dataset creation. We argue that in this stage, experts do not deal with ideal conditions, but there are inherent limitations which we conceptualised as epistemic differences and limits of labelling. We further argue that the external constraints influence how these inherent limitations manifest in situated collaborative domain settings.

The amount and type of data that reach the ground truth schema design is already shaped by the necessity of organisations to comply with regulatory standards. This has led the experts from ORG I and II to work with data that had limited representativeness from the start, further affected by the predefined purpose of use and geographical, demographic, and linguistic context for its collection and use. These had implications for the negotiations between data scientists, designers, and medical professionals on what "makes sense" to be labelled.

Domain negotiations that we observed, were grounded in epistemic differences that did not take place with symmetrically allocated roles, where the "separation of concerns" of each domain expertise is often negotiated against the tacit medical knowledge but where data scientists have the first say [32, 55, 62]. Having the development of AI models as the purpose of medical dataset creation, data scientists were positioned as the problem owners of the data creation processes. This further distanced the design of the labels from the medical domain experts and was manifested through misapprehensions about medical knowledge and practice. The tensions with the medical professionals often led to negotiations about what was medically important to be annotated versus what would lead to high-quality datasets from a data science perspective. At the same time, both of these standpoints had to correspond to the demands of the health-tech market.

We found that the externally imposed concerns, such as compliance with regulatory standards, the context of creation, and the intended use of the data, along with the commercial and operational pressures, condition the data collection and can affect ground-truth schema design. In fact, many crucial decisions and negotiations relevant to the final shape of the medical datasets take place during the stage of ground truth schema design. All three organisations under study were committed to developing AI systems in a responsible way. As such, the creation of high-quality training data was a crucial step. Yet, no matter how hard they tried to create representative, consistent, well-structured, high-quality data, the resulting datasets were already limited in different ways. We showed how these limits were predefined even before any data labelling occurred. The combination of external constraints that limit and structure data collection with the misapprehension of domain practice resulted in highly paid experts having to imagine and invent additional information to perform the tasks asked of them. A limited understanding of what is required for diagnosing various conditions from medical images could have consequences. Either new datasets would have to be created, which translates into a new data collection process, with all the regulatory constraints attached, or the labelling software would have to be more aligned with the existing professional practices following the guidance of expert annotators. Even where these issues were resolved, medical professionals annotated data based on their particular experience and tacit knowledge. This means that the geographical location of the experts affected what they expected to see in the data, showcasing that expertise does not account for the uneven distribution of diseases in different parts of the world.

6 LIMITATIONS AND FUTURE WORK

Our contribution builds on qualitative data from three organisations located in countries of the Global North. Creating medical AI datasets in different countries of the Global South may present different challenges and be influenced by a different set of factors that were not captured in our data. Further research is needed to better understand how medical AI data creation varies across different regions and cultures.

Our study focuses on only two medical areas: radiology and clinical trials. While we engaged with diverse types of medical data, creators of other medical datasets could face challenges unique and dependent on different types of medical specialisations. Future research should aim to explore the factors that influence the design of medical AI datasets across a wider range of medical specialisations to develop a more comprehensive understanding of the factors that influence it.

7 CONCLUSIONS

In this paper, we investigated the work of data scientists, medical professionals, and designers that takes place before the labelling of medical data. Building on the qualitative accounts of our ethnographic findings, our main contributions are:

- conceptualising five factors that influence the creation of medical datasets;
- disclosing how these factors condition the design of ground truth schemas;
- suggesting identified relationships amongst these factors;
- staging the design of the ground truth schemas as a highly contested, yet crucial step in the creation of medical datasets that precedes and conditions data annotation.

These overarching factors had a fundamental influence on the final shape of medical datasets created for AI use. First, the externally imposed constraints should be systematically taken into account during the entirety of the medical dataset creation processes, as these factors define the data collection and condition the design of the ground truth schemas. Second, we have exemplified the breadth of decisions taken before the annotation of medical data. Foundational decisions about the final shape of medical datasets take place during the design of a ground truth schema. Future endeavours in data science, law, and policy should consider this stage as crucial to achieving responsible medical AI.

ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to all of our participants, especially Dr. Elijah Kwasa, Dr. Edward Mwaniki, Dr. Marian Morris, Dr. Ruth Wanjohi, Dr. Mary Onyinkwa, Dr. Sayed Shahnur, and Dr. Samuel Gitau for their invaluable contributions and insightful input. Thank you for taking the time to engage with us and for your significant impact on our work.

REFERENCES

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019*. 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[2] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445736

[3] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (3 2015), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (6 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (12 2018), 587–604. https://doi.org/10.1162/tacl{_}a{_}00041

[6] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out - Classification and Its Consequences*. The MIT Press. https://mitpress.mit.edu/9780262522953/

[7] Rebekah R. Brown, Ana Deletic, and Tony H. F. Wong. 2015. Interdisciplinarity: How to catalyse collaboration. *Nature* 525, 7569 (9 2015), 315–317. https://doi.org/10.1038/525315a

[8] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2020. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 66 (12 2020), 101797. https://doi.org/10.1016/j.media.2020.101797

[9] K Charmaz. 2014. *Constructing Grounded Theory (2nd ed.)*.

