CHARLOTTE HÖGBERG

“This ground truth is muddy anyway”

Ground Truth Data Assemblages for Medical AI Development

Abstract
This article explores assemblages of ground truth datasets for the development of medical artificial intelligence (AI). By drawing from interviews and observations, I examine how AI experts developing medical AI relate to the referential truth basis of their work, their ground truths, as an epistemic concern. By addressing how datasets are assembled from different sources, and produced, augmented and synthesised, this study shows how ground truths are valued based on humanness, quality of medical expert judgements, temporality and technical qualities. Moreover, this article analyses truth practices as productive moments in AI development, the role of human expertise and the perceived strengths and limits of expert-based annotations. The valuations of ground truths shatter the image of medical classifications, and AI models, as stable neutral entities. Moreover, this article shows how valuations of ground truths encompass more than alignment with standardised expertise. To better understand the possibilities for medical AI to live up to ideals of accuracy, fairness, trustworthiness and transparency, we need more knowledge on assumptions, negotiations and epistemic concerns upon which medical AI is built.

Keywords: artificial intelligence, data, ground truth, medicine, epistemology

TWO MAIN QUESTIONS to consider with regards to the development of an artificial intelligence (AI) model is: On *the basis of what* does the algorithm “learn” the correct classification? And *what* is used to measure whether the algorithm’s prediction is accurate? What both of these questions amount to is, to a large extent, what choices have been made in the construction of the algorithm’s “ground truth”? (Jaton 2017). This article attends to ground truths as an epistemic concept, and the practices and value negotiations by which these truths are assembled as performative and productive moments in AI development. Through this, assumptions in AI development become visible, enabling an increased understanding of the conditions that are shaping AI’s possibilities to reach ideals such as accuracy, fairness, transparency and trustworthiness.

The term ground truth is used by AI developers and researchers to describe the referential datasets perceived as holding the “true” values of the phenomena that are computationally modelled for (e.g. Kang 2023). For example, a dataset that contains x-ray images and corresponding labels, describing whether there is a malign tumour present in the image, can be used as a ground truth for a cancer detection algorithm.

https://doi.org/10.37062/sf.62.27826

However, as Jaton (2017) argues, these perceived truths do not pre-exist but have to be constructed as datasets made to fit the task of the algorithm. This is in line with work in the social studies of science that emphasise data as inescapably local (Timmermans and Berg 1997; Loukissas 2019), and never actually raw (Gitelman 2013), as shaped by factors such as the place, organisation, time and means of collection and generation. Hence, digital data are not neutral mirrors of a natural world. Yet, as computational models to an increasing extent learn by examples and not by rules (e.g. Campolo and Schwerzmann 2023), data are used to make generalisations about complex phenomena.

This study intends to shed light upon how AI researchers deal with "ground truths" in the specific context of AI development for medical research and healthcare purposes. Within critical data studies and science and technology studies (STS), research has stressed matters such as the paradoxes, work and politics of data-driven healthcare (Hoeyer 2023; Avlona and Shklovski 2024; Bertelsen et al. 2024), the inheritance and historicity of digital medical data (Green and Svendsen 2021) and the role of algorithms in epidemiology such as in the case of enacting the zika pandemic (Lee 2021). Yet, research into the data related work of developers of medical AI remains scarce (Bertelsen et al. 2024).

With regards to the recent years' spurt of machine learning applications in medicine, and the hope that it will result in immense progress (Rajpurkar et al. 2022), there is a continuous need for insights into what assumptions are embedded in datasets and model constructions. While there is a growing body of literature concerned with the statistical content of large benchmark datasets for machine learning and issues such as a lack of representation, there is still relatively little social research that focuses on modes of construction of machine learning datasets and their function as informational infrastructures (Denton et al. 2021). This field, conceptualised by Thylstrup (2022) as critical dataset studies, is yet emerging within the broader scope of research on how data assemblages do work in the world (Kitchin and Lauriault 2018).

However, we need more empirical knowledge about ideas and practices concerning "ground truths" as a particular performative concept in AI development (Jaton 2023). Increased critical consideration can elucidate ground truth negotiations as a certain way of understanding AI models and their relation to medical phenomena, how knowledge-making is shaped by this process, and the sociology of truth in which people, artefacts and practices are involved (Jaton 2017; Henriksen and Bechmann 2020; Lebovitz, Levina & Lifshitz-Assaf 2021; Kang 2023; Zając et al. 2023). By drawing from qualitative empirical work, the aim of this article is thus to increase the knowledge of how researchers developing medical AI relate to the referential truth basis of their work, their ground truth, in terms of truth practices and valuations. In addition, it analyses ground truth as an epistemic concern in medical AI.

With an ambition to make visible the reasoning and practices around data, and specifically ground truth datasets in medical AI, this study shows how data are treated as a workable concern that has to be navigated by AI researchers. In this way, it analyses the role of: expert-based labelling; augmented and synthetic data; generalisation; and brittleness in the assemblage of ground truths for medical AI. Moreover, it brings ground truths to the fore as an emic concept, as well as shows how the valuation of ground truth qualities goes beyond the alignment with standardised expertise. The contribution of this article is both empirical, showing AI experts' reasoning and practices in relation to medical ground truths, and theoretical, by furthering the conceptual understanding of ground truth practices, negotiations and valuations, as performative elements in AI development. By this, I aim to contribute to STS, critical data studies and sociological perspectives on AI and information.

This article is structured as follows. First, I provide a background to the concept of ground truths and previous social research on ground truthing, followed by an introduction of the conceptual framework. Subsequently, I present the method used and the empirical findings. Lastly, the concluding discussion offers further analysis of the findings and this article's contribution.

Background

The concept of a "ground truth" has long been used in the fields of geology and meteorology, referring to the perceived reality of meteorological conditions by observed and registered measurements. In an ethnography of meteorological forecasters, Fine argues that they describe the search for ground truth as "focused on deciding 'what is real', given organisational demands to produce useful information" (Fine 2006:7). The concept is also used as a verb, ground truthing, accentuating the practised aspect of how ground truths are not given but are generated or assembled, and put into action. The concept of ground truths has been adopted by computer scientists and in the machine learning context. Also here, it is used as a referent to the "true values" of the modelled phenomena (Kang 2023), or in conjunction with a certain "mode of truth telling" pertaining to the ground truth "from which the algorithm generates its model of the world" (Amoore 2020:136). It can be describes as the repository from where machine learning models derive; as Kang (2023:3) argues: "it is literally where the truth and possibility of an algorithm are grounded."

