A Survey of Crowdsourcing in Medical Image Analysis

Ørting, S. N.; Doyle, A.; van Hilten, A.; Hirth, M.; Inel, O.; Madan, C. R.; Mavridis, P.; Spiers, H.; Cheplygina, V.

DOI
10.15346/hc.v7i1.1

Publication date
2020

Document Version
Final published version

Published in
Human Computation Journal

Human Computation (2020) 7:1:26

ISSN: 2330-8001, DOI: 10.15346/hc.v7i1.1

A Survey of Crowdsourcing in Medical Image Analysis

SILAS N. ØRTING✉, UNIVERSITY OF COPENHAGEN, COPENHAGEN, DENMARK

ANDREW DOYLE*, MCGILL CENTRE FOR INTEGRATIVE NEUROSCIENCE, MONTREAL, CANADA

ARNO VAN HILTEN*, ERASMUS MEDICAL CENTER, ROTTERDAM, THE NETHERLANDS

MATTHIAS HIRTH*, TECHNISCHE UNIVERSITÄT ILMENAU, ILMENAU, GERMANY

OANA INEL*, VRIJE UNIVERSITEIT AMSTERDAM, AMSTERDAM, THE NETHERLANDS, DELFT UNIVERSITY OF TECHNOLOGY, DELFT, THE NETHERLANDS

CHRISTOPHER R. MADAN*, UNIVERSITY OF NOTTINGHAM, NOTTINGHAM, UNITED KINGDOM

PANAGIOTIS MAVRIDIS*, DELFT UNIVERSITY OF TECHNOLOGY, DELFT, THE NETHERLANDS

HELEN SPIERS*, UNIVERSITY OF OXFORD, OXFORD, UNITED KINGDOM, ZOONIVERSE, UNIVERSITY OF OXFORD, OXFORD

VERONIKA CHEPLYGINA✉, EINDHOVEN UNIVERSITY OF TECHNOLOGY, EINDHOVEN, THE NETHERLANDS

✉ *silas@di.ku.dk, v.cheplygina@tue.nl*
** These authors contributed equally and are listed alphabetically by last name*

ABSTRACT

Rapid advances in image processing capabilities have been seen across many domains, fostered by the application of machine learning algorithms to "big-data". However, within the realm of medical image analysis, advances have been curtailed, in part, due to the limited availability of large-scale, well-annotated datasets. One of the main reasons for this is the high cost often associated with

producing large amounts of high-quality meta-data. Recently, there has been growing interest in the application of crowdsourcing for this purpose; a technique that is well established in a number of disciplines, including astronomy, ecology and meteorology for creating large-scale datasets. Despite the growing popularity of this approach, there has not yet been a comprehensive literature review to provide guidance to researchers considering using crowdsourcing methodologies in their own medical imaging analysis. In this survey, we review studies applying crowdsourcing to the analysis of medical images, published prior to July 2018. We identify common approaches and challenges and provide recommendations to researchers implementing crowdsourcing for medical imaging tasks. Finally, we discuss future opportunities for research and development within this emerging domain.

1. INTRODUCTION

The limited availability and size of labeled datasets for training machine learning algorithms is a common problem in medical image analysis (Greenspan et al., 2016; Litjens et al., 2017; Cheplygina et al., 2018). In several other fields, crowdsourcing – defined as the outsourcing of tasks to a crowd of individuals (Howe, 2006)– has been found effective for labeling large quantities of data. For example, in computer vision crowdsourcing has been used to annotate large datasets of images and videos with various tags (Kovashka et al., 2016), and online citizen science via platforms such as the Zooniverse has become well established across a number of academic domains including astronomy (Lintott et al., 2008), meteorology (Knapp et al., 2016) and ecology (Willi et al., 2019).

Due to the success of crowdsourcing, several researchers have recently applied these techniques to the annotation of medical images. Although such images present specific challenges, including absence of expertise of the crowd, several early papers such as (Mitry et al., 2013; Mavandadi et al., 2012; Maier-Hein et al., 2014a) have demonstrated promising results. Despite the growing interest, there has not been an overview of the work in this field. In this paper, we summarize existing literature on crowdsourcing in medical imaging.

This paper originated during the Lorentz workshop "Crowdsourcing in medical image analysis" in June 2018[1]. As participants of the workshop, we searched Google Scholar with the query "crowdsourcing AND (medical or biomedical)" and screened the results for papers focusing on the topic. Google Scholar was selected due to previous papers highlighting the poor indexing of the topic in databases and the high prevalence of crowdsourcing papers in conferences (Wazny, 2017). Additional papers were identified for inclusion by all authors, by examining the references and citations of selected papers. We did not exclusively focus only on journal papers, but included preprints and abstracts, as we realize that some studies may not have been prepared as scientific articles. However, we realize that such papers are more difficult to find, and thus there is still a degree of publication bias in our selection. Furthermore, we did not update our search query to include other relevant terms, such as "citizen science", which might have expanded the set of included papers.

We only included papers where the crowd was involved in the analysis of medical or biomedical

[1] https://web.archive.org/web/20190603151616/https://www.lorentzcenter.nl/lc/web/2018/967/info.php3?wsid=967&venue=Snellius Accessed Jun. 2020

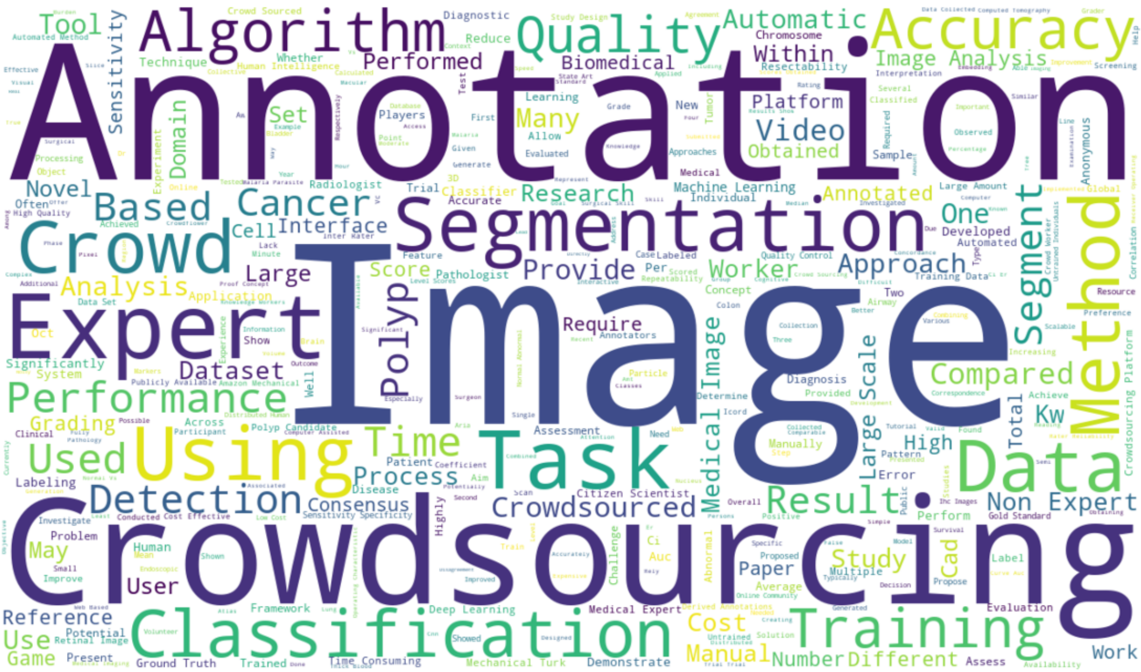

***Figure 1.** A word cloud of the abstracts of the surveyed papers.*

images, for example by annotating them. Our search strategy resulted in 57 papers. Key terms emerging from these studies are illustrated in Fig. 1. Five key dimensions were identified for discussion: the application involved, the type of interaction between the crowd and the images, the scale of the task (such as the number of images), the type of evaluation performed on the crowd annotations, and the results of the evaluation.

There are a number of surveys which are related to this work. However, they are quite different in scope:

- Ranard et al. (2014) survey crowdsourcing in health and medical research. They identify four tasks: problem solving, data processing, monitoring and surveying and cover 21 papers published until March 2013. In contrast, we only focus on papers where image analysis (i.e. data processing) is involved.
- Kovashka et al. (2016) survey crowdsourcing in computer vision. The surveyed papers focus on analysis of everyday/natural images. Only one of the 195 referenced papers ((Gurari et al., 2015a)) uses biomedical data.
- Wazny (2017) present a meta-review of crowdsourcing from 2006 to 2016. Similar to (Ranard et al., 2014), they take a more broad view of crowdsourcing. They review 48 existing review papers until August 2015, focusing on how each review categorizes the papers, for example by platform, size of crowd, and so forth.
- Alialy et al. (2018) is most similar to our survey, but only focuses on crowdsourcing in human

pathology. They do a systematic search with several steps, excluding conference papers or abstracts, and summarize seven papers. The coverage of literature is therefore much more limited than in this work.

The paper is organized as follows. In Sections 2 to 6 we summarize the reviewed papers according to the five dimensions we identified and in Section 7 we discuss overall trends, limitations, and opportunities for future research. A condensed overview of all papers and their properties according to the key dimensions is provided in Table 9 in the Appendix.

2. APPLICATIONS

There are a variety of crowdsourcing applications addressed in the surveyed papers. We group these applications by the type of task performed by the crowd, the biomedical content of the image, and the dimensionality of the images.

2.1. Type of task

Table 1 summarizes the different task types of the surveyed papers. An important task in medical image analysis is classification, and 42% of the surveyed papers focus on this task. Classification can refer to assigning a label to an entire image, such as diagnosing whether a chest CT image contains any abnormalities. Classification can also refer to assigning a label to a part of the image, for example, the type of abnormality located in a particular region of interest. Other types of labels include non-diagnostic labels such as image modality (de Herrera et al., 2014), visual attributes (Cheplygina and Pluim, 2018), and assessing the quality of the image (Keshavan et al., 2018). These three types of labels are based more on visual characteristics, and thus might be easier to provide than diagnostic labels without any medical training.

A further 39% of the papers focus on localization or segmentation. Typically the goal is to delineate the boundary of an entire healthy structure, or of an abnormality such as a lesion. The difference with how we define the classification task above is that instead of providing information about the image, the annotator has to modify the image, by providing positions or outlines. These tasks rely more on visual characteristics than classification tasks, and may be more easily explained to a non-expert crowd.

In 12% of the papers both classification and segmentation are addressed. Often this means that the annotator first has to indicate if the structure of interest is visible, and if yes, to locate it in the image.

Finally, 7% of papers request less standard tasks from their crowd. For example, Maier-Hein et al. (2015) focus on determining correspondence between pairs of images. Although this is a type of detection task, where the annotator has to locate points of interest in an image, it is also different since a point of reference is already provided. Another example is Ørting et al. (2017), where the annotator has to decide which image is more similar to a reference image. This is a type of classification problem, but again relying more on visual features than on prior knowledge.

