Brief Report

Two public chest X-ray datasets for computer-aided screening of pulmonary diseases

Stefan Jaeger[1], Sema Candemir[1], Sameer Antani[1], Yì-Xiáng J. Wáng[2], Pu-Xuan Lu[3], George Thoma[1]

[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; [2]Department of Imaging and Interventional Radiology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; [3]Department of Radiology, The Shenzhen No. 3 People's Hospital, Guangdong Medical College, Shenzhen 518020, China

Correspondence to: Stefan Jaeger. Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA. Email: stefan.jaeger@nih.gov.

Abstract: The U.S. National Library of Medicine has made two datasets of postero-anterior (PA) chest radiographs available to foster research in computer-aided diagnosis of pulmonary diseases with a special focus on pulmonary tuberculosis (TB). The radiographs were acquired from the Department of Health and Human Services, Montgomery County, Maryland, USA and Shenzhen No. 3 People's Hospital in China. Both datasets contain normal and abnormal chest X-rays with manifestations of TB and include associated radiologist readings.

Keywords: Tuberculosis (TB); computer-aided diagnosis; automatic screening; medical imaging; chest X-rays

Submitted Nov 15, 2014. Accepted for publication Nov 16, 2014.
doi: 10.3978/j.issn.2223-4292.2014.11.20
View this article at: http://dx.doi.org/10.3978/j.issn.2223-4292.2014.11.20

Introduction

Tuberculosis (TB) is a major global health threat (1,2). The advent of new powerful hardware and software techniques has triggered attempts to develop computer-aided diagnostic systems for TB detection (3). However, progress in the field has been hampered by the lack of publicly available radiographs for training machine learning algorithms for computer-aided diagnostic systems (4). In an effort to provide sufficient training data for the research community, allowing benchmark tests, the U.S. National Library of Medicine has made two datasets of postero-anterior (PA) chest radiographs available: the MC set and the Shenzhen set, which are announced here. Clinical readings are available for both sets. Furthermore, the MC set contains manually segmented lung masks for evaluation of automatic lung segmentation methods. Both datasets were de-identified by the data providers and were exempted from IRB review at their respective institutions. At NIH, the dataset use and public release were exempted from IRB review by the NIH Office of Human Research Projections Programs (No. 5357). Initial classification results for both sets have been presented in (3), and lung segmentation results for the MC set are shown in (5), which can serve as benchmarks for other researchers.

Montgomery County chest X-ray set (MC)

The MC set has been collected in collaboration with the Department of Health and Human Services, Montgomery County, Maryland, USA. The set contains 138 frontal chest X-rays from Montgomery County's Tuberculosis screening program, of which 80 are normal cases and 58 are cases with manifestations of TB. The X-rays were captured with a Eureka stationary X-ray machine (CR), and are provided in Portable Network Graphics (PNG) format as 12-bit gray level images. They can also be made available in DICOM format upon request. The size of the X-rays is either 4,020×4,892 or 4,892×4,020 pixels.

All image file names follow the same template: MCUCXR_####_X.png, where #### represents a 4-digit non-sequential numerical identifier, and X is either 0 for a normal X-ray or 1 for an abnormal X-ray. The clinical

reading for each X-ray is saved in a text file following the same format, except that the ending ".png" is replaced with ".txt". Each reading contains the patient's age, gender, and abnormality seen in the lung, if any. For example, a typical reading of an X-ray in the MC set has the following form:

Patient's Sex: F
Patient's Age: 031Y
cavitary nodular infiltrate in RUL; active TB

This is the X-ray reading of a woman who is 31 years old and has active TB, as indicated by cavitary nodular infiltrates in the right upper lobe.

For the MC set, we also make manual lung segmentations available, which we have used for our lung segmentation algorithm in (5), and which may serve as a publicly available reference dataset for similar experiments. We segmented the lungs under the supervision of a radiologist, following anatomical landmarks, such as the boundary of the heart, pericardium, and aortic arc. Furthermore, we outline the costophrenic angle, following the diaphragm boundary. Note that we exclude the area behind the heart and diaphragm. We draw an inferred boundary when the pathology is severe and affects the morphological appearance of the lungs. To save the binary lung mask, we use the same naming convention as used for the X-ray images and clinical readings. For each X-ray of the MC set, we save the corresponding binary lung mask separately for the left and right lung, in folders leftMask and rightMask, respectively.

Shenzhen chest X-ray set

The Shenzhen dataset was collected in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. The chest X-rays are from outpatient clinics and were captured as part of the daily hospital routine within a 1-month period, mostly in September 2012, using a Philips DR Digital Diagnost system. The set contains 662 frontal chest X-rays, of which 326 are normal cases and 336 are cases with manifestations of TB, including pediatric X-rays (AP). The X-rays are provided in PNG format. Their size can vary but is approximately 3K × 3K pixels.

All image file names follow the same template: CHNCXR_####_X.png, where #### represents a 4-digit numerical identifier, and X is either 0 for a normal X-ray or 1 for an abnormal X-ray. The clinical reading for each X-ray is saved in a text file following the same format, except that the ending ".png" is replaced with ".txt". Each reading contains the patient's age, gender, and abnormality seen in the lung, if any. For example, a typical reading of an X-ray in the Shenzhen set has the following form:

male 46yrs
bilateral PTB

This is the X-ray reading of a man who is 46 years old and is exhibiting signs of bilateral pulmonary tuberculosis (PTB).

Discussion

Both the MC and Shenzhen dataset can be requested via the contact on the following webpage: http://archive.nlm.nih.gov/repos/chestImages.php. We ask that requesters do not share the datasets outside of their research groups and organization, but encourage new requests to be forwarded directly to us. Further, we request that publications resulting from the use of this data attribute the source (National Library of Medicine, National Institutes of Health, Bethesda, MD, USA) and cite the following publications, which have used the data for classification experiments and lung segmentation (3,5).

Acknowledgements

This work is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

We thank Sonia Qasba, M.D., Medical Director, Tuberculosis Control Program, Montgomery County, Maryland, USA, for allowing us to make the MC set publicly available. We are grateful to Jonathan Musco, M.D., Department of Radiology, School of Medicine, University of Missouri-Columbia, Columbia, Missouri, USA, for verifying our lung segmentations for the MC set. We also appreciate the help of Michael Bonifant and Ellan Kim, who helped us validate the data.

Disclosure: The authors declare no conflict of interest.

References

1. World Health Organization. Global Tuberculosis Report, 2012.
2. Stop TB Partnership. The Global Plan to Stop TB 2011-2015. World Health Organization, 2011.

3. Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, Xue Z, Palaniappan K, Singh RK, Antani S, Thoma G, Wang YX, Lu PX, McDonald CJ. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 2014;33:233-45.
4. Jaeger S, Karargyris A, Candemir S, Siegelman J, Folio L, Antani S, Thoma G. Automatic screening for tuberculosis in chest radiographs: a survey. Quant Imaging Med Surg 2013;3:89-99.
5. Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, Karargyris A, Antani S, Thoma G, McDonald CJ. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans Med Imaging 2014;33:577-90.

Cite this article as: Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 2014;4(6):475-477. doi: 10.3978/j.issn.2223-4292.2014.11.20

