Accelerating Medical Labeling for Diagnostic AI: Add Non-Clinicians to Labeling Teams

Healthcare organizations can create a large repository of high-quality labeled medical images at speed and at scale for training AI algorithms by following best practices and by leveraging a data-labeling network of clinicians and non-clinicians.

Executive Summary

Artificial intelligence (AI)-driven diagnostic models based on medical images from X-rays, MRIs, and CT scans promise to improve outcomes and reduce costs with earlier detection and diagnosis. Their promise is such that AI-driven imaging and diagnostic start-ups were the largest group of all AI and machine learning (ML) start-ups in healthcare in 2019.[1] These AI/ML models learn how to diagnose by reviewing and recognizing recurring patterns associated with a vast amount of diverse and accurately labeled medical images.

Achieving this large set of labeled medical images is challenging to organizations for a variety of reasons. To comply with privacy regulations, either patients must

consent to their data being used for research purposes or organizations must de-identify the data. Data set creators must also obtain data from a wide range of patient populations so the data set is diverse and accurately reflects representative disease states and outcomes. Finally, many healthcare organizations assume that all medical images for data sets must be reviewed and labeled by specially trained physicians. This expensive and inefficient route to labeling data sets contributes to the current shortage of AI medical training data available to organizations today.

Many images may be processed by well-trained, non-clinician labelers accompanied by physician oversight. By using a combination of both clinically trained and non-clinician labelers, healthcare organizations can more quickly and more inexpensively create comprehensive, accurately labeled medical image data sets and hasten the use of AI for detection and diagnosis support. The keys to successfully using a diverse mix of labelers include understanding the problem that the AI model is meant to solve, evaluating the images against pre-defined requirements, and following a set of best practices.

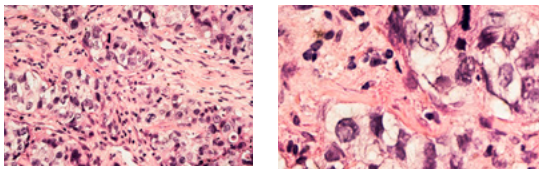
Diverse labelers for diverse data

AI/ML diagnostic models must learn from highly accurate training data sets containing a diverse set of diseases and outcomes. Faults in learning data sets may lead to faulty AI diagnostic models. Healthcare organizations often have tapped medical experts to review and label data to guard against inaccurate labels and ensure that a data set is meaningful. However, not all images need that level of expertise to be properly labeled. To determine what expertise is required, healthcare organizations must articulate the complexity of the question that the AI/ML model is meant to solve using these three criteria:

- **Annotation requirements.** Factors to consider include the grade of disease state, how many components need to be identified, and the level of detail required in the annotation, which influence the annotation method (e.g., semantic segmentation, bounding boxes).
- **Imaging modality.** These include X-rays, MRI images and CT scans. Some of these are easier to "read" than others. Within a modality, some images also may be of higher resolution than others.
- **Presentation of symptoms.** Some disease states present with distinct features, whereas others present more ambiguously. The answers that the healthcare organization requires for its AI/ML model determine how much clinical expertise that the labeling team needs to process the corresponding data set. Disease complexity is proportional to the clinical expertise needed for accurate data labeling (see Figure 1). A diagnostic model designed to merely detect potential tumors in a breast scan will need different data than a model that is meant to precisely stage the cancer. In the former case, non-clinicians trained to spot abnormalities may label images as normal or suspect. Clinicians can then review and stage the subset of suspect images.

Disease complexity: Proportional to clinical expertise for accurate data labeling

Example of a high-complexity case: Invasive Ductal Carcinoma

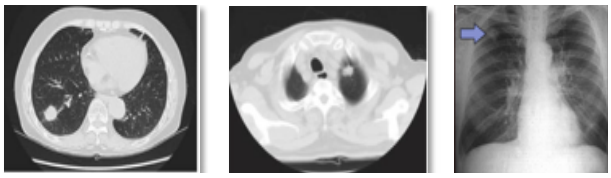


Current process that a pathologist undertakes to diagnose Invasive Ductal Carcinoma (IDC):

1. Identify tumor as part of a routine mammogram
2. Perform biopsy of the tumor
3. Analyze the biopsy
   - **Gland (acinus) formation**
     - Evaluate the percentage of the tumor that is made up of acini
   - **Size and shape of cells**
     - Compare diseased cell nuclei to benign breast epithelium
     - Evaluate the shape of the diseased cells compared to the benign breast tissue
   - **Mitosis rate**
     - Count the number of cells undergoing mitosis
4. Perform imaging exam (MRI, CT, or PET) for tumor details

Given the expertise required to recognize, annotate, and ultimately diagnose IDC, only a specialist would be able to provide high-quality annotations.