The perceived ground truth availability, the choice of method to make algorithms "learn", and the quality of the deriving models are seen as highly interdependent elements of machine learning (Siebert et al. 2020). In the context of medicine and healthcare, the ground truth datasets used for AI development generally can be described as containing knowledge objects in the form of representations of health statuses or other characteristics, by for example tabular data or images, paired with expert annotations and labels describing such matters as whether the image is depicting a tumour or not, or whether that specific individual suffered from a brain aneurysm or heart attack.

With regards to truth, a rich body of literature has contributed to our understanding of how scientific facts become stabilised and considered as truthful (e.g. Pinch and Bijker 1984; Latour 1987; MacKenzie 1990; Knorr-Cetina 1999; Daston and Gailson 2007). This literature has uncovered the social construction of technology and scientific facts, showing how they are shaped by actors, contexts and epistemic cultures (Knorr-Cetina 1999). Moreover, ideas about truth and objectivity have been problematised in previous research. Shapin (1994) argues that the basis for why scholarly claims are regarded as truths has changed over time. Scientific practices were in seventeenth-century England performed as part of genteel conduct, by which the word of a scientist was considered trustworthy based on ideas about a certain gentleman character seen as incapable of lying. Trust emerges as an important component of knowledge-making and what are considered as scientific truths about the natural world, and credibility of scientific facts, is now more closely tied to the credibility of the organisation that the scientist represents (Shapin 1994).

The role of the scientific experiment – the witnessing of it, the deriving data and reports – has also been examined as a practice that produces trust in scientific claims and justification of knowledge (Shapin & Schaffer 1985). Yet, ideas about truths are also tied to objectivity as an epistemic ideal. Daston & Gailson (2007) show how this ideal emerged as an epistemic virtue during the mid-nineteenth century and has been evolving, contingent on cultural and social ideas about accuracy and scientific community practices. Moreover, STS research has troubled the idealised distinction of science between dealing with the discovery of truth and technology and dealing with the application of truths, showing their intricate relationship within knowledge production (Pinch and Bijker 1984:402) and specifically in early variants of AI in terms of expert systems (Collins 1990; Forsythe 2001) and matters such as the politics and "truthiness" in risk prediction modelling (Amoore 2013; Weinkle and Pielke 2016). In addition, research has discussed the ideals of evidence-based medicine and shown the contingency and locality of medical standardisations, as constantly adapted to local needs (Timmermans and Berg 1997, 2003; Mackenzie et al. 2013), by which data is also collected and now increasingly used for statistical analysis and prediction purposes.

Previous social science studies of ground truthing emphasise it as the problematisation that is shaping algorithms, defining both their inputs and outputs (Jaton 2017), and suggests that internal and external factors impact the creation of ground truth schemas within the medical domain, through regulatory restrictions, commercial and operational pressure and epistemic differences (Zając et al. 2023). In a case study of a ground truthing project for personalised cancer immunotherapy, Jaton (2023) found that what it established came to be a contestable reference, rather than a undisputable "truth" due to non-stabilised measurement protocols. However, he argues that ground truths are a necessary condition to enable AI technologies in personalised medicine.

Earlier research also stresses the uncertainties pertaining to what expert knowledge is embedded in ground truths of AI. In a review of five machine learning (ML) tools prior to clincal implementation, Lebovitz, Levina & Lifshitz-Assaf (2021) depict how hospital managers questioned why the tools did not work as desired, leading managers from reviewing accuracy scores and ground truth labels, to evaluating the human experts' daily work of dealing with uncertainty and producing high-quality judgements. What they found was a disconnect where ML tools' ground truths incorporated expert know-what, but not the expert know-how that was important in clinical practice; yet, also how a focus on dissecting ground truths enabled ways to make sense of medical

AI and better understand the reason for unsatisfactory performance. This emphasises the need to increase our knowledge about reasoning and practices of ground truth data assemblages to better understand the implications of medical AI solutions. Previous work also shows how “truth practices” shape the making of AI for healthcare and work to reinvent truths and medical practices to elevate prevalent logics of decisions about patients, rather than discovering new truths through AI (Henriksen and Bechmann 2020). In this article, I aim to draw from and add to this empirical body of knowledge, but also consider how to conceptually analyse ground truth negotiations, narrations and practices of AI experts.

Conceptual framework

Machine learning and truth practices

In this article, the concept of ground truths is treated in an emic manner, based on how it is used by the AI and ML experts themselves, in their reasoning and practices. In line with the distinction made by Jaton (2021), I regard the researchers’ ML practices and assembling of ground truths as sociological practices. Drawing also from Mackenzie (2017), ML is considered a practice that involves both humans and machines. This entails a focus on the social construction of technology and the sociotechnical entanglement between humans and technology (Latour 2005). In that sense, medical AI is in this study understood as a technology made out of assemblages of human and non-human entities, since “algorithms are not autonomous technical objects, but complex sociotechnical systems” (Seaver 2018:378). Moreover, Seaver (2017) argues that we should regard algorithms *as* culture rather than *in* culture. In the context of medicine and AI research, the status of scientific facts is understood as formed within certain epistemic cultures and their material and discursive “epistemic machinery” (Knorr-Cetina 1999).

Based on these theoretical underpinnings, scientific facts and epistemic cultures are understood as co-constructed along with technologies. One part of this, and of relevance to ground truthing in medical AI development, is how it presents a way to make both scientific and engineering problems *doable* in practice (Fujimura 1987). More specifically in this article, the relayed and observed work of AI experts is approached as truth practices, which Henriksen and Bechmann (2020) outline as a type of multimodal, and multilevel, performance of truth, involving several actors, including engineers and medical specialists, with the use of different methods and sources of knowledge. In sum, truth is “performed within a network of different actors along with data, machines, and ML models” (Henriksen and Bechmann 2020:812).

(Un)stable classifications and valuations

From a sociological perspective, what most medical algorithms aim to perform is some sort of classification of health and (risk of) disease (although the tasks are not necessarily always conceptualised as classification by the narrower computer science terminology). In the case of medical AI, this includes the ontologies of a given disease, meaning that one has to decide which data and variables provide a basis for what the disease is and how it can be recognised. Mol argues for attending to the multiple ontologies that are assembled into diagnoses as a "coexistence of multiple entities that go by the same name" (Mol 2002:151). Mol's work shows that a medical condition can be construed as one phenomenon, while still being enacted as multiple: as a patient's level of pain when walking, a general practitioner's medical examination and anamnesis, the radiologist's interpretation of images, and the pathologist's analysis of the veins on the surgical table. This sensibility towards ontologies, in the shape of data, can be used as an analytical device to understand how medical conditions are turned into models and algorithms. This entails paying attention to the (in)stability of medical classifications (Bowker and Star 1999). It moreover acknowledges the ontological power (Mol 1999) that algorithms encompass as they perform worldling capacities with datasets as "classification engines" (Crawford 2021). In line with Mol's reasoning of multiple ontologies of medical conditions, Seaver (2017) emphasises algorithms as unstable multiples in themselves.

