# Quality Control Challenges in Crowdsourcing Medical Labeling

Miles Hutson*
hutson@google.com
Google Health
Palo Alto, California

Olga Kanzheleva*
okanzheleva@google.com
Google Health
Palo Alto, California

Caitlin Taggart*
ctaggart@google.com
Google Health
Palo Alto, California

Bilson J. L. Campana
bilson@google.com
Google Health
Palo Alto, California

Quang Duong
qduong@google.com
Google Health
Palo Alto, California

## ABSTRACT

Crowdsourcing has enabled the collection, aggregation and refinement of human knowledge and judgment, i.e. ground truth, for problem domains with data of increasing complexity and scale. This scale of ground truth data generation, especially towards the development of machine learning based medical applications that require large volumes of consistent diagnoses, poses significant and unique challenges to quality control. Poor quality control in crowdsourced labeling of medical data can result in undesired effects on patients' health. In this paper, we study medicine-specific quality control problems, including the diversity of grader expertise and diagnosis guidelines' ambiguity in novel datasets of three eye diseases. We present analytical findings on physicians' work patterns, evaluate existing quality control methods that rely on task completion time to circumvent the scarcity and cost problems of generating ground truth medical data, and share our experiences with a real-world system that collects medical labels at scale.

## CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Collaborative and social computing systems and tools**; • **Applied computing** → **Health care information systems**; *Life and medical sciences.*

## KEYWORDS

crowdsourcing, ground truth, data labeling, medical, healthcare, diagnosis, ophthalmology

**ACM Reference Format:**
Miles Hutson, Olga Kanzheleva, Caitlin Taggart, Bilson J. L. Campana, and Quang Duong. 2019. Quality Control Challenges in Crowdsourcing Medical Labeling. In *Proceedings of Data Collection, Curation, and Labeling for Mining and Learning Workshop at KDD '19 (KDD Workshop '19).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Equal contribution. Listed alphabetically.


*KDD Workshop '19, August 2019, Anchorage, AK, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The process of generating ground truth data, or *data labeling*, through human annotation is a major bottleneck in the development and deployment of machine learning applications. *Crowdsourcing* has become an important labor tool to address the issues of large data volume and work-human annotator matching [9]. *Quality control* in a crowdsourcing platform seeks to evaluate the quality of answers in order to reward or correct graders accordingly to ensure that graders generate high-quality, valid ground truth data. Time and financial costs for medical diagnoses are high, which restricts frequent performance re-evaluation. Graders may not meet requesters' desired quality bar because they lack the expertise to complete the task, attempt to game the system, or have misaligned motivations [2].

*Golden datasets* with known labels are often employed to control for quality [2]. Each label in a golden dataset is often generated by aggregating opinions from a panel of graders on the same input case. The final label can be the result of a simple majority vote among all graders' labels or multiple rounds of voting until a consensus is reached [2, 17]. These methods often incur high cost as they entail having the same input data labeled by multiple graders, sometimes repeatedly, or employing scarce specialist graders.

Alternative methods of identifying low quality graders often involve analyzing task-completion times. Cheng et al. [1] suggested the existence of a minimum time amount needed to complete a labeling task, although computing such time can be difficult. Once that minimum task-completion time is found, a grader completing a task in less time is a signal indicating low quality. However, as this minimum time estimation is an average for all graders, it cannot be directly applied to tasks with the intent of utilizing graders with varied levels of education and experience [1]. As medical graders are scarce for complex or difficult diagnoses, allowing grader diversity in skill, time availability, and compensation is critical for medical labeling at scale.

This study analyzes three novel datasets of eye diseases, produced from a real-world large-scale medical labeling tool, and presents the problems and measurements of quality control associated to them and other works. In section 2, we summarize our data. Next, in section 3 we present a qualitative and quantitative overview of factors that render medical data labeling and quality control uniquely challenging. In section 4, we examine alternative quality control methods in the absence of golden data and provide analyses that rely on task completion times and its derivatives for

identifying low performers. Finally, in section 5 we discuss our observations and suggest future works and improvements for medical labeling tools.

