On the Composition and Limitations of Publicly Available COVID-19 X-Ray Imaging Datasets

Beatriz Garcia Santa Cruz[2,1,✉], Jan Sölter[1], Matias Nicolas Bossa[1], and Andreas Dominik Husch[1]

[1] Luxembourg Centre for Systems Biomedicine
University of Luxembourg
7 Avenue des Hauts Fourneaux
4362, Esch-sur-Alzette
Luxembourg
beatriz.garcia@ext.uni.lu
{jan.soelter, matias.bossa, andreas.husch}@uni.lu

[2] National Dept. of Neurosurgery
Centre Hospitalier de Luxembourg
4, Rue Ernest Barble
Luxembourg

August 26, 2020

Abstract

Machine learning based methods for diagnosis and progression prediction of COVID-19 from imaging data have gained significant attention in the last months, in particular by the use of deep learning models. In this context hundreds of models where proposed with the majority of them trained on public datasets. Data scarcity, mismatch between training and target population, group imbalance, and lack of documentation are important sources of bias, hindering the applicability of these models to real-world clinical practice. Considering that datasets are an essential part of model building and evaluation, a deeper understanding of the current landscape is needed.

This paper presents an overview of the currently public available COVID-19 chest X-ray datasets. Each dataset is briefly described and potential strength, limitations and interactions between datasets are identified. In particular, some key properties of current datasets that could be potential sources of bias, impairing models trained on them are pointed out.

These descriptions are useful for model building on those datasets, to choose the best dataset according the model goal, to take into account the specific limitations to avoid reporting overconfident benchmark results, and to discuss their impact on the generalisation capabilities in a specific clinical setting.

***K*eywords** COVID-19 · machine learning · datasets · X-Ray · imaging · overview · bias · confounding

1 Introduction

Just a bit more than half a year has passed since the novel corona virus SARS-CoV-2 gained world wide attention and eventually developed to the global COVID-19 pandemic [1]. In course of the pandemic, diagnostics play a vital role in the management of cases and the distribution of potentially scarce resources, like hospital/ICU beds. Hence, there is an urgent necessity to create trustworthy tools for diagnosis and prognosis of the disease. While most of the people with COVID-19 infection do not develop pneumonia [2], the identification of induced COVID-19 pneumonia cases is essential.To this end, imaging studies such as

X-ray and CT are employed. Chest X-rays (CXRs), while less sensitive that CT, are widely available, fast, non-invasive, and relatively cheap tests to diagnose and monitor COVID-19 induced pneumonia [3]. Despite there is no single feature specifically associated with covid-19 induce pneumonia on a chest, its been associated with a combination of multifocal peripheral lung changes of ground glass opacity and/or consolidation, which are most commonly bilateral[2] .

Machine learning, and especially deep learning methods, promise to assist the radiologist in coherent diagnosis and interpretation of images [4, 5]. To this end, a remarkable amount of machine learning models has been proposed in this very short amount of time to tackle the problems of COVID-19 diagnosis, quantification and prognosis from X-Ray imaging [6, 7, 8]. But there is growing awareness in the community that the presence of different sources of bias significantly decreases overall generalisation of models, while model performances are overestimated in internal validation [9, 10, 11, 12].

The most common sources of bias are unknown confounders and selections bias. A confounder [13] is a variable that affects both, the potential predictor variable (the image in this case) and the outcome (mainly diagnosis and radiological findings). If we are not aware of confounders, and do not control for them, we can erroneously conclude that a given image feature is a strong predictor of the outcome, when in reality the association is spurious and it does not hold any more in a different setting, where the confounder takes different values. Selection bias [14] happens when subjects are not selected at random from the target population. If the image and the outcome affect the selection criteria a spurious association between image and outcome appears. As in the case of confounders, this association may not by present when the selection criteria are different and the model will misbehave. Other sources of bias [15], such as image noise, errors in outcome assignment or population prevalence imbalance could also lead to biased models. They can be mitigated in many cases, as far as the information to identify them is available. Deep learning models present the additional inconvenient, with respect to traditional regression models, that it is not always easy to control for confounders, but this issue is outside the scope of this paper.

In order to avoid or at least detect bias, it is important that datasets and models are well documented. In a recent review and critical appraisal for prediction models for diagnosis and prognosis of COVID-19 [16] all evaluated models were rated at high risk of bias. Authors arrived to the conclusion that they "do not recommend any of these reported prediction models for use in current practice". The most common causes of risk of bias in diagnostic models based on medical imaging was: lack of information to assess selection bias (such as how controls were selected or which patients had imaging) and lack of clear report of image annotation procedures and of quality control.

To raise awareness of such problems, we aim to give a comprehensive overview on current publicly available chest X-ray datasets, identifying strengths as well as limitations, such as most evident potential sources of bias. To this end, we tried to reconstruct the story behind each dataset in section 2 describing the way images were obtained, which additional information is provided, and how the corresponding labels have been generated. Thus this overview provides a more in depth description of datasets than previous works [17, 6, 7, 8]. Based on these findings, we discuss in section 3 the specific risks of bias posed by these datasets. Finally, we give some concluding remarks in section 4.

