RESEARCH-ARTICLE

In the Picture: Medical Imaging Datasets, Artifacts, and their Living Review

AMELIA JIMÉNEZ-SÁNCHEZ, IT University of Copenhagen, Copenhagen, Denmark

NATALIA ROZALIA AVLONA

SARAH DE BOER, Radboud University Medical Center, Nijmegen, Gelderland, Netherlands

VÍCTOR M CAMPELLO, University of Barcelona, Barcelona, Barcelona, Spain

AASA FERAGEN, Technical University of Denmark, Lyngby, Hovedstaden, Denmark

ENZO FERRANTE, National Council for Scientific and Technical Research, Buenos Aires, Provincia de Buenos Aires, Argentina

View all

Published: 23 June 2025

FAccT '25: The 2025 ACM Conference on Fairness, Accountability, and Transparency
June 23 - 26, 2025
Athens, Greece

https://doi.org/10.1145/3715275.3732035
ISBN: 9798400714825

In the Picture: Medical Imaging Datasets, Artifacts, and their Living Review

Amelia Jiménez-Sánchez
ITU
Copenhagen, Denmark
amji@itu.dk

Natalia-Rozalia Avlona
KU
Copenhagen, Denmark
rozalia.avlona@sund.ku.dk

Sarah de Boer
Radboudumc
Nijmegen, Netherlands
sarah.deboer@radboudumc.nl

Víctor M. Campello
UB
Barcelona, Spain
victor.campello@ub.edu

Aasa Feragen
DTU
Kongens Lyngby, Denmark
afhar@dtu.dk

Enzo Ferrante
CONICET & UBA
Buenos Aires, Argentina
eferrante@sinc.unl.edu.ar

Melanie Ganz
KU & Rigshospitalet
Copenhagen, Denmark
ganz@di.ku.dk

Judy Wawira Gichoya
Emory University
Atlanta, USA
judywawira@emory.edu

Camila Gonzalez
Stanford University
Stanford, USA
camgonza@stanford.edu

Steff Groefsema
RUG
Groningen, Netherlands
s.groefsema@rug.nl

Alessa Hering
Radboudumc
Nijmegen, Netherlands
alessa.hering@radboudumc.nl

Adam Hulman
AUH & AU
Aarhus, Denmark
adahul@rm.dk

Leo Joskowicz
HUJI
Jerusalem, Israel
josko@cs.huji.ac.il

Dovile Juodelyte
ITU
Copenhagen, Denmark
doju@itu.dk

Melih Kandemir
SDU
Odense, Denmark
kandemir@imada.sdu.dk

Thijs Kooi
Lunit
Seoul, Republic of Korea
email@thijskooi.com

Jorge del Pozo Lérida
ITU & Cerebriu A/S
Copenhagen, Denmark
jorgedelpozolerida@gmail.com

Livie Yumeng Li
AUH & AU
Aarhus, Denmark
livie.ymli@ph.au.dk

Andre Pacheco
UFES
Vitória, Brazil
apacheco@inf.ufes.br

Tim Rädsch
DKFZ & UHEI
Heidelberg, Germany
tim.raedsch@dkfz-heidelberg.de

Mauricio Reyes
UniBE
Bern, Switzerland
mauricio.reyes3@unibe.ch

Théo Sourget
ITU
Copenhagen, Denmark
tsou@itu.dk

Bram van Ginneken
Radboudumc & Plain Medical
Nijmegen, Netherlands
bramvanginneken@gmail.com

David Wen
Oxford University Hospitals
Oxford, United Kingdom
david.wen@nhs.net

Nina Weng
DTU
Kongens Lyngby, Denmark
ninwe@dtu.dk

Jack Junchi Xu
Copenhagen University Hospital & RAIT
Copenhagen, Denmark
jack.junchi.xu@regionh.dk

Hubert Dariusz Zajaç
KU
Copenhagen, Denmark
hdz@di.ku.dk

Maria A. Zuluaga
EURECOM
Sophia Antipolis, France
zuluaga@eurecom.fr

Veronika Cheplygina
ITU
Copenhagen, Denmark
vech@itu.dk

Abstract

Datasets play a critical role in medical imaging research, yet issues such as label quality, shortcuts, and metadata are often overlooked. This lack of attention may harm the generalizability of algorithms and, consequently, negatively impact patient outcomes. While existing medical imaging literature reviews mostly focus on machine learning (ML) methods, with only a few focusing on datasets for specific applications, these reviews remain static – they are published once and not updated thereafter. This fails to account for emerging evidence, such as biases, shortcuts, and additional annotations that other researchers may contribute after the dataset is published. We refer to these newly discovered findings of datasets as *research artifacts*. To address this gap, we propose a *living review* that continuously tracks public datasets and their associated research artifacts across multiple medical imaging applications. Our approach includes a framework for the living review to monitor data documentation artifacts, and an SQL database to visualize the citation relationships between research artifact and dataset. Lastly, we discuss key considerations for creating medical imaging datasets,

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732035

review best practices for data annotation, discuss the significance of shortcuts and demographic diversity, and emphasize the importance of managing datasets throughout their entire lifecycle. Our demo is publicly available at http://inthepicture.itu.dk/.

CCS Concepts

• **General and reference** → **Surveys and overviews**; **Evaluation**; • **Applied computing** → *Life and medical sciences*; • **Computing methodologies** → *Machine learning*.

Keywords

open data, data governance, healthcare, medical imaging, shortcuts, bias, research artifacts, living review

ACM Reference Format:

Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Sarah de Boer, Víctor M. Campello, Aasa Feragen, Enzo Ferrante, Melanie Ganz, Judy Wawira Gichoya, Camila Gonzalez, Steff Groefsema, Alessa Hering, Adam Hulman, Leo Joskowicz, Dovile Juodelyte, Melih Kandemir, Thijs Kooi, Jorge del Pozo Lérida, Livie Yumeng Li, Andre Pacheco, Tim Rädsch, Mauricio Reyes, Théo Sourget, Bram van Ginneken, David Wen, Nina Weng, Jack Junchi Xu, Hubert Dariusz Zając, Maria A. Zuluaga, and Veronika Cheplygina. 2025. In the Picture: Medical Imaging Datasets, Artifacts, and their Living Review. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3715275.3732035

1 Introduction

High-quality datasets are a key element to the development of machine learning (ML) models for medical imaging applications and, more generally, for healthcare. Such datasets are characterized by having diverse representation of patients, sufficient sample sizes, accurate labels or annotations, and comprehensive documentation. Failing to meet these requirements has a direct impact on a model's robustness and reliability [10, 23, 151, 167, 198], thereby affecting the model's clinical utility. Inaccurate or incomplete annotations can lead to models that make incorrect predictions. Insufficient or lack of documentation may miss information such as demographics or hospital scanner, leading to biased and inaccurate models [17, 125]. A lack of diversity limits the generalizability of the model across heterogeneous patient populations, whereas a sufficient sample size is necessary to ensure that the model can learn meaningful patterns and avoid overfitting. Ultimately, dataset quality is as important as the choice of the method to build the ML model.

Despite the critical importance of data, current efforts in medical imaging do not consider the evolving nature of datasets [73]. Such datasets often consist of two parts: images such as chest X-rays, and target labels such as lung diseases. However, additional evidence about these datasets, such as shortcuts, biases, or additional annotations, emerges over time, but is often not available in the original dataset documentation. We refer to these newly discovered aspects of datasets as *research artifacts*. Literature reviews, both in medical imaging and in ML more generally, primarily focus on ML models (see [11, 20, 105, 168, 195] for examples), with only a few addressing specific issues such as fairness in model predictions [18, 19, 151]. Some reviews summarize available datasets within a specific application such as dermatology [26, 186] or ophthalmology [91]. A recent work [45] assesses the documentation of publicly available magnetic resonance imaging (MRI), color fundus photography, and electrocardiogram datasets, yet carries out a static review that is reviewed and published once, without subsequent updates. Finally, datasets released by international competitions are often used after the competition for benchmarking state-of-the-art models [35]. Although the construction of these datasets is crucial for interpreting results, post-competition analyses are centered around model performance, limiting the discussion about the data to a snapshot of its main features (e.g., sample size or scanning devices). On a positive note, competitions do sometimes track the evolution of performance over time, recognizing the dynamic nature of both the datasets and methods. Overall, we believe that the medical imaging field is in need of dynamic or living reviews, inspired by efforts like living reviews of evidence on COVID-19 [192].

We aim to promote collaborative extensions of datasets and avoid snapshots or static datasets. The research we present was conducted in the context of a year-long, collaborative webinar, as well as an in-person workshop on current developments and challenges of medical imaging data. These events brought together a group of around 50 researchers from academia, industry, and clinicians, with backgrounds from ML to epidemiology and human-computer interaction, and research experience from 10+ countries in five continents. This paper is a synthesis of discussions during and after the webinars and workshop, which spanned topics across the entire dataset lifecycle, from creating medical imaging datasets to their use for validating algorithms to data governance, see Fig. 1. Our contributions are as follows:

- We present a proof of concept for a *living review* (publications and database, see Fig. 2) - a framework for enhancing dataset metadata through research artifacts, facilitating the discovery of emerging information about medical imaging datasets.
- We discuss key considerations for creating datasets, best practices for data annotation, significance of demographics and shortcuts, and data management throughout the dataset lifecycle.
- We release a demo of a living database of 16 datasets along with their connections and relationships to 24 research artifacts: shortcuts, annotations, and derivatives, for two medical image applications.
- We discuss how the research community can contribute to our living review, and invite them to do so.

2 Proposal: a living review of medical imaging datasets

Development of novel algorithms with state-of-the-art results is considered the more prestigious activity within the ML field [9, 158], leading to datasets often being considered "as-is" benchmarks. However, medical imaging datasets not only are foundational to ML development but they also evolve over time, both explicitly and implicitly. A dataset evolves *explicitly* if the data itself is updated, for example, due to errors. A dataset evolves *implicitly* as new evidence emerges, such as erroneous target labels, additional annotations, and shortcuts [123, 125]. For example, the CheXpert dataset [76] was initially published without the recommended datasheet [49]. A datasheet co-authored by some of the CheXpert authors was

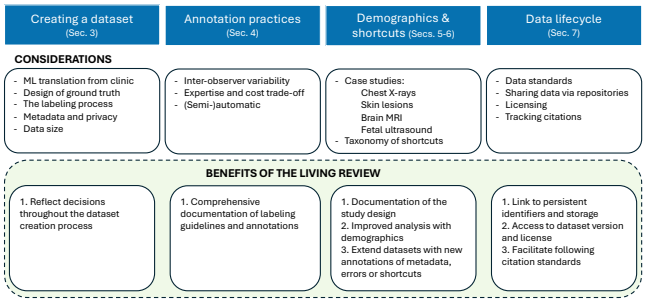

Figure 1: Benefits of our proposed living review framework (Sec. 2). Overview of considerations for creating a dataset (Sec. 3), data annotation practices and quality (Sec. 4), patient demographics (Sec. 5) and shortcuts (Sec. 6), and dataset lifecycle (Sec. 7).

later released [48], but this datasheet is not linked from the original dataset. By not incorporating such evidence into the original dataset, it is often not taken into consideration in subsequent research.

Traditional systematic reviews of medical imaging datasets, such as [26, 91, 99, 186], do not capture this evolution. We argue that we need a *living review* to keep track of novel open datasets, as well as emerging connections and issues in existing datasets. Living systematic reviews are a more recent development in meta-research, but they are crucial for rapidly evolving topics. Given the rapid developments in ML, it would be advantageous to adopt a similar framework for datasets. Our proposed work is different from these efforts, since [192] focuses on COVID-19 findings, not specifically on medical image datasets. [163, 164] focus on data extraction methods to aid doing systematic reviews, of these only [164] is a living review itself, which could provide inspiration for how to carry out our plans.

Here we outline our vision for the living review of open datasets, studying the relationships between the datasets and their research artifacts. We propose a framework of these relationships, as well as protocols for the maintenance of the living review.

Conceptualization. Our proposal for the living review consists of three parts, as illustrated in Fig. 2:

(1) an overarching living review publication,
(2) documentation of research artifacts via dataset-specific publications on Zenodo, which the overarching paper links to,
(3) a SQL database for exploring the links between the datasets and the research artifacts.

An "artifact", like a Greek amphora, is a produced object – the output of a process. We define a research artifact as any additional evidence related to a dataset, for example derived datasets like PruneCXR [69] and LongTailCXR [70], additional annotations like spurious correlations [25], segmentation masks [43], or compressed information in the form of embeddings [165]. Such artifacts are crucial as they provide the necessary context for understanding, reproducing and validating research findings. Moreover, computer science papers with shared artifacts receive about 75% more citations than those without [41], underscoring the critical role of artifacts in the visibility and impact of research.

We map the relation between the datasets and the research artifacts with the citation function (use, produce, extend, introduce,

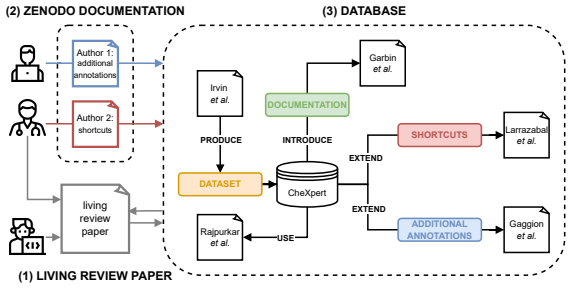

Figure 2: Our living review consists of three components: (1) an overarching living review paper, (2) documentation of research artifacts via publications on Zenodo linked by living review paper, and (3) a SQL database for exploring the links between the datasets and the research artifacts.

other). We exemplify this for the CheXpert dataset in Fig. 2, for more details and examples please see Tables C1-C2-C3.

Implementation. The overarching, continuously updated, living review publication will link the dataset-specific publications (peer-reviewed or preprints) where smaller teams can document datasets by building on existing guidelines [47, 154], and extending them by documenting resources that provide additional evidence related to responsible use of these datasets, such as demographic biases (Section 5) or shortcuts (Section 6). These datasheets could be hosted on Zenodo [36] which allows each datasheet to have its own Digital Object Identifier (DOI). To enhance interoperability and findability, we could package the documentation in a structured format like Croissant [3]. At regular intervals, we can update the overarching living publication, linking to the newly created datasheets.