2.2. Type of image

Medical images are acquired at vastly different scales and locations depending on the physiological measurement of interest. The imaging acquisition modality and strategy depends heavily on the scale of the anatomy of interest, and different technologies' expected contrast with surrounding tissues. Here we categorize the images by the type of structure that is being imaged, which narrows down the modality. We use the following categorization, also used in two recent surveys of medical

Table 1. Task types.

Task	Papers
Classify	(Albarqouni et al., 2016a),(Brady et al., 2014),(Brady et al., 2017),(Cheplygina and Pluim, 2018),(dos Reis et al., 2015),(Eickhoff, 2014),(Foncubierta Rodríguez and Müller, 2012),(Gur et al., 2017),(de Herrera et al., 2014),(Holst et al., 2015),(Huang and Hamarneh, 2017),(Keshavan et al., 2018),(Lawson et al., 2017),(Malpani et al., 2015),(Mavandadi et al., 2012),(McKenna et al., 2012),(Mitry et al., 2013),(Mitry et al., 2015),(Nguyen et al., 2012),(Park et al., 2016),(Park et al., 2017),(Smittenaar et al., 2018),(Sonabend et al., 2017),(Sullivan et al., 2018)
Segment	(Roethlingshoefer et al., 2017),(Boorboor et al., 2018),(Bruggemann et al., 2018),(Cabrera-Bean et al., 2017),(Chávez-Aragón et al., 2013),(Cheplygina et al., 2016),(Ganz et al., 2017),(Gurari et al., 2015b),(Heller et al., 2017),(Irshad et al., 2015),(Lee and Tufail, 2014),(Lee et al., 2016),(Lejeune et al., 2017),(Luengo-Oroz et al., 2012),(Maier-Hein et al., 2014a),(Maier-Hein et al., 2016),(O'Neil et al., 2017),(Park et al., 2018),(Rajchl et al., 2016),(Rajchl et al., 2017),(Sameki et al., 2016),(Sharma et al., 2017)
Segment + Click	(Della Mea et al., 2014),(Gurari et al., 2016),(Heim, 2018),(Irshad et al., 2017),(Leifman et al., 2015),(Mitry et al., 2016),(Timmermans et al., 2016)
Other	(Albarqouni et al., 2016b),(Maier-Hein et al., 2014b),(Maier-Hein et al., 2015),(Ørting et al., 2017)

imaging (Litjens et al., 2017; Cheplygina et al., 2018): brain, eye, heart, breast, lung, abdomen, histology/microscopy, multiple, other. Note that the histology/microscopy category includes tissue biopsied from different organs, as these images are similar in appearance. An overview of the surveyed papers and the image content in use is given in Table 2.

We compare the distribution of applications surveyed in this work with the two other surveys which used this categorization in Table 3. An interesting observation is that Litjens et al. (2017) and Cheplygina et al. (2018) have a similar distribution of applications despite surveying different topics: Litjens et al. (2017) covers deep learning, where a larger dataset is preferred, while Cheplygina et al. (2018) covers weakly supervised learning, where datasets are smaller in size. Given that crowdsourcing is often proposed as an alternative to weakly supervised learning, it is surprising that the current survey has a different distribution of papers.

Many of the papers in this survey are aimed at 2D images. The most common application is histopathology/microscopy with 28% of all the papers, followed by retinal images with 14% of the papers. Both applications are over-represented compared to Litjens et al. (2017) and Cheplygina et al. (2018). This overrepresentation in crowdsourcing studies may be because many retinal and microscopic images are acquired in 2D, which might be easier to use in a crowdsourcing study than 3D images.

Breast and heart images, which were already not well represented in the other two surveys, are almost absent in crowdsourcing studies. Both applications can be aimed at 2D or 3D images. However, perhaps due to lack of datasets or perceived difficulty of assessing these images, these applications are almost never considered for crowdsourcing.

Several other papers address applications where images are often 3D, such as the brain (9%) and the lungs (9%). Compared to Litjens et al. (2017) and Cheplygina et al. (2018), brain and lung images are underrepresented in crowdsourcing. This could be due to complexity of images or limitations in interfaces. One approach for dealing with 3D images is to select 2D parts of the original 3D images. For example, Ørting et al. (2017) and O'Neil et al. (2017) select axial slices. Cheplygina et al. (2016) shows patches of 2D projections in various directions in the image. Others circumvent the

***Table 2.* Application Domains.**

Domain	Papers
Abdomen	(Roethlingshoefer et al., 2017),(Heim, 2018),(Heller et al., 2017),(Maier-Hein et al., 2014a),(Maier-Hein et al., 2014b),(Maier-Hein et al., 2015),(Maier-Hein et al., 2016),(McKenna et al., 2012),(Nguyen et al., 2012),(Park et al., 2016),(Park et al., 2017),(Park et al., 2018),(Rajchl et al., 2017)
Brain	(Ganz et al., 2017),(Keshavan et al., 2018),(Rajchl et al., 2016),(Sonabend et al., 2017),(Timmermans et al., 2016)
Eye	(Brady et al., 2014),(Brady et al., 2017),(Lee and Tufail, 2014),(Lee et al., 2016),(Leifman et al., 2015),(Mitry et al., 2013),(Mitry et al., 2015),(Mitry et al., 2016)
Heart	(Gur et al., 2017)
Histo	(Albarqouni et al., 2016a),(Albarqouni et al., 2016b),(Bruggemann et al., 2018),(Cabrera-Bean et al., 2017),(Della Mea et al., 2014),(dos Reis et al., 2015),(Eickhoff, 2014),(Irshad et al., 2015),(Irshad et al., 2017),(Lawson et al., 2017),(Luengo-Oroz et al., 2012),(Mavandadi et al., 2012),(Sameki et al., 2016),(Sharma et al., 2017),(Smittenaar et al., 2018),(Sullivan et al., 2018)
Lung	(Boorboor et al., 2018),(Cheplygina et al., 2016),(Huang and Hamarneh, 2017),(O'Neil et al., 2017),(Ørting et al., 2017)
Multiple	(Foncubierta Rodríguez and Müller, 2012),(Gurari et al., 2016),(de Herrera et al., 2014),(Lejeune et al., 2017)
Other	(Chávez-Aragón et al., 2013),(Cheplygina and Pluim, 2018),(Gurari et al., 2015b),(Holst et al., 2015),(Malpani et al., 2015)

***Table 3.* Comparison of the distribution of applications in this survey and two other recent surveys in medical image analysis. Percentages are rounded to the nearest whole number.**

Application	This survey	Cheplygina et al. (2018)	Litjens et al. (2017)
Brain	9%	21%	18%
Eye	14%	4%	5%
Lung	9%	13%	14%
Breast	0%	6%	7%
Heart	2%	4%	7%
Abdomen	23%	14%	9%
Histo/Micro	28%	17%	20%
Multiple	7%	12%	4%
Other	9%	10%	16%

3D problem by presenting a video to the users where the entire image is displayed as a sequence of 2D frames (Boorboor et al., 2018). Only a few of the papers addressing 3D images, present images in 3D (Huang and Hamarneh, 2017; Sonabend et al., 2017).

The last type of data that is addressed is video, common for endoscopy and colonoscopy (both in the abdomen category). Several different approaches are used for presenting video data: 2D frames (Maier-Hein et al., 2014b, 2015, 2016; Heim, 2018; Roethlingshoefer et al., 2017), 3D renderings (Nguyen et al., 2012; McKenna et al., 2012), short video clips (Park et al., 2017), or longer videos that can be paused and annotated (Park et al., 2018).

Other applications of crowdsourcing include segmenting hip joints in 2D MRI (Chávez-Aragón et al., 2013), rating visual characteristics of dermatological images (Cheplygina and Pluim, 2018) and assessing surgical performance (Malpani et al., 2015; Holst et al., 2015). Two papers (Foncubierta Rodríguez and Müller, 2012; de Herrera et al., 2014) look at multiple applications, where the task is classifying image modality, rather than segmentation or diagnosis. A few papers address segmentation in multiple modalities: Gurari et al. (2016) focus on both natural and biomedical images, Lejeune et al. (2017) address segmentation across four medical applications.

2.3. Data availability

Next to categorizing the type of applications, we also examined whether the datasets used in these studies were publicly available. Out of 57 papers, at least 22 papers used at least one publicly available dataset. We only considered datasets as public when they were clearly identifiable as such, for example the paper described the dataset as openly available and contained a reference to a publication about the dataset, and/or a dataset website.

The public data sources that were used were:

- *AMIDA-13 challenge* https://ismi-amida13.grand-challenge.org/
- *EndoVis challenge* https://endovis.grand-challenge.org/
- *LIDC-IDRI database* https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
- *ISIC 2017 challenge* https://challenge.isic-archive.com/landing/2017
- *ImageCLEF medical dataset* https://www.imageclef.org/2016/medical
- *TCGA Portal* https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- *Healthy Brain Network Data Portal* http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/
- *BRATS challenge* http://braintumorsegmentation.org/ and others (different websites for different years)
- *BU-BIL segmentation repository* http://www.cs.bu.edu/fac/betke/BiomedicalImageSegmentation/
- *Human Protein Atlas* https://www.proteinatlas.org/

Anecdotally, in various cases the URL provided in the paper did not lead to the dataset in question. One reason was broken links, in this case we tried to find an up-to-date URL. Another reason was datasets which have multiple versions, such as the challenge datasets where data is released every year.

3. INTERACTION

An important aspect of crowdsourcing medical image annotations is task design. The interplay between the type of image data, the type of annotations that are needed and the available tools for

Table 4. Interaction Types.