2 Publicly Available datasets

In this section we present a collection of the currently available public datasets for chest X-ray that present clinical manifestations of COVID-19 pneumonia. In addition, some non-COVID-19 Chest X-ray dataset were also included because they have been utilised in models for pre-training [18] or to enrich datasets with control cases. Table 1 gives an overview on all presented datasets.

2.1 Radiological case reports of COVID-19

Different radiological associations are making efforts to collect and provide images and reports on the internet to facilitate knowledge transfer of radiologist. Among those, `www.eurorad.org`, `www.sirm.org`, `www.radiopedia.org`, `www.figure1.com` and `www.bsti.org.uk` provide images which are usually accompanied by a radiological description and varying depth of clinical information. While these great initiatives allow for fast transfers of information during this emergency times, the extraction of both, image and meta-information, is not straightforward. Fortunately, several initiatives such as Cohen/IEEE 8032 dataset (2.2.1) collect and present this information in a structured dataset format.

Table 1: Dataset collection for COVID-19 and non-COVID-19 Chest X-ray described in the paper. 2.2 Sections are COVID-19 oriented datasets, 2.3 are non-COVID19 oriented datasets and 2.4 correspond to compilation datasets

Section	Dataset name	COVID-19	Origin
2.2.1	Cohen/IEEE 8032 [1] [19]	✓	Global
2.2.1	Brixia-score-COVID-19 [2] [20]	✓	re-annotated subset from Cohen/IEEE 8032
2.2.1	GeneralBlockchain covid-19 [3]	✓	re-annotated from Cohen/IEEE 8032
2.2.2	agchung/Figure1 [4]	✓	Global
2.2.3	ML Hannover [5]	✓	Germany
2.2.4	agchung/Actualmed [6]	✓	Spain
2.2.5	Cancer Imagine Archive [7]	✓	USA
2.2.6	HM Hospitals [8]	✓	Spain
2.2.7	BIMCV-PADCHEST [9] [21]		Spain
2.2.7	BIMCV-COVID19 [10] -PADCHEST		Spain
2.2.7	BIMCV-COVID19+ [10] [22]	✓	Spain
2.3.1	CheXpert [11] [23]		USA
2.3.2	ChestXray-NIH [12] [24]		USA
2.3.2	RSNA Pneumonia Kaggle [13] [25]		re-annotated subset from ChestXray-NIH
2.3.2	ChestXray-NIH Google [14] [26]		re-annotated subset from ChestXray-NIH
2.3.3	Montgomery [15] [27]		USA
2.3.4	Shenzhen [16] [27]		China
2.3.5	UCSD-Guangzhou pediatric [17] [28]		China
2.3.6	MIMIC-CXR-JPG v2.0.0 [18] [29]		Israel
2.3.7	Indiana University/OpenI [19] [30]		USA
2.4.1	Kaggle COVID-19 radiography [20] [31]	✓	Global
2.4.2	V7 Darwin covid-19-chest-x-ray [21]	✓	Global
2.4.3	COVIDx [22] [32]	✓	Global