The SQL database consists of three tables (see Fig. C2): datasets, papers documenting their research artifacts, and the dataset-artifact relations. Our interactive demo http://130.226.140.142 (built with PostgreSQL, Python, Streamlit and the `st_link_analysis` package), supports dynamic dataset-artifact exploration. For additional technical details, see Section C. In the demo, we present 24 research artifacts related to 16 datasets, which include additional annotations, evidence of shortcuts, and derived datasets, focused on two applications: skin lesions and chest X-rays.

Maintenance. It is unrealistic to expect that we as authors of this work would be able to cover all areas of medical imaging, to maintain the living review in perpetuity, nor to ensure that all researchers are aware of its existence. We therefore propose a structure that promotes collaboration, incentivizes better practices around dataset documentation and citation, and regularly informs (new and existing) dataset users about new evidence. Further versions of the living review will require contributions from the community. During data-centric initiatives (in-person and online to include participants who are typically underrepresented at in-person events), we invite researchers to contribute to the living review. A limitation of the current proposal is that we do not yet integrate any mechanism for quality assessment, such as a data steward who will monitor the accuracy of the contributions.

3 Considerations for creating a dataset

Relevance to the living review. Researchers can access the recommended documentation to better understand decisions made throughout the dataset creation process, including the translation from the clinical problem, balancing metadata sharing with patient privacy, and considerations regarding dataset size.

Translation of clinical problem to ML problem. ML applications in healthcare are often related to specific parts of the treatment of a patient. Consequently, the development of a dataset will depend on translating these clinical challenges into ML problems that can be evaluated with ML metrics. This translation requires aligning expectations among various stakeholders in the multidisciplinary team while also establishing a utility criterion, for example, a reduction of workload of the radiology department [175, 196]. The right definition of the ML problem is essential for determining the necessary data (whether newly collected or already existing, such as from retrospective studies), designing an appropriate labeling process, assessing the usefulness of the proposed methods, and understanding the potential limitations of such evaluations from a clinical standpoint [149]. This step also helps to identify potential risks or pitfalls in the data creation process, where biases or shortcuts may be introduced. Importantly, we must remember that an ML dataset is always a *proxy*, and therefore achieving high performance on the ML task, such as detecting cancer, does not necessarily translate to desired outcomes, such as detecting cancer at earlier stages and reducing patient mortality.

Design of ground truth. Data is never truly raw or objective [38, 52, 198]. While we might hope for medical imaging data to be different, and only capture objective reality, this is not the case. Research shows that medical datasets are defined through the painstaking work of multidisciplinary teams within specific clinic, geographical, legal and socioeconomic contexts [115, 120, 198].

The processes shaping medical imaging datasets begin before any data is collected [198]. In particular, regulatory constraints determine what data can be collected and predetermine its purpose. However, issues arise during data acquisition, as data collection is often poorly defined, resulting in variable quality. In contrast, clinical trials follow more regulated protocols. Conducting such structured pilot studies can help refine and standardize consistent acquisition protocols. The context of creation and use direct the design of the datasets and can be accounted for through purposeful investigation of local assumptions and meanings embedded in the dataset. Similar to computer vision datasets [162], commercial and operational pressures hint at the fact that medical imaging datasets have their own politics and are created with a specific purpose, whether commercial or public. To this end, the development of a data management plan is increasingly required to streamline the data lifecycle process [73].

Impact of labeling process on data quality. The consequences of labeling decisions are better understood than those of earlier stages and include, among others, the clinical relevance of ML models [125, 126], the proliferation of social inequality and exclusion [98], and the impact on the performance of trained ML models [17]. To address those challenges, we must look into the processes, guidelines, and incentives of labeling [40], as well as the interpersonal and organizational structures of entities responsible for that work [115, 120]. For example, epistemic differences within multidisciplinary teams creating medical imaging datasets – such as misunderstandings regarding clinical terminology like "opacity" in chest X-rays, with different meanings across countries or lacking direct translations – can affect the clinical outcomes of developed models [198]. Overlooking design choices in medical datasets can lead to misalignment between ML systems and the real-world needs, (see also Section 4). Recognizing these datasets as constructed [145], not objective, is essential to their fair and equitable use in medical practice.

Trade-off between metadata and patient privacy. Metadata such as demographics is essential for evaluating the robustness and fairness of ML models. However, the need for detailed metadata often clashes with the imperative to protect patient privacy, as even anonymized data can carry risks of re-identification and misuse. We do not cover this in detail in this work, but various methods for federated and privacy-preserving ML have been developed, for example [86, 153].

Data size. ML systems are often expected to perform better with more data [87], as has been both observed in practice, and shown by statistical learning theory [183] wherein error tolerance, statistical dependency of the samples, data dimensionality, and model capacity all interact with the model performance. Datasets in medical imaging have grown from hundreds to thousands, with the largest public datasets like CheXpert [76], MIMIC-CXR [82], and Emory breast [78] with up to hundreds of thousands of patients. This contrasts with general computer vision datasets, which often contain millions of images and serve as the basis for many empirical findings. In industry, medical datasets are also scaling up to millions, offering advantages for model development.

With (smaller) public datasets, a common solution is to inject additional knowledge from another source into the dataset, such as domain knowledge provided by experts, or using data or representations from a different source. Transfer learning [20, 135] is therefore often used in medical imaging. A common approach is to fine-tune models pretrained on ImageNet [157], although recent results show this strategy is more sensitive to shortcuts [84] than if training on RadImageNet [114], a recent dataset with a million images from different radiological modalities.

Given the various factors contributing to the interaction between the data and the learning performance, there are no guarantees that larger datasets will necessarily result in better target models. For example, merging datasets from different sources can lead to shortcuts and biases [23, 169]. In this regard, dataset distillation [197] which aims to reduce the size of the data while maintaining or improving its representativeness, could lead to more robust algorithms, despite the "bigger is better" intuition.

4 Data annotation practices and quality

Relevance to the living review. Linked documentation of the annotation process (annotation guidelines + annotations) reflects observer variability and trade-offs between annotator expertise and cost.

Labeling tasks in medical imaging depend on the context of the disease, the task itself (classification, segmentation, etc) and how the target labels are acquired. For example, ground truth labels for lesions or nodules could be confirmed via biopsies, while other (gold standard) labels or annotations could be based on interpretation of the experts or other annotators. Here we discuss two important considerations which vary with the task: inter-observer variability and expertise-cost trade-offs.

Inter-observer variability. Establishing observer variability of the task at hand is crucial before data collection and annotation, as it helps set accuracy targets, determine the required number of annotators, estimate the needed data, and assess variability. Observer variability is measured by having multiple annotators curate a small set of representative examples and then calculating agreement for binary classification, or metrics like Dice score for segmentation [109].

Notably, observer variability varies greatly according to the task, anatomical structures and their sizes, image type, and expertise level [83]. For example, mostly healthy large organs, e.g., the lungs, the liver, and the brain, imaged on volumetric scans will have a low observer variability, while small structures, e.g., lung nodules, will show larger variability.

Observer variability is also influenced by inconsistent training across institutions (e.g., [6]) and insufficient guidelines for annotations [144]. Comprehensive labeling instructions [144], task-specific training [27], and a structured feedback loop can help mitigate this, but are not always feasible if the annotation task is not part of the clinical workflow, and therefore might be more often used in industry datasets.

Expertise and cost trade-offs. Selecting the right annotators for a task is another critical consideration [143], and given the growing demand for annotated data, careful thought must be given to matching the annotators' skills to the task's requirements, both in terms of accuracy and detail of the annotations needed, and domain-specific contexts, for example when diseases have different prevalence across countries.

Domain expertise of annotators can vary from clinicians to laypersons. *Expert*-annotated data can be collected from (retrospective) studies at hospitals, although some types of annotations needed for ML (such as granular annotations like image contours) may not be created as part of the clinical workflow. In such cases only weakly-labeled data might be available, and/or additional annotations might be created by other experts or graduate students for example. In industry, the commonly cited phrase "Garbage in - garbage out" prompts companies to allocate significant resources to ensure high quality annotations and curation, possibly hiring domain-specific experts through their customer base. This can result in datasets with hundreds of subcategories rather than a limited set of high-level categories, more granular annotations, and annotation workflows with various quality control measures.

Medical expertise is not always required for medically-related annotation tasks. For example, studies shown that laypersons are able to detect surgical instruments in laparoscopic images [108], and several other successful results with *crowdsourcing* in medical imaging have been reported [132], although often missing details about the annotation process. Additionally, annotations for training data typically do not require the same level of accuracy and detail as testing data used for evaluating the generalizability of the developed algorithms. This can allow using weakly-labeled training with scribbles [103, 106] or bounding boxes [24, 129], or in 3D data, leveraging sparse annotations only from a few slices within a volume [140]. Taken together, these strategies can allow employing novice annotators for annotating training data while reserving domain experts for curating the test set.

It is crucial to emphasize that despite the apparent advantages of crowdsourcing approaches, there is a lot of *hidden data work*, i.e., the labor involved in data collection and annotation is often invisible and undervalued [137, 158]. There are various reports of unethical practices, for example by tech companies, with regard to data workers who might be dealing with disturbing images or language. The motivations and conditions of workers annotating medical images might be different, but crowdsourcing studies often do not provide this information [132].

Expertise vs. cost trade-offs: (semi-)automatic approaches. Given the cost of annotation and the growing need for large datasets, various automated methods for annotation have been proposed. Computer vision methods can be used to extract additional features from the image, for example asymmetry, border irregularity or Fitzpatrick skin type in skin lesions [60, 148], and used as additional labels, for example via multi-task learning. Natural language processing (NLP) techniques have been proposed to extract diagnostic labels from unstructured clinical reports. However, this strategy has been shown to introduce labeling errors, for example in chest X-rays [123, 185]. Large language models (LLMs) have shown groundbreaking performance in the general language domain and are expected to unlock new possibilities for analyzing medical texts [174]. However, due to domain-specific challenges (language, contextual nuances, and the ambiguity inherent in medical terminology) their effectiveness is still limited, see for example studies for radiography or free-text CT or MR reports [39, 97, 119].

Hybrid approaches such as active learning [11, 44], active label correction [8] and hard sample mining (collecting additional "hard samples" from specific device manufacturers or rare disease subtypes) offer opportunities to combine the advantages of both human annotation and automated methods. While such methods are popular in literature, in academic papers the human annotators are sometimes simulated by giving the algorithm access to the existing labels in the data, while in industry, the (re)-labeling process is more often done with domain experts. This also allows industry datasets to be more dynamic than publicly available datasets. However, due to the proprietary nature, it is difficult to comment the cost vs. quality trade-offs, and what learnings public dataset creators can extract from this.

Final remarks. In general, it is crucial to remember that any annotation method, human or otherwise, will always be based on some assumptions. For example, methods which directly generate synthetic images and labels, are always based on some underlying data and will inherit its biases (and one could also question why the data needs to be generated explicitly if the data generating process is known). Without comprehensive documentation of the annotation process and its assumptions – which might not always

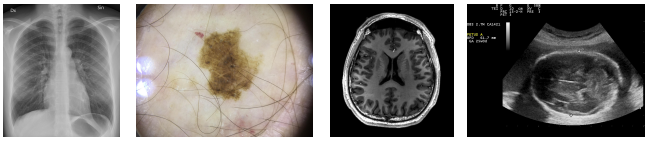

Figure 3: Case studies: (a) A normal posteroanterior chest radiograph of a patient with no visible signs of injury, showing letters that indicate left or right. (b) A malignant melanoma dermoscopic image from ISIC dataset, featuring dark corners and areas of intense brightness. (c) A cross-sectional T1-weighted MRI of a healthy human brain. (d) A fetal ultrasound image displaying the head of the fetus, where the text "BDP" refers to the head diameter, and "GA" indicates gestational age. The images (a), (c) and (d) are sourced from Wikimedia, and (b) from ISIC.

be explicit – can lead to failures when the assumptions do not hold for the data at hand. Another crucial point is that, for example, reporting bias will inevitably affect what we learn about about different annotation methods. A living review where authors of new studies (which might not be noticed due to factors like the publication venue) contribute to evidence around a specific dataset, could help reduce these problems.

5 Inside a dataset: demographics

Relevance to the living review. We present four case studies (chest x-rays, skin lesions, brain MRI and fetal ultrasound), each with unique properties. We propose addressing these factors in our living review, emphasizing the importance of the documentation of the study design and patient demographics.

The potential of overfitting to established benchmark datasets is a long-standing debate in the ML community [57, 71, 101]. While the past few years have seen a surge in public datasets, they typically lack demographic metadata about the subjects. This leads to a host of problems. If the data does not come with demographic information, researchers cannot assess whether the datasets and models trained on them have demographic biases. As a result, not only the resulting algorithms, but also what we learn about ML development, can be tainted by demographic biases without our knowledge. Equally important, research on algorithmic bias and fairness, which needs demographic metadata, has access to very few datasets. As a result, large amounts of research resources have been dedicated to a very small set of case studies, whose particularities – for better and for worse – drive the progress of the research field. Here we present four case studies, illustrated in Fig. 3, further clinically relevant details in Section B.

Case study 1: chest X-rays. Chest X-rays are the most commonly performed radiologic examinations worldwide [178], requiring significant expertise for accurate and meaningful interpretation [181]. ML revolutionized chest X-ray diagnosis, highlighted by the release of NIH-CXR14 dataset [185] and CheXNet's model claiming radiologist-level pneumonia detection [146]. However, these claims have been criticized for relying on shortcuts [81, 125], and for low inter-rater agreement [25]. Subsequently, additional datasets such as CheXpert [76], MIMIC-CXR [82], and PadChest [14] have become widely used in the research community.

With the primary purpose of diagnosing and/or monitoring pathologies, chest X-ray datasets often include demographic information like age, gender, sex, race and ethnicity to analyze potential biases[1]. While age and gender are typically included, (self-reported) race or ethnicity are only available in a few datasets, such as MIMIC-CXR and CheXpert [136]. Age is generally skewed toward older populations (PadChest median age 62, MIMIC-CXR largest group aged 60–80 [200]). Age can be predicted from chest X-rays [74], which can lead to favoring well-represented age groups. Even without significant gender imbalance, performance disparities between genders persist [96, 166, 167]. Balancing the data has proven ineffective, and studies have ruled out causes such as under-representation [96], physiological differences [188], and shortcut learning [81, 125, 130]. Less is known about the effect of label errors, which have different effects on diagnostic labels – in particular, the "no finding" label is known to often be associated with follow-up images of patients [124]. Performance disparities between racial groups favoring white individuals have been noted [166]. ML can infer protected attributes like race, despite this being a challenging task for human experts [51]. Recent research [53] suggests that fine-tuning on specialized datasets alone cannot reduce the influence of these protected attributes.

