Crowdsourcing labels for pathological patterns in CT lung scans: Can non-experts contribute expert-quality ground truth?

Alison Q. O'Neil[1], John T. Murchison[2], Edwin J. R. van Beek[3] and Keith A. Goatman[1]

[1] Toshiba Medical Visualization Systems Ltd., Edinburgh, UK
[2] Royal Infirmary of Edinburgh, Edinburgh, UK
[3] Clinical Research Imaging Centre, University of Edinburgh, Edinburgh, UK

Abstract. This paper investigates what quality of ground truth might be obtained when crowdsourcing specialist medical imaging ground truth from non-experts. Following basic tuition, 34 volunteer participants independently delineated regions belonging to 7 pathological patterns in 20 scans according to expert-provided pattern labels. Participants' annotations were compared to a set of reference annotations using Dice similarity coefficient (DSC), and found to range between 0.41 and 0.77. The reference repeatability was 0.81. Analysis of prior imaging experience, annotation behaviour, scan ordering and time spent showed that only the last was correlated with annotation quality. Multiple observers combined by voxelwise majority vote outperformed a single observer, matching the reference repeatability for 5 of 7 patterns. In conclusion, crowdsourcing from non-experts yields acceptable quality ground truth, given sufficient expert task supervision and a sufficient number of observers per scan.

1 Introduction

Crowdsourcing is gaining in popularity as a method for sourcing labels for the very large amounts of data required to train machine learning algorithms [7]. Previous experiments have shown that it is possible to use non-experts for cheaply and readily crowdsourcing medical imaging ground truth [14, 3], perhaps using gamification [11, 1], at least for reasonably straightforward problems.

This paper investigates whether it is feasible to commission non-experts to undertake a relatively specialist imaging annotation task — that of recognising and segmenting the pathological patterns which are seen in interstitial lung disease. To this end, a toy exercise was designed in which participants were recruited to annotate the *same* representative set of twenty scan slices. In order to render the task accessible to the layperson, we restricted it to be one of annotation rather than diagnosis. Each scan slice was provided with expert labels indicating the presence of the main patterns to be labelled, and participants were asked to annotate regions belonging to these patterns. These labels are usually noted in a radiology report; thus the objective was for the routine expert diagnosis to direct the non-expert in the rather time-consuming work of delineating the pathological regions. To assess performance, we quantitatively and qualitatively compared the annotations to those of an expert medical researcher (A.O.) and two experienced radiologists (J.M. and E.v.B.) respectively.

The contributions of this paper are as follows:

- To demonstrate how a specialist medical imaging ground truth task may be simplified such that a non-expert (given some basic training) performs comparably to an expert.
- To analyse which factors are predictive of good performance.
- To demonstrate how (and how many) non-expert observers should be assigned and combined for each scan in a real world crowdsourcing task, in order to improve label robustness.
- To provide practical recommendations for how this task might be better conducted in future.

2 Methodology

2.1 Ground truth for interstitial lung disease

Identification of the presence, volume and distribution of different pathological patterns is helpful for the diagnosis and prognosis of interstitial lung disease [8]. Training machine learning algorithms to recognise and segment such patterns requires large amounts of labelled data. Thus, for this paper, the ground truth exercise was to label regions representing each of the common lung disease patterns: *consolidation*, *emphysema*, *ground glass opacity (GGO)*, *ground glass opacity + reticulation*, *honeycombing*, *micronodules*, and *reticulation*. This is the same labelling system as used by Anthimopoulos *et al.* [2] for the same publicly available data [4], but with the addition of an emphysema class. Examples of these patterns are shown in Figure 1.

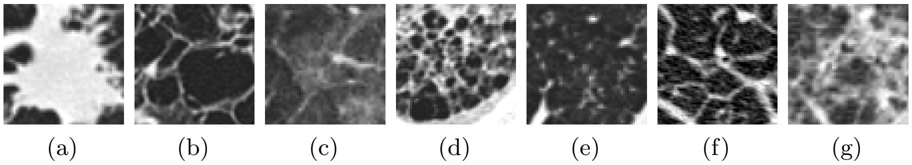


Fig. 1: Pathological lung patterns. a) *Consolidation* b) *Emphysema* c) *GGO* d) *GGO+Reticulation* e) *Honeycombing* f) *Micronodules* g) *Reticulation*

2.2 Data

Twenty computed tomography (CT) scan slices were selected from twenty different subjects in the MedGift ILD database [4]. The slices were chosen to span the range of disease labels, and each was labelled with one or two key patterns to be annotated by participants. Table 1 shows the pattern labels and medical diagnosis of each scan.