The different possibilities to shape ontology furthermore speaks to how data, sources, and expert judgements are valued in different ways. Valuation studies have offered a collection of lenses to make visible how and what values are ascribed to matters such as technology in practices and discourse. For example, Lee and Helgesson (2019) find a multivalence of algorithms in practice in biomedicine, different styles of valuation of algorithms and what configurations of algorithms and humans are considered as providing "good" bioscience and a good distribution of human versus technological agency. Styles in this conceptualisation centre on matters such as actors' articulations of problems, solutions and configurations, as analytical tools to examine the ambiguous role of algorithms (Lee and Helgesson 2019). Using these concepts, we can attend to how ground truths are valued in the development of medical AI, what is considered a "good ground truth" and how it is considered to contribute to "good" medicine and healthcare practice.

In sum, this study attends to truth and valuation practices in which ground truths are made as sociotechnical assemblages, built upon and performing medical classifications.

Method

One way to research algorithms and their role in society is to interview coders or conduct ethnography to uncover "the story behind the production of an algorithm and to interrogate its purpose and assumptions" (Kitchin 2017). This study focuses on a particular part of the "production" by attending to the reasoning and practices concerning ground truths for medical AI development. However, it does not extensively follow one specific algorithm but rather the views and practices of a group of experts in AI and ML as applied to medical research or healthcare. An "expert" is here defined as someone with institutional authority to construct reality, with knowledge that can be seen as having the potential to be hegemonial in organisations and fields of practice (Meuser and Nagel 2009:18–19). Thus, the expert is also identified by their professional role as a researcher, while expertise is still relational, acquired in practice, sociocultural conditioned and under negotiation (Meuser and Nagel 2009:18–19; Grundmann 2017).

This article draws from qualitative empirical work in the form of interviews and observations. In-depth, semi-structured, interviews were conducted with 15 researchers and doctoral students, in Denmark, Sweden and the Netherlands. Informants comprise a range from full professors to doctoral candidates and were identified through purposive sampling and snowball sampling. Most worked within publicly or privately funded research organisations, mainly AI centres at universities. However, two informants worked in research roles at commercial medical AI companies and four had shared research positions between universities and AI/engineering roles in hospitals or in commercial AI companies. This is in line with how engineering in academia collaborates with industry and domains of application. As the use of ML methods is becoming more widespread, disciplinary boundaries of those developing medical AI become more blurred. While most of my informants were active in the field of computer science, they had somewhat differing backgrounds. Some came from the fields of mathematics or epidemiology into AI research, or resided in more applied areas at universities, such as biomedical engineering, medical physics departments or in pathology. Due to the area of application, the informants in general published their research in both engineering and medical science journals.

In terms of AI technologies, most informants had experiences of working with several different technologies, such as convolutional neural nets, natural language processing; and with various types of data, including medical images (of brains, breasts, foetuses, hearts), sensor, tabular, and register data, and with aims such as disease/anomaly detection and prediction. All of them worked on AI solutions for medical applications and thereby share an epistemology mainly deriving from computer science, with much focus on developing or refining computational methods. For example, several were involved in improving computational anomaly or object detection, regardless of whether they were currently training an algorithm to detect tumours, pathology stains or cerebral infarcts in the images. Yet, to have medicine as the domains of application presents certain particularities at the borders between computer science methodologies and medical knowledge. They had a shared epistemological concern in how AI can be used to gain medical (and clinically useful) knowledge, and also strategic concern in how AI can be accepted and contribute to improved patient outcomes.

Six of the interviews were conducted online through a video conferencing tool, and one was conducted in person in a café at the premises of a medical university. Those remaining were conducted in person at the informants' place of work. In six cases, the interview was paired with short-term observations and demonstrations of the work conducted at the informant's lab. Multi-sited observations were also conducted at scholarly and intersectoral seminars and conferences focusing on AI in medicine and healthcare. While short-term ethnography has limitations in comparison to long-term onsite fieldwork, it can offer valuable insights into common practices (Pink and Morgan 2013). Transcripts and field notes were analysed by a grounded theory approach and thematically coded. Themes were identified by inductive analysis of reoccurring, contradicting or particular topics and narratives in the material (Ryan and Bernard 2003). When presenting the findings, pseudonyms are used to protect the integrity of informants. For this article, I focus a subset of the empirical results, where matters of ground truths became visible or articulated.

Truth becomings, limits and valuations

In this section I explore the question of how AI experts relate to their ground truths in medical AI development by a selection of empirical cases illustrating truth and valuation practices. The findings are presented in line with the identified themes but also have a processual meaning, starting from the acts of bringing ground truth datasets together, to the limits of certain characteristics of ground truths and the perils and hopes of acting without ground truths. Subsequently, I address the augmenting and synthesising of ground truths and how different qualities of ground truths are valued. As I attend to ground truth as an emic concept, this is also how I use the word *truth*, for example in the thematic sectioning, stressing both its emic flexibility and potential as analytical provocation. Moreover, when using the term "expert-based" in the context of medical AI, it refers to the judgement of the medical experts and not the AI experts. Yet, to begin with, ground truth datasets have to be assembled.

Bringing truths together

There is much that could be gained from early detection of anomalies in sonograms of human organs. Potentially, it can enable treatment, hinder adverse events and save lives. But sometimes sonographers miss signs on the screen, or there are not enough trained sonographers to consult. At an AI research centre, Johan is training algorithms to be able to detect anomalies in sonograms, by using convolutional neural networks for image analysis and object recognition. For this to be possible, the team needs to have a reference set of images that they can treat as depicting, versus not depicting, anomalies. In this case, it derives from data found in a medical register containing images and corresponding medical expert annotations. This judgement is what the team has to rely on, even if it is not without doubts:

> In terms of the anomalies … that is always the question of whether it was documented properly. Just because something is not there, doesn't mean that it wasn't discovered. It is always really hard to make this sort of assumption[s]. And it is still assumptions you need to make; to say, this is my ground truth.

In his statement, Johan shows how the establishment of a ground truth is a pragmatic positioning for the AI developer. Ground truths play a prominent role in the problematisation of medical AI, making medical problems doable (Fujimura 1987) for

AI development, by limiting what "my ground truth" is and what is to be found in the images, as well as when the algorithm succeeds or fails in the detection task. By bringing a dataset together and saying that these data hold the true values that we are modelling for, AI experts can use it in different ways. Sometimes it is seen as a separate dataset for testing and validating the algorithm's performance (how well it responds to "true" values) solely. Yet, ground truth can also be regarded as the whole dataset which you split into the parts needed for model development, one larger part that the algorithms can be trained on and smaller parts for testing and validation.