Case study 2: skin lesions. In recent years there has been significant growth in ML for dermatology [121], including tasks such as classification [133], segmentation [117], and lesion localization [110]. This surge may be largely explained by the availability of public datasets, such as Fitzpatrick17k [60], PAD-UFES-20 [134], and HIBA [152] – as well as the International Skin Imaging Collaboration (ISIC) [77], which aggregates over 490K images from datasets such as HAM10000 [177] and BCN 20000 [67].

Despite the progress, the under-representation of demographic groups in these datasets limits the generalizability of ML [26]. Skin lesions manifest differently across populations, influenced by factors like skin tone, genetic background, age and UV light exposure [187]. Nonetheless, most public datasets are limited in terms of geographic and skin tone diversity, with a predominance of lighter-skinned patients (Fitzpatrick I to III) [59, 186], potentially resulting in models that underperform for under-represented groups. For example, melanoma – the deadliest type of skin cancer – is much more prevalent in lighter-skinned individuals but can also affect those with darker skin tones, who might then be misdiagnosed [58].

Recent efforts to diversify skin lesion datasets, such as the inclusion of PAD-UFES-20 [134] and HIBA [152] in ISIC, have improved the representation of Latin American individuals, a region that was previously under-represented. However, there is still a strong lack of representation in terms of diversity of skin tone. Addressing these disparities is a global challenge that requires a collaborative effort from the research community, drawing on diverse perspectives and contributions from different backgrounds and regions worldwide.

[1]We recognize that both gender and sex, as well as race and ethnicity, are distinct and not binary. However, datasets often do not document which variable was collected and/or use the terms interchangeably. We recognize that there are complexities in diagnosis when individuals belong to multiple categories. When possible, we use the terms used by the original authors throughout this paper.

Case study 3: fetal ultrasound. Ultrasound is the fundamental imaging modality for antenatal care. The acquisition process consists of a physical examination with an ultrasound probe looking for specific 2D slices called standard planes. Several sources of variation affect the image quality, such as the maternal body mass index (BMI) or the experience of the sonographer acquiring the scan. Another important factor is the ethnicity, which is associated with variations in the normal fetal growth according to several multi-ethnic studies [33, 128, 170].

The few available datasets are the Fetal Planes DB [13] and data from the HC18 [180], FH-PS-AOP [79], and ACOUSLIC-AI [159] challenges, with none, to date, providing demographic information. More diverse datasets with detailed demographics, including BMI and ethnicity, image quality information, and expertise of the clinicians, are needed to develop fairer models and improve prenatal screening. This is true for low-, middle- and high-resource countries. In low and middle-income resource settings with poorer image quality due to portable devices this is crucial to reduce maternal complications and fetal mortality [189]. A recent Danish study [116] shows that across demographic subgroups, deep learning algorithms improve birth weight estimates from fetal ultrasound compared to clinical standard measurements extracted from those same ultrasound images – potentially because the clinical standard measurements are unable to use additional image content to make up for suboptimal ultrasound planes. Future studies should therefore look with care at how both image quality and study design affect performance across groups – for both ML and more traditional predictive approaches.

Case study 4: neuroimaging. MRI is the third most commonly performed imaging modality after CT and X-rays. Its superior soft tissue contrast enables detailed visualization of brain anatomy, making it ideal for detecting abnormalities, and most public MRI datasets for ML research focus on the brain [32]. Key application areas for brain MRI and notable datasets, include neurodegenerative diseases (OASIS [92, 94, 111, 112]), brain cancer (BraTS [5], LUMIERE [173], Ocana [127], and TCGA-GBM [161]), and stroke, (ISLES 2022 [66], ATLAS v2.0 [102], among others).

The conversion from the standard clinical imaging format, DICOM, to NIfTI or other formats is complex and error-prone [100] and depends on how different formats implement DICOM standards. Moreover, many public neuroimaging datasets undergo extensive preprocessing prior to release, often to standarize datasets for open challenges, and ensure that model evaluations focus on algorithm performance rather than the effects of preprocessing. These factors may explain why the neuroimaging community –unlike other disciplines– has made progress in advancing data-sharing practices by adopting standards and tools like BIDS [56], DataLad [63], NITRC.org [90] and OpenNeuro [113].

Demographic reporting in brain MRI is generally poor. For example, in US studies from 2010 to 2020, 77% report sex, but only 10% and 4% report race and ethnicity, respectively [172]. Studies show performance disparities in models by sex and race, with black females being most affected [31, 75]. Recent efforts to diversify brain MRI datasets include the addition of children (BraTS-PEDS [88]) and Sub-Saharan African populations (BraTS-Africa [2]). Future work should focus on preserving raw data authenticity, enhancing diversity, and more complete metadata.

6 Inside a dataset: shortcuts

Relevance to the living review. We suggest documenting and annotating new evidence related to errors or shortcuts in existing datasets.

Different terminology in the literature refers to shortcuts: confounders, spurious correlations, hidden stratification, etc. Shortcuts are decision rules that perform well on benchmark data but fail to transfer to challenging test cases, often with out-of-distribution data [50]. When shortcuts fail, they result in biases and misdiagnosis, for example chest pain in women as anxiety or heart burn [179], misconception that Black patients have high pain threshold [68], or delayed referrals for minority patients who may be judged as malingering or drug seeking [62]. Many shortcuts in medical imaging have been shown, often when the ML model memorizes irrelevant clinical characteristics, like the hospital where the patient was scanned [23]. Hence, when the shortcut is missing, the model performance drops. For example, the model can rely on proxies, such as chest drains for pneumothorax [81, 125], radiographic markers containing scanning locations for pneumonia (especially in the intensive care unit, where the prevalence of pneumonia is high) [199], dark corners or rulers for skin lesions [10], or patient positioning for COVID-19 [29].

As we develop more novel datasets and dataset derivatives such as masks, embeddings and foundation models that extract features from multiple datasets, feature visualizations have demonstrated that the embeddings show clear separation of sex and risk groups, showing that these models also encode these characteristics [53] and lead to bias. It is important to note that these may be challenging to evaluate because of the complexity of "model as a dataset" where the original dataset may not be available for inspection, making it difficult to know whether the shortcut learning occurs due to the data, the bias of the model, or bias introduced by a fine-tuning dataset. It is therefore imperative to interpret the model outputs carefully and audit the false positives and false negatives using domain expertise.

Taxonomy of shortcuts. Shortcuts can arise from spurious correlations (e.g. noise patterns or differences in intensity distributions across various scanners), demographic attributes, and their interactions. The Medical Imaging Contextualized Confounder Taxonomy (MICCAT) [84] (see Fig. A1) helps identify the shortcuts' origin and mitigation strategies. MICCAT extends beyond traditional demographic attributes to include a broader set of confounders that are domain- and context-specific confounders at both patient and environment levels.

Patient-level confounders include demographic attributes (sex/gender [1, 96], age [1], and race/ethnicity [51]) and anatomical factors related to organs or conditions (BMI, tissue/breast/bone density). While demographic factors are standard in bias analysis, anatomical variations may form subgroups where models underperform, and their identification often requires analysis beyond standard demographic characteristics [184].

Environment-level confounders include external and imaging confounders. External confounders involve visible elements like chest drains [81, 125], pen marks [191], patient positioning [29], or text overlays [104], creating localized artifacts. In contrast, imaging confounders result from the imaging process (acquisition devices or parameters, noise, or motion artifacts), leading to global artifacts that may be imperceptible to the human eye. Systematic variations in exposure setting for chest X-rays [95] or different characteristics of imaging equipment across centers [23] can introduce shortcuts, causing models to rely on acquisition-specific factors rather than genuine clinical factors.

Documenting what type of shortcut (patient or environment, external or internal) a dataset has in our living review would enable researchers to learn across applications. For example, researchers working on skin lesions could identify and learn from relevant studies and experiments from ophthalmology, even if these studies would not be findable with traditional search strategies.

7 Data lifecycle

Relevance to the living review. Our framework incentivizes researchers to adopt better dataset management practices. We recommend using persistent identifiers and storing, and versioning of datasets for reproducibility, and licensing and proper tracking for author attribution.

Effective research data management is crucial for reproducibility, re-usability, and efficiency. The data lifecycle consists not only of data acquisition and analysis, but rather of data acquisition (see Section 3), data organization and standardization, data and metadata annotation (see Section 4), data management and tracking during analysis, and ultimately, data maintenance. A comprehensive overview, with an example from neuroimaging, is provided in [122].

Data standards. Since research, and especially medical research data, is often collected for one purpose, but then reused for another, data standardization is crucial to enhance re-usability. Converting to established data standards directly after acquisition makes it easier to ensure transparency and provenance of the data. Part of standardizing the data is also often to add metadata and perform additional annotations. Here proper documentation of where/how the metadata and annotations were acquired are important to ensure the usefulness of the additional data.

Data standardization is not always easily achieved and very application dependent (see also Section 5). One example, the Brain Imaging Data Structure (BIDS) [141] exemplifies the role of data standards in neuroimaging research. By providing a consistent framework for organizing and describing datasets, BIDS has enabled large scale data sharing, reproducibility, and collaboration across the scientific neuroimaging community, facilitating development of interoperable tools and workflows [55], streamlining data analysis and reducing errors. This serves as an example for developing data standards for other medical imaging applications.

Sharing data via data repositories. Sharing data with the community is crucial for ensuring reproducibility, maximizing the scientific insights derived from the same set of participants, empowering researchers to reuse data, develop innovative analytical methods, refine scientific hypotheses, and conduct large-scale meta-analyses. Hosting services play a crucial role in advancing research and the development of ML models. Examples include repositories where researchers can access and share data (PhysioNet [54], Zenodo [36], OpenNeuro [113]) and host benchmarking competitions (Grand Challenge [16], Kaggle [85], HuggingFace [72]). Several of these are referred to as Community Contributed Platforms (CCPs).

Whenever a dataset is updated or changed, it should be versioned and the changes documented. While this functionality is available on some open data sharing platforms, it is not widely adopted. For example, HuggingFace provides a way to track the history of files and versions on their website, but does not explicitly version their datasets. Moreover, the issue of dataset version tracking is even more critical with copies of datasets on CCPs, where documentation, including pre-processing steps, is frequently lacking [80].

Finally, open source tools for decentralized data sharing and processing, such as DataLad [64], facilitate data management and analysis tracking, and cost less than industry-standard solutions. Data maintenance costs can be optimized by considering dataset use. In neuroimaging for example, only a small percentage (<10%) of datasets in large repositories are accessed after two years [139], thus rarely accessed data could be archived, i.e., stored in "cold" / low-cost storage, reducing the data maintenance costs.

Licensing. A concern for medical imaging datasets distributed on open repositories is their appropriateness for reuse. This has been partly addressed by the FAIR (Findable, Accessible, Interoperable, Reusable) principles, a data stewardship model for scientific data management and governance [190]. Nevertheless, compliance with FAIR is a complex process that requires a series of activities which are dependent on expertise and oversight of the researchers distributing the datasets, e.g., metadata release and licensing [80, 194].

A license is a standardized statement that specifies the permissible uses of the data and the associated constraints for the end user, with the most prevalent being the Creative Commons (CC) licensing suite. CC licenses allow specifying requirements for how to cite their attribution, as well as the conditions for the dataset reuse and potential re-sharing options for derivatives. Yet, there is still a lack of appropriate licenses for popular datasets [80, 107]. Possible reasons include the legal expertise of the dataset creators, the lack of auditing from the CCPs once the datasets are distributed, and the complexity of data protection laws, which often lead to confusion on how to license anonymized datasets which have already been approved in an initial release phase.

It is important to emphasize that adhering to the FAIR principles does not require fully releasing the data under open licenses – the data should be "as open as possible, as closed as necessary". Sensitive data can be shared under a Data Use Agreement (DUA), provided clear and precise guidelines are established for compliance with the DUA. As a step forward, conferences could offer workshops on data protection and licensing.

Tracking of dataset use. Tracking the use of datasets throughout publications and experiments is important for several reasons: giving credit to dataset providers and annotators, identifying changes to datasets due to errors or ethical concerns [138], and

ensuring fair comparisons of new methods. Without proper tracking, it becomes extremely difficult to identify which methods are impacted by dataset-related issues.

Sources like PubMed, open citation tools like OpenAlex [142] or Semantic scholar [42], and platforms like Papers with Code help track dataset use through citations and benchmark results. However, citation practices often fail to fully capture dataset use, as datasets are sometimes cited via URLs or footnotes [65, 171]. Some authors neglect citation altogether, and some datasets lack clear citation instructions [80]. Journals and conferences can improve citation tracking by requiring a "Data availability" section, a practice already adopted by some. Finally, current citation practices lack sufficient context to confirm that a dataset is actually used in the experiment, and not just cited as an example. Modifying the citation graph to be more fine-grained, as proposed in [12] and in our living review, would provide more context about dataset use.

8 Discussion and conclusions

We have comprehensively explored various factors that shape a medical imaging dataset, from translating the clinical problem into an ML problem to key considerations for creating high-quality datasets. These considerations include defining the ground truth, determining the necessary amount of data, and detailing the annotation process. We also examined the contents of datasets throughout four case studies, highlighting context-specific clinical factors, and describing a taxonomy of shortcuts. Furthermore, we discussed scientific data management and stewardship for the data lifecycle.

Our living review addresses these challenges by proposing a structure that encourages collaboration, better dataset documentation and citation practices, and regular updates on new evidence. Through a living review process, we facilitate reflection on decisions made throughout the dataset creation process, ensuring thorough documentation of annotation guidelines and annotations. The review promotes collaboration to extend open datasets through the addition of new annotations, and emphasizes the importance of including the study design and patient demographics. Our proposal considers persistent identifiers, dataset versioning, licensing, and ensuring proper citation standards, supporting researchers in adopting best practices for dataset management and accessibility.

We acknowledge several limitations of our proposed living review, including covering all areas of medical imaging, the effort for ongoing maintenance, and the challenge of ensuring awareness among all researchers. While we emphasize the importance of patient metadata, like demographics (see Section 5), we do not explore data privacy in detail.

