AI competitions as infrastructures of power in medical imaging

Luitse, D.; Blanke, T.; Poell, T.

DOI
10.1080/1369118X.2024.2334393
Publication date
2025
Document Version
Final published version
Published in
Information, Communication & Society

INFORMATION, COMMUNICATION & SOCIETY
2025, VOL. 28, NO. 10, 1735–1756
https://doi.org/10.1080/1369118X.2024.2334393

Routledge
Taylor & Francis Group

AI competitions as infrastructures of power in medical imaging

Dieuwertje Luitse, Tobias Blanke and Thomas Poell

Department of Media Studies, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

This article examines how platform-based AI competitions structure power relations in medical imaging research. It focuses on two leading platforms, Kaggle and Grand Challenge, which provide organisational as well as infrastructural support to run AI competitions. In dialogue with critical AI and platform studies research, we investigate how such competitions are organised – under which infrastructural conditions and by whom – and how this shapes processes of model production and evaluation. To address these concerns, we have collected data from 118 medical image AI competitions on Kaggle and Grand Challenge, organised between January 2017 and May 2022. In addition, a variety of platform boundary resources – platform documentation, competition descriptions, dataset descriptions, and competition leaderboards – have been gathered. The analysis of these materials shows, first, that *platforms* direct the AI development process by requiring substantial financial resources, defining which institutions can host a competition and under which conditions. Second, *competition organisers* define dataset diversity and the generalisability of models. As most datasets are constructed with data from hospitals in North America, Western Europe and China, the application of models to different geographical contexts is potentially limited. Finally, *competition participants* influence model development through the institutional, demographic, and disciplinary contexts in which they operate. Overall, the examination demonstrates the importance of critically interrogating the *entire* medical AI research pipeline, including the definition of research problems, the construction of datasets as well as model production and evaluation.

ARTICLE HISTORY
Received 24 February 2023
Accepted 26 February 2024

KEYWORDS
Artificial intelligence; AI competitions; medical imaging; research infrastructures; power

Introduction

Medical imaging is a vital research field for artificial intelligence (AI[1]), particularly since CTs or MRIs provide vast amounts of valuable health data (Prevedello et al., 2019). To stimulate development and facilitate the scientific evaluation of new models,[2] AI-based

CONTACT Dieuwertje Luitse d.luitse@uva.nl Department of Media Studies, University of Amsterdam, Turfdraagsterpad 9, Amsterdam 1012 XT, The Netherlands @dluitse

Supplemental data for this article can be accessed online at https://doi.org/10.1080/1369118X.2024.2334393

research in medical imaging is traditionally organised in a competitive manner. These competitions are typically platform-based and organised by consortia that issue dedicated research problems, focusing on specific tasks, such as breast cancer detection (TIGER–2022) or brain tumour segmentation (RSNAMICCAI–2021). Addressing such tasks, participants compete by developing and evaluating their machine-learning models with provided training and test datasets.[3] Teams are ranked on a leaderboard by the evaluation score of their (best) submission. These medical imaging AI competitions increasingly attract both public and commercial research organisations, which appears to impact this research field and the distribution of power within it (Maier-Hein et al., 2018). This paper examines how such platform-based competitions specifically structure power relations in medical image AI research.

Performance competitions have a long-standing tradition in machine-learning subfields like Computer Vision (CV) or Natural Language Processing (NLP). Competitions such as The Text REtrieval Conference (1992-ongoing) (e.g., Harman, 1993; Sparck Jones, 1995) have proven to be important for the development of new research approaches. Perhaps the most famous example of competitions playing a vital role in AI research is the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). Associated with the ImageNet dataset (Deng et al., 2009), this competition for image classification and object localization and detection ran between 2010 and 2017. In 2012, a team from the University of Toronto won the Challenge with a deep neural network-driven model that significantly outperformed all competing teams using graphics processing units (GPUs) (Krizhevsky et al., 2012). This moment is often marked as the turning point in AI research, shifting the field towards the currently dominant deep-learning paradigm. This fundamental shift would not have been possible without competitions, as it is heavily data-driven and requires large amounts of computing power (Dotan & Milli, 2020). It has, subsequently, accelerated the uptake of competitions and changed them towards ever larger datasets and new types of data.

The last decade has seen the emergence of specialised competition platforms such as Kaggle and Grand Challenge—two leading platforms in AI-based medical imaging. They provide organisational as well as infrastructural support (e.g., data storage, compute, Application Programming Interfaces (APIs), and Jupyter Notebooks) to run competitions initiated by public or commercial research organisations. Winning teams can expect prizes of over $1 million (Maier-Hein et al., 2018), usually provided by the organising institution or through competition sponsorships. While sharing some similarities, these platforms also differ in significant ways. Kaggle is a large commercial platform, which was acquired by Google in 2017 (Li, 2017). It regularly hosts competitions across various AI sub-domains, including medical image AI, and had over 10 million registered users in 2022. Grand Challenge is a much smaller specialised research community-driven platform with approximately 70,000 users. It was initiated by the Diagnostic Image Analysis Group at the Radboud University Medical Centre (RUMC) and is dedicated to competitions in medical imaging.

For medical image analysis, platform-hosted competitions constitute an important part of their overall research infrastructure, steering global research and development. We understand research infrastructures as 'sociotechnical systems that are centrally designed and controlled' (Plantin et al., 2018, p. 295), typically for developing new technologies, forms of knowledge, and scientific practice (Bowker & Star, 2000; Larkin, 2013).

In medical imaging, these infrastructures have been celebrated as 'accelerators for medical imaging innovations' (Bulten et al., 2022, p. 154), which enable research groups to collect and deploy valuable health data, come together around specific research problems, and compete for the best solutions. Models that perform well in these competitions are more likely to be adapted for clinical practice and can be seen to further steer the individual careers of winning participants through the publication of competition research papers (Reinke et al., 2021).

Relatively little is known, however, about how their platform-based infrastructures set the conditions for medical AI research, including development and evaluation. This is surprising given the growing concerns over the power dynamics in (medical) AI research in general and the development of machine-learning applications that increasingly amplify existing inequalities towards historically disadvantaged communities (Benjamin, 2019; Obermeyer et al., 2019; Thomas, 2021). So far, critical work has primarily focused on power imbalances in the construction of single research components such as medical imaging datasets (e.g., Oakden-Rayner, 2019; Rostamzadeh et al., 2022). While such elements are significant, *platform-hosted medical imaging competitions operate across the entire medical AI research pipeline*. They structure the definition of research problems, the construction of datasets as well as model production and evaluation by introducing tasks and metrics. It is therefore critical to understand the organisation of these competitions in their entirety, and the position key actors like platforms, organising institutions, and competition participants occupy within them.

This paper provides new insights through a critical study of medical imaging competitions on Kaggle and Grand Challenge between 2017 and 2022. Combining a critical AI and platform studies research approach, we investigate how competitions on Kaggle and Grand Challenge are organised – under which infrastructural conditions and by whom – and how this shapes processes of model production and evaluation. Pursuing this inquiry, we first situate our contribution within the critical literature on power differentials of power in AI research and development and on the infrastructural power of platforms. Second, we discuss our methodology and the materials we analysed to trace the organisation of platform-based medical imaging AI competitions. The remainder of the paper analyses the organisation of AI competitions, their power dynamics, and the implications for medical imaging.