[1] https://github.com/ieee8023/covid-chestxray-dataset
[2] https://github.com/BrixIA/Brixia-score-COVID-19#license-and-attribution
[3] https://github.com/GeneralBlockchain/covid-19-chest-xray-segmentations-dataset
[4] https://github.com/agchung/Figure1-COVID-chestxray-dataset
[5] https://github.com/ml-workgroup/covid-19-image-repository
[6] https://github.com/agchung/Actualmed-COVID-chestxray-dataset
[7] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226443
[8] https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version
[9] https://bimcv.cipf.es/bimcv-projects/padchest
[10] https://github.com/BIMCV-CSUSP/BIMCV-COVID-19
[11] https://stanfordmlgroup.github.io/competitions/chexpert/
[12] https://nihcc.app.box.com/v/ChestXray-NIHCC
[13] https://www.kaggle.com/nih-chest-xrays/data
[14] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview/description
[15] http://openi.nlm.nih.gov/imgs/collections/NLM-MontgomeryCXRSet.zip
[16] http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip
[17] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[18] https://physionet.org/content/mimic-cxr-jpg/2.0.0
[19] https://openi.nlm.nih.gov/imgs/collections/NLMCXR_png.tgz
[20] https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
[21] https://github.com/v7labs/covid-19-xray-dataset
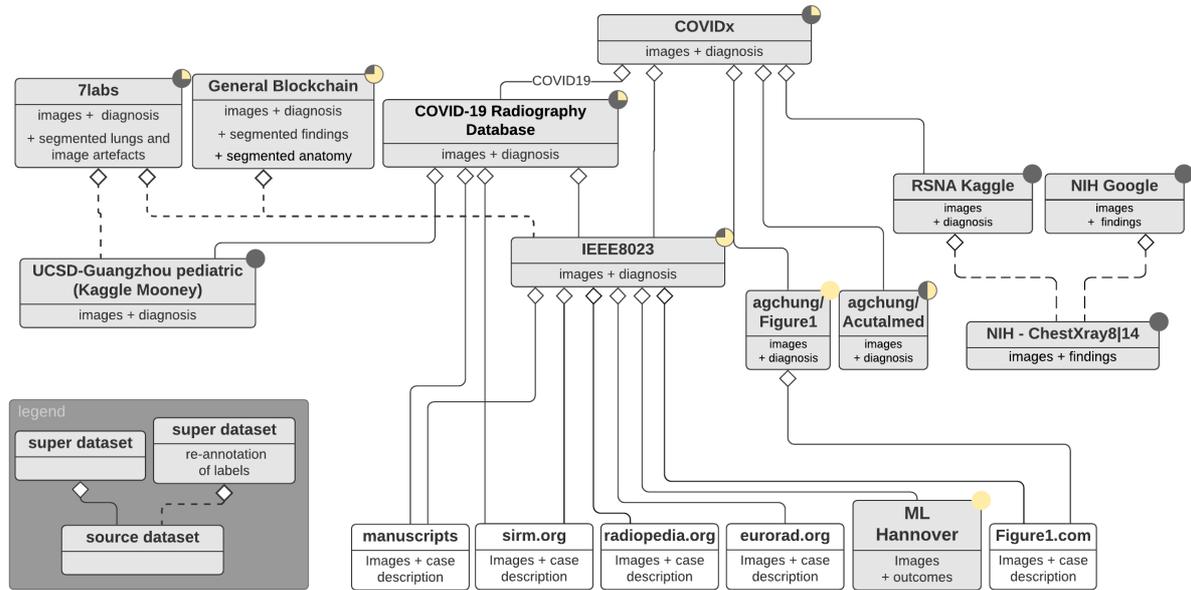[22] https://github.com/ieee8023/covid-chestxray-dataset

Figure 1: Overview on the relationship of interdependent datasets. Grey boxes represent datasets which can be downloaded as a single entity, whereas transparent boxes are general sources from where information has to be extracted manually (or with automated scraping tools). Diamond shapes indicate that the attached dataset includes (parts) of the connected dataset. Dotted lines describe the case when images of a source dataset are included, but the labels are discarded and new labels have been obtained in a re-annotation process. The coloring of the circle in the upper right corner of each datasets gives the rough proportion of cases with radiological findings of COVID-19 (from full yellow for datasets with only COVID-19 to full gray for datasets without any COVID-19 case.)

2.2 Datasets containing COVID-19 samples

2.2.1 Cohen/IEEE 8032

The Cohen/IEEE 8032 dataset [19] is a collection of cases "extracted from online publications, websites, or directly from the PDF", including the aforementioned case sharing websites (2.1) . Additionally, it incorporates the ML Hannover dataset (2.2.3). It provides a well curated tabular view on the image meta-data containing the source document of the image, the imaging origin (hospital, city or country) and information about scanning view for most images, as well as gender, age and a variety of clinical features and outcomes for some cases. Each image is assigned a diagnosis of respiratory disease, with a strong focus on COVID-19 (currently 450 out of 650) and very little cases of no finding (20).Additionally the dataset contains global severity scores for 100 images created in a post-hoc analysis of images according to a severity scheme [18].

Brixia-score-COVID-19 This dataset is the result of a post-hoc annotation process two in which two radiologist labelled images of a private dataset and about 200 images of the pubic Cohen/IEEE 8032 dataset with "BrixIA", i.e. localised severity scores [20].

GeneralBlockchain COVID-19 segmentations This dataset is a snapshot of the Cohen/IEEE 8032 (2.2.1) in which two radiologist created, in a post-hoc annotation process, segmentation masks for anatomical parts, medical devices (e.g. tubes and probes) and radiological findings (Ground glass opacities and Consolidations).

2.2.2 agchung/Figure1

This dataset contains images of 48 patients extracted from the aforementioned case sharing website `www.figure1.com` (2.1). For some patients, it contains additional information about sex, temperature, pO2 saturation, the scanning view as well as some clinical notes. Most of images are assigned one of the labels COVID-19, pneumonia or no finding.

2.2.3   ML Hannover

This dataset [33] from the Institute for Diagnostic and Interventional Radiology (Hannover, Germany), contains 240 chest-xray images (CR, DX) from 71 patients at different timepoints in the course of COVID-19. It contains metadata including scanning view (AP vs PA), patient master data, laboratory data and longitudinal information on admission, ICU-admission and death.