Learning from other communities. A medical imaging dataset is often a proxy for the actual healthcare problem and can lead to inaccurate diagnoses, but this is often not documented in the datasets or trained models, necessitating a living review of datasets as we propose. It is also clear that medical imaging can benefit significantly from other communities, for example clinical prediction research and epidemiology.

Data sharing in clinical prediction research can be improved. A recent review in ML for oncology reported that only 2 out of 46 studies shared their data [22], and data availability statements (reported in 76% of cases) often stated that data was available on request. However, this practice is frequently done merely to meet editorial requirements, rather than to provide easy access to data [156, 160], despite the introduction of the FAIR principles (see Section 7).

Algorithmic fairness was recently incorporated in the updated TRIPOD+AI statement, which provides guidance for reporting clinical prediction studies [21]. Researchers publishing clinical prediction models are specifically asked to report any methods used to assess and address model fairness. The reporting of open science practices (e.g., accessible study protocols, study registration, data and code sharing) is also included as a new focus in the updated statement.

Systematic reviews of clinical prediction studies often use the PROBAST checklist to assess bias risks and model applicability [118]. One fundamental point relevant to medical imaging is about appropriate inclusion and exclusion of participants. For example, including images of participants who already progressed to a severe disease state and would never be assessed with ML, might reduce generalizability of the model. This could explain why clinical prediction models are abundant in the literature, but much rarer in the healthcare sector [182].

Another question in PROBAST is whether predictors/images were similarly assessed for all participants – if not, shortcuts (see Section 6) can occur. Shortcuts are similar to confounding in epidemiology, where, for example, sex is associated with both access to treatment and disease prevalence. Developing a simple benchmark including only metadata could provide useful information on the magnitude of the images' additional predictive value. Confounding can also be dealt with stratification, but this can lead to smaller and smaller datasets, which is an issue given the already low sample sizes in prediction research [30]. While a list of potential shortcuts in the dataset documentation is helpful, it cannot replace domain expertise. However, in a survey of ML for (non-image) medical data, 35% of the studies did not have any authors with a medical affiliation [34], and one could speculate that it is even lower at ML conferences.

More generally in healthcare, the STANDING (STANdards for data Diversity, INclusivity and Generalizability) Together initiative [46], involving many clinicians, aims to reduce ML health inequalities by providing recommendations to increase transparency and generalizability of datasets. Consensus recommendations were developed in two parts by an international multidisciplinary team [4]: (1) guidance on how dataset creators should document their datasets, emphasizing transparency about the dataset's origin, composition, limitations, and potential biases, and (2) how data users can best utilize these datasets to minimize potential harm and ensure equitable performance across different population groups.

To conclude, the value of interdisciplinary collaborations and sharing best research practices across disciplines is crucial to develop clinically relevant applications that high quality datasets provide the basis for.

Future outlook. One of the core problems surrounding datasets are the current incentives that are rewarded and optimized for in ML. While we have a proposal for how to incentivize and maintain our living review, it will not fix all the problems since they require changes across many research fields.

Although more data-centric initiatives, such as the Datasets and Benchmarks Track at NeurIPS, are emerging, it is clear that more attention to data work is needed. Perhaps in this context it is worth mentioning that despite the importance of the topic and a diverse group of researchers, our own workshop proposal was initially rejected by two different venues. However, given recent progress, we remain hopeful data work will be more addressed and better acknowledged in the future, and we are in the process of applying for funding for related research and networking events, to ensure the continuity of this work.

Looking ahead, established peer-review processes in conferences could be adapted to support the publication and recognition of data artifacts, assigning them DOIs similar to standard papers. Conferences like Medical Imaging with Deep Learning (MIDL)[2] could invite data artifacts as papers for a special track, similar to the Datasets and Benchmarks track at NeurIPS. However, a more scalable approach would involve multiple conferences adopting a "rolling" review style, as has been adopted in the natural language processing community [150]. Another interesting option would be for conferences to invite collaborative contributions to selected datasets, similar to medical imaging competitions like those hosted at the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)[3]. In the existing scenario, the organizers select a dataset, competition participants develop algorithms, and after benchmarking, both organizers and many participants typically co-author a paper about the results. Contributions to a datasheet would, of course, not be trivial to "benchmark"; however, there would not need to be a "winner-takes-all" mentality. If different teams highlight different aspects of the datasets, their contributions could simply be combined.

With our work, we want to promote a cultural shift toward the responsible use of high quality datasets, created through collaboration with relevant stakeholders or clinical experts. We therefore invite everyone involved in and affected by healthcare ML to contribute to our living review and other dataset efforts.

Ethical considerations statement

We propose a living review framework to connect publicly available medical imaging datasets with associated research artifacts for the research community. Our proposal is based on publicly available data, and no additional private or sensitive data were collected. Researchers who want to contribute with new evidence to update our database should make sure that their annotation or documentation metadata is compliant with the General Data Protection Regulation (GDPR) in the European Union (EU), the EU AI Act, and other relevant national and international legislation governing data privacy. We briefly discuss considerations about the trade-off between metadata and patient privacy in Section 3, and considerations about FAIR principles and licensing in Section 7.

Adverse impact statement

The goal of our living review framework is to benefit the research community by improving the usability and accountability of publicly available medical imaging datasets. By providing insights into annotations, errors, and additional findings, our framework helps reduce the risk of misinterpretation or reliance on flawed data. However, without proper stewardship, moderation and ongoing maintenance, there is a risk that insufficiently validated datasets or artifact could be included. It is important to note that our proposal is a data/literature exploration tool for researchers, and should not directly be used for clinical decisions.

[2]https://2025.midl.io/

[3]https://conferences.miccai.org/2025/en/challenges.asp

Author Contributions

Authors, except for the first and last, are listed in alphabetical order. Contributions follow the CRediT author statement.

Conceptualization, Methodology, Writing - Original Draft & Final Draft: Veronika Cheplygina, Amelia Jiménez-Sánchez.

Data Curation, Visualization, Software: Amelia Jiménez-Sánchez.

Supervision, Funding Acquisition: Veronika Cheplygina.

Writing - Review & Editing: all authors contributed, with individual section contributions listed below in alphabetical order:

- *Introduction*: Veronika Cheplygina, Amelia Jiménez-Sánchez, Maria A. Zuluaga
- *Proposal: a living review of medical imaging datasets*: Veronika Cheplygina, Amelia Jiménez-Sánchez
- *Considerations for creating a dataset*: Veronika Cheplygina, Camila González, Steff Groefsema, Amelia Jiménez-Sánchez, Leo Joskowicz, Melih Kandemir, Hubert Dariusz Zając.
- *Data annotation practices and quality*: Veronika Cheplygina, Alessa Hering, Leo Joskowicz, Thijs Kooi, Tim Rädsch.
- *Inside a dataset: demographics*: Víctor M. Campello, Aasa Feragen, Jorge del Pozo Lérida, Andre Pacheco, Mauricio Reyes, Nina Weng, Jack Junchi Xu.
- *Inside a dataset: shortcuts*: Enzo Ferrante, Judy Wawira Gichoya, Dovile Juodelyte.
- *Data lifecycle*: Natalia-Rozalia Avlona, Sarah de Boer, Melanie Ganz-Benjaminsen, Bram van Ginneken, Amelia Jiménez-Sánchez, Théo Sourget.
- *Discussion*: Veronika Cheplygina, Adam Hulman, Amelia Jiménez-Sánchez, Livie Yumeng Li, David Wen.

Acknowledgments

This project has received funding from the Independent Research Council Denmark (DFF) Inge Lehmann 1134-00017B. The "In the Picture: Medical Imaging Datasets" workshop was funded by DFF and the Danish Data Science Academy (DDSA) with Grant ID 2024-2342. We extend our gratitude to the speakers and participants of the "Datasets through the Looking-Glass" webinar, who have helped to shape this research. EF gratefully acknowledges the support of the Google Award for Inclusion Research (AIR) Program. AJS was financed by the DFF grant MMC. AH is employed at Steno Diabetes Center Aarhus that is partly funded by a donation from the Novo Nordisk Foundation. AH is supported by a Data Science Emerging Investigator grant by the Novo Nordisk Foundation (NNF22OC0076725). TR was supported by a scholarship from the Hanns Seidel Foundation with funds from the Federal Ministry of Education and Research Germany (BMBF). DW received funding from the Jill and Herbert Hunt Scholarship, University of Oxford. MAZ is funded by TRAIN (ANR-22-FAI1-0003-02). We thank Freepick for the icons in Fig. 2.

References

[1] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt EJ Michels, Gerard Schouten, and Veronika Cheplygina. 2020. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. In *MICCAI LABELS workshop, Lecture Notes in Computer Science*, Vol. 12446. Springer, 183–192.

[2] M Adewole, JD Rudie, A Gbadamosi, et al. 2023. The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa). *arXiV preprint arXiV:2305.19369* (2023).

[3] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, et al. 2024. Croissant: A Metadata Format for ML-Ready Datasets. *arXiv preprint arXiv:2403.19546* (2024).

[4] Joseph E Alderman, Joanne Palmer, Elinor Laws, Melissa D McCradden, Johan Ordish, Marzyeh Ghassemi, Stephen R Pfohl, Negar Rostamzadeh, Heather Cole-Lewis, Ben Glocker, et al. 2025. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *The Lancet Digital Health* 7, 1 (2025), e64–e88.

[5] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo, Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K. Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. 2021. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv preprint arXiv:2107.02314* (2021). http://arxiv.org/abs/2107.02314

[6] Anjali Balagopal, Howard Morgan, Michael Dohopolski, Ramsey Timmerman, Jie Shan, Daniel F Heitjan, Wei Liu, Dan Nguyen, Raquibul Hannan, Aurelie Garant, et al. 2021. Psa-net: Deep learning–based physician style–aware segmentation network for postoperative prostate cancer clinical target volumes. *Artificial Intelligence in Medicine* 121 (2021), 102195.

[7] Imon Banerjee, Kamanasish Bhattacharjee, John L. Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N. Patel, Rakesh Shiradkar, and Judy Gichoya. 2023. "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation. *Journal of the American College of Radiology* 20, 9 (Sept. 2023), 842–851. doi:10.1016/j.jacr.2023.06.025

[8] Mélanie Bernhardt, Daniel C Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P Lungren, Aditya Nori, Ben Glocker, et al. 2022. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications* 13, 1 (2022), 1161.

[9] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM.

[10] Alceu Bissoto, Eduardo Valle, and Sandra Avila. 2020. Debiasing Skin Lesion Datasets and Models? Not So Fast. In *Computer Vision and Pattern Recognition (CVPR) Workshops*. 740–741.

[11] Samuel Budd, Emma C Robinson, and Bernhard Kainz. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* 71 (2021), 102062.

[12] Peter Buneman, Dennis Dosso, Matteo Lissandrini, and Gianmaria Silvello. 2021. Data citation and the citation graph. *Quantitative Science Studies* 2, 4 (2021), 1399–1422.

[13] Xavier P. Burgos-Artizzu, David Coronado-Gutierrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. 2020. *FETAL_PLANES_DB: Common maternal-fetal ultrasound images*. doi:10.5281/zenodo.3904280

[14] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vayá. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 66 (2020), 101797.

[15] Santiago Cepeda, Sergio García-García, Ignacio Arrese, Francisco Herrero, Trinidad Escudero, Tomás Zamora, and Rosario Sarabia. 2023. The Río Hortega University Hospital Glioblastoma dataset: A comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM). *Data in Brief* 50 (2023), 109617.

[16] Grand Challenge. 2025. A platform for end-to-end development of machine learning solutions in biomedical imaging. https://grand-challenge.org/. Accessed: 2025-01-17.

[17] Haihua Chen, Jiangping Chen, and Junhua Ding. 2021. Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability* 70, 2 (2021), 831–847.

[18] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science* 4 (2021), 123–144.

[19] Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering* 7, 6 (June 2023), 719–742. doi:10.1038/s41551-023-01056-8

[20] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis* 54 (2019), 280–296.

[21] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* (April 2024), e078378. doi:10.1136/bmj-2023-078378

[22] Gary S Collins, Rebecca Whittle, Garrett S Bullock, Patricia Logullo, Paula Dhiman, Jennifer A de Beyer, Richard D Riley, and Michael M Schlussel. 2024. Open science practices need substantial improvement in prognostic model studies in oncology using machine learning. *Journal of Clinical Epidemiology* 165 (2024), 111199.

[23] Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. 2023. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine Learning for Healthcare Conference*. PMLR, 110–127.

[24] Jifeng Dai, Kaiming He, and Jian Sun. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *International Conference on Computer Vision (ICCV)*. 1635–1643.

[25] Cathrine Damgaard, Trine Naja Eriksen, Dovile Juodelyte, Veronika Cheplygina, and Amelia Jiménez-Sánchez. 2023. Augmenting Chest X-ray Datasets with Non-Expert Annotations. *arXiv preprint arXiv:2309.02244* (2023).

[26] Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. 2021. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatology* 157, 11 (2021), 1362–1369.

[27] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–40.

[28] Jodi S. Dashe, Donald D. McIntire, and Diane M. Twickler. 2009. Maternal Obesity Limits the Ultrasound Evaluation of Fetal Anatomy. *Journal of Ultrasound in Medicine* 28, 8 (2009), 1025–1030. doi:10.7863/jum.2009.28.8.1025

[29] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* 3, 7 (2021), 610–619.

[30] Paula Dhiman, Jie Ma, Cathy Qi, Garrett Bullock, Jamie C Sergeant, Richard D Riley, and Gary S Collins. 2023. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Medical Research Methodology* 23, 1 (2023), 188.

[31] Mahsa Dibaji, Neha Gianchandani, Akhil Nair, Mansi Singhal, Roberto Souza, and Mariana Bento. 2023. Studying the Effects of Sex-Related Differences on Brain Age Prediction Using Brain MR Imaging. In *MICCAI Workshop on Clinical Image-Based Procedures*. Springer, 205–214.