Data, platform infrastructures, and power in AI research

Critical scholarship on power dynamics in AI research and development highlights the limits of bias- and ethics-oriented approaches and calls for more comprehensive analyses of *social practices* and *power relations* (e.g., Barabas et al., 2020; D'Ignazio & Klein, 2020). Dataset development has been of particular interest in this literature, showing how associated practices are influenced by global asymmetries and power dynamics (e.g., Miceli et al., 2022; Paullada et al., 2021). Following 'critical dataset studies,' (Thylstrup, 2022) datasets are considered key infrastructures, forming the underlying conditions upon which machine-learning research and development operates (Denton et al., 2021). They define which research problems are addressed, what is and what is not considered to be progress, providing the grounds upon which novel algorithmic techniques

are being developed, tested, and evaluated. A combination of a dataset and a corresponding evaluation metric summarise a machine-learning system's performance over a specific task as a single number or score (Raji et al., 2021) and set a so-called 'benchmark' for the competition-style of research in AI (Denton, 2022).

The impact of benchmarks on the organization of AI research has also been recognised by other authors, considering practices of benchmarking beyond dataset construction (e.g., Hutchinson et al., 2022; Raji et al., 2021). Grounded in a longer history of computing practices in computing such as the Common Task Framework (Donoho, 2017), benchmarking has become the dominant mode of machine-learning model evaluation over the past decade (Denton, 2022). According to Raji et al. (2021, p. 8), this has led to the development of entire infrastructures for 'competitive testing' allowing participants to continuously optimise their models, proposing small metric variations to improve performance. In the division of labour within competitions, competition organisers not only provide the necessary datasets, but also define the metrics to which models are being validated and tested. In other words, competition organisers are not only responsible for the development of the datasets, but they also take a central position in defining the standards by which models are produced and evaluated (Denton, 2022).

Building on these observations, this paper more closely investigates the specific power relations in competitions organised on two major competition platforms: Kaggle and Grand Challenge. Following work in platform and infrastructure studies (e.g., Plantin et al., 2018; Poell et al., 2021), these platforms can be understood *as large-scale computational infrastructures integrating a multitude of heterogeneous components, including data storage, compute power, APIs, and software tools that facilitate, aggregate, and govern interactions between complementors* (competition organisers and participants). Through their computational resources both Grand Challenge and Kaggle control infrastructural access, while preformatting or standardizing the activities for using their infrastructure.

We specifically draw on and contribute to the recent scholarship in platform studies, which examines how platforms exercise what can be called *infrastructural power* (e.g., Aradau & Blanke, 2022; Poell et al., 2021). In fact, following Foucault (1977, 1982), we can say that platforms as infrastructures do not simply possess a concentrated, universal hierarchical, one-directional form of power. We understand power through infrastructure as a set of dispersed relational forms, exercised 'as a way in which certain actions may structure the field of other possible actions' (Foucault, 1982, p. 791). It is an immanent multiplicity of relations that allow infrastructures to guide 'the possibility of conduct and putting in order the possible outcome' (Foucault, 1982, p. 789), thereby structuring its users' (subjects') horizon of possibilities for knowledge production.

To understand the 'microphysics' of infrastructural power we need to examine how infrastructures set standards and lead to standardization. Following Foucault (1977), we understand a microphysics of infrastructural power here as the network of mechanisms and micro-relations of power that materialises through infrastructure as a strategic assemblage of smaller elements to structure conduct of others and order possible outcomes. According to Bowker and Star (2000, pp. 14–15), standards can be understood as 'any set of agreed-upon rules for the production of (textual or material) objects' that span multiple communities of practice or sites of activity for knowledge production. Standards are enforced and deployed through infrastructures (e.g., platforms) and have consequences for how they can be used: they enable and condition certain types of work

and action or control ways of working and acting over others and may even operate as barriers for particular users or groups (Bowker & Star, 2000; Star, 1999). In so doing, infrastructural power is operated through standardization as it allows for 'screening out' (un)limited diversity (Star & Lampland, 2009) and instead set the conditions, structure, and foreclose possibilities for research (e.g., in medical imaging), how it is constrained, preserved, and compared.

Platforms like Kaggle and Grand Challenge set such technical standards, terms and conditions, and criteria for participation through a variety of so-called *boundary resources*, which include APIs, software development kits, developer terms, data, and software tools (Nieborg & Helmond, 2019). These boundary resources structure a network of power relations by enabling third parties such as competition organisers to gain attention for their research projects and crowdsource new approaches through platform infrastructures. At the same time, they allow these platforms to govern what participating actors *can* do and *how* they can do it within the competition environment, thereby possibly foreclosing other AI research directions and practices. From this perspective, we examine how Kaggle and Grand Challenge operate infrastructural boundary resources to provide new opportunities for research, while foreclosing others.

Well-funded actors – in terms of technical expertise, access to data, financial resources, and institutional connections – can play a central role in platform ecosystems (Poell et al., 2021). Building on platform boundary resources and infrastructures, these actors can define and marshal the participation of other actors in the ecosystem. A prominent example of such actors are the AI research divisions of large tech corporations such as Alphabet and Amazon, which have heavily invested in computing infrastructures (Luitse & Denkena, 2021). Also conceptualised as an emerging 'compute divide' (Ahmed & Wahed, 2020), this creates significant advantages in AI research and development for these corporations and a small set of elite universities such as Stanford or Carnegie Mellon, allowing these actors to set the standards in the field (Whittaker, 2021). Following this line of research, the present inquiry examines how AI-competition organisers not only provide datasets but also define the standards on which machine-learning models can be produced and evaluated. Hence, the paper critically examines both the infrastructural power of competition platforms and how key actors build on these research infrastructures to play a central role in the development of AI in medical imaging.

Studying competitions on Kaggle and Grand Challenge

Platform-hosted medical image AI competitions are complex entities that come with methodological challenges. First, the computational infrastructures of these competitions, like other platform infrastructures, are largely obfuscated, complicating empirical research. Second, competitions constantly evolve, and competition leaderboards are subject to change. Therefore, we only considered finished competitions for further analysis.

To overcome these challenges and understand how power dynamics on Kaggle and Grand Challenge enable and shape AI production and evaluation in imaging, we have developed a digital methods' approach that draws from data on platforms, competition organisers and participants from 118 medical image AI competitions on Kaggle and Grand Challenge, organised between January 2017 and May 2022. 15 competitions were organised on Kaggle, and 105 on Grand Challenge. In addition, for the two

platforms, we analysed a variety of boundary resources: (1) Kaggle and Grand Challenge platform documentation, including competition documentation, APIs, Notebooks and compute infrastructure; (2) competition descriptions, organisation, rules and evaluation pages; (3) dataset description pages; (4) competition leaderboards.