N	Diagnosis	Labels	N	Diagnosis	Labels
1	Idiopathic pulmonary fibrosis	E	11	Miliary tuberculosis	C, M
2	Idiopathic pulmonary fibrosis	H	12	Pulmonary Fibrosis	GR
3	Hypersensitivity pneumonitis	G, GR	13	Hypersensitivity pneumonitis	G
4	Miliary tuberculosis	M	14	–	H
5	–	E	15	Chronic eosinophilic pneumonia	R
6	Pulmonary fibrosis	R	16	Pulmonary tuberculosis	C
7	–	C	17	Hypersensitivity pneumonitis	R, GR
8	Cryptogenic organizing pneumonia	C, G	18	Hypersensitivity pneumonitis	G
9	Hypersensitivity pneumonitis	R, GR	19	Pulmonary fibrosis	E, GR
10	Hypersensitivity pneumonitis	H	20	Pulmonary fibrosis	H

Table 1: Scan diagnoses (3 unknown) and patterns to label (C=*Consolidation*, E=*Emphysema*, G=*GGO*, GR=*GGO+Reticulation*, H=*Honeycombing*, M=*Micronodules*, R=*Reticulation*)

2.3 Recruitment of participants

The exercise was completed by 34 volunteers from a company which makes medical imaging software. The participants have a variety of roles and levels of expertise, including junior scientists and software engineers, senior managers, and clinical experts. Entry and exit questionnaires were completed by all the participants. The entry questions were designed to ascertain each participant's level of experience, and the factors motivating their participation. The exit questionnaire gathered feedback on participants' experience of the exercise, and suggestions for improvement.

2.4 Annotation task

Prior to the annotation task, all participants received a one-hour long tutorial on interstitial lung disease and the patterns of interest (based on the Fleischner Society Glossary of Terms for Thoracic Imaging [5]), given by a biomedical sciences graduate (A.O.) who had recently attended a one-day hands-on training course on interstitial lung disease run by the British Institute of Radiology.

Participants were provided with the twenty pre-selected slices and asked to annotate patterns belonging to provided labels. Each participant annotated the images in a random order, to allow measurement of any training effect over the course of annotating the scans. Annotations were created using a tool that allowed users to draw polygonal regions of interest (ROIs) and assign a pattern class label to each ROI. The task was expected to take approximately two hours to complete. The use of online resources such as *Radiology Assistant* and *Google* was allowed and even encouraged, although collaboration between participants was prohibited.

3 Results

3.1 Evaluation of non-expert versus expert performance

Each annotation was scored by comparison to those of the reference annotator (A.O.) using Dice Similarity Score (DSC). The overall DSC was computed for each participant by weighting scans equally, and weighting patterns equally within a scan. Per-pattern DSC metrics were calculated for each participant by averaging over all examples of a pattern. In addition, the reference annotator repeated the annotations 10 days later to assess repeatability (the overall repeatability DSC was 0.806). Figure 2 summarises the results.