[32] Katharine A. Dishner, Bala McRae-Posani, Arka Bhowmik, Maxine S. Jochelson, Andrei Holodny, Katja Pinker, Sarah Eskreis-Winkler, and Joseph N. Stember. 2024. A Survey of Publicly Available MRI Datasets for Potential Use in Artificial Intelligence Research. 450-480 pages. Issue 2. doi:10.1002/jmri.29101

[33] J. C. Drooger, J. W. M. Troe, G. J. J. M. Borsboom, A. Hofman, J. P. Mackenbach, H. A. Moll, R. J. M. Snijders, F. C. Verhulst, J. C. M. Witteman, E. a. P. Steegers, and I. M. A. Joung. 2005. Ethnic Differences in Prenatal Growth and the Association with Maternal and Fetal Characteristics. *Ultrasound in Obstetrics & Gynecology* 26, 2 (2005), 115–122. doi:10.1002/uog.1962

[34] Andreas Ebbehoj, Mette Østergaard Thunbo, Ole Emil Andersen, Michala Vilstrup Glindtvad, and Adam Hulman. 2022. Transfer learning for non-image data in clinical research: a scoping review. *PLOS Digital Health* 1, 2 (2022), e0000014.

[35] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J Adler, Patrick Godau, Veronika Cheplygina, Michal Kozubek, Sharib Ali, et al. 2022. Biomedical image analysis competitions: The state of current participation practice. *arXiv preprint arXiv:2212.08568* (2022).

[36] European Organization For Nuclear Research and OpenAIRE. 2013. Zenodo. doi:10.25495/7GXK-RD71

[37] Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. 2024. Source-free unsupervised domain adaptation: A survey. *Neural Networks* (2024), 106230.

[38] Melanie Feinberg. 2017. A design perspective on data. In *Conference on Human Factors in Computing Systems (CHI)*, Vol. 2017-May. 2952–2963. doi:10.1145/3025453.3025837

[39] Matthias A Fink, Arved Bischoff, Christoph A Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, and Tim F Weber. 2023. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 308, 3 (2023), e231362.

[40] Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley. 1–164 pages. doi:10.1002/9781119306696

[41] Eitan Frachtenberg. 2022. Research artifacts and citations in computer systems papers. *PeerJ Computer Science* 8 (2022), e887.

[42] Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association* 106, 1 (2018), 145.

[43] Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Martina Aineseder, Diego H Milone, and Enzo Ferrante. 2023. CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *arXiv preprint arXiv:2307.03293* (2023).

[44] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*. PMLR, 1183–1192.

[45] Maria Galanty, Dieuwertje Luitse, Sijm H. Noteboom, Philip Croon, Alexander P. Vlaar, Thomas Poell, Clara I. Sanchez, Tobias Blanke, and Ivana Išgum. 2024. Assessing the documentation of publicly available medical image and signal datasets and their impact on bias using the BEAMRAD tool. *Scientific Reports* 14, 1 (Dec. 2024). doi:10.1038/s41598-024-83218-5

[46] Shaswath Ganapathi, Jo Palmer, Joseph E Alderman, Melanie Calvert, Cyrus Espinoza, Jacqui Gath, Marzyeh Ghassemi, Katherine Heller, Francis Mckay, Alan Karthikesalingam, et al. 2022. Tackling bias in AI health datasets through the STANDING Together initiative. *Nature Medicine* 28, 11 (2022), 2232–2233.

[47] Christian Garbin and Oge Marques. 2022. Assessing methods and tools to improve reporting, increase transparency, and reduce failures in machine learning applications in health care. *Radiology: Artificial Intelligence* 4, 2 (2022), e210127.

[48] Christian Garbin, Pranav Rajpurkar, Jeremy Irvin, Matthew P Lungren, and Oge Marques. 2021. Structured dataset documentation: a datasheet for CheXpert. *arXiv preprint arXiv:2105.03020* (2021).

[49] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[50] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. 2018. Generalisation in humans and deep neural networks. In *Neural Information Processing Systems (NeurIPS)*. 7538–7550.

[51] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. 2022. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health* 4, 6 (2022), e406–e414.

[52] Lisa Gitelman (Ed.). 2013. *"Raw Data" Is an Oxymoron*. MIT Press. Includes bibliographical references and index.

[53] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. 2023. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *EBioMedicine* 89 (2023).

[54] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (2000), e215–e220.

[55] Krzysztof J Gorgolewski, Fidel Alfaro-Almagro, Tibor Auer, Pierre Bellec, Mihai Capotă, M Mallar Chakravarty, Nathan W Churchill, Alexander Li Cohen, R Cameron Craddock, Gabriel A Devenyi, et al. 2017. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Computational Biology* 13, 3 (2017), e1005209.

[56] Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* 3 (6 2016). doi:10.1038/sdata.2016.44

[57] Alexej Gossmann, Aria Pezeshk, and Berkman Sahiner. 2018. Test data reuse for evaluation of adaptive machine learning algorithms: over-fitting to a fixed 'test' dataset and a potential solution. In *SPIE Medical Imaging*, Vol. 10577. SPIE, 121–132.

[58] Matthew Groh, Omar Badri, Roxana Daneshjou, Arash Koochek, Caleb Harris, Luis R Soenksen, P Murali Doraiswamy, and Rosalind Picard. 2024. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine* 30, 2 (2024), 573–583.

[59] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. 2022. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–26.

[60] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Computer Vision and Pattern Recognition (CVPR)*. 1820–1828.

[61] Hao Guan and Mingxia Liu. 2021. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering* 69, 3 (2021), 1173–1185.

[62] Adil H Haider, Eric B Schneider, N Sriram, Valerie K Scott, Sandra M Swoboda, Cheryl K Zogg, Nitasha Dhiman, Elliott R Haut, David T Efron, Peter J Pronovost, et al. 2015. Unconscious race and class biases among registered nurses: vignette-based study using implicit association testing. *Journal of the American College of Surgeons* 220, 6 (2015), 1077–1086.

[63] Yaroslav Halchenko, Kyle Meyer, Benjamin Poldrack, Debanjum Solanky, Adina Wagner, Jason Gors, Dave MacFarlane, Dorian Pustina, Vanessa Sochat, Satrajit Ghosh, Christian Mönch, Christopher Markiewicz, Laura Waite, Ilya Shlyakhter, Alejandro de la Vega, Soichi Hayashi, Christian Häusler, Jean-Baptiste Poline, Tobias Kadelka, Kusti Skytén, Dorota Jarecka, David Kennedy, Ted Strauss, Matt Cieslak, Peter Vavra, Horea-Ioan Ioanas, Robin Schneider, Mika Pflüger, James Haxby, Simon Eickhoff, and Michael Hanke. 2021. DataLad: distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software* 6 (7 2021), 3262. Issue 63. doi:10.21105/joss.03262

[64] Yaroslav O Halchenko, Kyle Meyer, Benjamin Poldrack, Debanjum Singh Solanky, Adina S Wagner, Jason Gors, Dave MacFarlane, Dorian Pustina, Vanessa Sochat, Satrajit S Ghosh, et al. 2021. DataLad: distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software* 6, 63 (2021), 3262.

[65] Nicholas Heller, Jack Rickman, Christopher Weight, and Nikolaos Papanikolopoulos. 2019. The role of publicly available data in miccai papers from 2014 to 2018. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 70–77.

[66] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. 2022. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data* 9, 1 (2022), 762.

[67] Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel C. F. Codella, Veronica Rotemberg, Allan C. Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, Susana Puig, Veronica Vilaplana, and Josep Malvehy. 2024. BCN20000: Dermoscopic Lesions in the Wild. *Scientific Data* 11, 1 (June 2024). doi:10.1038/s41597-024-03387-w

[68] Kelly M Hoffman, Sophie Trawalter, Jordan R Axt, and M Norman Oliver. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences* 113, 16 (2016), 4296–4301.

[69] Gregory Holste, Ziyu Jiang, Ajay Jaiswal, Maria Hanna, Shlomo Minkowitz, Alan C Legasto, Joanna G Escalon, Sharon Steinberger, Mark Bittman, Thomas C Shen, et al. 2023. How Does Pruning Impact Long-Tailed Multi-label Medical Image Classifiers?. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 663–673.

[70] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. 2022. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*. Springer, 22–32.

[71] Mahan Hosseini, Michael Powell, John Collins, Chloe Callahan-Flintoft, William Jones, Howard Bowman, and Brad Wyble. 2020. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews* 119 (2020), 456–467.

[72] HuggingFace. 2025. HuggingFace. https://huggingface.co/. Accessed: 2025-01-17.

[73] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering

and infrastructure. In *Fairness, Accountability, and Transparency (FAccT)*. 560–575.

[74] Hirotaka Ieki, Kaoru Ito, Mike Saji, Rei Kawakami, Yuji Nagatomo, Kaori Takada, Toshiya Kariyasu, Haruhiko Machida, Satoshi Koyama, Hiroki Yoshida, et al. 2022. Deep learning-based age estimation from chest X-rays indicates cardiovascular prognosis. *Communications Medicine* 2, 1 (2022), 159.

[75] Stefanos Ioannou, Hana Chockler, Alexander Hammers, Andrew P King, and Alzheimer's Disease Neuroimaging Initiative. 2022. A study of demographic bias in CNN-based brain MR segmentation. In *MICCAI Workshop on Machine Learning in Clinical Neuroimaging*. Springer, 13–22.

[76] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, Vol. 33. 590–597.

[77] The International Skin Imaging Collaboration (ISIC). 2024. ISIC archive. https://www.isic-archive.com/. Accessed: 2024-05-22.

[78] Jiwoong J Jeong, Brianna L Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, et al. 2023. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence* 5, 1 (2023), e220047.

[79] Bai Jieyun and Ou ZhanHong. 2023. *Pubic Symphysis-Fetal Head Segmentation and Angle of Progression*. doi:10.5281/zenodo.7851339

[80] Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Dovile Juodelyte, Théo Sourget, Caroline Vang-Larsen, Hubert Dariusz Zając, Anna Rogers, and Veronika Cheplygina. 2024. Copycats: the many lives of a publicly available medical imaging dataset. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. https://openreview.net/forum?id=X4KImMSIRq

[81] Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. 2023. Detecting Shortcuts in Medical Images - A Case Study in Chest X-rays. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.

[82] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6, 1 (2019), 317.

[83] Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. 2019. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology* 29 (2019), 1391–1399.

[84] Dovile Juodelyte, Yucheng Lu, Amelia Jiménez-Sánchez, Sabrina Bottazzi, Enzo Ferrante, and Veronika Cheplygina. 2025. Source Matters: Source Dataset Impact on Model Robustness in Medical Imaging. In *Applications of Medical Artificial Intelligence*, Shandong Wu, Behrouz Shabestari, and Lei Xing (Eds.). Springer Nature Switzerland, Cham, 105–115.

[85] Kaggle. 2025. Kaggle. https://www.kaggle.com/. Accessed: 2025-01-17.

[86] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima Jr, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, et al. 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* 3, 6 (2021), 473–484.

[87] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[88] Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Anna Zapaishchykova, Julija Pavaine, Lubdha M Shah, Blaise V Jones, Nakul Sheth, et al. 2024. Brats-peds: Results of the multi-consortium international pediatric brain tumor segmentation challenge 2023. *arXiv preprint arXiv:2407.08855* (2024).

[89] Kathryn E. Keenan, Jana G. Delfino, Kalina V. Jordanova, Megan E. Poorman, Prathyush Chirra, Akshay S. Chaudhari, Bettina Baessler, Jessica Winfield, Satish E. Viswanath, and Nandita M. deSouza. 2022. Challenges in ensuring the generalizability of image quantitation methods for MRI. 2820-2835 pages. Issue 4. doi:10.1002/mp.15195

[90] David N Kennedy, Christian Haselgrove, Jon Riehl, Nina Preuss, and Robert Buccigrossi. 2016. The NITRC image repository. *NeuroImage* 124 (2016), 1069–1073.

[91] Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. 2021. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health* 3, 1 (2021), e51–e66. doi:10.1016/S2589-7500(20)30240-5

[92] Lauren N. Koenig, Gregory S. Day, Amber Salter, Sarah Keefe, Laura M. Marple, Justin Long, Pamela LaMontagne, Parinaz Massoumzadeh, B. Joy Snider, Manasa Kanthamneni, Cyrus A. Raji, Nupur Ghoshal, Brian A. Gordon, Michelle Miller-Thomas, John C. Morris, Joshua S. Shimony, and Tammie L.S. Benzinger. 2020. Select Atrophied Regions in Alzheimer disease (SARA): An improved volumetric model for identifying Alzheimer disease dementia. *NeuroImage: Clinical* 26 (2020), 102248. doi:10.1016/j.nicl.2020.102248

[93] Rafsanjany Kushol, Pedram Parnianpour, Alan H. Wilman, Sanjay Kalra, and Yee Hong Yang. 2023. Effects of MRI scanner manufacturers in classification tasks with deep learning models. *Scientific Reports* 13 (12 2023). Issue 1. doi:10.1038/s41598-023-43715-5

[94] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medrxiv* (2019), 2019–12.

[95] Oran Lang, Doron Yaya-Stupp, Ilana Traynis, Heather Cole-Lewis, Chloe R Bennett, Courtney R Lyles, Charles Lau, Michal Irani, Christopher Semturs, Dale R Webster, et al. 2024. Using generative AI to investigate medical imagery models and datasets. *EBioMedicine* 102 (2024).

[96] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117, 23 (2020), 12592–12594.

[97] Bastien Le Guellec, Alexandre Lefèvre, Charlotte Geay, Lucas Shorten, Cyril Bruge, Lotfi Hacein-Bey, Philippe Amouyel, Jean-Pierre Pruvo, Grégory Kuchcinski, and Aghiles Hamroun. 2024. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiology: Artificial Intelligence* (2024), e230364.

[98] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical data curation for ai: An approach based on feminist epistemology and critical theories of race. In *AI, Ethics, and Society (AIES)*. AAAI/ACM, 695–703.

[99] Johann Li, Guangming Zhu, Cong Hua, Mingtao Feng, Basheer Bennamoun, Ping Li, Xiaoyuan Lu, Juan Song, Peiyi Shen, Xu Xu, et al. 2023. A systematic collection of medical image datasets for deep learning. *Comput. Surveys* 56, 5 (2023), 1–51.