Appendices I and II provide a full overview of the collected resources. The materials in Appendix II are organised per competition (Title) and sorted by hosting platform and timeframe to facilitate the analysis of competitions' organisation within the given period. For both Kaggle and Grand Challenge, data was systematically retrieved over the examined period. Because of the large numbers of competition participants, Grand Challenge leaderboard data for 11,847 participants was manually parsed. Kaggle data on leaderboards was computationally retrieved using the Kaggle API. Yet, due to platform-specific API constraints, the dataset of 3881 participants was sampled to (a) data from at least ten participants per competition; (b) data of participants that appear on more than one leaderboard (2763), as these active participants can be considered central to the development process; (c) randomly collected data of 1000 other participants to ensure diversity.

We have pursued the analysis of this data along three lines. First, we examined Kaggle and Grand Challenge's documentation to identify and compare how these *platforms* organise competitions. Exploring how Kaggle and Grand Challenge structure the way institutions can set up a competition, provides insight in how these platforms standardise and shape this process. Here, we also looked at how these platforms regulate several infrastructural components, ranging from data storage to model training and evaluation. On this infrastructural level, the analysis paid specific attention to the platforms' documentation on the computational infrastructures that are important for the training and evaluations of machine-learning models (Whittaker, 2021) as well as Kaggle and Grand Challenge's documentation of competition participation and submissions.

Drawing on the set of collected competition descriptions, organisation, and rules pages, the second level of inquiry concerned an investigation into the institutional backgrounds of *competition organisers*, the medical imaging datasets they provide, and the ways in which these actors define research problems. In addition, we considered the evaluation pages for all observed medical image AI competitions, to identify the metrics of model evaluation as defined by the organisers. This allowed us to gain a deeper understanding of the ways in which organising parties define the conditions for – and thereby direct – machine-learning model production by competition participants, as well as the evaluation of submitted models.

Finally, we further substantiated the first two levels of inquiry by using the collected Kaggle and Grand Challenge leaderboard data to map the distribution of *competition participants* in terms of geographical location and institutional background. This mapping provided insights in where well-performing AI models have been produced, as well as in the types of disciplinary backgrounds they build upon such as radiology, computer science, or biomedical engineering.

The infrastructural power of Kaggle and Grand Challenge

From a platform perspective, we find that both Kaggle and Grand Challenge, each in particular ways, structure the organisation of medical image AI competitions at multiple stages. First, the organisation of a competition for medical image analysis is conditioned

by the required financial resources. This is especially the case for Kaggle competitions. As a commercialised platform that operates across different domains, Kaggle requires competition organisers to construct their datasets and define a single machine-learning problem. The application is subsequently reviewed by the platform's dedicated 'competitions team' that decides whether a competition is eligible for Kaggle. If accepted, then the team consults organising institutions for data preparation (e.g., cleaning) and custom work in the competition design process (KG-2022a). Such custom engagement concerns redefining the machine-learning problem central to the competition in question, outlining competition-specific terms and rules, and setting the metrics for model evaluation.

All this work comes at a significant cost for competition organisers. In fact, the average budget for organising a Kaggle-hosted competition ranges from $85,000 to $200,000. While this budget estimate includes prize money, any 'special custom engagements' (KG-2022a) (e.g., defining evaluation metrics) may further increase it. Key variables include (1) the amount of Kaggle consultation for data preparation; (2) the number and value of prizes; and (3) the amount of custom work (e.g., on evaluation metrics) (KG-2022a). Grand Challenge, in turn, enables competitions at a minimum cost of €5000, plus significant variable costs for compute time and storage in the AWS cloud (GC-2022c). To calculate the variable budget, organisers are required to add a cost estimate to their competition application. If a competition ends up exceeding its estimated compute and storage budget, then the platform reserves its right to reduce the final costs, for example, by limiting the number of submissions allowed per participant (GC-2022c). Clearly, the significant costs for the organisation of Kaggle competitions limit who can apply for hosting a competition on this platform to the types of institutions that have the financial means to apply. While the barriers of entry for organising a Grand Challenge competition seem lower at first hand, its financial requirements for compute power and data storage through the AWS cloud can shape the organisation process as they set the conditions for a competition's scale, for example by regulating the number of models that can be submitted.

Second, Kaggle and Grand Challenge set the conditions of possibility for medical imaging competitions at the degree of data and task design. Grand Challenge requires organising actors to apply via an extensive request form (GC-2022a), evaluated by the platforms' team of reviewers at the RUMC. This team subsequently decides whether the proposal will be accepted and if so, in what way the platform will support the competition (GC-2022a). More specifically, it is through this questionnaire that Grand Challenge structures the organisation of medical imaging competitions from the ground up (Figure 1). The platform requires organisers to follow a set of pre-defined task types that can be conducted within a competition such as classification, detection, or segmentation. Furthermore, the platform structures competition organisation, standardised according to one of twenty predetermined human body structures. While most body structures are included in this pre-defined list (Figure 1), lower limb body structures such as upper and lower legs as well as feet remain excluded (only the knee is represented). Lastly, the platform pre-specifies the types of image data that can be used in a competition.

In conditioning the types of machine-learning tasks that can be defined when seeking to host a competition, the platform sets research directions towards topics such as the detection or segmentation of cancer – something that is commonly defined by a research

Task Types	(Body) Structures	(Data) Modalities
Classification	Colon (Abdomen)	CT
Detection	Kidney (Abdomen)	CT/MR
Generalisation	Liver (Abdomen)	Dermoscopy
Localization	Pancreas (Abdomen)	Endoscopy
Modelling	Spleen (Abdomen)	Fundus Photography
Prediction	Heart (Cardiac)	Histology
Reconstruction	Brain (Head and Neck)	Infrared Reflectance Imaging
Registration	Cranium (Head and Neck)	Mammography
Regression	Retina (Head and Neck)	MR
Segmentation	Teeth (Head and Neck)	OCT
Tracking	Knee (Lower Limb)	PET
	Cervix (Pelvis)	PET/CT
	Prostate (Pelvis)	PET/MR
	Skin (Skin)	Transmission Electron Microscopy
	Spinal Cord (Spine)	US
	Vertebral Column (Spine)	XR
	Lung (Thorax)	
	Breast (Thorax)	
	Hand (Upper Limb)	

Figure 1. Overview of Task types, (Body) Structures and (Data) Modalities that can be selected on Grand Challenge through the competition application form (GC-2022b).

community itself. This further solidifies its position as a critical research infrastructure in the field. The research communities' commitment to work within this ecosystem, consequently, risks narrowing the possibilities for alternative research types that do not fit the competition structure but may 'contribute to different understandings about AI [research] possibilities' (Luitse & Denkena, 2021, p. 11). Such excluded approaches involve, for example, various methods for public participation and consultation (Delgado et al., 2023; Young et al., 2023) in the production and evaluation of AI systems for medical imaging that rely on input from a larger consortium of actors, including those most affected by them.

