Deep Learning on medical images: ill-defined labels can be a good thing

Victor Fornasiero de Paiva∗1
1 Hospital Israelita Albert Einstein, Av. Albert Einstein 627, SP 05652-900, Brazil
Carlos Arruda Baltazar∗1
Andre Fonseca1,2
2 Universidade Federal do ABC, Av. dos Estados, 5001, SP 09210-580, Brazil
Maria Carolina Bueno da Silva†1
Birajara Soares Machado†1

Editors: Under Review for MIDL 2021

Abstract

It is usual to come across public available datasets that often provide label distributions that do not match the real world distributions (RWD) where the data came from when working on deep learning (DL) algorithms for medical images. On top of that, it is possible to observe that data curation strategies developed for dataset creation could be partially responsible for the distancing of distributions present on said medical image datasets. By data curation we mean all processes of cleaning and transformation of the acquired data following certain semantic conditions and domain-specific knowledge. Although it is expected that the curation stage modify the distribution of the data, we signal that one must tap gently into the cleaning and transformation steps of data curation in order to avoid rendering a dataset worthless from the perspective of supervised learning DL algorithms. In this study, we will explore the experiences we had on the creation of two chest x-ray (CXR) datasets and one DL model for pulmonary tuberculosis (TB) classification. By promoting the discussion of data curation not only as process to filter noises and artifacts related to the data itself, but as a multidisciplinary approach needed to reduce unwanted bias in the process of dataset creation while demanding fidelity to RWD, we hope more attention will be bestowed upon the current limitations of DL models trained with medical image datasets that fall short of maintaining real world applicability.

Keywords: deep learning, image analysis, chest x-rays, tuberculosis, data curation

1. Introduction

Deep learning (DL) models are one of the cut edge technologies for medical image analysis. When learning from complex and unstructured data, DL analysis has provided insights into pathology visualization across multiple modalities of medical images with potential applications in the healthcare context as a valuable tool for detection, segmentation, classification, among others. All of the aforementioned applications are based on expertise from a wide variety of fields which demands multi-disciplinary teams to achieve valuable and consistent results on projects such as computer-aided diagnosis (CAD) and clinical workflow optimization (Montagnon et al., 2020).

∗ Contributed equally
† Contributed equally

Deep Learning on medical images: ill-defined labels can be a good thing

DL techniques for pattern recognition in medical images became widespread in the last decade, highlighting enormous advances in automatic chest x-ray (CXR) analysis and generation of radiology reports. The DL approach can be considered the state of art method for those tasks but it is understandable that there can be inherent limitation to its appeal whereas the highest performance DL classifiers generally are the least explainable ones because of a series of ‘black box’ functions taking part on the decision making processes (Gunning and Aha, 2019). Additionally, the image features learned by DL models and the relevant radiology findings determined by specialists can be in disagreement, hampering the interface between the algorithm and healthcare professionals and ultimately leading to the conclusion that problems related to weak dataset curation methods are widening the gap between the real capabilities of DL models in the healthcare and the current initiatives for dataset creation (van Ooijen, 2019). Those issues have been strongly debated and novel techniques have been proposed to promote explainable DL models, providing support for the radiologists in the interpretation of DL outcomes and establishing a whole multidisciplinary process, which begins in data acquisition and curation stages.

Among infectious diseases, tuberculosis (TB) is the second deadliest worldwide. In 2019, the World Health Organization (WHO) estimated 10 million people with TB resulting in 1.4 million deaths. In Brazil, the reality of TB infection is close to the BRICS WHO recommendations include the use of chest radiography in TB detection and guidance. As there’s a significant lack of radiologists and an increasingly amount of exams in TB-prevalent locations, the development of CAD systems can be considered essential for TB screening and diagnosis (Lakhani and Sundaram, 2017).

Brazil has a public health system called the Unified Health Service (SUS). In this system, universal health coverage at different levels of assistance is provided free of charge to the population. Under the precepts of universality, integrality, equity, and community participation, SUS has made progress and has had setbacks in its implementation (Castro et al., 2019). The lack of financial resources, infrastructure and medical specialists limits the healthcare strategy in an unequal country (Fleury, 2011). In contrast, highly complex medical and hospital institutions located in large urban centers have high technology and resources to assist their patients (Castro et al., 2019). Thus, it is observed that the health system in Brazil is heterogeneous and has a multifaceted reality (Paim et al., 2011).