[100] Xiangrui Li, Paul S. Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. 2016. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods* 264 (5 2016), 47–56. doi:10.1016/j.jneumeth.2016.03.001

[101] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

[102] Sook Lei Liew, Bethany P. Lo, Miranda R. Donnelly, Artemis Zavaliangos-Petropulu, Jessica N. Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P. Simon, Julia M. Juliano, Anisha Suri, Zhizhuo Wang, Aisha Abdullah, Jun Kim, Tyler Ard, Nerisa Banaj, Michael R. Borich, Lara A. Boyd, Amy Brodtmann, Cathrin M. Buetefisch, Lei Cao, Jessica M. Cassidy, Valentina Ciullo, Adriana B. Conforto, Steven C. Cramer, Rosalia Dacosta-Aguayo, Ezequiel de la Rosa, Martin Domin, Adrienne N. Dula, Wuwei Feng, Alexandre R. Franco, Fatemeh Geranmayeh, Alexandre Gramfort, Chris M. Gregory, Colleen A. Hanlon, Brenton G. Hordacre, Steven A. Kautz, Mohamed Salah Khlif, Hosung Kim, Jan S. Kirschke, Jingchun Liu, Martin Lotze, Bradley J. MacIntosh, Maria Mataró, Feroze B. Mohamed, Jan E. Nordvik, Gilsoon Park, Amy Pienta, Fabrizio Piras, Shane M. Redman, Kate P. Revill, Mauricio Reyes, Andrew D. Robertson, Na Jin Seo, Surjo R. Soekadar, Gianfranco Spalletta, Alison Sweet, Maria Telenczuk, Gregory Thielman, Lars T. Westlye, Carolee J. Winstein, George F. Wittenberg, Kristin A. Wong, and Chunshui Yu. 2022. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data* 9 (12 2022). Issue 1. doi:10.1038/s41597-022-01401-7

[103] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*. 3159–3167.

[104] Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. 2024. Shortcut learning in medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 623–633.

[105] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60–88.

[106] Xiaoming Liu, Quan Yuan, Yaozong Gao, Kelei He, Shuo Wang, Xiao Tang, Jinshan Tang, and Dinggang Shen. 2022. Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images. *Pattern Recognition* 122 (2022), 108341.

[107] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv preprint arXiv:2310.16787* (2023).

[108] Lena Maier-Hein, Sven Mersmann, Daniel Kondermann, Sebastian Bodenstedt, Alexandro Sanchez, Christian Stock, Hannes Gotz Kenngott, Mathias Eisenmann, and Stefanie Speidel. 2014. Can Masses of Non-Experts Train Highly Accurate Image Classifiers? In *Medical Image Computing and Computer-Assisted*

Intervention (MICCAI). Springer, 438–445.

[109] Lena Maier-Hein, Annika Reinke, Evangelia Christodoulou, Ben Glocker, Patrick Godau, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A Riegler, et al. 2022. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv preprint arXiv:2206.01653* (2022).

[110] Sarmad Maqsood and Robertas Damaševičius. 2023. Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. *Neural Networks* 160 (2023), 238–258.

[111] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience* 22, 12 (2010), 2677–2684.

[112] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. 2007. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience* 19, 9 (09 2007), 1498–1507. doi:10.1162/jocn.2007.19.9.1498 arXiv:https://direct.mit.edu/jocn/article-pdf/19/9/1498/1936514/jocn.2007.19.9.1498.pdf

[113] Christopher J. Markiewicz, Krzysztof J. Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O. Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, Anita Jwa, and Russell Poldrack. 2021. The OpenNeuro resource for sharing of neuroscience data. *eLife* 10 (10 2021). doi:10.7554/eLife.71774

[114] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. 2022. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence* 4, 5 (2022), e210315.

[115] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.

[116] Kamil Wojciech Mikolaj, Anders Nymark Christensen, Caroline Amalie Taksøe-Vester, Aasa Feragen, Olav Bjørn Petersen, Manxi Lin, Mads Nielsen, Morten Bo Søndergaard Svendsen, and Martin Grønnebæk Tolsgaard. 2025. Predicting Abnormal Fetal Growth Using Deep Learning. *npj Digital Medicine, in press* (2025).

[117] Zahra Mirikharaji, Kumar Abhishek, Alceu Bissoto, Catarina Barata, Sandra Avila, Eduardo Valle, M Emre Celebi, and Ghassan Hamarneh. 2023. A survey on deep learning for skin lesion segmentation. *Medical Image Analysis* 88 (2023), 102863.

[118] Karel GM Moons, Robert F Wolff, Richard D Riley, Penny F Whiting, Marie Westwood, Gary S Collins, Johannes B Reitsma, Jos Kleijnen, and Sue Mallett. 2019. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of Internal Medicine* 170, 1 (2019), W1–W33.

[119] Pritam Mukherjee, Benjamin Hou, Ricardo B Lanfredi, and Ronald M Summers. 2023. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology* 309, 1 (2023), e231147.

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

[121] Maryam Naqvi, Syed Qasim Gilani, Tehreem Syed, Oge Marques, and Hee-Cheol Kim. 2023. Skin cancer detection using deep learning—a review. *Diagnostics* 13, 11 (2023), 1911.

[122] Guiomar Niso, Rotem Botvinik-Nezer, Stefan Appelhoff, Alejandro De La Vega, Oscar Esteban, Joset A. Etzel, Karolina Finc, Melanie Ganz, Rémi Gau, Yaroslav O. Halchenko, Peer Herholz, Agah Karakuzu, David B. Keator, Christopher J. Markiewicz, Camille Maumet, Cyril R. Pernet, Franco Pestilli, Nazek Queder, Tina Schmitt, Weronika Sójka, Adina S. Wagner, Kirstie J. Whitaker, and Jochem W. Rieger. 2022. Open and reproducible neuroimaging: From study inception to publication. *NeuroImage* 263 (2022), 119623. doi:10.1016/j.neuroimage.2022.119623

[123] Lauren Oakden-Rayner. 2020. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology* 27, 1 (2020), 106–112.

[124] Lauren Oakden-Rayner. 2025. Half a million x-rays! First impressions of the Stanford and MIT chest x-ray datasets. https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/. Accessed: 2025-01-07.

[125] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Conference on Health, Inference, and Learning (CHIL)*. ACM, 151–159.

[126] Lauren Oakden-Rayner, William Gale, Thomas A Bonham, Matthew P Lungren, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. 2022. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *The Lancet Digital Health* 4, 5 (2022), e351–e358.

[127] Beatriz Ocaña-Tienda, Julián Pérez-Beteta, José D Villanueva-García, José A Romero-Rosales, David Molina-García, Yannick Suter, Beatriz Asenjo, David Albillo, Ana Ortiz de Mendivil, Luis A Pérez-Romasanta, et al. 2023. A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data. *Scientific Data* 10, 1 (2023), 208.

[128] Keith K. Ogasawara. 2009. Variation in Fetal Ultrasound Biometry Based on Differences in Fetal Ethnicity. *American Journal of Obstetrics and Gynecology* 200, 6 (June 2009), 676.e1–676.e4. doi:10.1016/j.ajog.2009.02.031

[129] Youngmin Oh, Beomjun Kim, and Bumsub Ham. 2021. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*. 6913–6922.

[130] Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. 2024. Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods. In *MICCAI Workshop on Fairness of AI in Medical Imaging (FAIMI)*. Springer, 3–13.

[131] Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, et al. 2024. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *NPJ Digital Medicine* 7, 1 (2024), 124.

[132] Silas Nyboe Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. 2020. A survey of crowdsourcing in medical image analysis. *Human Computation* 7 (2020), 1–26.

[133] Andre GC Pacheco and Renato A Krohling. 2021. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE Journal of Biomedical and Health Informatics* 25, 9 (2021), 3554–3563.

[134] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. 2020. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief* 32 (2020), 106221.

[135] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.

[136] H Yi Paul, Tae Kyung Kim, Eliot Siegel, and Noushin Yahyavi-Firouz-Abadi. 2022. Demographic reporting in publicly available chest radiograph data sets: opportunities for mitigating sex and racial disparities in deep learning models. *Journal of the American College of Radiology* 19, 1 (2022), 192–200.

[137] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021).

[138] Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, J. Vanschoren and S. Yeung (Eds.), Vol. 1.

[139] Cyril Pernet, Claus Svarer, Ross Blair, John D Van Horn, and Russell A Poldrack. 2023. On the long-term archiving of research data. *Neuroinformatics* 21, 2 (2023), 243–246.

[140] Lena Philipp, Maarten de Rooij, John Hermans, Matthieu Rutten, Horst Karl Hahn, Bram van Ginneken, and Alessa Hering. 2024. Annotation-Efficient Strategy for Segmentation of 3D Body Composition. In *Medical Imaging with Deep Learning (MIDL)*.

[141] Russell A Poldrack, Christopher J Markiewicz, Stefan Appelhoff, Yoni K Ashar, Tibor Auer, Sylvain Baillet, Shashank Bansal, Leandro Beltrachini, Christian G Benar, Giacomo Bertazzoli, et al. 2024. The past, present, and future of the brain imaging data structure (BIDS). *Imaging Neuroscience* 2 (2024), 1–19.

[142] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs.DL]

[143] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Heller, Fabian Isensee, Annette Kopp-Schneider, and Lena Maier-Hein. 2025. Quality Assured: Rethinking Annotation Strategies in Imaging AI. In *European Conference on Computer Vision*. Springer, 52–69.

[144] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Schreck, A Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. 2023. Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence* 5, 3 (2023), 273–283.

[145] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

[146] P Rajpurkar. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225* (2017).

[147] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature Medicine* 28, 1 (2022), 31–38.

[148] Ralf Raumanns, Gerard Schouten, Max Joosten, Josien PW Pluim, Veronika Cheplygina, et al. 2021. ENHANCE (ENriching Health data by ANnotations

of Crowd and Experts): A case study for skin lesion classification. *Machine Learning for Biomedical Imaging* 1, December 2021 issue (2021), 1–26.

[149] Annika Reinke, Minu D. Tizabi, Carole H. Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, Matthew Blaschko, Florian Buettner, M. Jorge Cardoso, Jianxu Chen, Veronika Cheplygina, Evangelia Christodoulou, Beth Cimini, Gary S. Collins, Sandy Engelhardt, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Robert Haase, Fred Hamprecht, Daniel A. Hashimoto, Doreen Heckmann-Nötzel, Peter Hirsch, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, A. Emre Kavur, Hannes Kenngott, Jens Kleesiek, Andreas Kleppe, Sven Kohler, Florian Kofler, Annette Kopp-Schneider, Thijs Kooi, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, David Moher, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, M. Alican Noyan, Jens Petersen, Gorkem Polat, Susanne M. Rafelski, Nasir Rajpoot, Mauricio Reyes, Nicola Rieke, Michael Riegler, Hassan Rivaz, Julio Saez-Rodriguez, Clara I. Sánchez, Julien Schroeter, Anindo Saha, M. Alper Selver, Lalith Sharan, Shravya Shetty, Maarten van Smeden, Bram Stieltjes, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, Manuel Wiesenfarth, Ziv R. Yaniv, Paul Jäger, and Lena Maier-Hein. 2023. Common Limitations of Image Processing Metrics: A Picture Story. https://arxiv.org/abs/2104.05642

[150] ACL Rolling Review. 2025. Call for papers. https://aclrollingreview.org/cfp. Accessed: 2025-04-22.

[151] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. 2022. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications* 13, 1 (2022), 1–6.

[152] María Agustina Ricci Lara, María Victoria Rodríguez Kowalczuk, Maite Lisa Eliceche, María Guillermina Ferraresso, Daniel Roberto Luna, Sonia Elizabeth Benitez, and Luis Daniel Mazzuoccolo. 2023. A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. *Scientific Data* 10, 1 (2023), 712.

[153] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine* 3, 1 (2020), 1–7.

[154] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: development of a transparency artifact for health datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1943–1961.

[155] Alexandre Routier, Ninon Burgos, Mauricio Díaz, Michael Bacci, Simona Bottani, Omar El-Rifai, Sabrina Fontanella, Pietro Gori, Jérémy Guillon, Alexis Guyot, Ravi Hassanaly, Thomas Jacquemont, Pascal Lu, Arnaud Marcoux, Tristan Moreau, Jorge Samper-González, Marc Teichmann, Elina Thibeau-Sutre, Ghislain Vaillant, Junhao Wen, Adam Wild, Marie Odile Habert, Stanley Durrleman, and Olivier Colliot. 2021. Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. *Frontiers in Neuroinformatics* 15 (8 2021). doi:10.3389/fninf.2021.689675

[156] Anisa Rowhani-Farid and Adrian G Barnett. 2016. Has open data arrived at the British Medical Journal (BMJ)? An observational study. *BMJ Open* 6, 10 (2016), e011784.

[157] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[158] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[159] María Sofía Sappia. 2024. *ACOUSLIC-AI : Abdominal Circumference Operator-agnostic UltraSound Measurement.* doi:10.5281/zenodo.12697994

[160] Caroline J Savage and Andrew J Vickers. 2009. Empirical study of data sharing by authors publishing in PLoS journals. *PLoS One* 4, 9 (2009), e7078.

[161] L Scarpace, T Mikkelsen, S Cha, S Rao, S Tekchandani, D Gutman, JH Saltz, BJ Erickson, N Pedano, AE Flanders, et al. 2016. The cancer genome atlas glioblastoma multiforme collection (TCGA-GBM)(Version 4)[data set]. *The Cancer Imaging Archive* 10 (2016), K9.

[162] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[163] Lena Schmidt, Saleh Mohamed, Nick Meader, Jaume Bacardit, and Dawn Craig. 2023. Automated data analysis of unstructured grey literature in health research: A mapping review. *Research Synthesis Methods* (2023).

[164] Lena Schmidt, Ailbhe N Finnerty Mutlu, Rebecca Elmore, Babatunde K Olorisade, James Thomas, and Julian PT Higgins. 2021. Data extraction methods for systematic review (semi) automation: Update of a living systematic review. *F1000Research* 10 (2021).

[165] Andrew Sellergren, Atilla Kiraly, Tom Pollard, Wei-Hung Weng, Yun Liu, Akib Uddin, and Christina Chen. 2023. Generalized Image Embeddings for the MIMIC Chest X-Ray dataset. doi:10.13026/PXC2-VX69

[166] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *Pacific Symposium on Biocomputing*. World Scientific, 232–243.

[167] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* 27, 12 (2021), 2176–2182.

[168] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis* 88 (2023), 102802. doi:10.1016/j.media.2023.102802

[169] Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y Chen. 2024. The data addition dilemma. *arXiv preprint arXiv:2408.04154* (2024).