While dataset construction and task design for Kaggle-hosted competitions is left to the organising institutions, we find that this platform especially shapes the practices of competition organising institutions at the level of evaluation. Model evaluation,

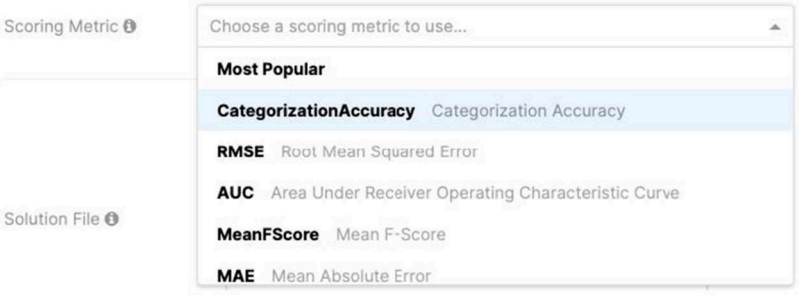

Figure 2. On Kaggle, competition organisers choose one evaluation metric from a predefined list. Custom metrics can be developed, but this comes with additional costs for the organising institution (KG-2022d).

Hutchinson et al. (2022, p. 1860) explain, is a 'system of arbitrary structure that takes a model as an input and produces outputs of some [numerical] form to judge the model' and evaluation design requires the selection of one or more evaluation metrics (e.g., accuracy) and a test dataset. Both the choice of metrics and the selection of test data are crucial for the machine-learning development pipeline as these entities influence how model performance is eventually measured and compared, ultimately deciding on what is considered a 'good' model for one particular task and which model is 'even better' than some alternative. Kaggle provides organisers "a list of the most popular evaluation metrics that will be useful for the majority of use cases" (KG-2022d) (Figure 2), despite it being difficult to choose one (Goodfellow et al., 2016). While organising actors with a larger financial budget are able to further customise their individual evaluation metrics with the Kaggle team (KG-2022a), the provision of this list allows the platform to standardise the metric selection process by limiting the options to a prestructured list organisers can choose from.

Hence, the role of Kaggle has become particularly dominant in standardising the metrics for measuring and comparing model performance in this competitive environment. This is a design task that otherwise requires careful and human-centred decision-making by multiple actors, including expert researchers (e.g., AI researchers specialised in healthcare and medical specialists), based on the intended model structure and use (Mitchell et al., 2019). Moreover, as Raji et al. (2021) point out, the focus on metrics for model evaluation on Kaggle leaves no room for organising institutions to mobilise alternate evaluation methods such as disaggregated analysis (Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019). This is unfortunate, as such methods have proven to be effective in the detection and mitigation of several demographic biases (e.g., toward race, ethnicity, gender, and age). This is a practice that should be standard for both dataset and model development for AI-based medical imaging (Garin et al., 2023).

Finally, both Kaggle and Grand Challenge condition model development in competitions through their platform boundary resources, regulating the access to compute power for training and testing models in competitions. For 50% of the Kaggle competitions under study, participants are required to use the integrated Kaggle (Jupyter) Notebooks infrastructure to build their models. Within these so-called 'Code Competitions,' model evaluation time on CPU or GPUs is set to a maximum number of hours (Figure 3).

This is a Code competition

Submissions to this competition must be made through Notebooks. In order for the "Submit to Competition" button to be active after a commit, the following conditions must be met:

- CPU Notebook <= 9 hours run-time
- GPU Notebook <= 6 hours run-time
- TPUs will not be available for making submissions to this competition. You are still welcome to use them for training models.
- No internet access enabled
- External data, freely & publicly available, is allowed. This includes pre-trained models.
- No custom packages enabled in kernels
- Submission file must be named `submission.csv`

Please see the Code Competition FAQ for more information on how to submit.

Figure 3. Example of Code Requirements provided by Kaggle within a competition (PANDA-2020).

Participants of a Grand Challenge competition are required to upload a file with their predictions on the test set (Type 1) or to submit their models in the form of a Docker container[4] (Type 2). In case of a Type 2 competition, the platform runs these algorithms on the private test set. Model running time, in turn, is dependent on the organisers' budget for computing power (GC-2022).

Considering that contemporary AI research approaches increasingly rely on large amounts of computing power (Dotan & Milli, 2020), such regulations around the use of compute infrastructures – like machine-learning datasets (Denton et al., 2021) – shape what novel algorithmic techniques can be developed and used to build models that are submitted in competitions (Hooker, 2020). In addition, it creates a research gap between well-funded organising institutions that can afford to provide more computing power to their participants for the training and testing of models in competitions which may lead to better performance results and competition organisers who do not have the financial means to do so. This extends previous research on boundary resources as governing instruments that shape platform activity (Nieborg & Helmond, 2019) towards (medical) AI research and development. Both Kaggle and Grand Challenge operate their boundary resources to control the use of compute infrastructures, thereby articulating directions of model development for AI-based medical imaging.

The role of competition organising institutions

At the level of competition organising institutions, our findings show that this group of actors' condition model development in competitions at several stages of the machine-learning development pipeline. Focusing on practices of task design first, organisers of

Grand Challenge are specifically drawn to design competitions around 3D segmentation tasks. 44% of the competitions hosted on this platform focused on semantic or instance segmentation of MRI or CT data.[5] For instance, a popular 'KiTS' competition (Heller et al., 2021) was organised on the segmentation of kidney tumours in abdominal CT imaging. This follows a longer tradition of the medical AI research community using its central position in the competition ecosystem to steer research in this particular direction. As Maier-Hein et al. (2018) show, 70% of all competitions organised prior to 2018 focused on segmentation tasks, defining the (semi-)automatic segmentation of medical imaging 'an important and popular research direction' (Heller et al., 2021, p. 3). Hence, considering the impact of competition in the field (Mendrik & Aylward, 2019), the focus on segmentation tasks is likely to receive more attention than others, not just within the community itself but also by industry actors looking for sponsors and partnerships.

Such public or industry partners and sponsors play an important role in shaping the collection and annotation of datasets (Denton, 2022). Successful competitions that attract many participants often provide large medical image datasets consisting of health data that is typically not publicly available (cf. ImageNet). To overcome this challenge, organising actors collaborate with each other or with industry partners to construct 'multicentre datasets' for model training, testing and, in some cases, post-competition internal validation. For example, the recent Kaggle-hosted competition on the diagnosis of prostate cancer (PANDA-2020) – the PANDA challenge – was organised by a consortium of researchers from Radboud University Medical Centre (the Netherlands), the Karolinska Institutet (Sweden) and Google Health (United States). The competition training and datasets of 12,625 whole-slide images originated from the two medical institutions listed above, as well as from Karolinska University Hospital (validation), two independent medical laboratories and one tertiary unspecified teaching hospital in the United States (validation). Data was annotated by pathology experts from the three countries (Bulten et al., 2022). This partnership allowed for building 'the largest publicly available dataset of prostate biopsies to date' (Bulten et al., 2022, p. 155) underpinning the PANDA challenge. The competition and its underlying EU-US multicentre dataset are claimed to facilitate the development of generalisable models 'to different patient populations across the world' (Bhattacharya et al., 2022). Its competition design, however, articulates a very particular understanding of dataset *diversity* – EU and US populations as representative of 'the world' – while data from patients in hospitals in e.g., Asian, Latin-American countries or African countries are not included. Such data are, of course, of critical importance to further explore the possibilities for model generalisability through situated accounts of model evaluation and testing outside of competition infrastructures.