## 2 DATA OVERVIEW

We focus on the applications of machine learning in assisting physicians to diagnose common eye conditions, though our analyses and observations for quality control should be applicable in other medical labeling problems outside of ophthalmology. We examine four separate labeling efforts designed for three eye conditions. Since this study's primary focus is the performance and consistency of graders, and these datasets pertain to ongoing and developing research, the eye conditions are left unnamed. Table 1 provides a brief summary.

| *Dataset* | *Classes* | *Images* | *Graders* | *Replication* | *Total Grades* |
|---|---|---|---|---|---|
| $A_0$ | 9 | 52,238 | 38 | 1 | 52,238 |
| $A_1$ | | 1,519 | 11 | 3 | 4,517 |
| $B$ | 4 | 891 | 9 | 9 | 6,749 |
| $C$ | 6 | 994 | 9 | 3 | 2,947 |

**Table 1: Summary of unnamed ophthalmology datasets. Dataset $A_0$ and $A_1$ are labeling tasks of the same eye disease.**

*Replication* indicates the number of ophthalmologists grading each image and classes is the number of diagnostic outcomes of interest for the dataset's disease. The total number of grades in replicated datasets is not simply $images \cdot replication$ due the changing direction of each dataset's associated research.

## 3 CHALLENGES IN HIGH-QUALITY IMAGE-BASED DIAGNOSES AT SCALE

With large amounts of highly consistent labeled data, deep learning [12] has recently been applied in developing very accurate classification applications in medical imaging fields, including breast cancer lymph node metastasis detection [14], pneumonia [15], and diabetic retinopathy (DR) [7, 8]. Labeling data at this large scale requires employing medical experts or professionals, often as graders on crowdsourcing platforms.

Low agreement rate among physicians has long been observed in all medical imaging fields, from grading DR [6], analyzing breast cancer specimens [4], and interpreting mammograms [5]. There have been various methods developed to improve diagnostic agreement and consistency. Physicians' interpretation variability can be narrowed by establishing reference standards or guidelines that include a rich corpus of example images [16]. Adjudication is a protocol that involves discussions among physicians examining the same patient case until a diagnosis consensus is reached. Adjudication has been shown to be valuable in the evaluation of these deep learning models [11]. However, the high cost of these approaches, both monetary and time-wise, hinders the creation of large golden datasets that form the foundation of many quality control solutions in crowdsourcing.

Varied grader practices and skill in interpreting images and disease grading guidelines produce disagreement that is well documented across many medical domains [13, 18]. These grading practices continue to diverge between images in the same labeling task, and even more so across medical domains. Even after adjusting for differences in graders' software usage in our datasets, we still observed different types of variability inherent to the tasks of medical data labeling and diagnosis.

### 3.1 Variability in Diagnostic Labels

Within ophthalmology, many factors contribute to diagnostic label variability, e.g. graders operating in clinics or reading centers [18] and diagnosing from digital or film-based images [13]. Analysis on data replicated to multiple graders shows high levels of disagreement in out datasets $A_1$, $B$, and $C$ at 64%, 27%, and 32%, respectively, of images with any disagreement. Table 2 shows the disagreement matrix of individual labels against the majority vote for each image of dataset $B$.

| | **Individual Labels** | | | |
|---|---|---|---|---|
| | *Ungradeable* | *Absent* | *Possible* | *Present* |
| Majority Label | | | | |
| Ungradeable | 144 | 52 | 10 | 1 |
| Absent | 94 | 5869 | 244 | 30 |
| Possible | 7 | 81 | 147 | 25 |
| Present | 1 | 7 | 10 | 27 |

**Table 2: Disagreement matrix for the dataset $B$, showing all grader's labels against the majority vote. The categories reflect a grading scale for the unspecified eye disease evaluated on retina images by multiple graders.**

The categories of grades correspond to the three severity levels of the disease, and graders may consider an image ungradeable if image quality issues prevent diagnosing severity, e.g. due to image blur or capturing an incorrect region. In the 15 cases which the majority vote indicated the presence of the disease, close to half had grades that indicated the disease was not present. These disagreements may be caused by the difficulty of the task, which motivates improved grading guidelines, or the performance of the grader. This ambiguity of grade quality in large disparity disagreements generalizes to all other datasets.