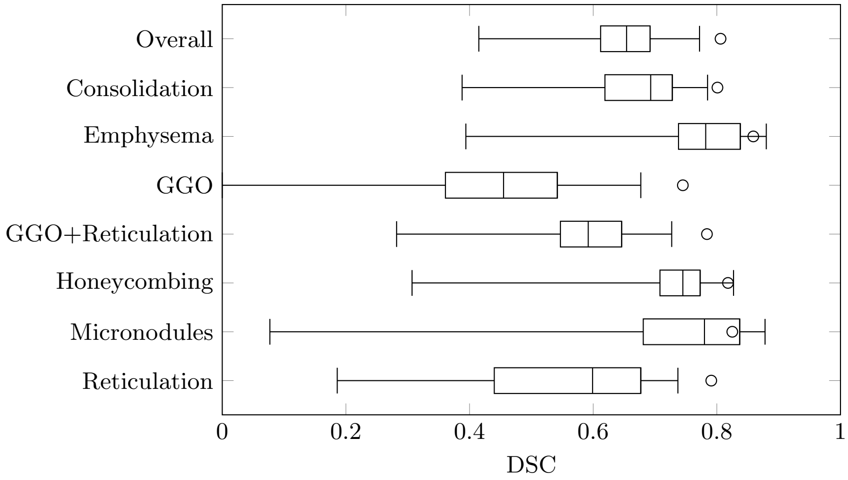


Fig. 2: The box plots indicate the median, upper and lower quartiles, and minimum and maximum DSC compared to the reference. The circles indicate the reference repeat scores.

There is clear variation in performance between classes, showing that some were more straightforward than others. It was known in advance that the distinction between e.g. *GGO*, *GGO + Reticulation*, and *Reticulation* might be open to interpretation. Also, there were a few cases of mistaken identity, with participants labelling vessels (pulmonary vessels and aorta) as pathology.

Following the exercise, interviews were held with two experienced pulmonary radiologists (J.M. and E.v.B.), who confirmed the veracity of the provided labels, and annotated the images with some obvious examples of each pattern. Figure 3 shows some qualitative results of four interesting cases, showing the radiologist and reference annotations overlaid on the results of the crowd.

It can be seen that for A (*Emphysema*) and B (*GGO*) in Figure 3, the range of variation of the crowd is comparable to the agreement (or disagreement) between the two radiologists. In each case, one radiologist is more sensitive and the other more specific for the given pattern, and the crowd approximately ranges between the two.