The power of partnerships in constructing datasets for model development also becomes visible in smaller competitions. A competition using colour fundus photography (CFP) (Airogs-2022), e.g., relied on large amounts of retinal data from EyePACS, a platform that provides 'end-to-end services and technology for blindness prevention programs' (Retinal Consulting Solution, n.d.). The organising team collaborated with EyePACS to acquire the large datasets necessary for training and evaluation. As the organisers of *AIROGS: Artificial Intelligence for RObust Glaucoma Screening* explain:

> [W]e used a large screening dataset of around 113,000 images from about 60,000 patients and approximately 500 different sites with a heterogeneous ethnicity. We splitted the data

> in a training set with about 101,000 gradable images […] and a closed test set with approximately 11,000 […] images. (Airogs 2022)

Considering the sensitivities of medical data and the regulatory restrictions around data sharing in medicine (Simko et al., 2022), it is usually difficult to construct such large image datasets. However, third-party collaboration allowed the organisers to proceed and host the competition on Grand Challenge, attracting 351 participants from 51 countries, and gathering solutions from 15 teams across 13 countries that eventually submitted a model during the final evaluation phase (De Vente, 2022). This enabled them to crowdsource a number of different approaches and move towards their aim of 'implementing robust AI for glaucoma screening in clinical practice' (De Vente, 2022).

Approaches to dataset diversity by competition organising institutions also shape model production and evaluation. As the majority of organisers are located in the United States, Western Europe and China (Figure 4), most datasets underlying the competitions included in our research, are constructed with data sourced from patients located in these respective geographic regions. Consequently, the data from hospitals in other parts of the world remains underrepresented across competitions, thereby inserting structural limitations, as well as racial and gender biases in these datasets (Paullada et al., 2021). This has a significant impact on how well the data and associated metrics of evaluation actually represent the particular task defined by the competition organisers (Raji et al., 2021; Thomas & Uminsky, 2022). In other words, it impacts what research claims can be made based on competition outcomes. Furthermore, model production in competitions is more likely steered into the direction of 'fine-tuning' for a specific competition task or overfitting towards a dataset to perform well, rather than focussing on providing a solution for a clinical problem that can be adapted and validated for different geographical contexts (Maier-Hein et al., 2018). This is further substantiated by the fact that details about patient consent for data collection are missing from the information on the organisation of competitions, revealing a particular 'politics of data reuse' (Thylstrup et al., 2022) in which patients clearly have little to no power. More generally, it is difficult to evaluate data diversity: dataset descriptions rarely further specify sociodemographic

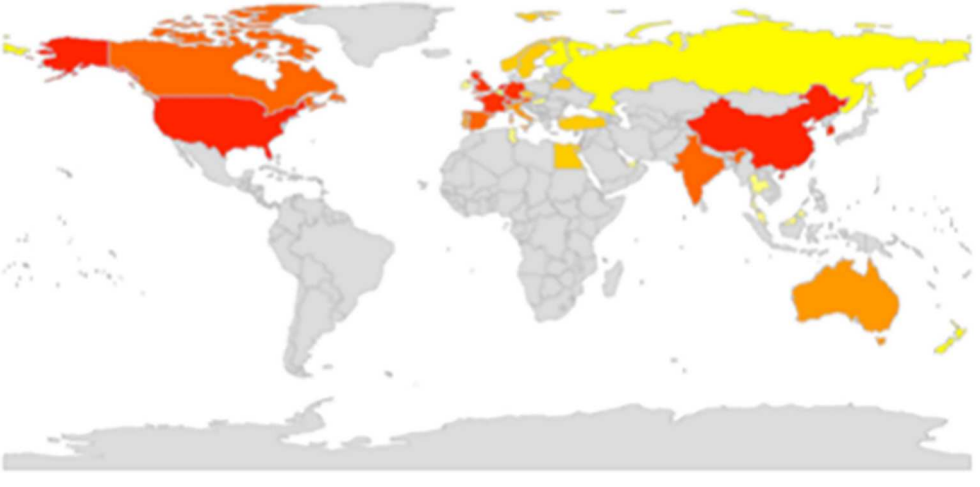

Figure 4. Distribution of competition organising institutions.

information in terms of gender and race, nor do they specify the reasons why a particular dataset was collected in the first place, or by whom it was annotated.

Competition organising institutions tend to approach dataset diversity primarily in a technical way. Descriptions of diversity focus on listing the variety of scanners (e.g., brand, scanner model) for imagery production as well as the diversity of image resolutions (ACROBAT-2022; CADA-AS-2022; autoPET-2021; SIIMISIC-2020; and SIIMACR-2019). This resonates with research by Scheuerman et al. (2021, p. 21), who observe that for computer vision datasets diversity 'was not commonly used in terms of human conditions such as race, gender or ability.'. Instead – and similar to our findings – diversity meant 'describing diverse instances of data' (p. 21). While technical information like image resolution is significant for medical image analysis, the 'technical' approach to diversity has consequences for model development by competition participants. Uncertainties concerning the sociodemographic diversity of training and test data very much impact participants' ability to detect and account for model bias against these properties during evaluation (Ganapathi et al., 2022).

Power relations among competition participants

The practices of competition participants are conditioned by Kaggle and Grand Challenge, as well as competition organising institutions. However, similar to dataset development (Denton et al., 2021), competitive machine-learning model development is also influenced by the demographic and disciplinary contexts in which participating teams operate. Research teams from different institutions across the globe have different research traditions that build upon a variety of 'disciplinary values' (Dotan & Milli, 2020) – values that ultimately play a role in guiding research approaches for model development.

From this participant perspective, we find that the global distribution of competing teams significantly differs between Kaggle and Grand Challenge and also depends on the focus of the various teams and their place within the machine-learning (research) community. In line with the platform's commercial nature, Kaggle competition participants are more likely to operate independently or are based at commercial companies such as large technology corporations like Amazon or NVIDIA, machine-learning startups, or dedicated (research) companies that focus on the development of specific applications. These companies are typically located in the United States, followed by India, Japan, Germany and China. This is illustrated in Figure 5, which shows the global distribution of Kaggle competition participants per 10 million inhabitants. The figure colours are mapped from red to yellow, with red representing a high frequency of participants for that particular country. Such industry involvement is particularly beneficial for the organising institution as Kaggle competitions attract many talented participants (Tauchert et al., 2020), resulting in an even greater number of (potentially) valuable submissions that can be used for further development.