There are few public available chest x-ray (CXR) datasets for pulmonary TB, and the majority of those contrast positive pulmonary TB CXRs with control cases, meaning DL models taking advantage of those datasets tend to learn how to screen for patterns likely related to pulmonary TB only when there are not other pathologies present. In the real world of medical images, a binary scenario where two classes carry little overlap between visual features is unlikely to take place, so it is cause of concern to have datasets match this description. The so-called ”noise” caused by introducing classes and/or labels that show similar distribution of visual features can arguably result in a richer dataset, which can be seen as a representation of a broader picture of the RWD where the data came from. By exploring the experiences we had on both the creation of two CXR datasets and a DL model for pulmonary TB classification, we hope this discussion bring more attention to the limitations of the current public medical image datasets for pulmonary TB, highlighting possible aspects associated with good habits to be adopted for the future dataset creation endeavors in this context.
2. Material and Methods

2.1. Data

In order to meet the heterogeneity of the Brazilian health system, we acquired data from a public reference center for the treatment of TB and a highly complex private institution. The two scenarios represent different realities and contribute to a greater wealth of data, including different pulmonary pathologies, with different radiographic stages and degrees of pulmonary impairment, as well as different exam qualities. The acquired data is not public available.

A trained team selected medical records of patients with confirmed diagnosis of pulmonary TB and/or other pulmonary comorbidities. The diagnostic and follow-up radiographs were not digitized, but printed. The team photographed these radiographs using a cell phone application designed by us to help facilitate the process of data acquisition and storage in a cloud structure. All radiographic views and media were included. The views were anteroposterior (AP), posteroanterior (PA), profile and apicolordotic, as discriminated in Table 1. The CXRs were stratified by media as: bond paper (25 examples from the public health service); x-ray film (4477 examples from the public health service and 385 examples from the private health service) and couche paper (299 examples from public health service).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Volumetry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA</td>
</tr>
<tr>
<td>Public health service</td>
<td>3131</td>
</tr>
<tr>
<td>Private health service</td>
<td>350</td>
</tr>
<tr>
<td>Total</td>
<td>3481</td>
</tr>
</tbody>
</table>

The two datasets acquired by our team encompasses not only the CXRs photographs, but also demographic data, such as: gender; age; ethnicity; educational level; and the SINAN number (a national system for compulsory notification of diseases and medical conditions developed by Brazilian health authorities). When used for reporting TB related cases, the notification system gathers data on smear, sputum culture, rapid molecular test, drug sensitivity test, chest radiography, HIV status, among others. The Public health service dataset is composed by 1140 female subjects with a mean age of 40 (SD = 20) years and 1650 male subjects with a mean age of 40 (SD = 20) years old. The private health service dataset is composed by 103 female subjects with a mean age of 46 (SD = 22) years and 176 male subjects with a mean age of 40 (SD = 22) years old.

For the modeling process, we used a subset from the private health service dataset containing 344 X-ray film examples and a subset from the public health service dataset; the later comprising 2717 PA CXRs from different medias (17 bond paper examples; 2609 X-ray film examples; and 132 couche paper examples), composed by 1650 male subjects with a mean age of 41 (SD = 40) years old and 1043 female with a mean age of 43 (SD = 53) years old. Each CXR photograph was carefully inspected and cropped to select only the printed area of the radiograph. Four thoracic radiologists analyzed the CXRs for patterns compatible with pulmonary TB involvement, such as: consolidation; cavity; opacity; nodule; granuloma; infiltrate; atelectasis; pleural effusion; pleural thickening; lung volume reduction; fibrosis and bronchiectasis. With this information, it was possible to gain
a broader perspective of pulmonary impairment on the two datasets. Finally, each CXR photograph was carefully cropped to select only the printed area of the radiograph.

Additionally, the development of a DL model for TB classification counted with data from three public CXR datasets and an open-access platform for sharing tuberculosis related data. The datasets were the Montgomery County chest X-ray set, the Shenzhen chest X-ray set (Jaeger et al., 2014), and the CheXpert dataset (Irvin et al., 2019). The TB Portals program was the open-access platform (Rosenthal et al., 2017).