Interaction	Papers
Click	(Bruggemann et al., 2018),(Cabrera-Bean et al., 2017),(Della Mea et al., 2014),(Huang and Hamarneh, 2017),(Lawson et al., 2017),(Lejeune et al., 2017),(Luengo-Oroz et al., 2012),(Park et al., 2018),(Rajchl et al., 2016)
Click + Compare	(Maier-Hein et al., 2014b),(Maier-Hein et al., 2015),(Maier-Hein et al., 2016)
Click + Draw	(Irshad et al., 2015)
Compare	(Ørting et al., 2017)
Draw	(Roethlingshoefer et al., 2017),(Boorboor et al., 2018),(Chávez-Aragón et al., 2013),(Cheplygina et al., 2016),(Ganz et al., 2017),(Gurari et al., 2015b),(Heller et al., 2017),(Lee and Tufail, 2014),(Lee et al., 2016),(Maier-Hein et al., 2014a),(O'Neil et al., 2017),(Sameki et al., 2016),(Sharma et al., 2017)
Rate	(Albarqouni et al., 2016a),(Albarqouni et al., 2016b),(Brady et al., 2014),(Brady et al., 2017),(Cheplygina and Pluim, 2018),(dos Reis et al., 2015),(Eickhoff, 2014),(Foncubierta Rodríguez and Müller, 2012),(Gur et al., 2017),(de Herrera et al., 2014),(Holst et al., 2015),(Keshavan et al., 2018),(Mavandadi et al., 2012),(McKenna et al., 2012),(Mitry et al., 2013),(Mitry et al., 2015),(Nguyen et al., 2012),(Park et al., 2016),(Park et al., 2017),(Smittenaar et al., 2018),(Sonabend et al., 2017)
Rate + Compare	(Malpani et al., 2015)
Rate + Draw	(Gurari et al., 2016),(Heim, 2018),(Irshad et al., 2017),(Mitry et al., 2016),(Sullivan et al., 2018),(Timmermans et al., 2016)
Rate + Draw + Click	(Leifman et al., 2015)

annotation, needs to be considered to successfully crowdsource annotations. A major component of task design is choosing how workers interact with the task. The type of interaction influences time per task and the required level of expertise and training, which ultimately translates into cost and quality. We identified four categories of interaction tasks across the studies surveyed:

– Rate an entire image
– Draw shapes to identify regions of interest
– Click on specific locations
– Compare two or more images

An overview of the surveyed studies is given in Table 4. Furthermore, we also observed that studies generally had crowds either create entirely new annotations on unlabeled data, or make responses based on pre-existing annotations, e.g., output from automated segmentation methods.

Rating entire images was the most common interaction and was the main task of 52% of the studies surveyed here. Ratings took many forms, identifying the presence/absence of specific visual features (Sonabend et al., 2017), counting number of cells (Smittenaar et al., 2018), assessing intensity of cell staining (dos Reis et al., 2015), or discriminating healthy samples from diseased (Mavandadi et al., 2012). Most commonly, crowd workers were asked to create new annotations (90% of rating tasks). Less commonly, crowd workers were asked to validate pre-existing annotations (14%). One study involved both validating pre-existing annotations and creating new ones (Heim, 2018), so the percentages do not sum to 100%. Existing annotations were the output of automated methods (Roethlingshoefer et al., 2017; Ganz et al., 2017; Gur et al., 2017) half of the time, and the crowdsourced annotations were used to identify instances with errors to be corrected.

Drawing a shape was the second most common task, comprising 38% of studies. Here crowd workers were asked to draw bounding boxes or segment outlines of structures of interest. Sometimes, this was only after identifying if a structure was present in the image or not (Heim, 2018). Similar

to rating images, drawing shapes was used as an interaction for both creating new annotations (90% of drawing tasks) and validating existing annotations (14%). In the case of evaluating existing annotations, drawing was used as a means to indicate the location of errors in segmentations produced by automated methods (Roethlingshoefer et al., 2017; Ganz et al., 2017).

Clicking on specific locations was the third most used interaction, occurring in 25% of studies. Clicking was only used to create new annotations such as identifying the precise location of specific cells, abnormalities, or artifacts within an image. The use of multiple clicks to outline a structure was considered a "drawing a shape" interaction. Selecting points was also used in pairs of video frames to determine the stereotactic correspondence of two video streams for follow-up 3D reconstruction (Maier-Hein et al., 2014b, 2015, 2016).

Comparing two or more images was the least used interaction, occurring in only 5 (9%) of studies. In all cases, comparisons were used to create new annotations, such as marking corresponding points in two consecutive video frames (Maier-Hein et al., 2015, 2016) or to choose which of two images was more similar to a target image (Ørting et al., 2017).

Overall, crowds were more often used to create new annotations, than to make judgments on existing annotations, which was done only in (Roethlingshoefer et al., 2017; Foncubierta Rodríguez and Müller, 2012; Ganz et al., 2017; Gur et al., 2017; de Herrera et al., 2014). Ratings and drawing of shapes can be used to obtain more detailed annotations than information already present in datasets. Clicking interactions are sometimes used to identify specific image features, but more commonly used to create bounding boxes or draw object boundaries. Evaluating existing annotations is always done with rating (Foncubierta Rodríguez and Müller, 2012; de Herrera et al., 2014) or drawing (Roethlingshoefer et al., 2017; Ganz et al., 2017) interactions. Different types of annotations were often collected based on the type of task information being sought, e.g., clicking to obtain counts and locations of specific image features (Cabrera-Bean et al., 2017; Della Mea et al., 2014) and drawing to determine segmentation boundaries (Gurari et al., 2015b; Chávez-Aragón et al., 2013). Ratings were a more general type of interaction and could be used to classify for whole-image level discrimination of image feature presence or category (Keshavan et al., 2018; de Herrera et al., 2014; McKenna et al., 2012) or to obtain estimates of feature certainty (dos Reis et al., 2015).

Rating and drawing interactions for existing annotations are usually chosen to speed up the annotation process, as verifying and correcting existing annotations is faster for crowd workers than annotating an image from scratch (Roethlingshoefer et al., 2017). Similarly, rating with predefined categories is preferred instead of free text input, which is prone to spelling mistakes and misunderstanding of the annotation task (Foncubierta Rodríguez and Müller, 2012). Timmermans et al. (2016) chose custom drawing instead of the most commonly used drawing of independent polygons to better identify the shapes of interest.

4. PLATFORM, SCALE AND WAGES

In this section we summarize the main meta parameters, and settings of crowdsourcing experiments. First, we classify the reviewed papers based on the type of platform used to perform the crowdsourcing experiments. Second, we report on the scale of the experiments where we consider the number of images annotated and the number of annotators per image. Table 5 summarizes these key aspects of the surveyed papers. Finally, we summarize the wages paid to crowd workers.

***Table 5.** Used platform and study scale in terms of completed annotations.*

	Scale			
Platform	L ($\geq$ 1000)	M (100 to 1000)	S (10 to 100)	XS ($\leq$ 10)
Paid	(Brady et al., 2017),(Foncubierta Rodríguez and Müller, 2012),(de Herrera et al., 2014),(Irshad et al., 2017)	(Eickhoff, 2014),(Gurari et al., 2015b),(Gurari et al., 2016),(Heim, 2018),(Irshad et al., 2015),(Maier-Hein et al., 2014a),(Maier-Hein et al., 2016),(McKenna et al., 2012),(Mitry et al., 2015),(Nguyen et al., 2012),(Ørting et al., 2017),(Park et al., 2016),(Park et al., 2017),(Park et al., 2018),(Sameki et al., 2016),(Irshad et al., 2017)	(Boorboor et al., 2018),(Brady et al., 2014),(Cheplygina et al., 2016),(Della Mea et al., 2014),(Ganz et al., 2017),(Holst et al., 2015),(Lee and Tufail, 2014),(Lee et al., 2016),(Maier-Hein et al., 2014b),(Maier-Hein et al., 2015),(Mitry et al., 2013),(Mitry et al., 2016),(Sharma et al., 2017)	
Volunteer	(dos Reis et al., 2015)		(Lawson et al., 2017),(Luengo-Oroz et al., 2012)	(Cabrera-Bean et al., 2017)
Experts			(Lejeune et al., 2017),(Sonabend et al., 2017)	
Students			(Cheplygina and Pluim, 2018)	
Custom	(Albarqouni et al., 2016b),(Rajchl et al., 2016),(Smittenaar et al., 2018),(Sullivan et al., 2018),(Timmermans et al., 2016)	(Albarqouni et al., 2016a),(Chávez-Aragón et al., 2013),(Gur et al., 2017),(Keshavan et al., 2018)	(O'Neil et al., 2017)	(Heller et al., 2017)
None	(Roethlingshoefer et al., 2017),(Mavandadi et al., 2012)	(Rajchl et al., 2017)		
NA	(Bruggemann et al., 2018)	(Bruggemann et al., 2018)		

Note: Some papers report multiple studies of different sizes. Therefore, these papers are listed more than once.

4.1. Crowdsourcing platforms

A potentially important factor that varies across the surveyed papers is the choice of platform for conducting crowdsourcing experiments. We classify the platforms into six categories: paid commercial marketplaces such as Amazon Mechanical Turk[2] and FigureEight (formerly known as CrowdFlower and acquired by Appen[3] in 2019), volunteers such as Zooniverse[4] and Volunteer Science[5], custom recruitment/platforms, lab participants, experts and simulation or no experiment at all. The most common choice is a commercial platform (55%). The second most common choice is a custom platform (23%) followed by a volunteer platform (10%). The remaining 8% were almost equally divided into the other categories with around 5% of all papers reporting prototypes or simulation studies. Around half of the papers we reviewed, namely 25 papers, motivate the choice of the platform and name some of their advantages and disadvantages.

The main reasons for choosing a paid commercial marketplace such as Amazon Mechanical Turk and FigureEight are both explicitly and implicitly mentioned: to be able to reach out to a large and diverse crowd (Boorboor et al., 2018; Brady et al., 2014; Bruggemann et al., 2018; Gurari et al., 2016; Irshad et al., 2015, 2017; Foncubierta Rodríguez and Müller, 2012); to be cost-efficient (Boorboor et al., 2018; Gurari et al., 2016; Irshad et al., 2015, 2017); to be time-efficient (Brady et al., 2014; de Herrera et al., 2014; Irshad et al., 2015, 2017) and to make use of in-place quality control mechanisms (Della Mea et al., 2014; Nguyen et al., 2012; Irshad et al., 2015, 2017; Della Mea et al., 2014). Amazon Mechanical Turk is preferred over other paid platforms because it allows requesters to add custom-built annotation interfaces (Cheplygina et al., 2016; Heim, 2018; Maier-Hein et al., 2015) and test their HITs in a sandbox (Heim, 2018). Similarly, researchers prefer the FigureEight platform because it is available in Europe (Della Mea et al., 2014) and because it provides an internal interface where they can set up the annotation task and distribute it to their network, without having to pay for the annotations (de Herrera et al., 2014; Foncubierta Rodríguez and Müller, 2012).

Volunteer or custom recruitment platforms are usually chosen to mitigate the disadvantages of paid commercial marketplaces. Such platforms enable users to build custom, lightweight applications, fine-tune interfaces and manipulate images for better annotations (Leifman et al., 2015; Albarqouni et al., 2016a; Heller et al., 2017; Keshavan et al., 2018; dos Reis et al., 2015). Furthermore, custom platforms allow requesters to mitigate privacy issues, as data remains in secure, centralized clinical repositories at any time (Gur et al., 2017; Heller et al., 2017). Sullivan et al. (2018) and Huang and Hamarneh (2017) chose to use a custom platform to be able to integrate gamification aspects that proved helpful to reduce the time needed for annotation, to motivate players and maintain them for longer annotation campaigns. One other major advantage of volunteer or custom platforms is the fact that they can reach out to contributors that are interested in supporting medical imaging research or science in general (Rajchl et al., 2016; dos Reis et al., 2015). Furthermore, such applications are usually much simplified, can be played at any time and on any device (Keshavan et al., 2018; dos Reis et al., 2015).