2.2.4   agchung/Actualmed

The Actualmed dataset contains images (CR, DX) of 215 patients. Each image is assigned a radiological diagnosis of "COVID19", "inconclusive" or "No finding". In addition it contains metadata of date and scanning view (AP vs PA). An analysis on the prevalence of COVID-19 findings within the different scanning views reveals a significantly higher prevalence of cases with COVID-19 cases in the AP scanning view.
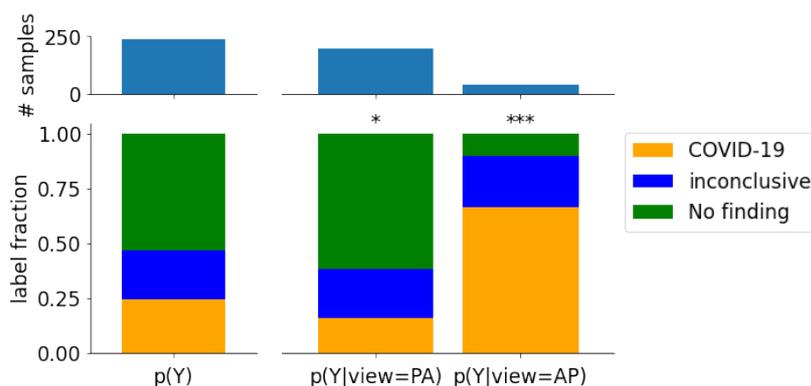
Figure 2: Top panel: Number of samples in the agchung/Actualmed dataset, for the full dataset (left) and for each scanning view (right). Bottom panel: fraction of associated labels in the full dataset (left) and conditioned on each scanning view (right). Stars mark the significance of the conditional distribution differing from the marginal distribution according to a χ^2 test.

2.2.5   Cancer Imagine Archive

This dataset was collected by the University of Arkansas for Medical Sciences (USA). It contains more than 200 chest X-ray studies for patients tested positive for COVID-19 including temporal data and key radiological findings. Moreover it provides extensive meta-information such as age, gender, race, several comorbidities and indicators of disease evolution such as ICU admission and mortality. Its focus lies on underrepresented rural populations in the USA.

2.2.6   HM Hospitals

This dataset contains more that 5000 chest X-ray for COVID-19 (PCR positive or pending) in DICOM format, hence it contains information about gender, view, modality and others. Additional longitudinal clinical information includes admission times, treatment, vital signs, laboratory results and disease progression information such as ICU admissions, discharge or death. According to dataset owners, the intended use is to build prediction models of disease evolution, epidemiological models and information on the response to the various treatments.

2.2.7   BIMCV

The Medical Imaging Databank of the Valencia Region (Banco digital de Imagen Medica de la Comunidad Valenciana) provides both a general chest x-rays dataset from pre-COVID19 times and a dataset specially with COVID-19 confirmed cases.

BIMCV-PADCHEST   The Padchest dataset [21] contains images (CR, DX, CT) of about 67000 patients together with their radiological reports from a single hospital (Hospital San Juan Hospital (Spain)) between 2009 and 2017. It contains extensive DICOM meta-data including sex, age and scanner manufacturer. DICOM

Scanning positions are given and grouped into 6 main classes (standard PA, standard L, AP vertical, AP horizontal, pediatric, and rib views) for images with position information available, otherwise a pre-trained CNN-model assigned automatically either one of the standard PA or L view.

To facilitate model training, the authors mapped the radiological reports into a hierarchical set of categories, i.e. "labels were extracted from the radiological reports, resulting in 22 differential diagnoses, 122 anatomic locations and 189 different radiological findings mapped onto standard Unified Medical Language System (UMLS)", including the presence of medical devices. This mapping of the reports was carried out for 27% of the dataset in a manual annotation process and completed for the rest of the data by a trained model.

BIMCV-COVID19-PADCHEST This is a subset of the BIMCV-PADCHEST dataset to four classes ("Control", "Infiltration without Pneumonia", "Pneumonia without Infiltration" and "Pneumonia with Infiltration"). It was created to facilitate (pre)-training of models related to COVID-19.

BIMCV-COVID19+ This dataset contains images (CR, DX, CT) of 1311 patients together with their radiological reports collected between February and April 2020 in 11 different hospitals of the Valencia region [22]. As within Padchest, 724 reports were manually mapped to the UMLS schema of the padchest dataset with two new categories (COVID19, COVID19-uncertain), and subsequently the remaining reports were mapped using a trained annotation model. Additional to the radiological reporting, the dataset contains for each patient the result of one or more PCR-test obtained after the imaging (usually within 5 days), whereby each patient had at least one positive PCR test. Furthermore in 10 images there are detailed ROI annotation for the radiological findings.