From the TB Portals program, we randomly selected 500 PA CXRs of pulmonary TB. 500 PA CXRs without radiology findings were randomly selected from the CheXpert dataset to serve as negative examples of pulmonary TB. The data from the Montgomery County chest X-ray set and the Shenzhen chest X-ray set were used in their entirety.

The data from the public and private health services comprised normal chest x-rays as well as chest x-rays with pulmonary changes other than TB. Considering the Public health service subset, there were cases of asthma; chronic obstructive pulmonary disease (COPD); other types of chronic obstructive pulmonary disease (OCOPD); atypical mycobacterial infections (AMI); bronchiectasis; pneumonia; silicosis; sarcoidosis; among others. In the Private health service dataset, there were cases of COVID-19; pneumothorax; emphysema; bronchopneumonia; lung cancer; and others. The presence of the five most common pathologies other than TB on the cases used for training were all together under 23.7% of the total number of examples from the public health service data and 16.6% of the total number of examples from the private health service data (see Table 2), which corresponds to 36.5% of the total number of negative examples used.

Table 2: The frequency of the five most common pathologies other than TB from the public and private health services subsets used for training the DL model for TB classification. COPD = chronic obstructive pulmonary disease; OCOPD = other types of chronic obstructive pulmonary disease; AMI = atypical mycobacterial infections.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pathology</th>
<th>Proportion</th>
<th>Number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public health service</td>
<td>Asthma</td>
<td>14.5%</td>
<td>380</td>
</tr>
<tr>
<td></td>
<td>COPD</td>
<td>3.1%</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>OCOPD</td>
<td>2.6%</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>AMI</td>
<td>2.2%</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Bronchiectasis</td>
<td>1.3%</td>
<td>34</td>
</tr>
<tr>
<td>Private health service</td>
<td>COVID-19</td>
<td>8.7%</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Pneumothorax</td>
<td>3.5%</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Emphysema</td>
<td>2.3%</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Bronchopneumonia</td>
<td>1.2%</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Lung cancer</td>
<td>0.9%</td>
<td>3</td>
</tr>
</tbody>
</table>
All selected examples for training were mapped for a binary scenario of presence of pulmonary TB. All of those with a confirmed diagnosis of pulmonary TB were mapped to a positive class (1) while all of those with a confirmed diagnosis of other pathology or without any radiology findings on the lung region were mapped to a negative class (0). The set was randomly split (stratified by class for each dataset) into a 80% training set, a 12% validation set (also called tuning set), and a 8% testing set. After the split process, the three sets were checked for the overlap of patient identifiers; the criterion being no overlap of patient identifiers between sets. This step was done to remove possibly similar visual features of CXRs from the same subject present on more than one set, ensuring all CXRs from any given subject to only be present in one of the sets. This procedure provided a 84% training set of 4123 cases (2587 TB (+); 1536 TB (−)), a 11% validation set of 534 cases (334 TB (+); 200 TB (−)), and a 5% testing set of 245 cases (111 TB (+); 134 TB (−)). Further details can be found in Figure 1.

2.2. Model

The binary cross-entropy loss function was used. A cost-sensitive learning method was adopted to provide a penalty to the majority class (Huang et al., 2016), where the inverse frequency of the positive class was used as a weight in the loss function for all positive examples, thus limiting the tendency to classify examples as part of the most frequent class (Zadrozny et al., 2003). The Lookaheads optimization algorithm (Zhang et al., 2019) with standard hyperparameter settings was used as the envelope to the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001. The DL model was trained for 1000 epochs with a learning rate decrease schedule of 50% at epochs 600, 800, and 900. Each batch of
examples for training consisted of 32 cases. The image resolution used was $512 \times 512$. The performance on the validation set was checked every 50 iterations.

Although a supervised learning method was adopted to train the image-wise classification aspect of the DL model, a weakly supervised learning method was used for pathology localization (Zhou, 2018). The framework in question was developed by Ye et al. (2020) and consists on the generation of feature maps following the class activation map (CAM) method introduced by Zhou et al. (2016) that are then passed through a series of operations and ultimately yield pixel-wise pathology probabilities inside the training loop.