However, such industry involvement simultaneously risks playing into the hands of commercial entities that may use the Kaggle competition model to attract machine-learning talent and clients. Participants in the top-performing teams of the *SIIM-ISIC Melanoma Classification* and *APTOS 2019 Blindness Detection* competitions, for example, are affiliated with H2O.ai, an open-source platform to build machine-learning applications for businesses, governments, nonprofits, or academic institutions. This company

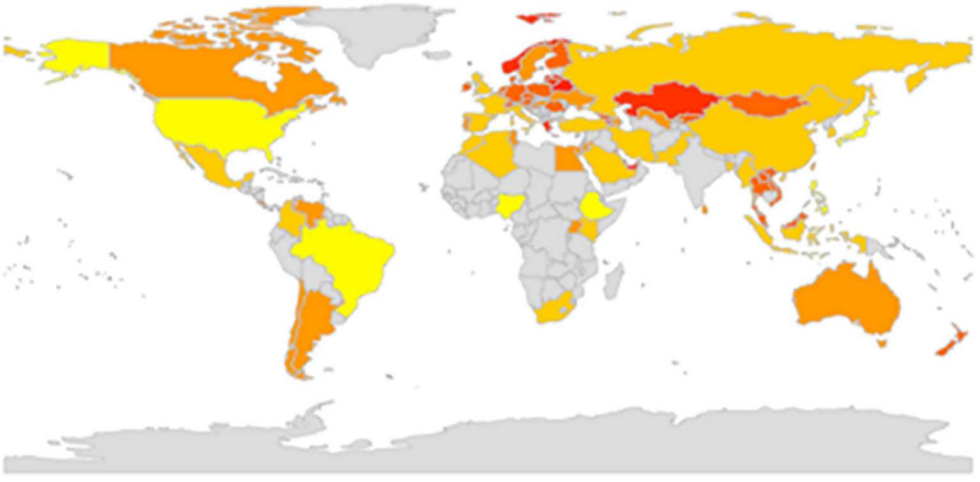

Figure 5. Distribution of competition participants for Kaggle competitions, scaled towards country population per 10 million inhabitants.

specifically advertises the number of so-called 'Grandmasters' (Kaggle Grandmasters | H2O.ai, n.d.) – a Kaggle title for participants that have 'consistently demonstrated outstanding performance' (KG-2022h). Similarly, a winning participant of the PANDA Challenge is affiliated with Rist, a Japan-based company that operates a team that 'allows for participation in Kaggle activities during work, based on the performance in Kaggle competitions.' (株式会社Rist | Rist Inc.) and gain visibility on the competition platform. Competition performance, in these examples, appears to function as a statistical measure to advertise a company's capabilities and attract clients or talent in a highly competitive market. As this environment is already increasingly captured by well-resourced large technology corporations and elite universities (Ahmed & Wahed, 2020), this risks further reducing the research opportunities for mid- and lower-tier academic institutions.

In addition, industry-embedded participants on Kaggle are rarely positioned in the health sector, although most medical imaging competitions are organised by medical institutions using sensitive and complex medical data. In contrast to the large number of hospital-affiliated computer scientists and biomedical engineers active on Grand Challenge (Figure 6), these Kaggle participants identify themselves as general 'data scientists' (Figure 7). This points to a scarcity of domain expertise among competition participants on this platform, shifting model development away from the medical imaging disciplines towards data science. Domain expertise is, however, crucial for the field (Thomas, 2021), especially to discover problems with annotation labels in medical imagery – an impactful data issue that is difficult to detect without a medical background, but one of the key sources of bias in AI-based medical imaging (Rostamzadeh, 2022). Considering the potential implications of (undesirable) data biases for model development in medicine (e.g., Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021), it is critical to detect such issues at multiple stages of the development pipeline, even though participating teams are not responsible for the construction of competition datasets.

Taken together, while medical imaging expertise is more present among competition participants on the Grand Challenge, we find that the geographic distribution of

Job Title	Number of Participants
Computer Science	1609
Biomedical Engineering	790
Electrical Engineering	544
Computer Engineering	219
Artificial Intelligence	188

Figure 6. Top 5 department titles Grand Challenge competition participants as reported by participants on their individual Grand Challenge profile pages.

participants (Figure 8) on this platform mirrors the global distribution of (public) investment in AI research and development. On Grand Challenge, the top two countries are China and Taiwan with the most participating teams on the platform. This is not surprising considering the Chinese ambition to 'become the world leader in AI by 2030' (Roberts et al., 2021, p. 59) and the Taiwanese '$517.5 million strategy to cultivate AI technology specialists and create an environment for AI scientific research' (Executive Yuan, 2018). Continuous participation in medical imaging competitions, therefore, seems to follow these state-led strategies. This has transformed competitions into powerful research infrastructures (Mendrik & Aylward, 2019; Reinke et al., 2021) and winning models in these competitions frequently lead to paradigm shifts in the field (e.g., AlexNet that won the 2012 ImageNet Challenge). Because of large state investments, research

Job Title	Number of Participants
Data scientist	1176
Machine Learning	227
Software Engineer	202
Learning Engineer	184
Data Science	177

Figure 7. Top 5 bigrams of job titles for Kaggle competition participants as reported by participants on their individual Kaggle profile pages.

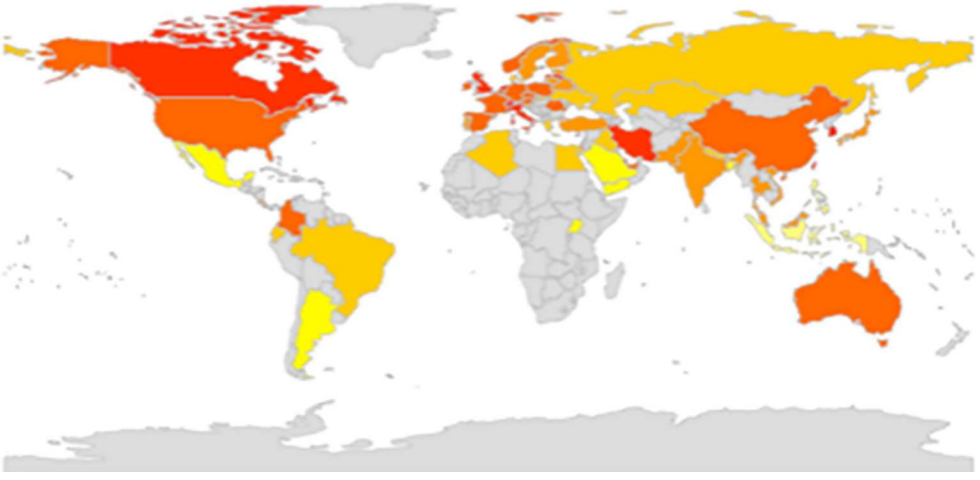

Figure 8. Distribution of competition participants for Grand Challenge competitions, scaled towards country population, per 10 million inhabitants.

groups have the financial means and the institutional connections to gain technical expertise and participate successfully in (prestigious) competitions, solidifying their position in the medical AI research and development ecosystem.

Conclusion

Examining the organisation of competitions through Kaggle and Grand Challenge, this article showed that these machine-learning development platforms enable, streamline, and scale the production and evaluation of AI models in medical imaging. They *facilitate* the participation in model development of large numbers of competing teams and research institutions from around the globe. At the same time, our analysis also suggests that the configuration of platforms, competition organisers, and participants *concentrates power* toward a small number of actors. This paradoxical combination of facilitating and opening fields of knowledge production, while simultaneously bringing about power concentrations can be seen in most sectors in which platforms intervene (e.g., Aradau & Blanke, 2022; Poell et al., 2021). For AI-based medical imaging, this has far-reaching consequences for *who defines the conditions under which AI research can be produced, on which population knowledge production takes place, and who reaps the benefits of AI development*. This can be observed on all three levels of our analysis.