[170] Line Sletner, Svein Rasmussen, Anne Karen Jenum, Britt Nakstad, Odd Harald Rognerud Jensen, and Siri Vangen. 2015. Ethnic Differences in Fetal Size and Growth in a Multi-Ethnic Population. *Early Human Development* 91, 9 (Sept. 2015), 547–554. doi:10.1016/j.earlhumdev.2015.07.002

[171] Théo Sourget, Ahmet Akkoç, Stinna Winther, Christine Lyngbye Galsgaard, Amelia Jiménez-Sánchez, Dovile Juodelyte, Caroline Petitjean, and Veronika Cheplygina. 2024. [Citation needed] Data usage and citation practices in medical imaging conferences. In *Medical Imaging with Deep Learing (MIDL), in press.*

[172] Elijah Sterling, Hannah Pearl, Zexuan Liu, Jason W. Allen, and Candace C. Fleischer. 2022. Demographic reporting across a decade of neuroimaging: a systematic review. 2785-2796 pages. Issue 6. doi:10.1007/s11682-022-00724-8

[173] Yannick Suter, Urspeter Knecht, Waldo Valenzuela, Michelle Notter, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. 2022. The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. *Scientific Data* 9 (12 2022). Issue 1. doi:10.1038/s41597-022-01881-7

[174] Hidde Ten Berg, Bram van Bakel, Lieke van de Wouw, Kim E Jie, Anoeska Schipper, Henry Jansen, Rory D O'Connor, Bram van Ginneken, and Steef Kurstjens. 2024. ChatGPT and generating a differential diagnosis early in an emergency department presentation. *Annals of Emergency Medicine* 83, 1 (2024), 83–86.

[175] Wen-Juan Tong, Shao-Hong Wu, Mei-Qing Cheng, Hui Huang, Jin-Yu Liang, Chao-Qun Li, Huan-Ling Guo, Dan-Ni He, Yi-Hao Liu, Han Xiao, Hang-Tong Hu, Si-Min Ruan, Ming-De Li, Ming-De Lu, and Wei Wang. 2023. Integration of Artificial Intelligence Decision Aids to Reduce Workload and Enhance Efficiency in Thyroid Nodule Management. *JAMA Network Open* 6, 5 (2023), e2313674. doi:10.1001/jamanetworkopen.2023.13674

[176] Pai-Jong Stacy Tsai, Matthew Loichinger, and Ivica Zalud. 2015. Obesity and the Challenges of Ultrasound Fetal Abnormality Diagnosis. *Best Practice & Research Clinical Obstetrics & Gynaecology* 29, 3 (April 2015), 320–327. doi:10.1016/j.bpobgyn.2014.08.011

[177] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5, 1 (2018), 1–9.

[178] UN UNSCEAR et al. 2000. Sources and effects of ionizing radiation. *United Nations Scientific Committee on the Effects of Atomic Radiation* (2000).

[179] Marly van Assen, Ashley Beecy, Gabrielle Gershon, Janice Newsome, Hari Trivedi, and Judy Gichoya. 2024. Implications of bias in artificial intelligence: considerations for cardiovascular imaging. *Current Atherosclerosis Reports* 26, 4 (2024), 91–102.

[180] Thomas L. A. van den Heuvel, Dagmar de Bruijn, Chris L. de Korte, and Bram van Ginneken. 2018. *Automated Measurement of Fetal Head Circumference Using 2D Ultrasound Images.* doi:10.5281/zenodo.1327317

[181] Bram Van Ginneken, BM Ter Haar Romeny, and Max A Viergever. 2001. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging* 20, 12 (2001), 1228–1241.

[182] Florien S van Royen, Karel GM Moons, Geert-Jan Geersing, and Maarten van Smeden. 2022. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *European Respiratory Journal* 60, 3 (2022).

[183] V.N. Vapnik. 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* 10, 5 (1999), 988–999. doi:10.1109/72.788640

[184] My von Euler-Chelpin, Martin Lillholm, Ilse Vejborg, Mads Nielsen, and Elsebeth Lynge. 2019. Sensitivity of screening mammography by density and texture: a cohort study from a population-based screening program in Denmark. *Breast Cancer Research* 21, 1 (Oct. 2019), 111. doi:10.1186/s13058-019-1203-3

[185] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR)*. 2097–2106.

[186] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, et al. 2022. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health* 4, 1 (2022), e64–e74.

[187] David Wen, Andrew Soltan, Emanuele Trucco, and Rubeta N Matin. 2024. From data to diagnosis: skin cancer image datasets for artificial intelligence. *Clinical and Experimental Dermatology* (2024), llae112.

[188] Nina Weng, Siavash Bigdeli, Eike Petersen, and Aasa Feragen. 2023. Are Sex-Based Physiological Differences the Cause of Gender Bias for Chest X-Ray Diagnosis?. In *MICCAI Workshop on Clinical Image-Based Procedures*. Springer, 142–152.

[189] World Health Organization (WHO). 2025. "Fact sheets: maternal mortality". https://www.who.int/news-room/fact-sheets/detail/maternal-mortality. Accessed: 2024-08-20.

[190] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[191] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology* 155, 10 (2019), 1135–1141.

[192] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Banafsheh Arshi, Vanesa Bellou, Marc MJ Bonten, et al. 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj* 369 (2020).

[193] Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. 2020. MRI manufacturer shift and adaptation: Increasing the generalizability of deep learning segmentation for MR images acquired with different scanners. *Radiology: Artificial Intelligence* 2 (2020), 1–10. Issue 4. doi:10.1148/ryai.2020190195

[194] Xinyu Yang, Weixin Liang, and James Zou. 2024. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on HuggingFace. In *International Conference on Learning Representations (ICLR)*.

[195] Xin Yi, Ekta Walia, and Paul Babyn. 2019. Generative adversarial network in medical imaging: A review. *Medical Image Analysis* 58 (2019), 101552.

[196] Sung Hyun Yoon, Sunyoung Park, Sowon Jang, Junghoon Kim, Kyung Won Lee, Woojoo Lee, Seungjae Lee, Gabin Yun, and Kyung Hee Lee. 2023. Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload. *European Radiology* 34, 2 (2023), 1094–1103. doi:10.1007/s00330-023-10124-1

[197] Ruonan Yu, Songhua Liu, and Xinchao Wang. 2023. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[198] Hubert Dariusz Zając, Natalia Rozalia Avlona, Finn Kensing, Tariq Osman Andersen, and Irina Shklovski. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Conference on AI, Ethics, and Society (AIES)*. 351–362.

[199] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine* 15, 11 (2018), e1002683.

[200] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. 2022. Improving the fairness of chest X-ray classifiers. In *Conference on Health, Inference, and Learning (CHIL)*. PMLR, 204–233.

A Shortcuts

Machine learning models are prone to rely on spurious correlations to make predictions, which are usually easier to detect than the genuine disease patterns, as explained in Section 6. Such shortcuts can be well-localized objects in the image – e.g., mechanical ventilation tubes or pacemakers, which are often present in patients with certain diseases– or more global ones, like specific noise patterns or intensity distributions associated to a certain acquisition setting or device brand. Demographic attributes like sex/gender, age or ethnicity can also lead to shortcut learning, disproportionately impacting historically underserved subgroups [7], especially when datasets are highly imbalanced. To categorize the different types of shortcuts, we adopt the Medical Imaging Contextualized Confounder Taxonomy (MICCAT) [84], see Fig. A1, which we believe can be useful to understand how spurious correlations arise and how they can be mitigated. MICCAT extends beyond traditional demographic attributes, such as sex/gender, age, and ethnicity, to include a broader set of confounders that are domain- and context-specific. These confounders encompass patient-level and environment-level factors.

Patient-level confounders include both demographic attributes and anatomical confounders. Demographic attributes, such as gender [1, 96], age [1], and ethnicity [51], represent standard factors typically considered in bias analysis. Anatomical confounders, on the other hand, refer to physical or disease-related characteristics specific to organs or conditions. Examples include body mass index, tissue density, breast density, and bone density, as well as combinations of these factors. Such anatomical variations may define subgroups where models underperform, and their identification often requires analysis beyond standard demographic characteristics [184].

Environment-level confounders include both external and imaging confounders. External confounders arise from physical or virtual elements within the image, such as chest drains [81, 125], pen marks near skin lesions [191], patient positioning [29], or text and measurement calipers [104]. These elements typically produce localized artifacts visible to the human eye. In contrast, imaging confounders result from variations in the imaging process, including differences in equipment brands, scanner types, acquisition parameters, noise, or motion artifacts. Such confounders generally create global artifacts that affect the entire image and may not be perceptible to the human eye. Systematic differences, such as variations in exposure settings for chest X-rays [95] or distinctive characteristics of imaging devices used across different medical centers [23], can unintentionally introduce shortcuts for machine learning models. Rather than learning clinically relevant features, models may instead rely on these acquisition-specific factors, which are associated with disease labels but do not represent true underlying medical conditions [131].

B Clinical case studies

We present four case studies, illustrated in Fig. 3, with key points highlighted in Table C4.

Case study 1: chest X-rays. Chest X-rays are the most commonly performed radiologic examinations worldwide [178], requiring significant expertise for accurate and meaningful interpretation [181]. The advent of artificial intelligence is a game changer for

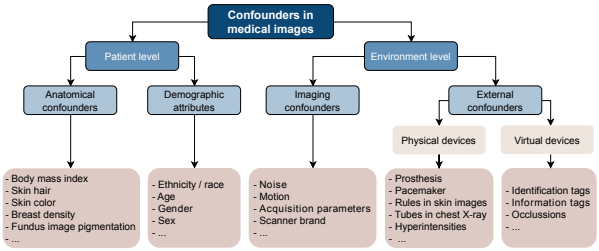

Figure A1: MICCAT: Medical Imaging Contextualized Confounder Taxonomy. Reproduced from [84] with permission of the authors.

automatic chest X-ray diagnosis, marked by the release of the large-scale NIH-CXR14 dataset [185]. This milestone spurred a wave of studies, including claims of machine learning models achieving radiologist-level performance, exemplified by CheXNet's pneumonia detection model [146]. However, these claims that have been criticized for relying on shortcuts, such as the use of chest drains for pneumothorax classification [81, 125], and for demonstrating low inter-rater agreement for pathologies like pneumonia [25]. The latter may be attributed to the fact that pneumonia is a differential diagnosis requiring clinical information – laboratory tests, physical examinations, symptoms and signs, etc. – that extends beyond chest X-ray images [14]. Subsequently, additional datasets such as CheXpert [76], MIMIC-CXR [82], and PadChest [14] have become widely used in the research community.

With the primary purpose of diagnosing and/or monitoring treatment of various pathologies, those chest X-ray dataset often include associated demographic information such as age, gender/sex, and race/ethnicity, which enables further analysis of potential disparities and biases in automatic chest X-ray diagnosis. Among these key demographic attributes, age and sex/gender are commonly included in most datasets, whereas race/ethnicity is available in only a few, such as MIMIC-CXR and CheXpert [136].

Chest X-ray datasets are generally skewed toward older populations, with PadChest's median age at 62, and MIMIC-CXR's largest group aged 60–80 [200]. Studies have shown that age can be predicted from chest X-rays [74], raising concerns about unintended information leakage favoring well-represented age groups. Widely used chest X-ray datasets show no significant gender distribution imbalance (e.g., NIH-CXR14: {M:56.5%, F:43.5%}, CheXpert: {M:58.7%, F:41.3%}). However, performance disparities between male and female groups in disease classifiers persist [96, 166, 167], with unclear causes. Balancing the data has proven ineffective. Studies have ruled out hypothesis like under-representation [96], physiological differences (e.g., the presence of breasts) [188] and shortcut learning (e.g., support devices like chest drains [81, 125, 130]).

Race and ethnicity are important demographic attributes, but are rarely included in chest X-ray datasets. Among 23 reviewed datasets [136], only MIMIC-CXR and CheXpert report self-reported race, with predominantly white populations accounting for 60.7% and 56.4% of samples, respectively. Performance disparities between racial groups favoring white individuals have been noted [166]. Studies show deep learning can infer protected attributes like race

from chest X-rays, despite this being a challenging task for human experts [51]. This raises critical concerns about the potential use of protected features in the decision-making process, which could lead to unfair biases. Recent research [53] suggests that transfer learning alone cannot confirm the influence of protected attributes; instead combining methods like test-set resampling, multitask learning, and model inspection provides offer better insights within neural network feature representations.

Case study 2: Skin lesions. Over the past few years, the use of computer vision algorithms in dermatology has experienced significant growth [121]. This surge may be largely explained by the availability of publicly accessible skin lesion datasets – such as Fitzpatrick17k [60], PAD-UFES-20 [134], and HIBA [152] – as well as the International Skin Imaging Collaboration (ISIC) –, which aggregates different datasets such as HAM10000 [177] and BCN 20000 [67], in an archive – the ISIC Archive[4]) that contains more than 490K skin lesion images. Dermoscopy and clinical photography are common imaging modalities in the development of ML models that address tasks such as classification [133], segmentation [117], and lesion localization [110].

Despite the progress made in the field, the under-representation of demographic groups in skin lesion datasets limits the generalizability of ML models [26]. Skin lesions may manifest differently across populations, influenced by factors like skin tone, genetic background, age and UV light exposure [187]. Nonetheless, most publicly available datasets are limited in terms of geographic and skin tone diversity, predominantly featuring of patients with Fitzpatrick skin types I to III [59, 186]. This lack of diversity may lead to biased models that underperform for under-represented groups. For example, melanoma – the deadliest skin cancer – is much more prevalent in fair-skinned individuals but can also occur in those with darker skin tones, which may be misdiagnosed due to the limited representation in training data [58].

Recent efforts to diversify skin lesion datasets, such as the inclusion of PAD-UFES-20 [134] and HIBA [152] datasets in the ISIC archive, have improved the representation of Latin American individuals. A region that was previously under-represented in the ISIC archive. However, there is still a strong lack of representation in terms of diversity of skin tone. In this context, addressing these disparities is a global challenge that requires a collaborative effort from the research community, drawing on diverse perspectives and contributions from different backgrounds and regions worldwide. Such initiatives are important to enhance fairness and promote equitable AI solutions for skin lesion analysis, working toward technologies that better serve patients across diverse demographic groups.