### 3.2 Variability in Grading Time

While different graders are expected to spend varying amounts of time on their assigned tasks, such variability is particularly pronounced in medical labeling. In particular, the very long-tailed distributions of task-completion time across all graders for dataset $A_0$, depicted as the black line in Figure 1, and other datasets, illustrate large inherent differences among tasks of the same labeling task. Moreover, different graders exhibit different work patterns: figure 1 highlights how seven graders, with the most grades, in dataset $A_0$ differ in time spent on a majority of their tasks. As graders are assigned tasks uniformly randomly, tasks of similar difficulty levels demonstrably took varying amounts of time from these seven graders.

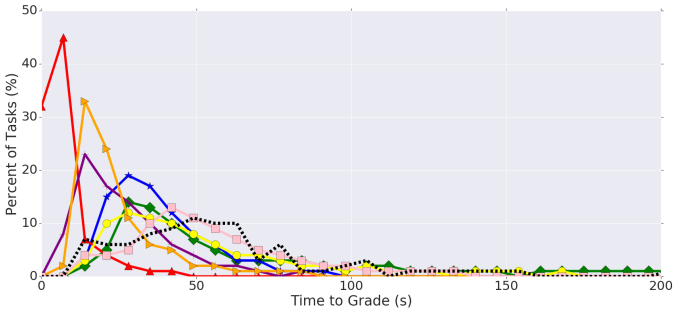


**Figure 1: Task-completion times of the seven graders with the most grades in the dataset $A_0$, along with average times in black-dashed.**

## 4 ASSESSING GRADER PERFORMANCE IN THE ABSENCE OF GOLDEN DATA

The aforementioned variability in grading practices, diagnostic outcomes and labeling efforts poses serious additional challenges to the creation of golden datasets. In the absence or scarcity of expensive golden data, medical labeling requesters often gravitate towards time-based quality control methods.

### 4.1 Time-based Assessment

A potential indicator of graders' performance is their grading speed. Assuming a majority of skilled graders have similar grading pace, grading speed outliers may be a signal of a low or high quality grader. To demonstrate this variety in grading speed we computed for each grader their median grading time per *session*, during which any two consecutive tasks are less than twenty minutes apart. Figure 2 displays median grading times over the progression of labeling dataset $A_0$ for the seven graders that labeled the most images in dataset $A_0$.

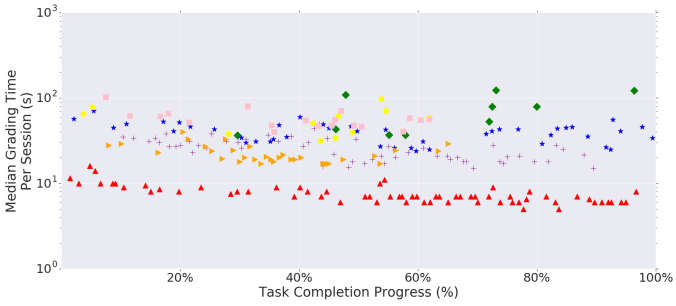


**Figure 2: Median grading time per session over the progression of labeling dataset $A_0$. The red-triangle series captures a known low quality grader.**

In dataset $A_0$, graders had been independently evaluated by interleaving images with known *golden* labels. The grader plotted in red triangle, grader $G_r$, was found to be low quality as they failed to produce a sufficient number of correct labels on these golden images. Noticeably in Figure 2, $G_r$ is on average 6 times faster than the second fastest grader. We also examine median grading