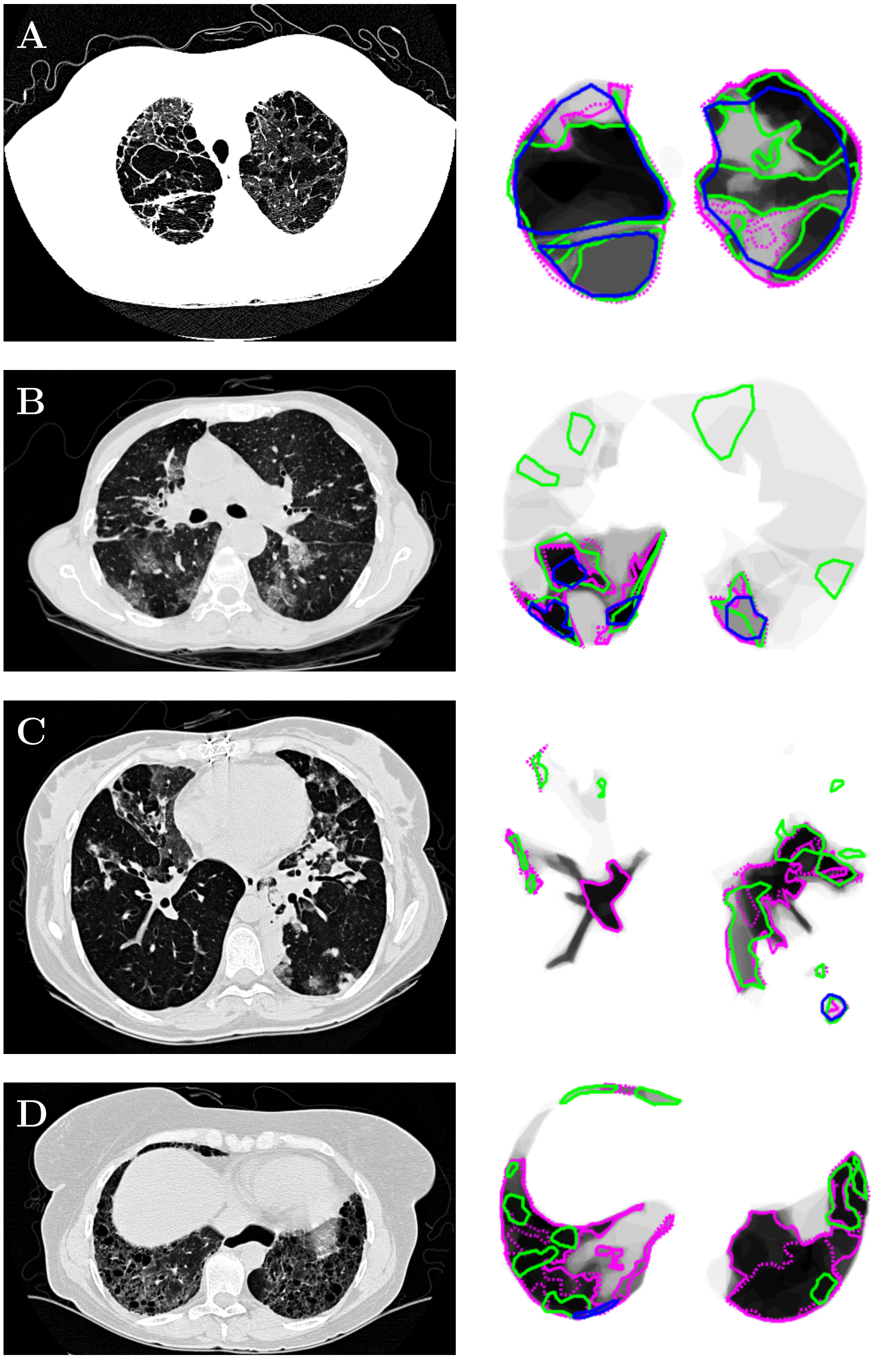


Fig. 3: Some example results: A) Emphysema B) GGO C) Consolidation D) Honeycombing. Scan slices are shown on the left and annotations are shown on the right. The greyscale background is proportional to the number of participants who annotated the label i.e. white = no annotations and black = all 34 annotations. The reference results are shown in magenta (dotted line for the repeat). The radiologists' annotations are shown in blue and green.

Examples C and D illustrate where improvements could be made. In C (consolidation), it is difficult to distinguish vessels from consolidation. It can be seen that the radiologists were cautious with their labelling compared with the reference, who outlined both vessel and consolidation where they were adjacent and therefore not separable. The crowd generally followed the philosophy of the reference, but some of the crowd confused what is definitely vessel with consolidation. In D (honeycombing), both radiologists were stricter on the definition of honeycombing than the reference, and both raised the differential diagnosis with bronchiectasis. Honeycombing and bronchiectasis lie on a spectrum [12], and the bronchiectasis label was not included in our labelling system.

In summary, it was observed that in many cases the variability of the crowd matched the variability between the two radiologists, and this variability was reflective of underlying ambiguity in the pattern definition — or the ambiguity of the boundary between patterns such as *GGO* versus *GGO + reticulation*. However, in future the whole volume should be provided to the annotator rather than single slices, such that vessels can be better tracked and distinguished from consolidation (with appropriate teaching examples). We should also consider adding further labels such as bronchiectasis and fibrosis (fibrosis not illustrated here).

3.2 Factors predicting performance

None of the participants had specific prior experience of interstitial lung disease images. However, it was predicted that there may be a correlation between prior imaging experience and performance, particularly if insufficient training was provided for the task. Participants rated their level of experience with medical imaging data, from level 0 (little to none), to level 4 (clinical researcher). Figure 4 shows a plot of performance versus experience level. There is no significant correlation, suggesting that adequate guidance was provided for this task. Further, it was hypothesised that a training effect might be observed, however no correlation was measured between the scan ordering (randomised between participants) and each participant's performance.

Conversely, there is a weak correlation between the time spent on the task and performance (see Figure 4). The times shown are self-reported estimates. It is likely that those participants who performed better took time to do more research and/or took more care with their annotations. Visible annotation behaviour (number of regions, number of polygon vertices, rate of polygon vertices) was also analysed and found to exhibit no correlation with performance.