An analysis about prevalence of COVID-19 radiological findings with regard to the DICOM "series description" shows a significant prevalence difference between the varying acquisition protocols. The least prevalence is observed in the "Erect PA" view, whereas the highest prevalence is present ind the "AP horizontal" and "Thorax AP" view.
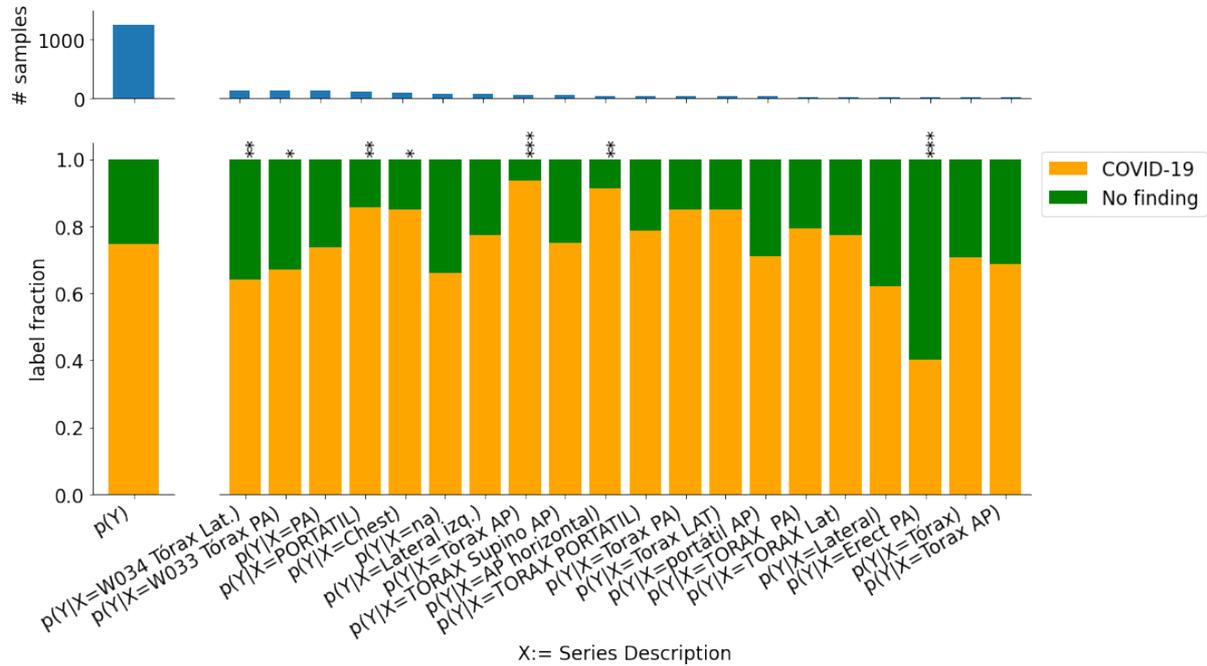

Figure 3: Top panel: Number of total samples in the full BIMCV-COVID19+ dataset (left) and in subsections partiioned by the DICOM metadata entry of Series Description (right). Bottom panel: Fraction of radiological findings in the full BIMCV-COVID19+ dataset (left) and when conditioned on the Series Description (right). Stars mark the significance of the conditional distribution differing from the marginal distribution according to a χ^2 test.

2.3 Chest X-ray datasets without COVID-19 samples

2.3.1 CheXpert

This large datasets [23] released at the beginning of 2019 includes 224,316 chest radiography from 65,240 patients retrospectively collected from Stanford hospital. For fourteen classes of radiological findings associated with the most common respiratory diseases a rule-based processing of the reports generated labels of presence, absence, or uncertainty thereof. Additionally this dataset comprises a dedicated testset, with 500 images annotated by consensus of eight radiologist.

2.3.2 NIH

ChestXray-NIH This dataset [24] contains 120k images of 30k patients from the clinical PACS database at National Institutes of Health Clinical Center. In its original version, all images were labelled with originally eight (ChestXray8) and in the current version fourteen (ChestXray14) different findings from thoracic pathology. The labels have been extracted from the corresponding radiological reports by Natural Language Processing. The underlying reports are not publicly available, but the dataset contains meta-information about gender, age and view position. Additionally there are annotated bounding boxes for all findings in about 1000 images.

RSNA Pneunomia Kaggle This dataset [25] is a subset of 30k images from the ChestXray-NIH dataset with an enrichment of images with a Pneunomia related diagnosis. In a well defined annotation process a team of radiologist annotated areas of Lung Opacity with bounding boxes solely based on the information present in the image.

ChestXray-NIH Google This is another subset of 18k images from the ChestXray-NIH dataset with extra labels [26]. In another well designed annotation process radiologist assigned each image only with "access to the patient age and image view (PA/AP), but not to additional clinical or patient data" the presence/absence of four findings: pneumothorax, opacity, nodule/mass and fracture.