First, *platforms* set the conditions of possibility for the AI development process by requiring substantial financial resources, defining which institutions can and cannot take part in hosting a competition and setting the rules and regulations for competition organisation. The Kaggle platform operates significant hosting fees that raise the barriers to entry. Grand Challenge, in turn, conditions the cost and scale of competitions by relying on AWS. Additionally, we found that both Kaggle and Grand Challenge structure the competition design process at the stages of (1) dataset construction (Kaggle & Grand Challenge); (2) task types and subject focus areas for research (Grand Challenge); and (3) setting evaluation metrics (Kaggle). Consequently, while datasets remain important components that set the

basic conditions for AI research and development in terms of problem definition, articulation of progress, and development of novel algorithmic techniques (Denton, 2022; Thylstrup, 2022), the platforms that host performance competitions play a key role in this process too. This raises concerns about the position of the research community and its ability to explore new techniques for model development or 'alternative methods of evaluation' (e.g., Raji et al., 2021; Thomas & Uminsky, 2022) outside the realm of benchmarking. Such alternatives are currently excluded within this competitive ecosystem.

Second, *competition organising institutions* direct model production by focusing competitions on image segmentation tasks, partly to attract attention from major industry actors developing clinical applications. Successful competitions are particularly reliant on powerful public and industry partnerships for the construction of large-scale multicentre datasets consisting of data from different geographic areas. Through such datasets, competition organisers define dataset diversity and the generalisability of machine-learning models, even though the underlying data often represents a limited number of patient populations from only a few hospitals and countries. Most datasets are constructed with data from hospitals in North America, Western Europe, and China, limiting the possibility of the application of models to other geographical contexts (Maier-Hein et al., 2018). These observations align with Denton's (2022) argument that competition organisers not only develop competition datasets, but also they exercise power by further standardising model production and evaluation with respect to a given machine-learning task and they frame what knowledge can be derived from those processes.

Finally, *competition participants* influence model production through the institutional, demographic, and disciplinary contexts in which they operate. We found that participation in Kaggle is particularly driven by aspirations to attract talent in the highly competitive AI industry. Strikingly, such industry-driven teams mainly consist of data scientists who do not have specific medical imaging domain expertise. We note that the absence of domain expertise risks further shifting model production away from the medical imaging disciplines towards general data science – a concerning development considering that medical know-how is critical (Thomas, 2021). Participation in medical AI competitions is also driven by national AI strategies by governments, such as the Chinese and Taiwanese. These strategic state efforts steer financial resources and institutional connections to a handful of university-affiliated teams to develop their technical expertise and participate successfully in (prestigious) competitions.

Considering the increasing concerns about the social inequalities medical AI systems may produce (e.g., Obermeyer et al., 2019; Thomas, 2021), our observations underline the importance of critical research into machine-learning development, particularly as the medical domain comes with a set of 'unique ethical, legal, and regulatory challenges' (Markus et al., 2021, p. 1). Pursuing such research, it is vital to cover the *entire* research pipeline, including model evaluation. It is only by considering the entire research and development infrastructure that risks of bias, inequality, and power concentration can be addressed and mitigated.

Notes

1. 'AI' encompasses various approaches rooted in different research traditions and disciplines. Using AI in this paper, we refer to 'narrow AI' deep-learning systems.

2. We understand machine-learning models as algorithmic assemblages that aggregate data, the model architecture, and the learning algorithm into a processual workflow to solve a particular task.
3. Training data refers to the set of data instances used to train a model by fitting its parameters to perform a defined task. Test data is subsequently leveraged to evaluate the model after training and measure its ability to generalise to data that the model has not seen during training (Scheuerman et al., 2021).
4. A Docker container is a standard software unit that packages all code, libraries and additional files to quickly run a software application (e.g., a machine learning model) from one computing environment to another.
5. Semantic segmentation refers to the detection of multiple objects within a single category (e.g., human) Instance segmentation requires the identification of individual objects within a specific category (e.g., each individual person) (Minaee et al., 2022).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the University of Amsterdam Research Priority Area Artificial Intelligence for Health Decision-Making.

Notes on contributors

Dieuwertje Luitse is a Ph.D. candidate in Data Bodies at the University of Amsterdam, department of Media Studies. Her research critically examines the power, ethics and politics of data, AI system development and infrastructures in healthcare. This project is part of an interdisciplinary research focus area on AI for Health Decision-making [Email: d.luitse@uva.nl]

Tobias Blanke is a University Professor of Artificial Intelligence and Humanities at the University of Amsterdam and the Institute for Logic, Language, and Computation [Email: t.blanke@uva.nl].

Thomas Poell is a Professor of Data, Culture & Institutions at the University of Amsterdam. He is the co-author of Platforms and Cultural Production (Polity, 2021) and The Platform Society (Oxford University Press, 2018) [Email: t.poell@uva.nl].

ORCID

Dieuwertje Luitse http://orcid.org/0000-0003-0652-3315
Tobias Blanke http://orcid.org/0000-0003-0956-987X
Thomas Poell http://orcid.org/0000-0002-3664-2588

References

Ahmed, N., & Wahed, M. (2020). The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. arXiv:2010.15581.arXiv

Aradau, C., & Blanke, T. (2022). *Algorithmic reason: The new government of self and other* (1st ed.). Oxford University Press.

Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020). Studying up: reorienting the study of algorithmic fairness around issues of power. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 167–176. https://doi.org/10.1145/3351095.3372859

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Bhattacharya, I., Khandwala, Y. S., Vesal, S., Shao, W., Yang, Q., Soerensen, S. J. C., Fan, R. E., Ghanouni, P., Kunder, C. A., Brooks, J. D., Hu, Y., Rusu, M., & Sonn, G. A. (2022). A review of artificial intelligence in prostate cancer detection on imaging. *Therapeutic Advances in Urology*, *14*, 1–31. https://doi.org/10.1177/17562872221128791

Bowker, G. C., & Star, S. L. (2000). *Sorting things out*. MIT Press.

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M.B., Evans, A.J., van der Kwast, T., Allan, R., Humphrey, P.A., Grönberg, H., Samaratunga, H., … Eklund, M. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, *28*(1), 154–163. https://doi.org/10.1038/s41591-021-01620-2

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, *81*, 77–91.

Delgado, F., Yang, S., Madaio, M. (2023). The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23. https://doi.org/10.1145/3617694.36232

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Denton, E. (2022). CITP seminar Emily Denton—data, power, AI ethics … . Retrieved 7 April, from https://www.youtube.com/watch?v=VsgxV-LKUb0.