3.3 Crowdtruthing in the real world: Assigning and combining multiple observers

The previous results have shown the range in annotation quality between observers. It is likely that more consistent results could be achieved by combining annotation results from multiple observers, and this is true also of expert annotations, since human error or variations in pattern interpretation might be identified and

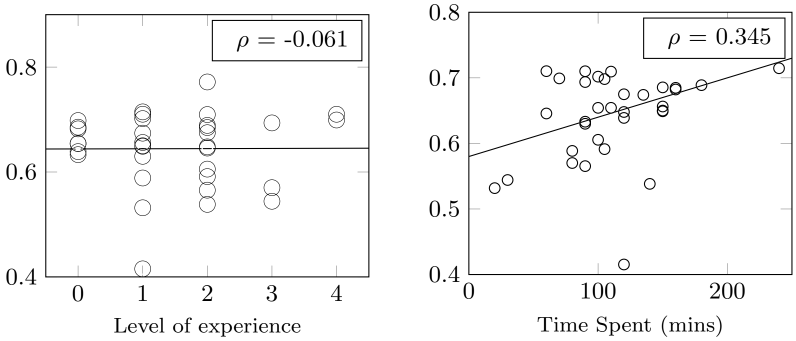
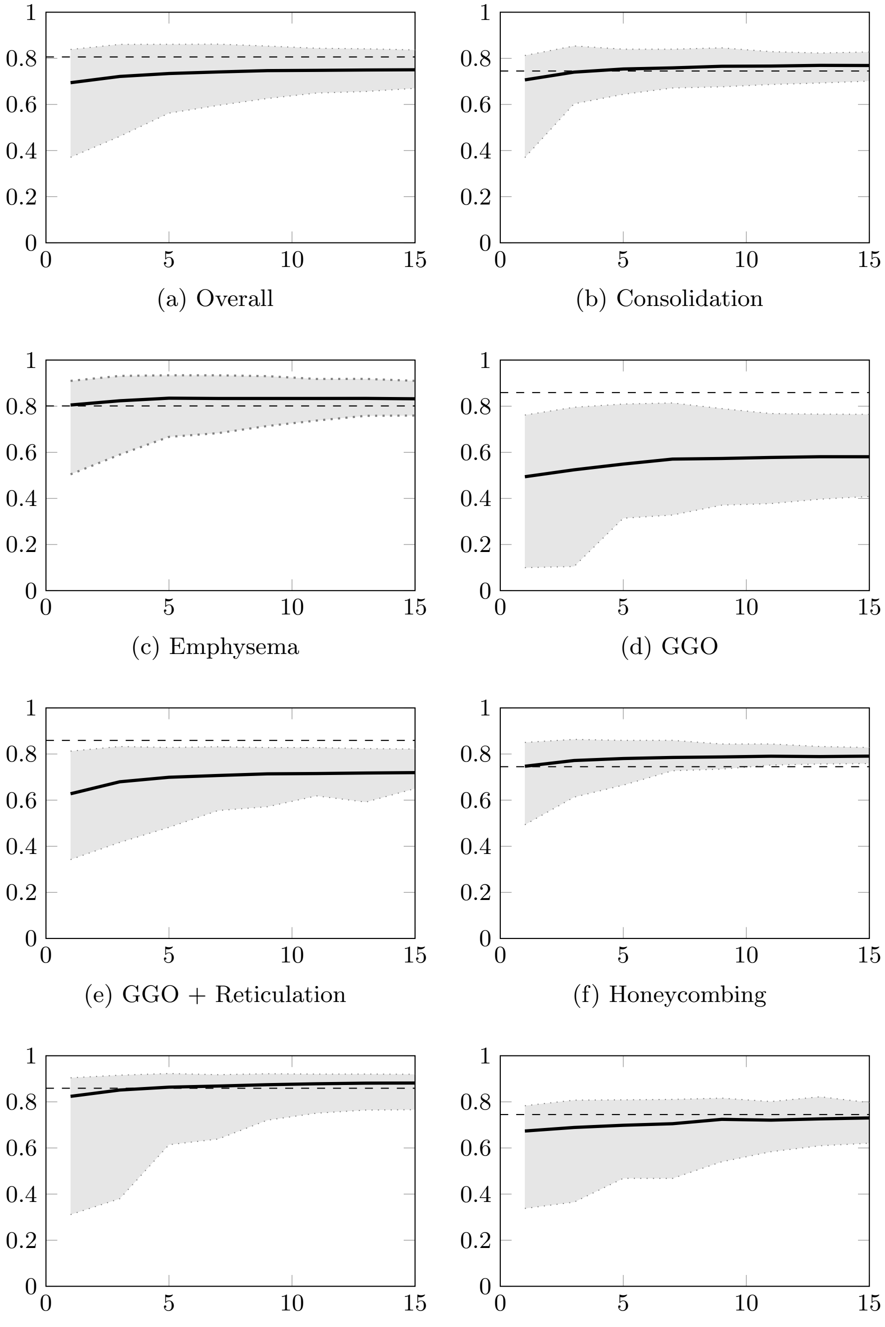