2.3.3 Montgomery

This pre-COVID-19 pandemic dataset (2014) contains 138 frontal chest images collected in the department of health and human services, Mongomery Country, Maryland, USA. To each image, a short radiological report and a disease diagnosis is assigned (58 images with tuberculosis manifestations and 80 controls), as well as a lung segmentation annotation automatically generated under the supervision of a radiologist using anatomical landmarks [34]. The images themselves contain written markings of the scanning view (AP and PA) and there is additional metadata about gender and age. The intended used of the dataset is to boost of computer-aided diagnosis for pulmonary diseases with focus on TB [27].

2.3.4 Shenzhen

This dataset, released together with the Montgomery dataset contains images collected in collaboration the Shenzhen Hospital in China. It contains 662 frontal X-ray, of which 326 are normal and 336 contain TB manifestations. Additionally metadata includes sex, age and a short radiological description [27].

2.3.5 UCSD-Guangzhou pediatric

This dataset contains more than 5000 chest X-ray images from children (AP view) selected from retrospective cohorts of pediatric patients of 1-5 year old from Guangzhou Women and Children's Medical Center, Guangzhou [28]. In an annotation process by two experts all images were assigned a diagnosis of viral/bacterial Pneumonia or Normal. No further meta-data is available.

2.3.6 MIMIC-CXR-JPG v2.0.0

This large pre-COVID-19 dataset comprise 377,110 chest x-rays associated with 227,827 de-identified imaging studies sourced from the Beth Israel Deaconess Medical Center. Images are provided with 14 labels derived from two natural language processing tools (NegBio and CheXpert) applied to the corresponding free-text radiology reports. [29].

2.3.7 Indiana University Chest X-rays/OpenI

This dataset from the Indiana University was created to provide a publicly available searchable database and comprises of 7500 Chest X-rays [30]. It includes view information, radiological reports including main findings and impressions. In a well documented post-hoc annotation process the reports have been mapped to localised findings, e.g. "Cicatrix/lung/base/left".

2.4 Compilation datasets

2.4.1 Kaggle COVID-19 radiography database

The Kaggle COVID-19 radiography database [31] is compiled from different datasets. It contains COVID19 cases form the aforementioned Cohen/IEEE 8032 (2.2.1) as well as cases extracted from the sirm.org website and from 43 publications. Furthermore it incorporates the data of the UCSD-Guangzhou pediatric dataset (2.3.5) providing control cases and cases of viral pneumonia except COVID19. Thus the dataset contains the 3 labels COVID19, normal and viral pneumonia. The compiled dataset itself does not include any meta-information, except the image dataset source.

2.4.2 V7 Darwin covid-19-chest-x-ray-dataset

This is another dataset that is a compilation of the UCSD-Guangzhou pediatric dataset (2.3.5) for non-COVID-19 and the Cohen/IEEE 8032 for COVID-19 cases (2.2.1). From a downstream annotation process it includes for all images extra lung segmentation masks and ignore masks marking radiologist's drawings (e.g. arrows) and medical devices overlapping the lungs.

2.4.3 COVIDx

The COVIDx dataset [32] is compiled from different sub-datasets. It contains the aforementioned Cohen/IEEE 8032 (2.2.1), angchung/Actualmed (2.2.4) and angchung/Figure1 (2.2.2) dataset and the COVID-19 cases of the COVID-19 radioagraphy database (2.4.1) as well as Pneumonia and Normal cases from the RSNA Kaggle Pneumonia dataset (2.3.2). Thus each image is assigned one of the labels COVID-19, Pneumonia or Normal. As these sub-datasets contain common sources, the authors take care to remove duplication with the same source URL. Nonetheless one should still acknowledge that there is the risk of case duplication due to incoherent source descriptions, e.g. it is possible to find occasional cases with suspicious similar metadata (Table 2)

Table 2: Example of two cases in the COVIDx dataset which are highly similar accroding to the available metadata

patientid	sex	age	temperature	notes
150	M	28	39.1	28M previously fit and well, not on any regular medications, presented with a 6 day Hx of fever, non-productive cough and SOB for the last 4 days. His symptoms started as sore throat and coryzal symptoms 8 days prior to his presentation and he reported contact with a friend with similar symptomatology. [...]
COVID-00003a	M	28	39.1	28M previously fit and well, not on any regular medications, presented with a 6 day Hx of fever, non-productive cough and SOB for the last 4 days. His symptoms started as sore throat and coryzal symptoms 8 days prior to his presentation and he reported contact with a friend with similar symptomatology. [...]

3 Discussion

Medical Imaging Models, especially Convolutional Neural Networks (CNNs), are known not only to learn underlying diagnostic features, but also to exploit confounding image information. For example, it was shown that the acquisition site, regarding both the hospital system and the specific department within a hospital, can be predicted with very high accuracy ($> 99\%$) [11]. If disease prevalence is associated with the acquisition site, as it is often the case, this can be a strong confounder. Thus, in any composite dataset having separated sub-datasets for COVID-19 and control cases, the dataset is completely confounded with the group label and therefore it is difficult to isolate the disease effect from dataset effect, making learning almost impossible and posing a high risk of overestimating prediction performance. Indeed it has been observed that, by training on different COVID-19 and non-COVID-19 dataset combinations, the "deep model specialises not in recognising COVID features, but in learning the common features [of the specific datasets]" [35]. Eventually, a CNN model is able to identify the source dataset with a high accuracy ($> 90\%$) solely from the image border region containing no pathology information at all [12].