The Convolutional Neural Network (CNN) used was the DenseNet-121 (Huang et al., 2017). A batch normalization strategy was adopted (Ioffe and Szegedy, 2015) as well as a drop-out rate of 20% on the fully-connected layers (Srivastava et al., 2014). The data augmentation regime applied to images consisted of spatial changes (random rotation; translation; distance changes on the z-axis) and color jitter changes (random brightness values; contrast values; saturation values).

### 3. Results and Discussion

The model performance for pulmonary TB detection was evaluated on the test set and its subsets. The accuracy was determined by the standard threshold 0.5 and through an optimized cut-off that achieves the maximum F1-score value, the harmonic mean of precision and recall, used to assess classifiers with different class prevalences (Lipton et al., 2014). Considering the global test set, Figure 2 (left plot) demonstrates the binary classification ROC curve with corresponding area-under-curve AUC= 0.94 and standard accuracy Acc= 0.85.

In table 3 is listed the model accuracy values for all datasets and decision thresholds addressed in this work. The optimized cut-off provides higher accuracy in almost all test data classification, due to the best precision-recall balance obtained by the F1 measure, considered useful for ranking models and analyzing the classifier robustness (Forman and Scholz, 2010).

![Figure 2: ROC curve, AUC, the standard accuracy for the test set, and the confusion matrices for the test set, considering the standard and optimized cut-offs.](image)
Table 3: Model accuracy on the test set stratified by dataset and accuracy following specific cut-off values (standard and optimized).

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Accuracy Standard cut-off (0.5)</th>
<th>Accuracy Optimized cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>Public health service</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>TB Portals subset</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>CheXpert subset</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Private health service</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Shenzhen</td>
<td>0.80</td>
<td>0.92</td>
</tr>
<tr>
<td>Montgomery</td>
<td>0.78</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Figure 2 (right plot) depicts the pulmonary TB classification correctness following the two thresholds under consideration. The optimized cut-off classification provides lower false-negatives cases than the standard approach. This optimized classifier based on the balanced trade-off between precision and recall is more efficient to analyze heterogeneous data. Further results can be seen in 3.

4. Conclusion

The performance of DL algorithms trained under a supervised learning approach highly depend on careful analysis of data acquisition protocols as well as thoughtful data annotation and curation stages. Even state-of-the-art DL models with excellent performance on fixed benchmark datasets can confer non-robust results when presented to RWD from the same modality of data it was trained on (Shen et al., 2017; Munappy et al., 2019). The necessity to rethink specific aspects related to the creation of medical image datasets is crucial to facilitate the initiatives of DL algorithms that are expected to be robust and capable of learning representations and abstractions that occur within a radiological spectrum. The current limitations observed in some of the public available medical image datasets must not discourage DL initiatives in the healthcare, but signal that although DL model architecture and optimization technicalities are important, the attention to the data can be considered the most delicate yet important aspect of a successful approach.

This paper focused on exploring the experiences we had as a team working on the creation of two medical image datasets as well as a DL model for pulmonary TB classification. The discussion on the quality of annotation and data curation strategies for the datasets being published is essential to better guide current and future endeavors in the use of DL algorithms. There are highly prevalent diseases worldwide that could benefit from robust DL models as supporting tools for diagnosis and monitoring of treatment response. As the current bottleneck of DL applied to medical images can be considered the data (from the perspective of amount and quality), we hope the present work serve as an example of how ill-defined labels can be considered to enrich the data when present in parsimony.

Acknowledgments

We thank the Brazilian Ministry of Health.
Deep Learning on medical images: ill-defined labels can be a good thing

Figure 3: The images correspond to chest radiographs of individuals with confirmed TB in posteroanterior view, with pulmonary changes marked with bounding boxes and the respective heatmaps. The 'score' value present in the images is related to the model output, where score values closer to zero mean no patterns related to pulmonary TB were found and scores closer to one mean that patterns likely related to pulmonary TB were found on the CXR. Image a, obtained from a private hospital, shows consolidation in the right pulmonary apex. In b there are pulmonary opacities associated with fibro-retractable streaks and pulmonary volumetric reduction in the upper left field. This radiograph was obtained from the Shenzhen dataset (Jaeger et al., 2014). In c there are pulmonary opacities, some of them confluent in consolidations, sparse throughout the left lung and in the upper and middle fields on the right. This radiograph was obtained at a public hospital and was printed on couche paper.
References