[10] Haihua Chen, Jiangping Chen, and Junhua Ding. 2021. Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability* 70, 2 (6 2021), 831–847. https://doi.org/10.1109/TR.2021.3070863

[11] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. 2019. How to develop machine learning models for healthcare. *Nature Materials* 18, 5 (5 2019), 410–414. https://doi.org/10.1038/s41563-019-0345-0

[12] Adele Clarke. 2005. *Situational Analysis*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America . https://doi.org/10.4135/9781412985833

[13] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23, 2 (3 2016), 304–310. https://doi.org/10.1093/jamia/ocv080

[14] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. (12 2021). http://arxiv.org/abs/2112.04554

[15] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 227–236. https://doi.org/10.1145/3514094.3534187

[16] Virginia Dignum. 2020. Responsibility and artificial intelligence. In *Oxford Handbook of Ethics of AI*, Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.). Oxford University Press, Chapter 11, 215–231.

[17] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery 2022 36:6* 36, 6 (9 2022), 2074–2152. https://doi.org/10.1007/S10618-022-00854-Z

[18] Melanie Feinberg. 2017. A Design Perspective on Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Vol. 2017-May. ACM, New York, NY, USA, 2952–2963. https://doi.org/10.1145/3025453.3025837

[19] Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing*. John Wiley & Sons, Inc., Hoboken, NJ, USA. 1–164 pages. https://doi.org/10.1002/9781119306696

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (12 2021), 86–92. https://doi.org/10.1145/3458723

[21] R. Stuart Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. 2021. "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies* 2, 3 (11 2021), 795–827. https://doi.org/10.1162/QSS{_}A{_}00144

[22] Lisa Gitelman. 2013. *"Raw Data" Is an Oxymoron*. MIT Press. 9–10 pages. https://nyuscholars.nyu.edu/en/publications/raw-data-is-an-oxymoron

[23] James Thomas Patrick Decourcy Hallinan, Mengling Feng, Dianwen Ng, Soon Yiew Sia, Vincent Tze Yang Tiong, Pooja Jagmohan, Andrew Makmur, and Yee Liang Thian. 2022. Detection of Pneumothorax with Deep Learning Models: Learning From Radiologist Labels vs Natural Language Processing Model Generated Labels. *Academic Radiology* 29, 9 (9 2022), 1350–1358. https://doi.org/10.1016/J.ACRA.2021.09.013

[24] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer Vision and Conflicting Values: Describing People with Automated Alt Text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 543–554. https://doi.org/10.1145/3461702.3462620

[25] Anne Henriksen and Anja Bechmann. 2020. Building truths in AI: Making predictive algorithms doable in healthcare. *Information Communication and Society* 23, 6 (2020), 802–816. https://doi.org/10.1080/1369118X.2020.1751866

[26] Charles Hill, Rachel Bellamy, Thomas Erickson, and Margaret Burnett. 2016. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 162–170. https://doi.org/10.1109/VLHCC.2016.7739680

[27] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. (5 2018). http://arxiv.org/abs/1805.03677

[28] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, Vol. 20. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3419249.3420149

[29] Rory Horner. 2020. Towards a new paradigm of global development? Beyond the limits of international development. *Progress in Human Geography* 44, 3 (6 2020), 415–436. https://doi.org/10.1177/0309132519836158

[30] Eduard Hovy and Julia Lavid. 2010. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *INTERNATIONAL JOURNAL OF TRANSLATION* 22, 1 (2010).

[31] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[32] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 560–575. https://doi.org/10.1145/3442188.3445918

[33] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (1 2019), 590–597. http://arxiv.org/abs/1901.07031

[34] Saahil Jain, Akshay Smit, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Effect of Radiology Report Labeler Quality on Deep Learning Models for Chest X-Ray Interpretation. (4 2021). http://arxiv.org/abs/2104.00793

[35] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. (1 2019). http://arxiv.org/abs/1901.07042

[36] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (7 2021), 695–703. https://doi.org/10.1145/3461702.3462598

[37] Dana Li, Lea Marie Pehrson, Lea Tøttrup, Marco Fraccaro, Rasmus Bonnevie, Jakob Thrane, Peter Jagd Sørensen, Alexander Rykkje, Tobias Thostrup Andersen, Henrik Steglich-Arnholm, Dorte Marianne Rohde Stærk, Lotte Borgwardt, Kristoffer Lindskov Hansen, Sune Darkner, Jonathan Frederik Carlsen, and Michael Bachmann Nielsen. 2022. Inter- and Intra-Observer Agreement When Using a Diagnostic Labeling Scheme for Annotating Findings on Chest X-rays: An Early Step in the Development of a Deep Learning-Based Decision Support System. *Diagnostics 2022, Vol. 12, Page 3112* 12, 12 (12 2022), 3112. https://doi.org/10.3390/DIAGNOSTICS12123112