Besides acquisition site, the demographic characteristics of populations can also be a strong confounder. Datasets that take cases from the UCSD-Guangzhou pediatric dataset (2.3.5) as non-COVID-19 examples (maximum age 5y) pose the risk that models will associate anatomical features of age with the diagnosis since, for example in the Cohen/IEEE 8032 (2.2.1) dataset, the minimum age is 20y (mean 54y). If controlling for confounders is already difficult in deep learning models, normalising the images in such a wide and disjoint range of ages seems an impossible task.

But one not only has to be wary in composite datasets, also single source datasets are not free of potential confounders and other sources of bias. The classical example would be a different imaging protocol depending on the patient's health status. For example, the PA erect view is the preferred imaging view in general, but if the patient is not able to leave the bed it is much more common to do an AP view image. Indeed, in two exemplary datasets that contain labels of radiological findings, one can observe a significant change in prevalence of COVID-19 cases depending on the scanning view (see Figures 3 and 2).

Another confounding factor might be the presence of medical devices like ventilation equipment or ECG cables, which allows a model to associate images with patient treatment instead of disease status. For example, for the NIH ChestXray14 dataset (2.3.2), a critical evaluation has shown that "in the pneumothorax class, [...] 80% of the positive cases have chest drains. In these examples, there were often no other features of pneumothorax" [36]. Datasets which provide additional annotations on the presence of medical devices (e.g. "BIMCV", "7labs", "General Blockchain") facilitate a risk analysis on this confounding effect and also enable mitigation strategies in training.

In general, one has to distinguish between labels which have been annotated by taking only the image itself into account, and labels which have been generated by a different source, i.e. from another diagnostic method like CT or PCR. Unfortunately, radilogical reports done in a in clinical routine are a mixture of both. Radiologist are often aware of the patients clinical context and this information is reflected in the reports, since they are done with the aim of communicating information between different doctors. For example, has been shown for the NIH ChestXray14 dataset (2.3.2) that, in a substantial fraction of images, the associated finding extracted from the reports can not be confirmed by a post-hoc assessment of the the images [36].

Biases arise more easily when outcome labels and prediction model intended application is not clearly defined. If the model objective is to find radiological manifestation of the disease in the images that are not necessarily apparent to the radiologist naked eye, the labels should be the best possible diagnostic assessment obtained by any diagnostic test that doesn't include image information from same modality. For example, a perfectly righteous goal could be to determine whether a feature observed in CT, but not visible in XR, could be detected by subtle signals that ML models can identify. In contrast, if the goal is to reproduce radiological findings (for example, to save radiologist time) the label should be radiological annotations assessed by an independent clinician that has no information except for the image. Otherwise the risk of bias increases sensibly and the generalisation ability is compromised, because we can not really understand where the key information is coming from, what the model is learning, and what the possible sources of bias are. In this sense, it's worth noting that a couple of datasets do provide such annotations solely derived from the images (RSNA Kaggle, NIH Google, General Blockchain, 7labs).

4 Concluding Remarks

We identify serious limitations in most, if not all, currently available datasets. It is urgently needed that more images from larger and better datasets are made publicly available. Dataset owners should make an

effort to improve documentation about whole dataset building process to increase significantly the dataset value and the quality of models trained on them. For example: there should be a clear statement of dataset intended use, and explicit warning of common misuse cases; label definition and generating procedure should be reported in detail, so that other researchers can verify accuracy of label assignments and evaluate the utility and adequacy to the problem at hand; finally, datasets should contain cohort characteristics and subject selection criteria information, in order to evaluate the risk of selection bias and to check if the training and target population has similar characteristic.

Contrary to classical statistical models or standard machine learning methods, deep learning models are highly complex systems that may have several building steps. For reasons that range from avoiding overfitting to reducing memory and computation needs, some parts of the models are sometimes pre-trained, using specific dataset, and then kept fixed as other parts of the model are fintuned later. Quality standards of datasets used to pre-train these building blocks may not necessarily be as high as the ones for finetuning the final model, and some of these datasets that are deemed close to useless for training a serious medical diagnostic tool may be perfectly appropriate for pre-trainig steps. One only has to be cautious not including those subjects in the cross-validation loops of the final model.

Although dataset quality is the most important requirement for a medical diagnostic system to be reliable, other aspects of the model building are also prone to biases. Adherence to transparent practices, such as the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) reporting guideline [37], and assessing risk of bias with PROBAST (Prediction model Risk Of Bias ASsessment Tool) [38], would be a starting point. However extensions of these guidelines are required in order to be fully applicable to deep learning systems [16].