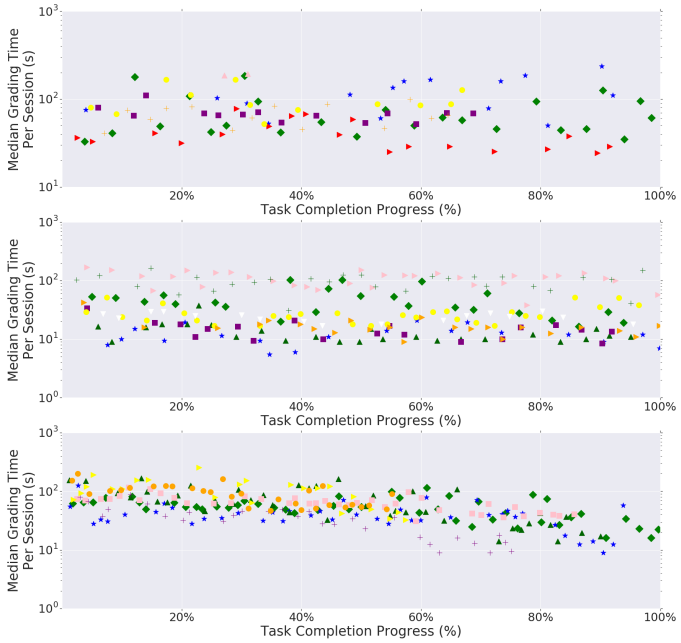


**Figure 3: Median grading time per session over the progression of labeling datasets *top)* $A_1$, *middle)* $B$ and *bottom)* $C$.**

times per session in datasets $A_1$, $B$ and $C$ in figure 3. Unlike dataset $A_0$, there was no golden label dataset available, since labeling cost and complexity were prohibitive for interleaving golden images into these datasets. As a result, additional signals are needed to retroactively assess the quality of graders as well as the effectiveness of grading time as an indicator for grader quality.

### 4.2 Estimates of Grader Quality

All of the aforementioned datasets that lack golden data, $A_1$, $B$ and $C$, have multiple graders labeling each image. This enables the usage of expectation maximization (EM) [3] to estimate graders' error rates, or their confusion matrices, which form our baseline measure for grader quality. Each grader's confusion matrix value $\pi^k_{i,j}$ gives the probability that grader $k$, when presented with a task of true diagnostic label $i$, will label it $j$. The EM algorithm is run on all images on datasets with replication $A_1$, $B$, and $C$. The resulting confusion matrices, $\pi$, form the foundation for computing *soft cost* to capture grader quality [10].

Given that grader $k$ assigns label $j$ to an image, the corresponding *soft label* vector $\mathbf{soft}^k(j)$ (i.e. posterior estimate), derived from $\pi^k_{*,j}$, is the best possible probability estimate for the true class of the image over all possible classes. A grader's soft label vector allows one to account for the grader's systemic biases as well as true label priors [10].

Given a grader's soft label vector, Ipeirotis et al. [10] provides us with the computation of a grader's expected soft cost $\mathbf{Cost}(\{\mathbf{soft}^k(j)\}_j)$ given their soft label vectors over all classes, weighted by the probability of grader $k$ assigning label $j$. For generalizability, we assume the cost of each classification error to be uniform.

Figure 4 shows example soft cost matrices for graders $G_0$ and $G_1$ in $B$. Grader $G_0$ labeled questions 'Possible' 4.9% of the time. Of those, 64% of the time the label should be switched to 'Absent', and 2.5% of the time the label should be switched to 'Ungradable', and 33% of the time the label should stay 'Possible', as demonstrated in our soft cost matrix.

Grader $G_1$ achieved perfect precision when they label 'Present,' but only labeled 'Present' 0.1% of the time. The soft cost matrix suggests that $G_1$ should have been less conservative in their labeling: 19% of their 'Possible' labels should be switched to 'Present', and 56% of their 'Ungradable' labels were assigned to images whose image quality are sufficiently good for other graders to label.

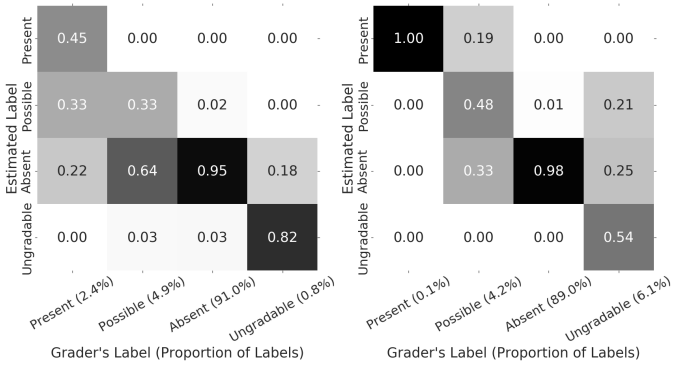


**Figure 4: Example soft cost matrices for graders *left)* $G_0$ and *right)* $G_1$ in $B$. A grader's distribution of assigned labels is added to the column labels.**