Among disadvantages of chosen paid commercial marketplaces we find the following. Brady et al. (2014) mention that built-in qualifications in Amazon Mechanical Turk such as "Photo Moderation

[2]https://www.mturk.com Accessed Jun. 2020

[3]https://www.appen.com Accessed Jun. 2020

[4]https://www.zooniverse.org Accessed Jun. 2020

[5]https://volunteerscience.com Accessed Jun. 2020

Master" are not useful for the medical imaging analysis task conducted in the paper, as there were only a few annotators available. Furthermore, this increased both the cost and the time to complete the experiments (Brady et al., 2014; McKenna et al., 2012). Another disadvantage of Amazon Mechanical Turk mentioned by Brady et al. (2017) is the fact that workers could potentially use automated scripts to accept or reserve large amounts of HITs at a time. As a consequence, the time needed to complete each HIT can not be computed reliably anymore. Cheplygina et al. (2016) note that the integration of custom-built annotation interfaces in Amazon Mechanical Turk, although useful, is costly and time-consuming for novice users of the platform. Mitry et al. (2016) mention that the integrated annotation tool in Amazon Mechanical Turk only allowed rectangles to be drawn, while this could affect users' ability to capture more irregular regions of interest. Another major disadvantage of Amazon Mechanical Turk is the fact that the platform was not always available for requesters outside US (Maier-Hein et al., 2015). Regarding FigureEight, the major disadvantages are the difficulty of setting up gold questions for annotation tasks that involve drawing and segmentation (Della Mea et al., 2014) and insufficient settings to control for the number of annotations that should be performed for each image (Albarqouni et al., 2016a).

While volunteer and custom platforms could attract a large pool of participants due to their advertisement campaigns, large drop-off in user participation can be seen over time (Smittenaar et al., 2018; dos Reis et al., 2015). In one case, advertisement campaigns only attracted neuroscientists from the social network of the requesters thus limiting the diversity of the annotators (Keshavan et al., 2018). Furthermore, many participants seem to annotate only a few samples (dos Reis et al., 2015).

4.2. Scale

We summarize the scale of the crowdsourcing experiments in terms of number of images annotated, number of images per task, and number of annotations per image.

Number of images: We classify the number of images into four categories: very small (less than 10 annotated images), small (10 to 100 annotated images), medium (100 to 1000 annotated images) and large (more than 1000 annotated images). Column #I in Table 9 shows an overview of the exact number of images annotated in each paper included in the review. The large majority of reviewed papers (70%) report small and medium scale experiments, while a smaller part report large experiments (22%) or very small experiments (5%). However, in around 3% of the reviewed papers, the scale of the experiments is not reported.

Number of images per task: In total, 25 out of the 57 papers included in this survey report on the number of images per task, 25% use one image per task, 7% use five images per task, 5% use ten images per task, while the other 7% use between 3 and 84 images per task. Irshad et al. (2015) and Irshad et al. (2017) also mention that out of the five images in the task, four images are actually unlabeled data, while one image is a gold question.

Number of annotations per image: We divide the number of annotations per image into two categories: a single annotator per image (5%) or multiple annotators per image (63%). Surprisingly, for 33% of surveyed papers the number of annotations per image is not reported nor can it be inferred. Column #Ann/I in Table 9 shows an overview of the exact number of annotations per image, as reported in each paper reviewed.

Overall, the experiments using a single annotator per image involve either simulations or locally recruited, volunteer-based annotators that are not remunerated. The number of annotators per image

for experiments using multiple annotators per image ranges from 2 to 5000. However, the majority (66%) of these experiments use between 5 to 25 annotators per image.

4.3. Annotators Wage

We classify the wage given to annotators into six different categories: a few dollars per hour, less than or equal to $0.10 per annotation, more than $0.10 per annotation, volunteers (no monetary incentive), not specified (if we have no information about compensation) and none (if no actual experiment or recruitment took place).

More than a third (34%) of papers did not specify anything about wage. In 34% of papers the wage was less than or equal to $0.10, in 24% of papers crowds where volunteers with no monetary incentive, in 5% of papers the wage was more than $0.10, and in 3% of papers the wage was an hourly payment of a few dollars per hour.

Overall, very few and mainly the papers that mention an hourly payment considered crowd worker wages in relation to minimum wage rules and regulations.

5. EVALUATION

In this section we describe how the crowdsourced annotations are evaluated. This is done via two strategies: ensuring sufficient quality of annotations by preprocessing and estimating the utility of the crowd annotations for the task at hand. Although the two strategies are closely related and should be considered jointly when designing crowdsourcing experiments, it is informative to consider them separately here.

The first strategy is closely related to the field of quality control in crowdsourcing. Numerous approaches exist to tackle this, starting from simple majority voting and worker filtering to sophisticated statistical and machine learning methods that consider workers' specific skills, task difficulty and clarity of task descriptions. The second strategy is more domain-specific, as different tasks may have different levels of tolerance for errors.

5.1. Preprocessing of annotations

Preprocessing of annotations broadly covers what is done to the crowdsourced annotations prior to using them for their intended purpose. It includes filtering individual annotations and/or aggregating annotations. The majority (84%) of the surveyed papers perform some form of preprocessing.

5.1.1. Filtering individuals

One way to filter annotations, is to remove annotations made by "poorly performing" annotators. Most crowdsourcing platforms offer a rating score for workers that provides an estimate of their performance, based on their percentage of previously approved tasks. This score is used in 16% of surveyed papers to filter workers prior to assigning tasks. A related approach, used in 12% of surveyed papers, is to exclude workers that fail a test task prior to the actual tasks. A refinement of this, used in 23% of surveyed papers, is to integrate separate test tasks in the tasks and exclude workers that fail the tests. Park et al. (2018), for example, added a smiley face to colonoscopy videos to ensure attention.

Another common filtering approach for individual workers, used in 23% of surveyed papers, is comparing annotations to gold standard annotations. In this case, tasks with known gold standard annotations, are injected into the regular working process. A worker's correspondence with the gold standard can then be used to estimate individual worker performance. In contrast to platform

Table 6. Filtering mechanisms.

Filtering	Papers
Before	(Heim, 2018),(Holst et al., 2015),(Lee et al., 2016),(McKenna et al., 2012),(Nguyen et al., 2012),(Ørting et al., 2017)
Before/during	(Albarqouni et al., 2016a),(Gurari et al., 2015b),(Irshad et al., 2015),(Irshad et al., 2017),(Mavandadi et al., 2012),(Park et al., 2018)
During	(Boorboor et al., 2018),(Della Mea et al., 2014),(dos Reis et al., 2015),(Foncubierta Rodríguez and Müller, 2012),(Keshavan et al., 2018),(Leifman et al., 2015),(Luengo-Oroz et al., 2012),(Malpani et al., 2015)
After	(Chávez-Aragón et al., 2013),(Cheplygina et al., 2016),(O'Neil et al., 2017),(Park et al., 2016)
Before/after	(Gurari et al., 2016),(Mitry et al., 2015),(Mitry et al., 2016)

Table 7. Aggregation mechanisms.

Aggregation	Papers
Mean/median	(Cheplygina et al., 2016),(Cheplygina and Pluim, 2018),(Della Mea et al., 2014),(dos Reis et al., 2015),(Ganz et al., 2017),(Lejeune et al., 2017)
Majority	(Eickhoff, 2014),(Gurari et al., 2016),(Maier-Hein et al., 2016),(Mitry et al., 2016),(Nguyen et al., 2012),(O'Neil et al., 2017),(Park et al., 2017)
Majority/weighted	(Albarqouni et al., 2016a),(Brady et al., 2017),(Gurari et al., 2015b),(Heim, 2018),(Irshad et al., 2017),(Malpani et al., 2015),(Park et al., 2016)
Weighted	(Albarqouni et al., 2016b),(Keshavan et al., 2018),(Smittenaar et al., 2018)
Other	(Cabrera-Bean et al., 2017),(Holst et al., 2015),(Huang and Hamarneh, 2017),(Leifman et al., 2015),(Luengo-Oroz et al., 2012),(Maier-Hein et al., 2014b),(Maier-Hein et al., 2015),(McKenna et al., 2012),(Ørting et al., 2017),(Sameki et al., 2016),(Sharma et al., 2017),(Sullivan et al., 2018)

scores and unrelated test tasks, this approach assesses worker performance on the specific task, allowing more fine-grained worker selection. The filtering mechanisms used in the surveyed papers are summarized in Table 6.

5.1.2. Aggregating results

One of the main benefits of crowdsourcing is the fast and cost-effective collection of a large number of annotations. This allows aggregating annotations to reduce noise in the individual annotations.

Majority voting is widely used due to its computational and conceptual simplicity, and was found in 23% of the papers. In the context of medical image analysis, majority voting is applied to annotations, ratings, and also to aggregate slices of images. Heim (2018), for example, used crowdsourcing for organ segmentation in computed tomography scan. Multiple organ outlines are collected via an online tool and pixel-wise majority voting is applied to improve the accuracy of the segmentation. In the case of numerical ratings, mean and median statistics are also used in 12% of the papers to determine a final annotation. For example, Cheplygina et al. (2016) used the median to aggregate the areas of the annotations created by individual workers. A more sophisticated version of the majority vote uses additional information about the general quality of workers. This information can be derived if workers perform multiple tasks or if gold standard data is available. Weighted voting is used in 16% of surveyed papers, for example, Keshavan et al. (2018) used the XGBoost algorithm to estimate annotator weights, and Brady et al. (2017) estimated the weights of the annotators as the probability that an annotator is correct while taking task difficulty into account. The aggregation mechanisms used in the surveyed papers are summarized in Table 7.

5.2. Evaluating annotations

Evaluating how well crowd annotations solve the intended purpose is most commonly (79% of surveyed papers) done by directly comparing crowdsourced annotations to a gold standard. In about 16% of surveyed papers crowd annotations are used for training a machine learning method, and the performance of the machine learning method used to indirectly evaluate annotations. The remaining 5% have no evaluation of how well annotations solve the intended purpose.

The gold standard originates from different sources. In about 25% of surveyed papers, the gold standard is based on a single expert, in about 37% the gold standard is based on multiple experts, and in the remaining papers the number of experts is not reported or no expert gold standard is used. Using a gold standard based on a single expert can be problematic since experts often disagree on all but the most trivial tasks. However, only 3 of 21 papers that use multiple experts consider how well experts agree.

Expert-based gold standards are generally not obtained from experts performing exactly the same task as the crowd. In several cases the only difference in expert and crowd tasks is due to differences in user interface, e.g. a clinical workstation for experts and a web interface for crowds. As long as the fundamental task is the same (e.g. count cells) and the user interface has not been dramatically changed we consider the expert and crowd tasks to be the same. Using this definition, about 40% of the papers use the same task and about 40% use a different task. In the remaining 19% it is either not reported or no expert gold standard is used.