Fig. 4: Factors predicting performance. Level of expertise and time spent are plotted against DSC compared to the reference. Correlation coefficients are shown (Spearman's rank and Pearson's for the first and second plots respectively).

corrected. In a real world crowdsourcing exercise, some questions would thus arise. How many observers should be assigned to each scan? How are their annotations best combined to give an annotation of predictable and reasonable quality?

To investigate this, different odd numbers of observers between one and fifteen were combined using majority vote at each voxel. For each number of observers, 200 combinations were randomly drawn from the 34 annotations, after omitting the few cases where the annotation was zero i.e. the participant had forgotten or was unable to label the key pattern. As in earlier DSC computations, the problem is simplistically treated as binary (i.e. a one-vs-all approach taken when evaluating each pattern), even where more than one pattern was labelled in a scan. The graphs in Figure 5 show the median, minimum and maximum values, both overall and for each pattern, averaged across the twenty scans.

In summary, multiple observers give a better result than a single observer. The median increases and the range in DSC metrics narrows increasingly as more observers are added, with little improvement beyond the $k = 9$ observer. Note that the minimum, maximum and median converge at the limit of $n = 34$ observers, where there is just one possible combination of observers. For 5 of 7 patterns, the median DSC matches the repeat DSC and the range converges whilst $k \ll n$, showing that when sufficient observers are combined, the limit of accuracy is reached. For *GGO* and *GGO + Reticulation*, combination of multiple observers does not bring the crowd into agreement with the reference, suggesting that observers generally had a different idea to the reference for where the threshold between ground glass opacity and healthy tissue lies. STAPLE [15] methods were also tried (results not shown), initialised using both uniform (0.99999) and learnt rater sensitivities and specificities (learnt from the first ten scans and applied to the second ten), and STAPLE gave worse results than the majority vote. This is in line with what other authors have found [10, 9].



(a) Overall

(b) Consolidation

(c) Emphysema

(d) GGO

(e) GGO + Reticulation

(f) Honeycombing

(g) Micronodules

(h) Reticulation

Fig. 5: Graphs showing the number of observers (x-axis) versus the reference DSC (y-axis) for the consensus (combined) annotation, for different pathological patterns. The solid lines indicate the median and the grey shading indicates the span from minimum to maximum (figures are the *mean* minimum, median and maximum across all scans). The dashed lines indicate the reference repeatability score.

4 Discussion

Overall, the crowd performed well relative to the reference segmentations, with some observers for some patterns matching the reference repeatability. Where there was variation, this was often indicative of genuine ambiguity between patterns. The greater range of disagreement for e.g. ground glass opacity compared to emphysema in this exercise has been observed by other authors measuring agreement between radiologists [13]. In fact, the combined annotations displayed as greyscale values in Figure 3 could be interpreted as probabilities associated with the respective labels, and even used as soft labels for a machine learning algorithm in line with the "dark matter" idea promoted by Hinton *et al.* [6]. Note that agreement both between non-experts and between radiologists would be increased with a more stringent ground truth protocol (this might involve e.g. prescribing a Hounsfield Unit range for ground glass opacity).

Experiments regarding combination of observers showed that multiple observers outperformed a single observer. For many patterns, when sufficient observers are combined, the median DSC matches the reference repeatability DSC and the DSC range converges around the reference repeatability DSC, showing that the limit of accuracy is reached. Improvements as discussed earlier (additional teaching for distinguishing normal anatomy such as vessels from pathology, provision of three-dimensional context, additions to the labelling system, a more stringent ground truth protocol), should both raise the repeatability DSC and reduce the number of observers required to achieve a consistent result.

In conclusion, given sufficient expert task supervision and a sufficient number of observers per scan, crowdsourcing with non-experts can yield ground truth fit for use in image analysis algorithms.