## 4.3 Evaluating Time-Based Methods Against Estimated Grader Quality

Given our hypothesized relationship between a grader's grading time and quality, we tested the relationship between statistics of grading time distribution and their soft cost. First, we combined all replicated datasets ($A_1$, $B$, $C$) by linearly normalizing the individual grading times in each dataset to a scale between 0 and 1. We then checked the p-value of correlation between their soft cost and each of the mean, standard deviation, and median of graders' times. Correlations were statistically insignificant (at level 0.05) with p values of 0.53 (mean), 0.13 (standard deviation), and 0.87 (median).

Next, we performed linear regression (with a bias term) on the combined datasets. The features for said regression were as above: mean, standard deviation, and median. The correlation between predicted soft cost values produced by these regressors and actual soft cost values was statistically insignificant (p-value of 0.11).

Finally, we performed linear regression (with a bias term) on the individual normalized datasets. The features for regression were the same. P-values corresponding to the correlations between predicted soft cost values and actual soft cost values are 0.02 ($A_1$), 0.16 ($B$), and 0.11 ($C$). Figures 5 depicts the relationships between grading time's median and standard deviation (x and y axes) and soft cost (color shade).

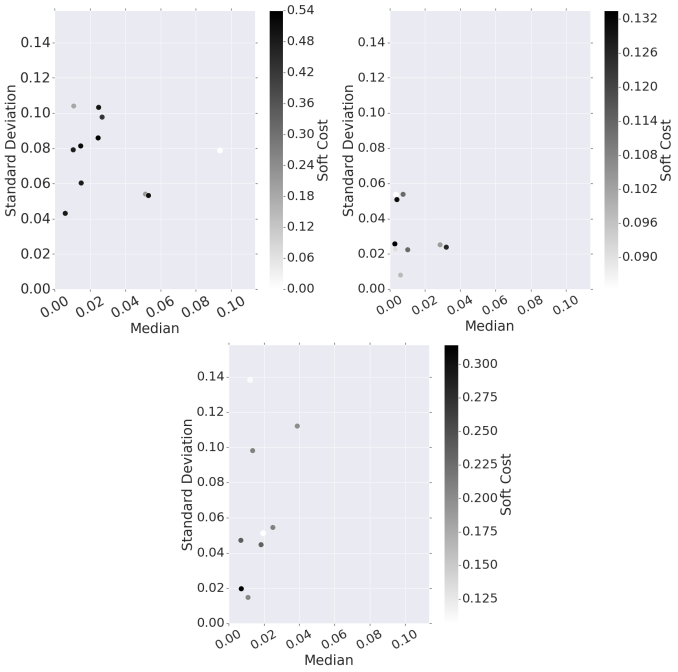


**Figure 5: A comparison of grading time's statistics and soft cost for *top-left)* $A_1$, *top-right)* $B$, and *bottom)* $C$.**

## 5 DISCUSSIONS

Grading time appears to be a useful signal in detecting low-quality graders in both datasets $A_0$ and $A_1$ from the same eye condition: a low-quality grader $G_r$ was identified due to their unusually high speed in $A_0$ (figure 2), and grader soft cost is found to be inversely correlated with grading time's statistics in $A_1$ (figure 5). Note that $G_r$ was only present in $A_0$but not $A_1$. Outside of the aforementioned condition, there is insufficient evidence to confirm the role of grading time's statistics in determining grader quality, as described in Section 4.3. A number of factors other than grader quality may influence grading time. For example, higher quality images may need less time to interpret, resulting in short grading time. An optometrist likely has less experience in grading when compared to a retina specialist, but only in certain cases does that difference translate to additional grading time.

While the EM-generated soft cost provides us with an estimate of grader quality, the approach's replication requirement limits it to only datasets with replication. Even in such datasets with replication, computing grading time's statistics demands much less data than running EM. As a result, one may be able to compute and use grading time for quality estimation much earlier in the lifetime of a labeling effort.

Comparing variations in grading time across tasks is a promising first step towards identifying useful proxy metrics for label quality. More data with golden labels will be needed to further explore methods that solve this difficult problem. Furthermore, additional types of data, such as grader interactions with the medical labeling tool, may provide higher granularity and fidelity signals per session to assess quality.