There are several reasons for asking crowds to perform a different task than what experts have done for the gold standard. Some papers use a simplified version of the expert task in order to make the task easier or more suitable as a small self-contained task. For example, ranking relative performance in pairs of surgical videos instead of grading performance in each (Malpani et al., 2015); assessing visual similarity of images instead of classifying disease patterns (Ørting et al., 2017); refining segmentation proposals instead of performing a full segmentation (Maier-Hein et al., 2016); annotating polyps in a single frame instead of in a full video (Park et al., 2017) or counting stained cells instead of classifying disease status (Irshad et al., 2017). Other papers focus on changing the user interface, such as Lejeune et al. (2017) who used an eye tracker for segmentation instead of a mouse, or Albarqouni et al. (2016b) and Mavandadi et al. (2012) who changed the user interface to support gamification strategies.

In a few papers, the evaluation is focused on variation in annotations. For example, Lee and Tufail (2014) and Lee et al. (2016) evaluate annotations in terms of inter-rater reliability; and Heller et al. (2017),Huang and Hamarneh (2017),Leifman et al. (2015), and Sonabend et al. (2017) compare individual annotations to aggregated annotations. Measuring variability of annotations it not directly useful for evaluating the correctness of annotations. However, annotation variability is essential when evaluating how much the crowdsourced annotations can be trusted. Additionally, variation provides an indirect measure of correctness. Large variation can indicate that annotations are often wrong, while small variation indicates that annotations are often correct or the task has been designed such that annotators are consistently wrong. All evaluation used in the survey papers are aggregated in Table 8.

Table 8. Evaluation of Annotations.

Gold Standard	Direct	Indirect
Multiple Experts	(Albarqouni et al., 2016b),(Boorboor et al., 2018),(Bruggemann et al., 2018),(Gurari et al., 2015b),(Gurari et al., 2016),(Heim, 2018),(Holst et al., 2015),(Irshad et al., 2015),(Keshavan et al., 2018),(Lawson et al., 2017),(Leifman et al., 2015),(Luengo-Oroz et al., 2012),(Maier-Hein et al., 2014b),(Malpani et al., 2015),(Mitry et al., 2013),(Mitry et al., 2015),(Mitry et al., 2016),(Smittenaar et al., 2018)	(Albarqouni et al., 2016a),(Gur et al., 2017),(Ørting et al., 2017)
One Expert	(Cheplygina et al., 2016),(Della Mea et al., 2014),(dos Reis et al., 2015),(Eickhoff, 2014),(Foncubierta Rodríguez and Müller, 2012),(Ganz et al., 2017),(Irshad et al., 2017),(Mavandadi et al., 2012),(O'Neil et al., 2017),(Park et al., 2016),(Park et al., 2017),(Park et al., 2018)	(Maier-Hein et al., 2014a),(Rajchl et al., 2016)
Other	(Chávez-Aragón et al., 2013),(Heller et al., 2017)	(de Herrera et al., 2014)
Unknown	(Brady et al., 2014),(Brady et al., 2017),(Cabrera-Bean et al., 2017),(Huang and Hamarneh, 2017),(Maier-Hein et al., 2015),(McKenna et al., 2012),(Nguyen et al., 2012),(Sameki et al., 2016),(Sullivan et al., 2018)	(Cheplygina and Pluim, 2018),(Lejeune et al., 2017),(Maier-Hein et al., 2016)
Na	(Lee and Tufail, 2014),(Lee et al., 2016),(Sharma et al., 2017),(Sonabend et al., 2017)	

Note: Some paper are excluded from this table as they did not report the method of generating the gold standard data nor the type of comparison. These paper are: (Roethlingshoefer et al., 2017),(Rajchl et al., 2017),(Timmermans et al., 2016)

6. RESULTS AND RECOMMENDATIONS

Here, we provide an overview of the primary results and recommendations emerging from the papers examined in this review. Complementary to the topics discussed in Section 5, we consider how effective the application of crowdsourcing to medical image analysis is, and provide recommendations to ensure data quality.

6.1. How effective is the application of crowdsourcing to medical image analysis?

The vast majority of studies examined in this review found crowdsourcing to be a valid approach for data production. Crowdsourcing of medical image analysis was noted to be an accurate approach (Lawson et al., 2017) that can produce large quantities of annotations needed to solve high-throughput problems requiring human input (Irshad et al., 2015; dos Reis et al., 2015; Lee and Tufail, 2014; Maier-Hein et al., 2014b). Crowdsourcing can be used to create new annotations or make existing data more robust, both cheaper and faster than annotation by medical experts (Rajchl et al., 2016; Holst et al., 2015; Gurari et al., 2016; Eickhoff, 2014; Park et al., 2017).

Although the relative efficacy of crowdsourcing applied to medical image analysis will be dependent on the complexity of the task, the papers examined here show crowdsourcing to be an effective methodology across a wide variety of applications, including objective assessment of surgical skill (Malpani et al., 2015), emphysema assessment (Ørting et al., 2017), polyp marking in virtual colonoscopy (Park et al., 2018), identification of chromosomes (Sharma et al., 2017) and biomarker discovery in immunohistochemistry data (Smittenaar et al., 2018). Notably, only one project stated that crowdsourcing could not always be applied effectively to the studied task ("it is very difficult and maybe even impossible to entirely outsource the task of labelling mitotic figures in histology

images to crowds" (Albarqouni et al., 2016a)).

Rather than comparing the absolute performance of the crowd to experts or to algorithms, it might be worth considering their relative benefits. For example, crowds were particularly useful for rare classes (Sullivan et al., 2018), which are often difficult cases for algorithms. Another situation where crowds can be useful is identifying data that is missing from the gold standard provided by experts, see for example (Luengo-Oroz et al., 2012). Benefits of combining crowds with algorithms were also demonstrated by (Albarqouni et al., 2016a; Keshavan et al., 2018; Sharma et al., 2017).

6.2. Recommendations to ensure data quality

The papers examined in this review included suggestions to improve the quality of data produced through crowdsourcing. These suggestions focused on refining the task design, crowdsourcing platform and post-processing of annotations. We summarize these recommendations here.

6.2.1. Task design

As discussed, crowdsourcing has been applied effectively to many medical imaging applications. However, careful study design remains necessary to ensure generation of data of sufficient quality.

The selection and design of an appropriate crowdsourcing task is central to project success. Effort should be made to make the task simple and unambiguous (Rajchl et al., 2016; Gurari et al., 2016), and to present study data appropriately (McKenna et al., 2012). For unavoidably challenging tasks, crowdsourcing may still provide useful data, for instance, through enabling a rapid first-pass evaluation of large scale data sets (Della Mea et al., 2014; Park et al., 2017). Particularly challenging tasks may be made tractable through gamification (Albarqouni et al., 2016b) or careful reframing of the task, e.g. crowdsourcing of emphysema assessment was made possible through reframing the task as a question of image similarity (Ørting et al., 2017). Alternatively, it may be possible to achieve the desired data quality simply through asking a larger cohort of crowd workers to perform each task per data point. Gurari et al. (2016) give an interesting example of task design where quality and speed of crowdsourced segmentations in natural images are increased by flipping images, suggesting that familiarity with an image can be detrimental.

Besides the technical and methodological challenges, also ethical considerations have to be addressed in this design phase. The workers should be informed of the visual content of the task, e.g. surgery images, before the first image is shown. Further, an appropriate wage should be provided. An appropriate wage is often hard to determine as it depends on various factors, like the home country of the workers, platform in use, and complexity of the task. However, these factors should be considered and reported in publications.

6.2.2. Crowdsourcing platform

The choice of crowdsourcing platform can influence study cost and completion time, as well as the size and demographics of the crowd. Furthermore, different platforms offer distinct features which may influence the quality of data produced. For example, Heller et al. noted that user interface features, such as zoom and intuitive controls, can increase data quality. Contingent on the complexity of the task and interface design, training materials should be provided, as this can improve results (McKenna et al., 2012). However, this is not always necessary - in some cases minimal (Brady et al., 2014) or no training (Ganz et al., 2017) was required.

6.2.3. Post-processing

Post-processing of annotations is recommended to improve annotation quality by removing annotations from poorly performing workers. Alternatively, if multiple workers annotate the same data it is possible to improve annotation quality by aggregating annotations.

The surveyed papers propose a variety of criteria for filtering individual annotations. For example, time spend on task (O'Neil et al., 2017), expected shape of segmentation (Cheplygina et al., 2016; Chávez-Aragón et al., 2013), correlation with other workers' results (Sharma et al., 2017; Chávez-Aragón et al., 2013) and correlation to experts annotations or ground truth (Sameki et al., 2016; Keshavan et al., 2018; Irshad et al., 2017, 2015; Foncubierta Rodríguez and Müller, 2012). However, due to the lack of comparisons between different filtering approaches, the only clear recommendation from these works is to use some form of filtering.

Nguyen et al. (2012) found that filtering unreliable workers did not have a significant influence when annotations from multiple workers are aggregated. However, aggregating without taking individual performance into account might not be the best approach. Malpani et al. (2015) compared different aggregation methods, and found that weighted voting, with weights based on self-reported confidence scores, improved results compared to simple majority voting. Similarly, Irshad et al. (2015) found that aggregating segmentations from 3-5 workers, using weights based on consensus and worker trust scores, improved performance over using single worker annotations. Further, Cheplygina and Pluim (2018) found that disagreement between workers was predictive of melanoma diagnosis in skin lesions, suggesting that simple aggregation, such as majority voting or mean statistics, might not be the best approach.

7. DISCUSSION

In this section we discuss the trends, limitations and opportunities within crowdsourcing in medical imaging.

7.1. Trends

As discussed in Section 2, crowdsourcing is applied to a variety of medical images, however, it is most commonly applied to histology or microscopy images. The trend for crowdsourcing of this image type may be due to the ease of which these (typically 2D) images can be incorporated into a crowdsourcing or citizen science project. Alternatively, the microscopy images examined in these papers may have not been derived from a patient, and would therefore not require the consent of an individual to use for crowdsourcing purposes.

The most common crowd task is rating entire images. This is somewhat surprising, given that we would expect such tasks to rely more on prior knowledge than other crowdsourced tasks, such as drawing outlines of objects. Again, this trend might be facilitated due to the ease with which rating images can be incorporated in existing platforms.

Most crowdsourcing studies are set up on commercial platforms, followed by custom platforms. Each image is annotated by multiple crowd workers, who typically receive less than $0.10 per annotation. On the one hand, this low reimbursement might be a product of researchers trying to optimize the total number of annotations given a particular budget. On the other hand, it could be a lack of awareness of what appropriate compensation should be (Hara et al., 2017).

A surprising finding is that, often, important details about the crowd and their compensation are

missing. Besides missing details in terms of crowd compensation, we find missing details regarding the number of requested annotations per unit. While for some of the surveyed papers, we could infer an approximation of the number of annotations gathered per unit by checking the scale of the experiment and the total amount of annotations gathered, for at least a third of the surveyed papers (33%) this was not possible due to a lack of detail when describing the crowdsourcing experimental methodology.