The pragmatic positioning is what makes Johan speak of the *assumptions* you need to make, regardless of potential flaws or incompleteness. He has to assume that the labels that have been assigned to the images in the database by medical experts hold true, if he wants to use the labels for training in supervised or semi-supervised learning and for validating the performance of his creation. One particularity of the ground truths assembled for medical AI development is their enactment of how to measure and diagnose medical phenomena by the inclusion and exclusion of different types of data, and methodological choices, which steers how AI can be used to detect, predict or treat medical conditions.

In another room, Christian is going through painted segments of slices of a brain on his screen. Next to the images, he has a window open where he sees the lines of code. As he puts it, he is an expert in training algorithms, not in the anatomy of the human brain. He describes the laborious process of recruiting neurologists and having them sit and literally paint all areas of the brain upon each imaged slice, a process by which the sociomaterial aspects of ground truth assemblage becomes evident. The aim of this is to construct a dataset by which the algorithm can learn brain segmentation and to subsequently make it possible to evaluate whether the algorithm is identifying and demarcating the right area of the brain. This can, inevitably, have severe effects. To complete the painting of one brain could take a neurologist a whole day of work. It is a time consuming and expensive set-up, Christian complains, but he argues that it is worth it to achieve a ground truth dataset that is as trustworthy and accurate as possible.

These are two examples of how datasets with human medical expert labels work as ground truths for the AI expert. The assemblages of ground truths also show how they are the products of a co-constitutive shaping of truths, through data collection, generation and the medical expert's labelling, combined with the AI developers' reinforcement of it as truth claims, used by models and in validation by organisations.

The limits of expert-based ground truths

On the top floor of an AI research centre, we are sitting in a room talking about Aksel's visionary project to use natural language processing (NLP) on a wide range of data, from demographic and health registers, medical records and so forth, to identify risk factors for disease. For their computer models to be able to learn about characteristics that could be risk factors, they need longitudinal data that describes which individuals, with what characteristics, developed a certain condition. In that regard, they have to rely on the diagnosis labels assigned by clinicians, working as their ground truth for model training and evaluation.

> [W]e are working a lot with diagnosis codes, and even there, most doctors ... so there are thousands of diagnoses. And the doctors they use maybe one hundred of them … […] we know with the ground truths that some of the diagnoses are inaccurate … because they are created by humans.

The humanness of expert-based labelling that Aksel refers to is depicted as an aspect that makes ground truths more trustworthy, yet it is also something considered as making AI vulnerable to flawed judgements. The complexity of assembling data for AI development emerges in these instances. Those developing AI models have to make assumptions about whether data and labels are valid referents to the real world. Informants show the ways in which human labelling is principal, and whilst they point to the precarity of having to depend on human expertise in terms of for example potential uneven quality, insufficient documentation and expertise, it is what they have to work with to have a real-world comparison.

At a university, Lars works within biomedical technologies, developing algorithms that can interpret data (signals) and make health predictions, but also algorithms that have a more technical purpose, for example removing noise to improve algorithmic interpretation of heart signals and "get rid of diagnostic interferences". When in his office, we discuss where data come from to make these health predictions, and how expert judgements are particularly essential for medical ground truths, in comparison with many other domains where it is easier for the AI developer or laymen to annotate, or review annotations, of ground truth data. What the doctor says is what is treated as truth:

> It has long been like that, for some signal or some images or something that yes, the ground truth is what the doctor has said about this image, then whether it is a sufficiently detailed description or whether it was a rough sorting or something like that, but it is that. It's been a ground truth, but it's also the case that as long as it's the ground truth, the machine can't be better than what the doctor was then, and maybe they're not doing it 100% right ... And somewhere, so if you're talking to a computer engineer here, maybe they're looking for a better truth, that is, where you can kind of say yes, but then we want to know more.

The argument behind this reasoning is that, by sticking to expert judgements as ground truths, you cannot find what is not labelled. This sets a clear limitation when using them for knowledge discovery. There is an idea of AI as being able to surpass human abilities, but allowing this to its full extent could mean that expert-based ground truths would not be considered enough for training and evaluating AI. As Lars expresses it, they will no longer do:

> We have lived in a time where the ground truth has been that here we have, in our industry then, measured things and here we have someone who looked at it. But, but that … I think the machines will gain up on that. It won't do to train on, you have to train on something else which is better than it because otherwise you learn to do the same mistakes we had before.

Some AI developments are described as having no ground truth and to be acting in the absence of truth. If no labelled data are available, are too hard to access, or present too great a risk to collect, there are still different ways to approach developing medical AI for these phenomena. One option is to use unsupervised learning, in which the algorithm learns without guidance by pre-assigned labels. This is the case in some exploratory research lead by the hope of AI discovering something that the medical experts cannot see. Still, without a ground truth, the question is how to validate AI performance.

Aksel describes how they in his project try to "filter out noise" rather than deciding what is important for the task of the model. One reason for this is the aim of discovering new risk factors for disease and "all these hidden things". This leads Aksel to suggest that they operate without a "strictly defined" ground truth, as he says: "we are building a model without knowing the truth, just trying to get a good representation." But what is a *good* representation?

Generating truths

It is argued that also augmented, or even synthetic, data can form a ground truth by plausibly representing the statistical properties of a real-world phenomenon, without corresponding with actual real-world referents. In some regards, simulation studies are seen as having the perfect "known" truth as fully constructed data, made with the aim of its being a total representation without any potentially false negatives. To some degree, a synthesised ground truth introduces validity, especially technically, yet in other ways, it introduces new uncertainties. One informant argues that it is useful with synthetic data in some scenarios, to enlarge datasets or get more samples of specific subgroups. Still, to review all correlation matrices and output for all potential variables would be impossible, he argues, concluding: "I think in me there would always be a doubt that okay, maybe by data, the synthetic data reflects these and these variables in the real data really well. But I'm not sure how well it reflects the other ones."

Based on the informants' reasoning and practices, they seem to find inevitable limitations in having to rely upon expert-made ground truths, one being the limitation on knowledge discovery if models are based on, or evaluated against, what is already known about for example the risk factors or early signs of a disease. Yet, synthetic data are seen as inheriting this limitation, as Johan expresses it:

> [S]o the problem is with augmented or synthetic data, you're not getting anything that you don't already have, that makes sense. So, if there's one sort of anomaly in the brain that you just don't know about, you're not going to get that through data augmentation or through some sort of generated models or something, but on the other hand is really useful to just take what you have and make it more diverse.