5 Acknowledgements

Many thanks to Phil Tolland who developed the software for the ground truth collection tool, and to all of the employees at Toshiba Medical Visualization Systems who took part in this study: Allan Barklie, Erin Beveridge, Antony Brown, Gerald Chau, Alasdair Corbett, Ross Davies, Matt Daykin, Ben Docherty, Venkatesh Gaddam, Keith Goatman, Marta Guarisco, Joseph Henry, Corné Hoogendoorn, Pia Kullik, Aneta Lisowska, Steve Magness, Craig Matear, James Matthews, Chris McGough, Haritha Miryala, Brian Mohr, Costas Plakas, Ian Poole, Marco Razeto, Faye Riley, Matt Shepherd, Simeon Skopalik, Andy Smout, Ken Sutherland, Paul Thomson, Phil Tolland, John Tough, Aidan Wellington and Gavin Wheeler.

References

1. Shadi Albarqouni, Stefan Matl, Maximilian Baust, Nassir Navab, and Stefanie Demirci. Playsourcing: A novel concept for knowledge creation in biomedical research. In *International Workshop on Large-Scale Annotation of Biomedical Data*

and Expert Label Synthesis (LABELS), volume 10008 of *Lecture Notes in Computer Science*, 2016.

2. Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 35(5):1207–1216, 2016.
3. Veronika Cheplygina, Adria Perez-Rovira, Wieying Kuo, Harm A W M Tiddens, and Marleen de Bruijne. Early experiences with crowdsourcing airway annotations in chest CT. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS)*, volume 10008 of *Lecture Notes in Computer Science*, 2016.
4. Adrien Depeursinge, Alejandro Vargas, Alexandra Platon, Antoine Geissbuhler, Pierre Alexandre Poletti, and Henning Müller. Building a reference multimedia database for interstitial lung diseases. *Computerized Medical Imaging and Graphics*, 36(3):227–238, 2012.
5. David M. Hansell, Alexander A. Bankier, Heber MacMahon, Theresa C. McLoud, Nestor L. Müller, and Jacques Remy. Fleischner society: Glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722, 2008.
6. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Neural Information Processing Systems*, 2014.
7. Mokter Hossain and Ilkka Kauranen. Crowdsourcing: A comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):1753–8297, 2015.
8. S M Humphries, K Yagihashi, J Huckleberry, B H Rho, J D Schroeder, M Strand, M I Schwarz, K R Flaherty, E A Kazerooni, E J R van Beek, and D A Lynch. Idiopathic pulmonary fibrosis: Data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. *Radiology*, 2017.
9. Thomas Robin Langerak, Uulke A van der Heide, Alexis N T J Kotte, Max A Viergever, Marco van Vulpen, and Josien P W Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, 2010.
10. Koen Van Leemput and Mert R Sabuncu. A cautionary analysis of STAPLE using direct inference of segmentation truth. In *MICCAI Proceedings*, volume 8673 of *Lecture Notes in Computer Science*, pages 398–406, 2014.
11. Miguel Angel Luengo-Oroz, Asier Arranz, and John Frean. Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, 14(6), 2012.
12. Sara Piciucchi, Sara Tomassetti, Claudia Ravaglia, Christian Gurioli, Carlo Gurioli, Alessandra Dubini, Angelo Carloni, Marco Chilosi, Thomas V Colby, and Venerino Poletti. From traction bronchiectasis to honeycombing in idiopathic pulmonary fibrosis: A spectrum of bronchiolar remodeling also in radiology? *BMC Pulmonary Medicine*, 16(87), 2016.
13. Margaret L Salisbury, David A Lynch, Edwin J R van Beek, Ella A Kazerooni, Junfeng Guo, Meng Xia, Susan Murray, Kevin A Anstrom, Eric Yow, Fernando J Martinez, Eric A Hoffman, and Kevin R Flaherty. Idiopathic pulmonary fibrosis: The association between the adaptive multiple features method and fibrosis outcomes. *American Journal of Respiratory and Critical Care Medicine*, 195(7), 2017.
14. Dmitrij Schlesinger, Florian Jug, Gene Myers, Carsten Rother, and Dagmar Kainmuller. Crowdsourcing image segmentation with aSTAPLE. *arXiv*, 2017.
15. Simon K Warfield, Kelly H Zhou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7), 2004.