Crowdsourced annotations are generally processed prior to evaluating how well the annotations solved the intended purpose. Simply excluding workers based on platform scores or a single test task is not as popular as continuously monitoring worker performance. 61% of the surveyed papers aggregate annotations from multiple crowd workers. This is most commonly done by simple majority voting, but some papers use estimates of task difficulty and/or worker performance to obtain a weighted aggregation.

The most common approach to evaluating the quality of preprocessed annotations is by comparing to an expert defined gold standard. A smaller set of papers use the annotations to train a machine learning method and evaluate the performance of the trained method.

The studies we reviewed almost unanimously conclude that crowdsourcing is a viable solution for medical image annotation, which may seem unexpected given the complexity of medical imaging as a field in general. There might be several possible reasons for the lack of negative results. One is researchers selecting tasks which they already expect to be suitable for crowdsourcing. Another reason is publication bias, with papers demonstrating negative results having less chance of being published, which is also an issue in computer vision (Borji, 2018).

7.2. Limitations

There are a number of limitations in the way that the current studies are being conducted. There is generally a lack of clarity in the reporting of experimental design and evaluation protocols. Additionally, ethical questions regarding worker compensation, image content and patient privacy are rarely discussed, but seem crucial to address. In several papers the study design appears to be ad-hoc. Characteristics such as the platform, number of annotators, how the task is explained and so forth, are not always motivated, or even described. This creates difficulties in understanding what leads to a successful crowdsourcing study and increases the barrier for researchers who have not used crowdsourcing before. The studies which do examine such factors are often conducted on a single application, making it difficult to generalize lessons learned to other applications. Detailed documentation of experiments is a crucial factor for ensuring reproducible science and essential for replication studies.

Another problem is the evaluation of results. The quality of crowdsourced annotations is generally estimated by comparing directly to expert annotations. However, variation in both expert defined gold standard and crowd annotations are not systematically accounted for, making it difficult to assess if crowd annotations are actually good enough. When using annotations to train machine learning methods, noisy crowd annotations might not be a problem if handled by the method. However, variation in annotations should still be investigated in this case. A related problem is using expert annotations to filter crowd annotations, which would not be possible for real unlabelled data, thus leading to overly optimistic results.

Overall, the surveyed papers reported successful results. However, from our personal experience

and discussions with other researchers, it is non-trivial to setup a crowdsourcing project for medical images. Due to the lack of negative results, the current literature does not inform researchers inexperienced with crowdsourcing about the main considerations of such a project. Furthermore, very few articles report on pilot experiments which aim to calibrate and identify the optimal crowdsourcing parameter settings such as the number of annotators per image.

There are important ethical issues which are largely not mentioned in the papers we surveyed. First of all, details about compensation are often missing, whereas this can have an important effect on the crowd (Hara et al., 2017). Furthermore, what is reasonable compensation in one country, may be too low for another country due to different cost of living. How to set the compensation fairly is an open issue that researchers should consider in their work.

Another ethical concern is whether it is possible and/or appropriate to share images with the crowd. Some images (for example of surgery) may be traumatic to view or unsuitable for children, which is more unique to the medical domain than other areas where crowdsourcing is applied e.g. astronomy or ecology projects. Another issue is sharing images from the perspective of patient consent, which is an issue that must be considered case by case.

7.3. Opportunities

Several papers discuss directions they want to take in further research. One of the popular directions is increasing the role of machine learning. Several papers not using machine learning plan to do so in future (Brady et al., 2017; Cheplygina and Pluim, 2018; Sullivan et al., 2018). Papers that already use machine learning discuss improvements to their algorithms or crowd-algorithm combinations (Sharma et al., 2017; Sameki et al., 2016).

Related to the above, tailoring the tasks to individual workers is another possibility. The rating score given to workers by platforms only reflects an overall completion rate, and might be artificially high because employers tend to rate the majority of the tasks positive and apply a filtering afterwards. Considering worker scores on different task types could help to make a better selection of workers beforehand.

Another strategy discussed as future work is the use of gamification. Several papers by Luengo-Oroz et al. (2012),Mavandadi et al. (2012), Albarqouni et al. (2016b), and Sullivan et al. (2018) have explored this idea citing increased motivation of annotators. While Luengo-Oroz et al. (2012),Mavandadi et al. (2012), and Albarqouni et al. (2016b) have task-specific games, Sullivan et al. (2018) take a more task-independent approach of a mini-game within an existing, larger game. This could be an opportunity for many other researchers, without the need to design a game from scratch. Finally, annotating images at a festival as presented by Timmermans et al. (2016) could be an interesting direction.

Beyond the opportunities that the papers discuss as future research, we see a number of other future directions for the community as a whole. Perhaps the most important future direction is openly sharing our experiences with crowdsourcing, including failures. Due to publication bias, current papers may not reflect the performance and difficulties encountered in a typical crowdsourcing project.

More generally, there is an opportunity to create a set of guidelines for crowdsourcing medical imaging studies. Rather than relying on ad-hoc choices, researchers could then make informed de-

cisions about the platform, reward of the annotators and other variables. For example, the European Citizen Science Initiative has a selection of guides for performing citizen science studies[6]. A further opportunity is to interact more with other fields where crowdsourcing has been in use longer, and to see which of their best practices are also applicable to medical imaging.

Interacting with workers could both improve projects and help establish guidelines. Workers have created communities (e.g. Reddit[7], Facebook) and discussion boards[8] for some platforms. Chandler et al. (2014) found that 28% $\pm$ 5% of the workers on Mechanical Turk read discussion boards and blogs related to Mechanical Turk. The topics of conversations, in order of frequency, are: pay, gratification, completion time, difficulty, how to successfully complete, purpose and the requesters' reputation. These forums are a valuable source for researchers for gathering information, measuring opinions and getting feedback on improving their project. This is particularly important because high throughput workers are more likely to discuss tasks (Chandler et al., 2014). This subgroup (less than 10 % of the workers do more than 75% of the work (Hara et al., 2017)) is likely to have experience with similar tasks (Chandler et al., 2014), and interaction with these workers may result in various improvements such as improvements of the user interface as observed by Bruggemann et al. (2018).

Next to image analysis, crowdsourcing could also be a way to collect, rather than curate, data to improve medical knowledge. This could vary from donating your own medical images, such as MedicalDataDonors[9] to contributing experiences about rare diseases. Since such initiatives do not focus on image analysis we did not include them in this survey, however the work by Ranard et al. (2014); Wazny (2017) may be good starting points for readers interested in these topics.

ACKNOWLEDGMENTS

The authors would like to thank eScience-Lorentz grant 2018 and Ms Gerda Filippo (Lorentz center) for their support in organizing the workshop where this paper was conceived. We thank the researchers who responded to our preprint for their valuable comments and suggestions. Silas Ørting was supported by the Danish Council for Independent Research (DFF) and the Netherlands Organization for Scientific Research (NWO). Matthias Hirth was supported by Deutsche Forschungsgemeinschaft (DFG) under Grants HO4770/2. The authors alone are responsible for the content. Oana Inel was supported by the IBM PhD Fellowship program.

8. REFERENCES

Albarqouni, S, Baur, C, Achilles, F, Belagiannis, V, Demirci, S, and Navab, N. (2016)a. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Transactions on Medical Imaging* 35, 5 (May 2016), 1313–1321.

Albarqouni, S, Matl, S, Baust, M, Navab, N, and Demirci, S. (2016)b. Playsourcing: a novel concept for knowledge creation in biomedical research. In *Deep Learning and Data Labeling for Medical Applications*. Springer, 269–277.

Alialy, R, Tavakkol, S, Tavakkol, E, Ghorbani-Aghbologhi, A, Ghaffarieh, A, Kim, S. H, and Shahabi, C. (2018). A review on the applications of crowdsourcing in human pathology. *Journal of pathology informatics* 9 (2018).

[6]https://ecsa.citizen-science.net/blog/collection-citizen-science-guidelines-and-publications Accessed Jun. 2020

[7]https://www.reddit.com/r/TurkerNation/ Accessed Jun. 2020

[8]https://www.mturkforum.com Accessed Jun. 2020

[9]http://www.medicaldatadonors.org Accessed Jun. 2020

Boorboor, S, Nadeem, S, Park, J. H, Baker, K, and Kaufman, A. (2018). Crowdsourcing lung nodules detection and annotation. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, Vol. 10579. International Society for Optics and Photonics, 105791D.

Borji, A. (2018). Negative results in computer vision: A perspective. *Image and Vision Computing* 69 (2018), 1–8.

Brady, C. J, Mudie, L. I, Wang, X, Guallar, E, and Friedman, D. S. (2017). Improving consensus scoring of crowdsourced data using the Rasch model: development and refinement of a diagnostic instrument. *Journal of medical Internet research* 19, 6 (2017).

Brady, C. J, Villanti, A. C, Pearson, J. L, Kirchner, T. R, Gup, O, and Shah, C. (2014). Rapid grading of fundus photos for diabetic retinopathy using crowdsourcing. *Investigative Ophthalmology & Visual Science* 55, 13 (2014), 4826–4826.

Bruggemann, J, Lander, G. C, and Su, A. I. (2018). Exploring applications of crowdsourcing to cryo-EM. *Journal of structural biology* 203, 1 (2018), 37–45.

Cabrera-Bean, M, Pages-Zamora, A, Diaz-Vilor, C, Postigo-Camps, M, Cuadrado-Sánchez, D, and Luengo-Oroz, M. A. (2017). Counting Malaria Parasites with a two-stage EM based algorithm using crowsourced data. In *Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2283–2287.

Chandler, J, Mueller, P, and Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46, 1 (2014), 112–130.

Chávez-Aragón, A, Lee, W.-S, and Vyas, A. (2013). A crowdsourcing web platform-hip joint segmentation by non-expert contributors. In *Medical Measurements and Applications Proceedings (MeMeA), 2013 IEEE International Symposium on*. IEEE, 350–354.

Cheplygina, V, de Bruijne, M, and Pluim, J. P. (2018). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *arXiv preprint arXiv:1804.06353* (2018).

Cheplygina, V, Perez-Rovira, A, Kuo, W, Tiddens, H, and de Bruijne, M. (2016). Early experiences with crowdsourcing airway annotations in chest CT, In Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS). *Large-scale Annotation of Biomedical data and Expert Label Synthesis* (2016), 209–218.

Cheplygina, V and Pluim, J. P. W. (2018). Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 105–111.

de Herrera, A. G. S, Foncubierta-Rodríguez, A, Markonis, D, Schaer, R, and Müller, H. (2014). Crowdsourcing for medical image classification. *Swiss Medical Informatics* 30 (2014).