As becomes visible from this quote, augmenting or synthesising data is not perceived as solving the limitations of expert-based ground truths. However, it could potentially serve other functions. In Aksel's NLP work, he sees synthetic data as something that would solve many issues. The datasets would be "proper" and big enough, possible to share and work on with any computer, but also offer a ground truth that is *realistic enough*:

> … cause in the end we are not trying to predict on a single person, we are trying to make a model that works on a larger population. So, there synthetic data works quite well since we don't … like this ground truth is muddy anyway, so with synthetic data we can generate something that is shareable and there might be some flaws but we don't really care cause on a bigger scale it is realistic, that is all we care about, right?

The statement that the ground truth is "muddy anyway" should not be regarded as a dismissal of the validity of the research and model development, but rather in line with what Jaton (2017) and Kang (2023) stress as the pragmatic perspective that developers have towards what they conceptualise as their ground truth. This means that it is not actually considered as an absolute factual truth but as a way of finding a workable truth basis. However, it does point to what several of the informants express or imply in their work, that not all ground truths are considered equally valid or valuable. Here the valuation practices of ground truthing for medical AI emerge.

Valuations of brittle truths

In the previous quote by Lars, he referred to engineers looking for a *better* truth. When ground truths for medical AI are assembled, the researchers perform valuations of what are the most accurate and trustworthy data sources and human experts. For example, one of the informants stress how they consider the pathologists' judgement as a more reliable data source than the radiologists' reading of images from the same case.

In some of the empirical encounters, the ground truth was perceived as a non-issue, solely taken as a given, as "the facts". In general, when it was generated by data collection conducted by the researchers themselves, as for example sensor data by devices, it was seen as unproblematic. The expert-based ground truths, however, often presented the potential issue of interobserver and intraobserver variability, meaning that the medical assessment (as in cancer versus not cancer) can vary between different observers or in repeated assessments by the same observer. The risk of flawed expert judgements, or experts with different levels of accuracy, is hard to control for. In their establishments of ground truths, the informants perform valuations of data by which some hospitals or expert groups are deemed as providing more reliable diagnoses or assessments than others, enabling the creation of models the developers perceive as more trustworthy.

In his work to make multimodal prediction models of breast cancer risk, Niclas describes how they value data sources against each other in the quest for establishing the best ground truth. As they rely on hospital data, he argues that their models only get as good as hospitals perform. He refers to studies showing that one hospital could be four times as accurate in their judgement in comparison to another facility, in extreme cases.

> So that's something that we're working a lot on trying to understand, which ones are the better hospitals? As it is often elite hospitals, like university hospitals, that provide the best quality, we try to select them because it is more likely that there is better ground truth with them, so to speak, and so we train the model on that and so on. But there is no really effective way to get around that [problem], so everything goes via the hospital quality that is available at the time, so it's always … it's a challenge.

One aspect subjected to valuations of ground truths is that of temporality, which is sometimes a challenge when assembling medical ground truths for AI. In the case of breast cancer prediction, when considering image data from mammography screening exams, Niclas and his team are not valuing the ground truth at the point of screening as highly as the outcome five years later. This is as the algorithm is supposed to detect early signs of cancer in the images, and hopefully even earlier than the radiologists are able to detect it. To know whether the algorithm missed something in the image, it has to be evaluated against a later ground truth. This is to some extent the case for all risk prediction and with regards to medicine and health, where many of the conditions of the human body that the algorithms are supposed to grapple with are not static entities but evolving biological processes. This suggests another issue with expert-based medical expert ground truths from for example medical records or one-point human annotation: they are one event and one judgement, fixed in time. Growing cancers that need to be detected as early as possible, and risks that need to be mitigated, encompass predictions with a long arch of time series, multiple events and complexities. In these cases, it also shows how AI developers sometimes have to deal with multiple ontologies, by which medical conditions are enacted in several different ways by expert judgement datapoints, multimodality, and yet is identified by other ontological limits which have been turned into numbers for analysis by computational models. For them, cancer can be a range of pixel values. Epistemic uncertainties emerge when AI experts master computational methods but not how to, by themselves, visually review for example medical images for disease detection.

Aside from valuations in terms of better or worse sources, the incompleteness of ground truths is by some depicted as a constant worry. One of these worries is the incomplete patient trajectory. In most cases, the ground truth is one snapshot of events or one time series, without all the parallel (documented or undocumented) timelines of events that have an impact on people's health. As Johan stated earlier, just because something is not in the data, it does not mean it was not discovered, as this is a matter of both proper documentation and access to all relevant data sources:

> And a lot of times you don't get ground truth at all, like for example with the brain anomalies. So, usually like the only ground truth you have is, you have an examination and then later on you have another examination. You can basically test that our hypothesis holds up so. That same patient, how did that patient develop? In reality, we only have these sort of data points that we get from whatever medical records we have.

Valuations of ground truths also take into account technical traits, such as size and representation of technical properties (for example images from different imaging equipment). There is however a factor or perceived brittleness or robustness of documentation, absences of (correct) labels, but also the multiple ontologies of medical classifications. In other words, ground truths could be *leaky*. The concept of data leakage is used in ML to describe when ground truth data unintentionally leaks into the training set so that the algorithm is trained on the exact same data on which it is later evaluated, a statistical *faux pas*. Yet, the ground truths could be argued to be also leaky in terms of omitting accurate representations. Still, the risk of this posing a problem is perceived as different for different diagnoses. As one of the informants gave as an example, a heart attack might potentially be hard to miss to register in the medical records, but if someone has unmedicated type 2 diabetes it is much more likely to go undocumented. What the brittleness of truths put into question is also to what extent the truth of one (dataset) can be the truth of many?

Can the truth be generalised?

I am sitting next to Hanna in her office. We are looking through a set of images, and their corresponding data annotations, coming from the local hospital with which her project collaborates. The task for the model she is developing is to detect cancerous tumours in the images, guided by expert labelling. As she has worked in several projects concerning medical images and disease detection models, she says that it is not really ideal only having data from one hospital for training and evaluation. The hospital in this case uses one specific imaging equipment, and images from different vendors could have somewhat different technical qualities. The risk is that the algorithm will learn to separate diseased from healthy samples only in images from one particular vendor. Yet, there are possible workarounds to decrease that risk. To make models that can work on data from all hospitals or manufacturers, the informants consider it necessary to augment data, add noise to, or synthesise data in order to build a referential dataset seen as a better representation of the phenomena modelled. The initial ground truth dataset is not perceived as able to speak the truth for all cases. The truth as it is already known has to be expanded or else it will be hard to argue for its clinical validity.