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2), 1–14. https://doi.org/10.1177/20539517211035955

De Vente, C. (2022). Coen de Vente AIROGS challenge report: AI models can be used for glaucoma screening, but do they know when they cannot? Retrieved May 10, 2022, from https://coendevente.com/project-airogs.html

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

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26* (4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Dotan, R., & Milli, S. (2020). Value-laden disciplinary shifts in machine learning. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 294. https://doi.org/10.1145/3351095.3373157

Executive Yuan, R. O. C. Taiwan. (2018). *AI innovation: Grand strategy for a small country*. Retrieved December 14, 2022, from https://english.ey.gov.tw/News3/9E5540D592A5FECD/edadb735-e6a6-43e1-ac93-1959602bb3ec

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. Vintage.

Foucault, M. (1982). The subject and power. *Critical Inquiry*, *8*(4), 777–795. https://doi.org/10.1086/448181

Ganapathi, S., Palmer, J., Alderman, J. E., Calvert, M., Espinoza, C., Gath, J., Ghassemi, M., Heller, K., Mckay, F., Karthikesalingam, A., Kuku, S., Mackintosh, M., Manohar, S., Mateen, B. A., Matin, R., McCradden, M., Oakden-Rayner, L., Ordish, J., Pearson, R., … Liu, X. (2022). Tackling bias in AI health datasets through the STANDING together initiative. *Nature Medicine*, *28*(11), 2232. https://doi.org/10.1038/s41591-022-01987-w

Garin, S. P., Parekh, V. S., Sulam, J., & Yi, P. H. (2023). Medical imaging data science competitions should report dataset demographics and evaluate for bias. *Nature Medicine*, *29*(5), 1038–1039. https://doi.org/10.1038/s41591-023-02264-0

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Harman, D. (1993). Overview of the first TREC conference. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 36–47. https://doi.org/10.1145/160688.160692

Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., …

Weight, C. (2021). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical Image Analysis*, *67*, 101821. https://doi.org/10.1016/j.media.2020.101821

Hooker, S. (2020). *The Hardware Lottery*. arXiv:2009.06489 [Cs]. arXiv.

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., & Prabhakaran, V. (2022). Evaluation gaps in machine learning practice. *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1859–1876. https://doi.org/10.1145/3531146.3533233

Kaggle. (n.d.). 株式会社Rist | Rist Inc. Retrieved December 12, 2022, from https://www.rist.co.jp/en/kaggle/

Kaggle Grandmasters | H2O.ai. (n.d.). Retrieved December 13, 2022, from https://h2o.ai/company/team/kaggle-grandmasters/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Larkin, B. (2013). The politics and poetics of infrastructure. *Annual Review of Anthropology*, *42*(1), 327–343. https://doi.org/10.1146/annurev-anthro-092412-155522

Li, F.-F. (2017, 9 March). Welcome Kaggle to Google Cloud [Blog]. Google Cloud Blog. https://cloud.google.com/blog/products/gcp/welcome-kaggle-to-google-cloud

Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, *8*(2), 1–14. https://doi.org/10.1177/20539517211047734

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., … Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, *9*(1), 1–13. https://doi.org/10.1038/s41467-018-07619-7

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Mendrik, A. M., & Aylward, S. R. (2019). A framework for challenge design: Insight and deployment challenges to address medical image analysis problems. arXiv:1911.08531.

Miceli, M., Posada, J., & Yang, T. (2022). Studying Up machine learning data: Why talk about bias when We mean power? *Proceedings of the ACM on Human–Computer Interaction*, *6*(34): 1–14.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Nieborg, D. B., & Helmond, A. (2019). The political economy of Facebook's platformization in the mobile ecosystem: Facebook Messenger as a platform instance. *Media, Culture & Society*, *41*(2), 196–218. https://doi.org/10.1177/0163443718818384

Oakden-Rayner, L. (2019). Exploring large scale public medical image datasets. arXiv:1907.12720.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Poell, T., Nieborg, D., & Duffy, B. E. (2021). *Platforms and cultural production*. Polity.
Prevedello, L. M., Halabi, S. S., Shih, G., Wu, C. C., Kohli, M. D., Chokshi, F. H., Erickson, B. J., Kalpathy-Cramer, J., Andriole, K. P., & Flanders, A. E. (2019). Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, *1*(1), e180031. https://doi.org/10.1148/ryai.2019180031
Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. arXiv:2111.15366 [Cs].arXiv.
Raji, I. D., & Buolamwini, J. (2019). Actionable auditing. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. https://doi.org/10.1145/3306618.3314244
Reinke, A., Tizabi, M. D., Eisenmann, M., & Maier-Hein, L. (2021). Common pitfalls and recommendations for grand challenges in medical artificial intelligence. *European Urology Focus*, *7*(4), 710–712. https://doi.org/10.1016/j.euf.2021.05.008
Retinal Consulting Solution. (n.d.). EyePACS. Retrieved May 15, 2022, from https://www.eyepacs.com/solution
Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society*, *36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2
Rostamzadeh, N., Mincu, D., Roy, S., Smart, A., Wilcox, L., Pushkarna, M., Schrouff, J., Amironesei, R., Moorosi, N., & Heller, K. (2022). Healthsheet: Development of a Transparency Artifact for Health Datasets. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1943–1961. https://doi.org/10.1145/3531146.3533239
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y
Scheuerman, M. K., Hanna, A., & Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW2), 317.
Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), 2176. https://doi.org/10.1038/s41591-021-01595-0
Simko, A., Garpebring, A., Jonsson, J., Nyholm, T., & Löfstedt, T. (2022). Reproducibility of the Methods in Medical Imaging with Deep Learning. arXiv:2210.11146. arXiv.
Sparck Jones, K. (1995). Reflections on TREC. *The Second Text Retrieval Conference (TREC-2)*, *31*(3), 291–314.
Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326
Star, S. L., & Lampland, M. (2009). Reckoning with standards. In M. Lampland, & S. L. Star (Eds.), *Standards and their stories: How quantifying, classifying, and formalizing practices shape everyday life* (pp. 3–34). Cornell University Press.
Tauchert, C., Buxmann, P., & Lambinus, J. (2020). *Crowdsourcing data science: A qualitative analysis of organizations' usage of Kaggle competitions*.
Thomas, R. (2021, 4 January). Medicine's machine learning problem. *Boston Review*. https://www.bostonreview.net/articles/rachel-thomas-medicines-machine-learning-problem/
Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. *Patterns (New York, N.Y.)*, *3*(5), 100476.
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. https://doi.org/10.1177/01634437211060226
Thylstrup, N. B., Hansen, K. B., Flyverbom, M., & Amoore, L. (2022). Politics of data reuse in machine learning systems: Theorizing reuse entanglements. *Big Data & Society*, *9*(2), 1–10. https://doi.org/10.1177/20539517221139785

Whittaker, M. (2021). *Distinguished speaker series with an inaugural theme: AI and Inequalities: Creating change: From ethics to organizing: Getting serious about AI*. Hariri Institute for Computing, Boston University. https://www.youtube.com/watch?v=_BzU0bD0Ics

Young, M., Akinrinade, I., Calderon, A., Lara, R., Loeb, E., & Onnekikami, T. (2023, October 18). Shaping AI systems By shifting power. *Data & Society: Points*. https://medium.com/datasociety-points/shaping-ai-systems-by-shifting-power-ee95f7c3edf9