Della Mea, V, Maddalena, E, Mizzaro, S, Machin, P, and Beltrami, C. A. (2014). Preliminary results from a crowdsourcing experiment in immunohistochemistry. In *Diagnostic pathology*, Vol. 9. BioMed Central, S6.

dos Reis, F. J. C, Lynn, S, Ali, H. R, Eccles, D, Hanby, A, Provenzano, E, Caldas, C, Howat, W. J, McDuffus, L.-A, Liu, B, and others, . (2015). Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine* 2, 7 (2015), 681–689.

Eickhoff, C. (2014). Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Gamification for Information Retrieval (GamifIR)*. ACM, 53–56.

Foncubierta Rodríguez, A and Müller, H. (2012). Ground truth generation in medical imaging: a crowdsourcing-based iterative approach, In ACM Multimedia workshop on Crowdsourcing for Multimedia. *Workshop on Crowdsourcing for Multimedia, ACM Multimedia* (2012), 9–14.

Ganz, M, Kondermann, D, Andrulis, J, Knudsen, G. M, and Maier-Hein, L. (2017). Crowdsourcing for error detection in cortical surface delineations. *International journal of computer assisted radiology and surgery* 12, 1 (2017), 161–166.

Greenspan, H, Van Ginneken, B, and Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1153–1159.

Gur, Y, Moradi, M, Bulu, H, Guo, Y, Compas, C, and Syeda-Mahmood, T. (2017). Towards an efficient way of building annotated medical image collections for big data studies. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 87–95.

Gurari, D, Sameki, M, and Betke, M. (2016). Investigating the influence of data familiarity to improve the design of a crowdsourcing image annotation system. In *Human Computation (HCOMP)*.

Gurari, D, Theriault, D, Sameki, M, Isenberg, B, Pham, T. A, Purwada, A, Solski, P, Walker, M, Zhang, C, Wong, J. Y, and others, . (2015)b. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In *2015 IEEE winter conference on applications of computer vision*. IEEE, 1169–1176.

Gurari, D, Theriault, D, Sameki, M, Isenberg, B, Pham, T. A, Purwada, A, Solski, P, Walker, M, Zhang, C, Wong, J. Y, and Betke, M. (2015)a. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In *Winter Conference on Applications of Computer Vision, (WACV)*. 1169–1176.

Hara, K, Adams, A, Milland, K, Savage, S, Callison-Burch, C, and Bigham, J. (2017). A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. *arXiv preprint arXiv:1712.05796* (2017).

Heim, E. (2018). *Large-scale medical image annotation with quality-controlled crowdsourcing*. Ph.D. Dissertation. German Cancer Research Center (DKFZ).

Heller, N, Stanitsas, P, Morellas, V, and Papanikolopoulos, N. (2017). A Web-Based Platform for Distributed Annotation of Computerized Tomography Scans. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 136–145.

Holst, D, Kowalewski, T. M, White, L. W, Brand, T. C, Harper, J. D, Sorensen, M. D, Truong, M, Simpson, K, Tanaka, A, Smith, R, and others, . (2015). Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. *Journal of endourology* 29, 10 (2015), 1183–1188.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1–4.

Huang, M and Hamarneh, G. (2017). SwifTree: Interactive Extraction of 3D Trees Supporting Gaming and Crowdsourcing. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 116–125.

Irshad, H, Montaser-Kouhsari, L, Waltz, G, Bucur, O, Nowak, J, Dong, F, Knoblauch, N. W, and Beck, A. H. (2015). Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. In *Pacific Symposium on Biocomputing*. World Scientific, 294–305.

Irshad, H, Oh, E.-Y, Schmolze, D, Quintana, L. M, Collins, L, Tamimi, R. M, and Beck, A. H. (2017). Crowdsourcing scoring of immunohistochemistry images: Evaluating Performance of the Crowd and an Automated Computational Method. *Scientific Reports* 7 (2017), 43286.

Keshavan, A, Yeatman, J, and Rokem, A. (2018). Combining citizen science and deep learning to amplify expertise in neuroimaging. *bioRxiv* (2018), 363382.

Knapp, K. R, Matthews, J. L, Kossin, J. P, and Hennon, C. C. (2016). Identification of Tropical Cyclone Storm Types Using Crowdsourcing. *Monthly Weather Review* 144, 10 (09 2016), 3783–3798. DOI:http://dx.doi.org/10.1175/MWR-D-16-0022.1

Kovashka, A, Russakovsky, O, Fei-Fei, L, and Grauman, K. (2016). Crowdsourcing in Computer Vision. *Foundations and Trends in Computer Graphics and Vision* 10, 3 (2016), 177–243.

Lawson, J, Robinson-Vyas, R. J, McQuillan, J. P, Paterson, A, Christie, S, Kidza-Griffiths, M, McDuffus, L.-A, Moutasim, K. A, Shaw, E. C, Kiltie, A. E, and others, . (2017). Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays. *British journal of cancer* 116, 2 (2017), 237.

Lee, A. Y, Lee, C. S, Keane, P. A, and Tufail, A. (2016). Use of Mechanical Turk as a MapReduce framework for macular OCT segmentation. *Journal of ophthalmology* 2016 (2016).

Lee, A. Y and Tufail, A. (2014). Mechanical Turk based system for macular OCT segmentation. *Investigative Ophthalmology & Visual Science* 55, 13 (2014), 4787–4787.

Leifman, G, Swedish, T, Roesch, K, and Raskar, R. (2015). Leveraging the crowd for annotation of retinal images. In *International Conference of the Engineering in Medicine and Biology Society (EMBC)*. IEEE, 7736–7739.

Lejeune, L, Christoudias, M, and Sznitman, R. (2017). Expected exponential loss for gaze-based video and volume ground truth annotation. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 106–115.

Lintott, C. J, Schawinski, K, Slosar, A, Land, K, Bamford, S, Thomas, D, Raddick, M. J, Nichol, R. C, Szalay, A, Andreescu, D, Murray, P, Vandenberg, J, Murray, P, Berg, J. v. d, Sullivan, B, Wood, C, Iliff, M, Bonney, R, Fink, D, Kelling, S, Huss, J, Orozco, C, Goodale, J, Wu, C, Batalov, S, Vickers, T, Valafar, F, Su, A, Khatib, F, Cooper, S, Tyka, M, Xu, K, Makedon, I, Popovic, Z, Baker, D, Players, F, McGonigal, J, Cooper, S, Khatib, F, Treuille, A, Barbero, J, Lee, J, Beenen, M, Leaver-Fay, A, Baker, D, Popovic, Z, Players, F, Khatib, F, Dimaio, F, Cooper, S, Kazmierczyk, M, Gilski, M, Krzywda, S, Zabranska, H, Pichova, I, Thompson, J, Popovic, Z, Jaskolski, M, Baker, D, Ahn, L, Dabbish, L, Ahn, L, and Dabbish, L. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey ? *Monthly Notices of the Royal Astronomical Society* 389, 3 (Sept. 2008), 1179–1189. DOI:http://dx.doi.org/10.1111/j.1365-2966.2008.13689.x

Litjens, G, Kooi, T, Bejnordi, B. E, Setio, A. A. A, Ciompi, F, Ghafoorian, M, van der Laak, J. A, Van Ginneken, B, and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

Luengo-Oroz, M. A, Arranz, A, and Frean, J. (2012). Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of medical Internet research* 14, 6 (2012).

Maier-Hein, L, Kondermann, D, Roß, T, Mersmann, S, Heim, E, Bodenstedt, S, Kenngott, H. G, Sanchez, A, Wagner, M, Preukschas, A, and others, . (2015). Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences.

International Journal of Computer Assisted Radiology and Surgery 10, 8 (2015), 1201–1212.

Maier-Hein, L, Mersmann, S, Kondermann, D, and others, . (2014)b. Crowdsourcing for reference correspondence generation in endoscopic images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 349–356.

Maier-Hein, L, Mersmann, S, Kondermann, D, Bodenstedt, S, Sanchez, A, Stock, C, Kenngott, H. G, Eisenmann, M, and Speidel, S. (2014)a. Can Masses of Non-Experts Train Highly Accurate Image Classifiers? In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 438–445.

Maier-Hein, L, Ross, T, Gröhl, J, Glocker, B, Bodenstedt, S, Stock, C, Heim, E, Götz, M, Wirkert, S, Kenngott, H, and others, . (2016). Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 616–623.

Malpani, A, Vedula, S. S, Chen, C. C. G, and Hager, G. D. (2015). A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *International journal of computer assisted radiology and surgery* 10, 9 (Sept. 2015), 1435–47.

Mavandadi, S, Dimitrov, S, Feng, S, Yu, F, Sikora, U, Yaglidere, O, Padmanabhan, S, Nielsen, K, and Ozcan, A. (2012). Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. *PLoS ONE* 7, 5 (2012).

McKenna, M. T, Wang, S, Nguyen, T. B, Burns, J. E, Petrick, N, and Summers, R. M. (2012). Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence. *Medical image analysis* 16, 6 (2012), 1280–1292.

Mitry, D, Peto, T, Hayat, S, Blows, P, Morgan, J, Khaw, K.-T, and Foster, P. J. (2015). Crowdsourcing as a Screening Tool to Detect Clinical Features of Glaucomatous Optic Neuropathy from Digital Photography. *PLoS ONE* 10, 2 (2015), 1–8.

Mitry, D, Peto, T, Hayat, S, Morgan, J. E, Khaw, K.-T, and Foster, P. J. (2013). Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of Images in the EPIC Norfolk Cohort on behalf of the UKBiobank Eye and Vision Consortium. *PLoS ONE* 8, 8 (2013), e71154.

Mitry, D, Zutis, K, Dhillon, B, Peto, T, Hayat, S, Khaw, K.-T, Morgan, J. E, Moncur, W, Trucco, E, and Foster, P. J. (2016). The accuracy and reliability of crowdsource annotations of digital retinal images. *Translational vision science & technology* 5, 5 (2016), 6–6.

Nguyen, T. B, Wang, S, Anugu, V, Rose, N, McKenna, M, Petrick, N, Burns, J. E, and Summers, R. M. (2012). Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology* 262, 3 (2012), 824–833.

O'Neil, A. Q, Murchison, J. T, van Beek, E. J, and Goatman, K. A. (2017). Crowdsourcing Labels for Pathological Patterns in CT Lung Scans: Can Non-experts Contribute Expert-Quality Ground Truth? In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 96–105.

Ørting, S. N, Cheplygina, V, Petersen, J, Thomsen, L. H, Wille, M. M. W, and de Bruijne, M. (2017). Crowdsourced emphysema assessment. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*. Springer, 126–135.

Park, J. H, Mirhosseini, S, Nadeem, S, Marino, J, Kaufman, A, Baker, K, and Barish, M. (2017). Crowdsourcing for identification of polyp-free segments in virtual colonoscopy videos. In *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, Vol. 10138. International Society for Optics and Photonics, 101380V.

Park, J. H, Nadeem, S, Marino, J, Baker, K, Barish, M, and Kaufman, A. (2018). Crowd-assisted polyp annotation of virtual colonoscopy videos. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, Vol. 10579. International Society for Optics and Photonics, 105790M.