When Lars discusses the disease prediction models that he works on, he emphasises that in the end, they are statistical tools that can work statistically well at group level, but when you come down to the level of one individual, it is "just one data point" and the prediction might not hold true anymore. He says "and then somewhere it's the journey of life", describing it as documented in quality registers or medical records.

> … that now I got this thing, and now I got this and now I have this medicine like that, but it's also very complex because this data, it won't be ... It won't be such homogeneous groups because everyone gets different medications and then everyone stresses differently and eats different things and takes different risks and thus, so it's very, very difficult to get clean databases here. You probably have to admit that.

This complexity of life raises issues in predicting health outcomes, as in Lars's work. They need to make algorithms and models that can work on data from all hospitals or manufacturer and on all unseen patients, hence to build a truth that is as generalisable as possible. This speaks both for the situatedness of data, and the work of constructing ground truths that are perceived as valid, representative and capable of enabling the learning of AI tools across contexts. As such, it shows the complexity and value negotiations in the truth practices of AI experts.

Discussion

In this study, the truth practices and the valuations of ground truths emerge as performative, productive acts in medical AI development, by the reasoning and practices of the AI experts. When ground truths are brought together, this article shows how the AI experts have to rely on expert-based judgements to enable AI development, even when these require much work or when doubting its completeness. The humanness of expert-based ground truths emerges as somewhat of a double-edged sword. It is seen as an aspect that makes technical solutions more trustworthy, by not being solely computational but based on the knowledge of medical experts. Yet, this is also perceived as setting technical and epistemic limitations on development when aiming for AI to surpass human capabilities. As ground truths are already, in some aspects, seen as "muddy", there is a hope among some informants that synthetic data can work equally well or better than more traditional sources of ground truths, while introducing uncertainties of how deriving models should be validated. Yet, synthetic data are not perceived as solving the limitations on knowledge discovery, as these are seen as being inherited when data is synthesised.

In this study, the styles of valuation (Lee and Helgesson 2019) of ground truths are performed with regards to at least four identified aspects of quality. First, *humanness*, which is instilling trust by its closeness to trained, accepted medical judgement and clinical practice while also being devalued in relation to the other identified aspects. Second, *quality of human medical expertise*, as elite data sources are believed to encompass the best medical knowledge and hence provide the most accurate labels. Third, *temporality*, as enabling prediction capabilities in models, but also in, fourth, *technical qualities*, much due to their ability to support generalisation of ground truths across cases, contexts and equipment. These aspects are in line with other research about how certain organisations and characteristics are indicative of trusted truths (Shapin 1994) or as representing objective factual knowledge (Daston and Gailson 2007), now in the form of elite hospitals, the quality of human judgement and technical specifications.

The importance ascribed to ground truths as *the facts* can furthermore be regarded as part of what Campolo and Schwerzmann (2023) call "artificial naturalism", in which the example-based authority is established between data and (in this case, medical) norms. To a large extent, it is the ground truth that enacts (Mol 2002) the medical condition in the development of medical AI, by limiting what counts as the reality of the disease. Moreover, ground truthing can be understood as articulation work (Fujimura 1987), pulling the production of datasets, training and evaluation strategies together to make medical problems *doable* for AI development. Human expert labels are seen as able to instil trust and provide measurement for comparison, and by this reasoning, ground truth also works as a scientific credibility device in epistemic cultures (Knorr-Cetina 1999) of AI development in medicine. If the ground truth is not seen as good enough, it is harder to get the application clinically implemented and accepted in the clinic. However, a "good" ground truth from a development perspective is not necessariy what a good ground truth looks like for hospital managers or other stakeholders (Lebovitz, Levina & Lifshitz-Assaf 2021). This is as a fully "known" ground truth, such as by computational simulation, offer perhaps the greatest technical possibilities (Siebert et al. 2020), but from other perspectives, what is most highly valued by some of the informants and their collaborators are data deriving as closely as possible from medical expertise and the clinical floor. In this regard, different valuations and truth practices make visible the tension between the trust in expert judgements and the value of increased generalisability and technological advancement, especially considering the goal of surpassing human experts' performance. Among informants, there are differences in terms of their valuation of synthetic data's ability or potential to function as a ground truth for medical AI. The idea that AI experts are looking for a truth that is better than expert-based labels, to be able to surpass them, corresponds with Henriksen and Bechmann's (2020:804) findings that "labels born out of the practical assessment of patients are not regarded as usable targets for model training when the goal is to leverage the existing classification logic that healthcare practitioners use".

Ground truths play a prominent role in the making of AI, but by that also in the computational making of the medical phenomenon AI is aiming to grasp. By Mol's (2002) conceptualisation, the ground truthing enacts the disease in a certain way in how data and labels are assembling ontologies into a model of the medical phenomena. Through a focus on ground truths, we can further the discussion of what ontologies are, and are not, included in the medical ground truths for AI development, and what the impact of these inclusions and exclusions could be. The focus on evidence-based medicine and clinical data as a gold standard for reviewing medical knowledge (Timmermans and Berg 2003) moreover suggests a need for more discussion about what expertise is included in medical AI models (Lebovitz, Levina & Lifshitz-Assaf 2021). In Jaton's (2023) case of the construction of a medical ground truth, it was found to constitute a contested reference set due to the lack of a standardised basis for measurements, yet regardless of that, it continued to be used since it managed to reach certain quality standards and robustness for the making of specific AI technologies. This also suggests that the valuation of ground truth qualities encompasses more than alignment with standard expertise.

The reflexivity of the informants, with regards to the ability to enable generalisation, suggests a need to attend to notions about data inheritance (Green and Svendsen 2021), which suggests that historical medical data is representative of future populations and individuals previously "unseen" by the algorithm. This data dependency instils a normative order (Campolo and Schwerzmann 2023), still there are other normative orders that are posed by the algorithms but also normativities enacted in ground truth practices (Lee and Björklund Larsen 2019). Moreover, ground truthing plays a great part in the worldling capacity of algorithms, while never escaping the instability of medical categories. In that sense, the perceived brittleness of ground truths shatter the perception of medical standards and classifications, as well as medical AI tools, as stable neutral entities (Bowker and Star 1999). By this understanding, we need to acknowledge ground truths as sociotechnical entanglements and as inevitably limited in some sense. As Jaton argues:

> These ground-truthing processes engage people, efforts, and resources. Yet, in principle, the products of these processes (i.e. ground-truth datasets) remain limited, arbitrary, and socio-culturally oriented. Consequently, algorithms—as devices that approximate relationships among ground-truth datasets—*are* also limited, arbitrary and socio-culturally oriented. (Jaton 2023:803)

This resonates with the notion that a ground truth never should be taken as an absolute quantification or datafication of a phenomenon residing in the "real-world"; as emphasised by Kang, it implies "not necessarily a representation of 'reality,' but rather the translatability of a problem of interest, which allows it to be legible to and expressed in the language of mathematics" (Kang 2023:3–4). Or as suggested by Fine (2006), ground truthing is an organisational practice of crafted measures and verification, including the production of predictive claims and strategies for measuring them.