[38] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray. 1993. The Unified Medical Language System. *Yearbook of Medical Informatics* 02, 01 (8 1993), 41–51. https://doi.org/10.1055/s-0038-1637976

[39] Marry L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 22, 3 (2012), 276–282. https://doi.org/10.11613/BM.2012.031

[40] Teresa Mendonca, Pedro M. Ferreira, Jorge S. Marques, Andre R. S. Marcal, and Jorge Rozeira. 2013. PH2 - A dermoscopic image database for research and benchmarking. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5437–5440. https://doi.org/10.1109/EMBC.2013.6610779

[41] Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (5 2022). https://doi.org/10.1145/nnnnnnn

[42] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (1 2022), 1–14. https://doi.org/10.1145/3492853

[43] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020), 1–25. https://doi.org/10.1145/3415186

[44] Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. Documenting Data Production Processes: A Participatory Approach for Data Work. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (11 2022), 510. https://doi.org/10.1145/3555623

[45] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 161–172. https://doi.org/10.1145/3442188.3445880

[46] Stefania Milan and Emiliano Treré. 2019. Big Data from the South(s): Beyond Data Universalism. *Television & New Media* 20, 4 (5 2019), 319–335. https://doi.org/10.1177/1527476419837739

[47] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[48] 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. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356

[49] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445402

[50] Elizamary de Souza Nascimento, Iftekhar Ahmed, Edson Oliveira, Marcio Piedade Palheta, Igor Steinmacher, and Tayana Conte. 2019. Understanding Development Process of Machine Learning Systems: Challenges and Solutions. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 1–6. https://doi.org/10.1109/ESEM.2019.8870157

[51] Luke Oakden-Rayner. 2019. Exploring large scale public medical image datasets. (7 2019). http://arxiv.org/abs/1907.12720

[52] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning* (2 2020), 151–159. https://doi.org/10.1145/3368555.3384468

[53] 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*, Vol. 2015-April. ACM, New York, NY, USA, 3147–3156. https://doi.org/10.1145/2702123.2702298

[54] Dennis Reidsma and Jean Carletta. 2008. Reliability Measurement without Limits. *Computational Linguistics* 34, 3 (9 2008), 319–326. https://doi.org/10.1162/coli.2008.34.3.319

[55] David Ribes, Andrew S Hoffman, Steven C Slota, and Geoffrey C Bowker. 2019. The logic of domains. *Social Studies of Science* 49, 3 (June 2019), 281–309. https://doi.org/10.1177/0306312719849709 Publisher: SAGE Publications Ltd.

[56] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and Explanation in Computational Notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Vol. 2018-April. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173606

[57] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M 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*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445518

[58] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–37. https://doi.org/10.1145/3476058

[59] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A Step Toward More Inclusive People Annotations for Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 21. ACM, New York, NY, USA, 916–925. https://doi.org/10.1145/3461702.3462594

[60] Cathrine Seidelin, Yvonne Dittrich, and Erik Grönvall. 2018. Data Work in a Knowledge-Broker Organisation: How Cross-Organisational Data Maintenance shapes Human Data Interactions. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference, HCI 2018*. https://doi.org/10.14236/ewic/HCI2018.14

[61] Susan Leigh Star and Anselm Strauss. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work* 8, 1-2 (1999), 9–30. https://doi.org/10.1023/A:1008651105359/METRICS

[62] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3491102.3517537

[63] Holger Voormann and Ulrike Gut. 2008. Agile corpus creation. *Corpus Linguistics and Linguistic Theory* 4, 2 (11 2008), 235–251. https://doi.org/10.1515/CLLT.2008.010/MACHINEREADABLECITATION/RIS

[64] Laura Trajber Waisbich, Supriya Roychoudhury, and Sebastian Haug. 2021. Beyond the single story: 'Global South' polyphonies. *Third World Quarterly* 42, 9 (9 2021), 2086–2095. https://doi.org/10.1080/01436597.2021.1948832

[65] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–24. https://doi.org/10.1145/3359313

[66] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2019. ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases. In *Advances in Computer Vision and Pattern Recognition*. 369–392. https://doi.org/10.1007/978-3-030-13969-8{_}18

[67] Hubert D. Zając. 2022. Designing ground truth for Machine Learning - conceptualisation of a collaborative design process between medical professionals and data scientists. *Proceedings of 20th European Conference on Computer-Supported Cooperative Work* (2022). https://doi.org/10.48340/ecscw2022{_}p04

[68] 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). https://doi.org/10.1145/3392826