Case study 3: fetal ultrasound. Ultrasound is the fundamental imaging modality for antenatal care. The acquisition process consists of a physical examination with an ultrasound probe of the pregnant woman's womb looking for specific two-dimensional slices, called standard planes. After the location of one of these planes, the clinical expert often performs a series of annotations on the image to extract measurements of several structures of interest like the fetal head, femur, heart or abdomen, the placenta or the maternal cervix, among others, that are essential for the assessment of the pregnancy prognosis and the fetal well-being and growth. Several sources of variation might affect the image quality, such as the maternal body mass index (BMI) or the experience of the sonographer acquiring the scan. Maternal obesity, in particular, results in a higher difficulty to complete a full survey and a decreased visualization of fetal anatomies, being the face and the heart the most difficult anatomies to observe [28, 176]. One would expect detection accuracy disparities across varying BMI values, which could result in unfair predictions if not accounted for.

Another important factor is the ethnicity, which is associated with variations in the normal fetal growth according to several multi-ethnic studies [33, 128, 170]. Thus, growth charts based on image biomarkers, such as the head circumference or the femur length, need to be adapted to the local population.

With all this evidence, ML models trained with one cohort will inevitable exhibit biases toward a specific population subgroup, emphasizing the importance of multi-cohort and multi-variate analysis. Unfortunately, few or no evidence is reported on the effect of these biases for ML models. One of the reasons might be the scarcity of existing open fetal ultrasound datasets, with none, to date, providing demographic information. The few available datasets are the Fetal Planes DB [13] and data from the HC18 [180], FH-PS-AOP [79], and ACOUSLIC-AI [159] challenges. Overall, the analysis of fetal ultrasound imaging is either conducted by teams with access to their own local cohort or constrained to the few existing open datasets and tasks, making the multi-cohort analysis very difficult. Hence, there is a strong need for more diverse datasets with detailed demographic information, including BMI and ethnicity, to develop fair and unbiased models. Among others applications, this could allow for the implementation of robust ML-based tools in low-resource settings through portable devices to facilitate prenatal screening and try to reduce the pressing challenge of fetal mortality in low- and middle-income countries.

Case study 4: brain MRI. MRI is the third most commonly performed imaging modality after CT and X-rays. Its superior soft tissue contrast enables detailed visualization of brain anatomy, making it significantly more sensitive and specific for detecting abnormalities within the brain. This capability explains why most public MRI datasets available for AI research focus on the brain [32]. Key application areas for brain MRI and AI, along with some of the most notable datasets, include neurodegenerative diseases, represented by the OASIS Brains project datasets for Alzheimer's disease [92, 94, 111, 112]; brain cancer, with datasets like the BraTS Challenge [5], LUMIERE [173], Ocana [127], RHUH-GBM [15], and TCGA-GBM [161]; and stroke, with ISLES 2022 [66], ATLAS v2.0 [102], among others. The neuroimaging community has made significant progress in advancing data-sharing practices by adopting standards like BIDS [56] and employing tools like DataLad [63] for versioning and tracking. Representative platforms for hosting neuroimaging datasets include NITRC.org [90] and OpenNeuro [113], with the latter adhering to the BIDS standard and hosting over 700 brain MRI datasets, as identified through our web scraping analysis. Additionally, tools for converting legacy datasets to these formats are actively being developed [155].

[4]https://www.isic-archive.com/

The conversion from the standard clinical imaging format, DICOM, to NIfTI or other formats is complex and error-prone [100] and depends on how different formats implement DICOM standards. Several practical considerations drive the conversion. From a data science perspective, the DICOM standard's slice-wise storage of imaging volumes is less efficient for processing compared to volumetric formats like NIfTI. Moreover, most data science tools and libraries lack robust support for direct manipulation of DICOM files, favoring instead the use of formats designed for streamlined computational workflows. While these conversions simplify data processing pipelines, they often lead to an under-appreciation of the extensive clinical metadata embedded in DICOM files. This metadata is crucial because it captures details such as field strength, acquisition parameters, and vendor-specific variations—factors known to significantly influence the generalizability of deep learning models. Studies have demonstrated that neglecting these elements can reduce model robustness across different imaging settings [89, 93, 193], which is also reflected in the high interest in the field of domain adaptation for medical imaging applications [37, 61].

Many publicly available neuroimaging datasets undergo extensive preprocessing prior to release. Often they are standardized for open challenges, ensuring model evaluations focus on algorithm performance rather than preprocessing effects. While this approach has facilitated the comparison of methods within open challenges, it has also contributed to a disconnect between these standardized datasets and the complexities of real-world clinical data. This disconnect may inadvertently hinder technological advancements aimed at improving preprocessing techniques themselves. Skull stripping, also referred to as brain extraction, is a preprocessing step specific to neuroimaging that involves removing the skull and extracranial structures. It is particularly useful in tasks where these structures might lead to false-positive detections or oversegmentation. Additionally, skull stripping is often employed as a privacy mechanism, especially in structural MRI, to eliminate facial features. However, if performed improperly, it can result in the unintended removal of brain tissue, posing significant challenges in neuroimaging tasks where lesions are located near the brain's periphery, such as meningiomas. Intensity normalization, while commonly used to address the non-quantitative nature of MRI, is among the most destructive preprocessing steps, as the original intensity values cannot be recovered post-normalization. This irreversible transformation can limit the utility of datasets for certain applications. Examples of datasets shared with minimal preprocessing are the ISLES 2022 dataset [66], which was provided in nearly raw form, with skull stripping performed solely for anonymization purposes and data shared in NIfTI format, and the ISLES 2024 dataset, which was shared in NIfTY format as well as anonymized DICOMS. Such datasets offer valuable opportunities to study and address the challenges associated with preprocessing in a more authentic representation of real-world data.

Finally, reporting of demographics in brain MRI is generally very poor. For example, [172] analyzed demographic reporting in MR neuroimaging studies in the US over the past decade and found that biological sex was reported in 77% of studies, while race and ethnicity were reported in only 10% and 4%, respectively. Recent efforts have assessed the impact of demographic factors on model performance. In [31], disparities were found in brain age prediction models across sex subgroups and datasets. Similarly, [75] found race bias was more pronounced than sex bias on CNN-based MR segmentation, with Black females being the most affected subgroup.

In conclusion, MRI remains a cornerstone imaging modality for neuroimaging research, driven by its unparalleled ability to visualize brain anatomy with high sensitivity and specificity. The wealth of publicly available brain MRI datasets has enabled significant advances in AI applications, but these datasets are often limited by extensive preprocessing and a lack of inclusion of key demographic information. Preprocessing steps, while facilitating standardized comparisons in open challenges, can obscure the challenges inherent in real-world data and hinder progress in improving preprocessing pipelines themselves. Additionally, the underutilization of rich metadata in DICOM files and the complexities of format conversions further highlight the need for tools and practices that bridge the gap between clinical imaging and AI research. Efforts to address demographic disparities in neuroimaging datasets are beginning to emerge, as seen in recent efforts like the addition of children (BraTS-PEDS [88]) and Sub-Saharan African populations (BraTS-Africa [2]) to BraTS. Moving forward, the neuroimaging community must prioritize practices that maintain the authenticity of raw imaging data, incorporate diverse populations, and leverage metadata more effectively to maximize the potential of AI in medical imaging.

C Living review

In Section 2, we propose a proof of concept for the faliving review to keep track of medical imaging datasets and their related research artifacts. The resources that we consider as research artifacts are broadly classified into two categories, see Table C1:

- resource roles: describe the purpose the artifact serves in the research process, generally categorized into material or method, and
- resource types: describe the format or content of the artifact, regardless or its role.

We map the relation between the datasets and the research artifacts with the citation function (use, produce, extend, introduce, other), see Table C2. We show research artifacts and their citation function for CheXpert dataset in Table C3.

Our living review consists of three parts, as illustrated in Fig. 2:

(1) an overarching living review publication,
(2) documentation of research artifacts via dataset-specific publications on Zenodo, which the overarching paper links to,
(3) a SQL database for exploring the links between the datasets and the research artifacts.

Our demo is available at http://130.226.140.142, and a screenshot is shown in Fig. C1.

Technical details. We collect datasets, papers, and their relations into a SQL database. We visualize the database with DBeaver as shown in Fig. C2. The database comprise of three tables: one for datasets, another for papers, and a third one for datasets usages. The datasets table includes details such as the dataset ID, name and modality of the dataset. The papers table contains information on the ID, BibTeX key name, DOI, and arXiv link. The datasets usages table stores the ID, paper_ID, dataset_ID, annotations, shortcuts,

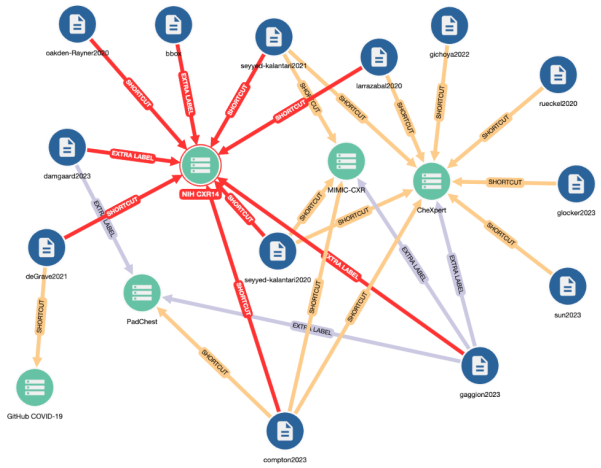

Figure C1: A screenshot of our demo living review. We highlight the relations for NIH-CXR14 dataset.

and the taxonomy to which the confounder belongs. Example of annotations include chest drains, segmentation masks, and bounding boxes. Example of shortcuts include dark corners and demographic attributes. Our demo is built with PostgreSQL, Python, Streamlit and the `st_link_analysis` package.

Resource role	Resource type	Description
Material	Dataset	corpus, image sets, etc.
	Additional labels	segmentation masks, demographic information, report of errors, etc.
	Shortcuts	resource finds evidence of errors in the annotations or shortcuts in the dataset.
Method	Tool	toolkit, software, system, etc.
	Code	codebase, library, API, etc.
Mixed	Mixed	citations referring to multiple resources

Table C1: List of resource roles and resource types.

Citation function	Description
Use	Used in the citing paper's research
Produce	First produced or released by the citing paper's research
Extend	Used in the citing paper's research but are improved, upgraded, or changed to work for other problems in the course of the research
Introduce	The resources or the related information (*e.g.*, background, applications) are introduced
Other	The citation does not belong to the above categories

Table C2: List of citation functions.

Research artifact	Citation function	Resource type	Example
Irvin et al. [76]	Produce	Dataset	CheXpert dataset
Rajpurkar et al. [147]	Use	-	-
Garbin et al. [48]	Introduce	Documentation	Datasheet [49] for CheXpert
Larrazabal et al. [96]	Extend	Shortcut	Gender bias
Gaggion et al. [43]	Extend	Additional annotation	Segmentation masks

Table C3: Example of research artifacts related to CheXpert dataset.

Case study	Tasks and Datasets	Annotation practices	Demographics	Shortcuts	Data lifecycle
1: Chest X-rays	T: Classification of chest X-ray diseases D: NIH-CXR14, CheXpert, PadChest, MIMIC-CXR	Experience radiologists *vs.* automatic NLP extraction	Self-reported race/ethinicity Skewed towards elder patients	Chest drains, hospital scanner, radiographic markers	
2: Skin lesions	T: Classification, segmentation and lesion localization. D: ISIC, HIBA, PAD-UFES-20, HAM10000, Fitzpatrick17k		Skin type, genetic background Recent additions of demographics with HIBA and PAD-UFES-2	Dark corners, hair, patches, rulers, ink marking/staining	
3: Fetal ultrasound	T: 2D plane localization, fetal biometrics estimation D: Fetal Planes DB, HC18, ACOUSLIC-AI, FH-FS-AOP	Experience of the sonographer might affect the image quality	None to date		
4:Neuroimaging	T: Neurogenerative diseases and stroke detection. D: OASIS, LUMIERE, BraTS, Ocana, TCGA-GBM, ATLAS v2.0, ISLES 2022		Poorly reported: in US studies (2010–2020), 77% report sex, but only 10% report race and 4% ethnicity. Recent inclusion of children and Sub-Saharan African populations in BraTS		Standards (e.g. BIDS, preprocessing, conversion from DICOM to NIfTI format) Data sharing platforms like Datalad, NITRC.org, OpenNeuro

Table C4: Summary highlighting various aspects of the four case studies. T: tasks; D: datasets. Note: this table is not exhaustive but summarizes key points discussed in the manuscript.

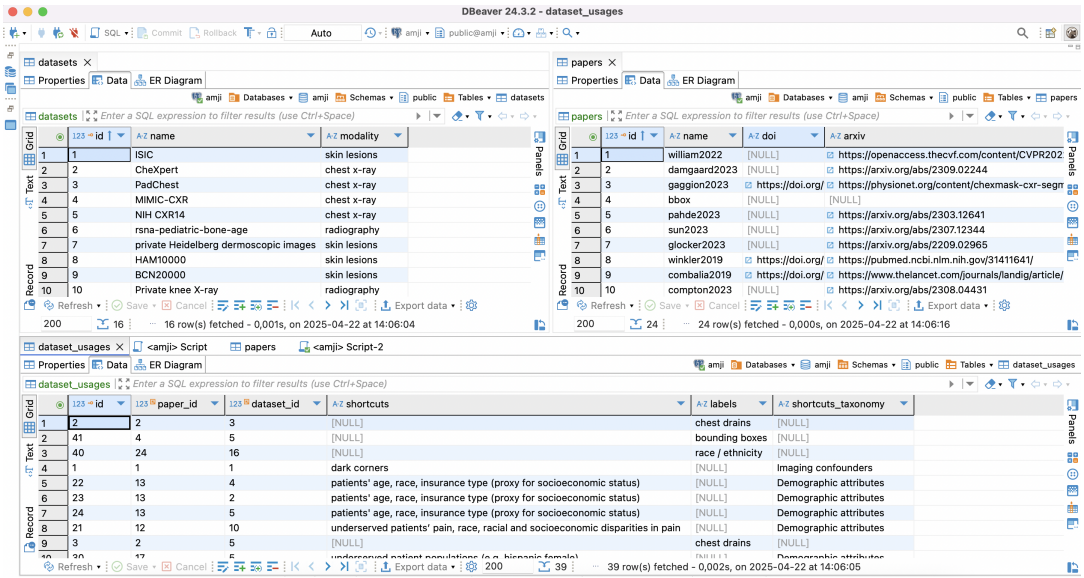

Figure C2: A screenshot of our SQL database in DBeaver.