Truth practices also seem to work as a way for AI experts to grapple with the medical condition that their model's tasks involve. Epistemic differences can impact the creation of ground truth schemas (Zając et al. 2023), a finding also echoed by this study. The informants are experts in training algorithms, and not in medical assessment of sonograms or brain segmentations. In the border between medical and AI knowledge, informants speak of expertise at large in terms of divisions into different domains (Ribes et al. 2019). AI experts mostly can be regarded as what D'Ignazio and Klein (2020) call “strangers in the dataset”. They have to make sense of, or arrange for, the labelling of medical conditions to teach models to grasp structures and learn to generalise, while they themselves are usually “outsiders” to the clinical practice. Instances of critical reflexivity regarding the conditioning of their ground truths show how AI experts have to work pragmatically from a place of uncertainty to find ways to best represent the medical phenomenon and make models that can be utilised across contexts.

These truth practices can to some extent be construed as foremost reinventing truths and elevating already prevalent logics in healthcare through the modelling of AI (Henriksen and Bechmann 2020). Yet, I would suggest that several informants of this study express reflexive negotiations in relation to their ground truths. To some extent, the AI experts are held hostage by external demands and conditions, in the same regard as Fine’s forecasters, by having to rely on medical expert judgements, hospitals, data infrastructures, medical classifications and equipment, while at the same time having to argue for the validity and accuracy of their own models. Fine (2006) argues that the forecasters of his study are almost hostage to mechanical claims beyond their control, and part of organisational practices and ideals that can also represent tensions:

> Facts, seemingly objective, become claims that are locally produced through organizational choice, but by being seemingly objective, they serve to create a hegemonic zone, preventing questioning. It is surely unfair to suggest that all that exists is a patina of truth, but adherence to the mechanical claims constitute a bureaucratic strategy, evident when the otherwise taken-for-granted mechanically produced truth is challenged by lived experience. (Fine 2006:7)

In a similar manner, and with regards to AI, Jaton argues that we get the algorithms of our ground truths and the “*ground truths of our organizations and metrological equipment”* (Jaton 2023:803). I would suggest that ground truths encompass even more, involving an array of people, organisations, equipment, standards and expertise.

Conclusions

This study shows how ground truths are productive epistemic enactments necessary for the development of medical AI, and something that the developers have to carefully navigate. Bates (2018) stresses data friction as something to foster rather than overcome. In this study, the making, reasoning about, and valuation of ground truth data shows a high reflexivity and awareness amongst the informants about the limits and pragmatic positioning they have to adopt towards their ground truths. These articulations are in need of more attention and consideration, to better understand the impact of new technologies. As argued by Lee and Helgesson (2019:680) “if we fail to acknowledge the divergent valuations of technology in situated settings, we risk becoming blind to actors’ struggles to work with automation and algorithms”.

In the empirical instances of dealing with ground truth assemblages, I show how they are shaped by matters such as medical classification, information and data infrastructures in terms of for example medical registers; moreover, how they are enabled and constrained by organisational as well as technological and scientific ideals. This suggests that we need to pay attention to ground truth assemblages as practices, narratives, artefacts and devices in knowledge production. Through the reasoning and practices of those developing AI, this article shows how datasets are assembled from different sources, and produced, augmented and synthesised. This addresses the role of human expertise, perceived strengths and limits of expert-based annotations, (in)stability of medical classifications, and sometimes, data friction.

These results can inform how developer practices fit together with grand calls for action, such as that medical AI should be fair, trustworthy or transparent. To be able to distinguish the possibilities of making such matters doable in practice, we need more knowledge on the processes going into the making of AI and the perspectives of those that are developing the technologies. What is perceived as a *good* ground truth? And what challenges or tensions are there in the practices in pursuit of that truth? We can continue the troubling of algorithms as stable, transparency as binary, and fairness as residing only in the algorithm (Lee 2021), by paying greater attention to assemblages of ground truths, and how they are valued and put to work in the development of AI.

References

Amoore, L. (2013) *The politics of possibility: Risk and security beyond probability.* Duke University Press.

Amoore, L. (2020) *Cloud ethics. Algorithms and the attributes of ourselves and others.* Duke University Press.

Avlona, N.-R. & I. Shklovski (2024) "Torquing patients into data: Enactments of care about, for and through medical data in algorithmic systems", *Information, Communication & Society* 27 (4):735–757.

Bates, J. (2018) "The politics of data friction", *Journal of Documentation* 74 (2):412–429.

Bertelsen, P.S., C. Bossen, C. Knudsen & A. Pedersen (2024) "Data work and practices in healthcare: A scoping review", *International Journal of Medical Informatics* 184: 105348.

Bowker, G.C. & S.L. Star (1999) *Sorting things out : classification and its consequences.* MIT Press.

Campolo, A. & K. Schwerzmann (2023) "From rules to examples: Machine learning's type of authority", *Big Data & Society* 10 (2): 20539517231188725.

Collins, H.M. (1990) *Artificial experts : Social knowledge and intelligent machines.* MIT Press.

Crawford, K. (2021) *Atlas of AI : power, politics, and the planetary costs of artificial intelligence.* Yale University Press.

Daston, L. & P. Gailson (2007) *Objectivity.* New York: Zone Books.

Denton, E., A. Hanna, R. Amironesei, A. Smart & H.Nicole (2021) "On the genealogy of machine learning datasets: A critical history of ImageNet", *Big Data & Society* 8 (2):20539517211035955.

D'Ignazio, C. & L.F. Klein (2020) *Data feminism.* MIT Press.

Fine, G.A. (2006) "Ground truth: Verification games in operational meteorology", *Journal of Contemporary Ethnography* 35 (1):3–23.

Forsythe, D.E. (2001) *Studying those who study us: An anthropologist in the world of artificial intelligence.* Stanford University Press.

Fujimura, J.H. (1987) "Constructing `do-able' problems in cancer research: Articulating alignment", *Social Studies of Science* 17 (2):257–293.

Gitelman, L. (2013) Raw data is an oxymoron. *Infrastructures Series.* MIT Press. Cambridge, 1 Online Resource (203 Seiten).