Park, J. H, Nadeem, S, Mirhosseini, S, and Kaufman, A. (2016). C^2A: Crowd consensus analytics for virtual colonoscopy. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 21–30.

Rajchl, M, Koch, L. M, Ledig, C, Passerat-Palmbach, J, Misawa, K, Mori, K, and Rueckert, D. (2017). Employing Weak Annotations for Medical Image Analysis Problems. *arXiv preprint arXiv:1708.06297* (2017).

Rajchl, M, Lee, M. C, Schrans, F, Davidson, A, Passerat-Palmbach, J, Tarroni, G, Alansary, A, Oktay, O, Kainz, B, and Rueckert, D. (2016). Learning under Distributed Weak Supervision. *arXiv preprint arXiv:1606.01100* (2016).

Ranard, B. L, Ha, Y. P, Meisel, Z. F, Asch, D. A, Hill, S. S, Becker, L. B, Seymour, A. K, and Merchant, R. M. (2014). Crowdsourcing: harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine* 29, 1 (Jan. 2014), 187–203.

Roethlingshoefer, V, Bittel, S, Kenngott, H, Wagner, M, Bodenstedt, S, Ross, T, Speidel, S, and L, M.-H. (2017). How to Create the Largest In-Vivo Endoscopic Dataset. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*.

Sameki, M, Gurari, D, and Betke, M. (2016). ICORD: Intelligent Collection of Redundant Data ? A Dynamic System for Crowdsourcing

Cell Segmentations Accurately and Efficiently. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1380–1389.

Sharma, M, Saha, O, Sriraman, A, Hebbalaguppe, R, Vig, L, and Karande, S. (2017). Crowdsourcing for chromosome segmentation and deep classification. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 786–793.

Smittenaar, P, Walker, A. K, McGill, S, Kartsonaki, C, Robinson-Vyas, R. J, McQuillan, J. P, Christie, S, Harris, L, Lawson, J, Henderson, E, and others, . (2018). Harnessing citizen science through mobile phone technology to screen for immunohistochemical biomarkers in bladder cancer. *British journal of cancer* 119, 2 (2018), 220.

Sonabend, A. M, Zacharia, B. E, Cloney, M. B, Sonabend, A, Showers, C, Ebiana, V, Nazarian, M, Swanson, K. R, Baldock, A, Brem, H, and others, . (2017). Defining glioblastoma resectability through the wisdom of the crowd: a proof-of-principle study. *Neurosurgery* 80, 4 (2017), 590–601.

Sullivan, D. P, Winsnes, C. F, Åkesson, L, Hjelmare, M, Wiking, M, Schutten, R, Campbell, L, Leifsson, H, Rhodes, S, Nordgren, A, and others, . (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology* 36, 9 (2018), 820.

Timmermans, B, Szlávik, Z, and Sips, R.-J. (2016). Crowdsourcing ground truth data for analysing brainstem tumors in children. In *Belgium Netherlands Artificial Intelligence Conference (BNAIC)*.

Wazny, K. (2017). Crowdsourcing ten years in: A review. *Journal of global health* 7, 2 (2017).

Willi, M, Pitman, R. T, Cardoso, A. W, Locke, C, Swanson, A, Boyer, A, Veldthuis, M, and Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution* 10, 1 (2019), 80–91. DOI: http://dx.doi.org/10.1111/2041-210X.13099

APPENDIX - DATA AND OVERVIEW OF REVIEWED PAPERS

***Table 9.** Summary of the surveyed papers. A question mark "?" indicates the information was not found in the paper. A longer summary and the short summary presented in this table are available for download via Figshare, https://doi.org/10.6084/m9.figshare.9751850.v1.*

	Section 2		Section 3	Section 4					Section 5				
Paper	Task	Domain	Interaction	I	#I	#Ann/I	Platform	Reward	Filtering	Aggregation	Comparison	Gold standard
(Albarqouni et al., 2016a)	classify	histo	rate	M	550	10	custom	unknown	before/during	majority/weighted	indirect	multiple experts
(Albarqouni et al., 2016b)	other	histo	rate	L	10000	3	custom	volunteers	none	weighted	direct	multiple experts
(Roethlingshoefer et al., 2017)	segment	abdomen	draw	L	10000	?	none	none	none	none	na	na
(Boorboor et al., 2018)	segment	lung	draw	S	80	10	paid	low	during	none	direct	multiple experts
(Brady et al., 2014)	classify	eye	rate	S	16/4	10/500	paid	unknown	none	none	direct	?
(Brady et al., 2017)	classify	eye	rate	L	1200	10	paid	low	none	majority/weighted	direct	?
(Bruggemann et al., 2018)	segment	histo	click	M/L	190/3446	10/5	paid	low	none	none	direct	multiple experts
(Cabrera-Bean et al., 2017)	segment	histo	click	XS	2	5000	volunteer	volunteers	none	other	direct	?
(Chávez-Aragón et al., 2013)	segment	other	draw	M	274	≈10	custom	volunteers	after	none	direct	other
(Cheplygina et al., 2016)	segment	lung	draw	S	90	10	paid	unknown	after	average	direct	one expert
(Cheplygina and Pluim, 2018)	classify	other	rate	S	100	6	students	volunteers	none	average	indirect	?
(Della Mea et al., 2014)	s+c	histo	click	S	13	10	paid	unknown	during	average	direct	one expert
(dos Reis et al., 2015)	classify	histo	rate	L	180172	?	volunteer	volunteers	during	average	direct	one expert
(Eickhoff, 2014)	classify	histo	rate	M	569	≈21	paid	low	none	majority	direct	one expert
(Foncubierta Rodríguez and Müller, 2012)	classify	multiple	rate	L	300000	?	paid	unknown	during	none	direct	one expert
(Ganz et al., 2017)	segment	brain	draw	S	100	10	paid	low	none	average	direct	one expert
(Gur et al., 2017)	classify	heart	rate	M	900	?	custom	unknown	none	none	indirect	multiple experts
(Gurari et al., 2015b)	segment	other	draw	M	235	7	paid	low	before/during	majority/weighted	direct	multiple experts
(Gurari et al., 2016)	s+c	multiple	rate+draw	M	405/370	5	paid	low	before/after	majority	direct	multiple experts
(Heim, 2018)	s+c	abdomen	rate+draw	M	364	10	paid	low	before	majority/weighted	direct	multiple experts
(Heller et al., 2017)	segment	abdomen	draw	XS	1	?	custom	unknown	none	none	direct	other
(de Herrera et al., 2014)	classify	multiple	rate	L	17002	?	paid	volunteers	none	none	indirect	other
(Holst et al., 2015)	classify	other	rate	S	12	50	paid	low	before	other	direct	multiple experts
(Huang and Hamarneh, 2017)	classify	lung	click	?	?	?	custom	unknown	none	other	direct	?
(Irshad et al., 2015)	segment	histo	click+draw	M	810/455	1/1,2,3	paid	unknown	before/during	none	direct	multiple experts
(Irshad et al., 2017)	s+c	histo	rate+draw	M/L	380/5338	10/3	paid	unknown	before/during	majority/weighted	direct	one expert
(Keshavan et al., 2018)	classify	brain	rate	M	722	?	custom	volunteers	during	weighted	direct	multiple experts
(Lawson et al., 2017)	classify	histo	click	S	10	≈26	volunteer	hourly	none	none	direct	multiple experts
(Lee and Tufail, 2014)	segment	eye	draw	S	18	2	paid	low	none	none	direct	na
(Lee et al., 2016)	segment	eye	draw	S	61	≈2,3	paid	low	before	none	direct	na
(Leifman et al., 2015)	s+c	eye	rate+draw+click	?	?	?	custom	volunteers	during	other	direct	multiple experts
(Lejeune et al., 2017)	segment	multiple	click	S	73	?	experts	unknown	none	average	indirect	?
(Luengo-Oroz et al., 2012)	segment	histo	click	S	28	≈200	volunteer	volunteers	during	other	direct	multiple experts
(Maier-Hein et al., 2014a)	segment	abdomen	draw	M	120	10	paid	unknown	none	none	indirect	one expert
(Maier-Hein et al., 2014b)	other	abdomen	click+compare	S	100	≈7	paid	unknown	none	other	direct	multiple experts
(Maier-Hein et al., 2015)	other	abdomen	click+compare	S	100	≈400	paid	low	none	other	direct	?
(Maier-Hein et al., 2016)	segment	abdomen	click+compare	M	300	?	paid	unknown	none	majority	indirect	?
(Malpani et al., 2015)	classify	other	rate+compare	M	360	?	?	hourly	during	majority/weighted	direct	multiple experts
(Mavandadi et al., 2012)	classify	histo	rate	L	9394	?	none	unknown	before/during	none	direct	one expert
(McKenna et al., 2012)	classify	abdomen	rate	M	600	≈20	paid	low	before	other	direct	?
(Mitry et al., 2013)	classify	eye	rate	S	100	≈20	paid	low	none	none	direct	multiple experts
(Mitry et al., 2015)	classify	eye	rate	M	127	≈20	paid	low	before/after	none	direct	multiple experts
(Mitry et al., 2016)	s+c	eye	rate+draw	S	84	≈64	paid	low	before/after	majority	direct	multiple experts
(Nguyen et al., 2012)	classify	abdomen	rate	M	268	20	paid	low	before	majority	direct	?
(O'Neil et al., 2017)	segment	lung	draw	S	20	34	custom	volunteers	after	majority	direct	one expert
(Ørting et al., 2017)	other	lung	compare	M	360	≈10	paid	low	before	other	indirect	multiple experts
(Park et al., 2016)	classify	abdomen	rate	M	299	≈18	paid	low	after	majority/weighted	direct	one expert
(Park et al., 2017)	classify	abdomen	rate	M	163	≈15	paid	unknown	none	majority	direct	one expert
(Park et al., 2018)	segment	abdomen	click	M	562	?	paid	unknown	before/during	none	direct	one expert
(Rajchl et al., 2016)	segment	brain	click	L	5000	1	custom	volunteers	none	none	indirect	one expert
(Rajchl et al., 2017)	segment	abdomen	?	M	150	1	none	none	none	none	na	na
(Sameki et al., 2016)	segment	histo	draw	M	270	5	paid	low	none	other	direct	?
(Sharma et al., 2017)	segment	histo	draw	S	50	?	paid	low	none	other	direct	na
(Smittenaar et al., 2018)	classify	histo	rate	L	≈8300	[5,25]	custom	volunteers	none	weighted	direct	multiple experts
(Sonabend et al., 2017)	classify	brain	rate	S	20	?	experts	unknown	none	none	direct	na
(Sullivan et al., 2018)	classify	histo	rate+draw	L	65596	?	custom	volunteers	none	other	direct	?
(Timmermans et al., 2016)	s+c	brain	rate+draw	L	1171	≈3	custom	volunteers	none	none	na	na