We hope our overview will help modellers to choose the appropriate dataset for their modelling needs while, at the same time, raise awareness on biases to look out for while training models. We also encourage everyone to validate models and report benchmarking results on a possibly small but very well curated external dataset, which is carefully selected to represent the real clinical use case as close as possible. Of course, in a best case scenario such gold standard test datasets would be publicly available to transparently foster progress in the field.

Acknowledgement

References

[1] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). *International Journal of Surgery*, 2020.

[2] Joanne Cleverley, James Piper, and Melvyn M Jones. The role of chest radiography in confirming covid-19 pneumonia. *bmj*, 370, 2020.

[3] Rowa Aljondi and Salem Alghamdi. Diagnostic value of imaging modalities for covid-19: Scoping review. *Journal of Medical Internet Research*, 22(8):e19673, 2020.

[4] Garry Choy, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E Samir, Oleg S Pianykh, J Raymond Geis, Pari V Pandharipande, James A Brink, and Keith J Dreyer. Current applications and future impact of machine learning in radiology. *Radiology*, 288(2):318–328, 2018.

[5] Morgan P McBee, Omer A Awan, Andrew T Colucci, Comeron W Ghobadi, Nadja Kadom, Akash P Kansagra, Srini Tridandapani, and William F Auffermann. Deep learning in radiology. *Academic radiology*, 25(11):1472–1480, 2018.

[6] Afshin Shoeibi, Marjane Khodatars, Roohallah Alizadehsani, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Ali Khadem, Delaram Sadeghi, Sadiq Hussain, Assef Zare, et al. Automated detection and forecasting of covid-19 using deep learning techniques: A review. *arXiv preprint arXiv:2007.10785*, 2020.

[7] Md Islam, Fakhri Karray, Reda Alhajj, Jia Zeng, et al. A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). *arXiv preprint arXiv:2008.04815*, 2020.

[8] Muhammad Ilyas, Hina Rehman, and Amine Naït-Ali. Detection of covid-19 from chest x-ray images using artificial intelligence: An early review. *arXiv preprint arXiv:2004.05436*, 2020.

[9] Charlotte Soneson, Sarah Gerster, and Mauro Delorenzi. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS one*, 9(6):e100335, 2014.

[10] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. *arXiv preprint arXiv:2002.02497*, 2020.

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

[12] Gianluca Maguolo and Loris Nanni. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *arXiv preprint arXiv:2004.12823*, 2020.

[13] Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statist. Sci.*, 14(1):29–46, 02 1999.

[14] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[15] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[16] Laure Wynants, Ben Van Calster, Marc MJ Bonten, Gary S Collins, Thomas PA Debray, Maarten De Vos, Maria C Haller, Georg Heinze, Karel GM Moons, Richard D Riley, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj*, 369, 2020.

[17] Md Fahimuzzman Sohan. So you need datasets for your covid-19 detection research using machine learning? *arXiv preprint arXiv:2008.05906*, 2020.

[18] Joseph Paul Cohen, Lan Dao, Paul Morrison, Karsten Roth, Yoshua Bengio, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, Marzyeh Ghassemi, Haifang Li, et al. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *arXiv preprint arXiv:2005.11856*, 2020.

[19] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[20] Alberto Signoroni, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, and Davide Farina. End-to-end learning for semiquantitative rating of covid-19 severity on chest x-rays. *arXiv 2006.04603*, 2020.

[21] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, page 101797, 2020.

[22] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco Garcia, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

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

[24] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[25] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[26] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.

[27] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[28] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[29] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

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

[31] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandaker Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al-Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *arXiv preprint arXiv:2003.13145*, 2020.

[32] Zhong Qiu Lin Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images, 2020.

[33] Hinrich B. Winther, Hans Laser, Svetlana Gerbel, Sabine K. Maschke, Jan B. Hinrichs, Jens Vogel-Claussen, Frank K. Wacker, Marius M. Höper, and Bernhard C. Meyer. COVID-19 Image Repository. *figshare.com*, 5 2020.

[34] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

[35] Enzo Tartaglione, Carlo Alberto Barbano, Claudio Berzovini, Marco Calandri, and Marco Grangetto. Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data. *arXiv preprint arXiv:2004.05405*, 2020.

[36] Luke Oakden-Rayner. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112, 2020.

[37] Karel GM Moons, Douglas G Altman, Johannes B Reitsma, John PA Ioannidis, Petra Macaskill, Ewout W Steyerberg, Andrew J Vickers, David F Ransohoff, and Gary S Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Annals of internal medicine*, 162(1):W1–W73, 2015.

[38] Robert F Wolff, Karel GM Moons, Richard D Riley, Penny F Whiting, Marie Westwood, Gary S Collins, Johannes B Reitsma, Jos Kleijnen, and Sue Mallett. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58, 2019.