Green, S. & M.N. Svendsen (2021) "Digital phenotyping and data inheritance", *Big Data & Society* 8 (2): 20539517211036799.

Grundmann, R. (2017) "The problem of expertise in knowledge societies", *Minerva* 55 (1):25–48.

Henriksen, A. & A. Bechmann (2020) "Building truths in AI: Making predictive algorithms doable in healthcare", *Information, Communication & Society* 23 (6):802–816.

Hoeyer, K. (2023) *Data paradoxes the politics of intensified data sourcing in contemporary healthcare.* MIT Press.

Jaton, F. (2017) "We get the algorithms of our ground truths: Designing referential databases in digital image processing", *Social Studies of Science* 47 6):811–840.

Jaton, F. (2021) *The Constitution of Algorithms : Ground-truthing, programming, formulating.* MIT Press.

Jaton, F. (2023) "Groundwork for AI: Enforcing a benchmark for neoantigen prediction in personalized cancer immunotherapy", *Social Studies of Science* 53 (5):787–810.

Kang, E.B. (2023) "Ground truth tracings (GTT): On the epistemic limits of machine learning", *Big Data & Society* 10 (1): 20539517221146122.

Kitchin, R. (2017) "Thinking critically about and researching algorithms", *Information, Communication & Society* 20 (1):14–29.

Kitchin, R. & T. Lauriault (2018) "Toward critical data studies: Charting and unpacking data assemblages and their work", 3–20 in J. Thatcher, J. Eckert & A. Shears (Eds) *Thinking big data in geography: New regimes, new research.* University of Nebraska Press.

Knorr-Cetina, K. (1999) *Epistemic cultures : How the sciences make knowledge.* Harvard University Press.

Latour, B. (1987) *Science in action : How to follow scientists and engineers through society.* Harvard University Press.

Latour, B. (2005) *Reassembling the social: An introduction to actor-network-theory.* Oxford University Press.

Lebovitz, S., N. Levina & H. Lifshitz-Assaf (2021) "Is AI ground truth really true? The dangers of traning and evaluating AI tools based on experts' know-what", *MIS Quarterly* 45 (3):1501–1525.

Lee, F. (2021) "Enacting the pandemic: Analyzing agency, opacity, and power in algorithmic assemblages", *Science & Technology Studies* 34 (1):65–90.

Lee, F. & L. Björklund Larsen (2019) "How should we theorize algorithms? Five ideal types in analyzing algorithmic normativities", *Big Data & Society* 6 (2):2053951719867349.

Lee, F. & C.-F. Helgesson (2019) "Styles of valuation: Algorithms and agency in high-throughput bioscience", *Science, Technology, & Human Values* 45 (4):659–685.

Loukissas, Y.A. (2019) *All data are local: Thinking critically in a data-driven society.* MIT Press.

Mackenzie, A. (2017) *Machine learners : Archaeology of a data practice.* MIT Press.

Mackenzie, A., C. Waterton, R. Ellis, E. Frow, R. McNally, L. Busch & B. Wynne (2013) "Classifying, constructing, and identifying life: Standards as Ttransformations of 'The Biological'", *Science, Technology, & Human Values* 38 (5):701–722.

MacKenzie, D.A. (1990) *Inventing accuracy : A historical sociology of nuclear missile guidance.* MIT Press.

Meuser, M. & U. Nagel (2009) "The expert interview and changes in knowledge production", 17–42 in A. Bogner, B. Littig & W. Menz (Eds) *Interviewing experts.* London: Palgrave Macmillan.

Mol, A. (1999) "Ontological Politics: A Word and Some Questions", The Sociological Review 47(1):74–89

Mol, A. (2002) *The body multiple : Ontology in medical practice.* Duke University Press.

Pinch, T.J. & W.E. Bijker (1984) "The social construction+ of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other", *Social Studies of Science* 14 (3):399–441.

Pink, S. & J. Morgan (2013) "Short-term ethnography: Intense routes to knowing", *Symbolic Interaction* 36 (3):351–361.

Rajpurkar, P., E. Chen, O. Banerjee & E. Topol (2022) "AI in health and medicine", *Nature medicine* 28 (1):31–38.

Ribes, D., A. Hoffman, S. Slota & G. Bowker (2019) "The logic of domains", *Social Studies of Science* 49 (3):281–309.

Ryan, G.W. & H.R. Bernard (2003) "Techniques to identify themes", *Field Methods* 15 (1):85–109.

Seaver, N. (2017) "Algorithms as culture: Some tactics for the ethnography of algorithmic systems", *Big Data & Society* 4(2): 2053951717738104.

Seaver, N. (2018) "What should an anthropology of algorithms do?" *Cultural Anthropology* 33 (3):375–385.

Shapin, S. (1994) *A social history of truth : Civility and science in seventeenth-century England.* University of Chicago Press.

Shapin, S. & S. Schaffer (1985) *Leviathan and the air-pump : Hobbes, Boyle, and the experimental life : including a translation of Thomas Hobbes, Dialogus physicus de natura aeris by Simon Schaffer.* Princeton University Press.

Siebert, J., L. Joeckel, J. Heidrich, K. Nakamichi, K. Ohashi, I. Namba, R. Yamamoto & M. Aoyama (2020) "Towards Guidelines for Assessing Qualities of Machine Learning Systems", 17–31 in M. Shepperd, F. Brito e Abreu, A. Rodrigues da Silva & R. Pérez-Castillo, R. (Eds) *Quality of information and communications technology.* Cham: Springer.

Thylstrup, N.B. (2022) "The ethics and politics of data sets in the age of machine learning: Deleting traces and encountering remains", *Media, Culture & Society* 44 (4):655–671.

Timmermans, S. & M. Berg (1997) "Standardization in action: Achieving local universality through medical protocols", *Social Studies of Science* 27 (2):273–305.

Timmermans, S. & M. Berg (2003) *The gold standard. The challenge of evidence-based medicine.* Temple University Press.

Weinkle, J. & R. Pielke (2016) "The truthiness about hurricane catastrophe models", *Science, Technology, & Human Values* 42 (4):547–576.

Zajac, H., N. Avlona, T. Andersen, F. Kensing & I. Shklovski (2023) "Ground truth or dare: Factors affecting the creation of medical datasets for training AI". *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, August 8–10, 2023, Palais des congrès de Montréal, Montreal, QC, Canada, pp. 351–362. Association for Computing Machinery.

Author presentation

Charlotte Högberg is a PhD Student in Technology and Society at Lund University, Sweden.
Contact: charlotte.hogberg@lth.lu.se