## REFERENCES


[1] Justin Cheng, Jaime Teevan, and Michael S. Bernstein. 2015. Measuring Crowdsourcing Effort with Error-Time Curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1365 – 1374. https://doi.org/10.1145/2702123.2702145

[2] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1, Article 7 (Jan. 2018), 40 pages. https://doi.org/10.1145/3148148

[3] A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 20–28. http://www.jstor.org/stable/2346806

[4] Joann G. Elmore, Gary M. Longton, Patricia A. Carney, Berta M. Geller, Tracy Onega, Anna N. A. Tosteson, Heidi D. Nelson, Margaret S. Pepe, Kimberly H. Allison, Stuart J. Schnitt, Frances P. OâĂŹMalley, and Donald L. Weaver. 2015. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA* 313, 11 (03 2015), 1122–1132. https://doi.org/10.1001/jama.2015.1405 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2203798/joi150021.pdf

[5] Joann G. Elmore, Carolyn K. Wells, Carol H. Lee, Debra H. Howard, and Alvan R. Feinstein. 1994. Variability in Radiologists' Interpretations of Mammograms. *New England Journal of Medicine* 331, 22 (1994), 1493–1499. https://doi.org/10.1056/NEJM199412013312206 arXiv:https://doi.org/10.1056/NEJM199412013312206

[6] Alan D Fleming, Keith A Goatman, Sam Philip, Gordon J Prescott, Peter F Sharp, and John A Olson. 2010. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *British Journal of Ophthalmology* 94, 12 (2010), 1606–1610. https://doi.org/10.1136/bjo.2009.176784 arXiv:https://bjo.bmj.com/content/94/12/1606.full.pdf

[7] Rishab Gargeya and Theodore Leng. 2017. Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology* 124 7 (2017), 962–969.

[8] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316, 22 (12 2016), 2402–2410. https://doi.org/10.1001/jama.2016.17216 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2588763/joi160132.pdf

[9] Panagiotis G. Ipeirotis. 2010. Analyzing the Amazon Mechanical Turk Marketplace. *XRDS* 17, 2 (Dec. 2010), 16–21. https://doi.org/10.1145/1869086.1869094

[10] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 64 – 67. https://doi.org/10.1145/1837885.1837906

[11] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* 125, 8 (2018), 1264 – 1272. https://doi.org/10.1016/j.ophtha.2018.01.034

[12] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444. https://doi.org/10.1038/nature14539

[13] Helen K Li, Larry D Hubbard, Ronald P Danis, Adol Esquivel, Jose F Florez-Arango, Nicola J Ferrier, and Elizabeth A Krupinski. 2010. Digital versus film fundus photography for research grading of diabetic retinopathy severity. *Investigative ophthalmology & visual science* 51, 11 (2010), 5846–5852.

[14] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Gregory S. Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. 2017. Detecting Cancer Metastases on Gigapixel Pathology Images. *CoRR* abs/1703.02442 (2017). arXiv:1703.02442 http://arxiv.org/abs/1703.02442

[15] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *CoRR* abs/1711.05225 (2017). arXiv:1711.05225 http://arxiv.org/abs/1711.05225

[16] Michael C. Ryan, Susan Ostmo, Karyn Jonas, Audina Berrocal, Kimberly Drenser, Jason Horowitz, Thomas C. Lee, Charles Simmons, Maria Ana Martinez-Castellanos, R. V Paul Chan, and Michael Chiang. 2014. Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2014 (2014), 1902–1910.

[17] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. crowdEEG: A Platform for Structured Consensus Formation in Medical Time Series Analysis. In *Symposium: Workgroup on Interactive Systems in Health at the the 2019 CHI Conference on Human Factors (CHI '19)*.

[18] Ingrid U Scott, Neil M Bressler, Susan B Bressler, David J Browning, Clement K Chan, Ronald P Danis, Matthew D Davis, Craig Kollman, Haijing Qin, Diabetic Retinopathy Clinical Research Network Study Group, et al. 2008. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal bevacizumab for diabetic macular edema. *Retina (Philadelphia, Pa.)* 28, 1 (2008), 36.