Example of a low-complexity case: Pulmonary Nodules



Current process to diagnose a malignant pulmonary nodule:

1. Perform imaging exam (X-ray and CT)
   - **Appearance of nodule**
     - Inhomogeneous versus homogeneous density
     - Thick versus thin walls
   - **Location of nodule**
     - Upper lung versus other locations in the lung
   - **Shape of nodule**
     - Margin shape – Irregular or spiculated margins
2. Discuss medical history (smoking history & family history of cancer)
3. Perform follow-up imaging exam (X-ray and CT)
4. Perform biopsy to determine if malignant

Given the relatively straightforward characteristics that are evaluated, the level of domain expertise needed to annotate the nodules is relatively low.

Figure 1

By starting with a small foundational team, organizations can perfect their labeler selection and quality assurance (QA) processes before scaling to larger projects. The foundational team should consist of physicians, other clinicians and non-physicians, so that it can adapt and flex to label any initial disease state. For a labeling team made up of primarily non-clinical labelers, QA will be built into the processes along with enhanced training and frequent audits of labelers' work. For an AI/ML model that requires labeling by highly trained physicians, QA likely will involve fewer audits but require more time spent establishing the "ground truth" data (see Figure 2).

Differing considerations for low-, medium-, and high-complexity cases

	High-complexity cases	Medium-complexity cases	Low-complexity cases
Description	Complex cases require advanced clinical expertise and often have substantial variation between cases.	With physician oversight, other clinicians and non-clinicians can be trained to identify the disease.	Non-clinicians can be trained to locate easily identifiable diseases with training and physician oversight.
Labeler mix	Physicians	Mix of physicians, other clinicians, and non-clinicians	Non-clinicians with some physician oversight
Considerations	1. Complex data labeling annotation types (e.g., grading level of diagnoses, semantic segmentation) 2. Advanced imaging modalities (e.g., MRIs, mammography) 3. Ambiguous symptoms of the disease that are not consistent across cases 4. Minimal or no clinical training needed; likely fewer audits from QA perspective 5. Some training on annotation best practices potentially needed	1. Straightforward data labeling annotation types (e.g., bounding boxes, locating centroids, polygon outlines) 2. Easier imaging modality (e.g., black and white, defined shapes) 3. Distinct patterns of symptoms 4. Some training on annotation best practices potentially needed 5. Likely fewer audits for clinical accuracy	1. Straightforward data labeling annotation types (e.g., bounding boxes, locating centroids, polygon outlines) 2. Easier imaging modality (black and white, defined shapes, etc.) 3. Distinct patterns of symptoms 4. Minimal or no training needed on annotation best practices 5. Substantial clinical training and subsequent quality assurance and audits necessary

Figure 2

Basing the labeling team's required clinical expertise QA oversight level on the needs of the AI/ML model provides organizations with a scalable, and time- and cost-effective approach. It allows organizations to evaluate and label multiple disease states concurrently by reducing reliance on already-busy physicians. This accelerates the creation of an ImageNet repository of medical data.

Best practices

Healthcare organizations must support their diverse labeling teams with best practices in data pre-processing, training and quality control (see Figure 3). This helps ensure that the resulting labeled data sets meet expectations and are appropriate for AI/ML diagnostic model training.

Ensuring high-quality results from non-clinical labelers

Figure 3

Data preprocessing. Data preprocessing sets up the project for success by ensuring that the data is properly prepared for the data labelers and by establishing quality parameters and the requirements for the training set. The size and depth of the training set needs to be selected based on statistical significance and the complexity of the proposed problem. An experienced statistician or data scientist is critical to this step. They should advise on the validity of the labeled data and help maintain the project metrics as it scales. In preprocessing, it is also important to ensure that the images are in similar formats with adequate resolution. This preparatory step helps increase efficiency and improves the quality and accuracy of labels by allowing labelers to make precise observations and more consistent annotations.

Ground truth. The ground truth data set consists of a subset of data points identified and labeled by an expert. This data set becomes the gold standard benchmark to which all other labelers' work is compared to monitor the quality of labels and help ensure the project's ultimate success.

Once defined, organizations may use the ground truth data set to identify high-quality labelers and to continuously monitor their performance throughout the project's lifecycle. Error-seeding, in which random errors are distributed in the image stream, can be based off the ground truth data set to measure how well labelers compare in quality to the domain expert.

Training. Regardless of whether the team is composed entirely of physicians or includes non-clinicians, extensive training at the project onset prepares all labelers, sets expectations and safeguards the integrity of the labels. In the initial training, the medical images and labels are broken down to a granular, easy-to-spot level and the labeling requirements are described in a series of simple, clear steps.

Ongoing training helps ensure that labelers consistently understand and apply requirements. Industry best practices for creating training materials include leveraging deep domain expertise to define clear labeling criteria. Additionally, cultural competency is a key training component because

diagnoses or levels of severity may differ among countries. For example, a mild case of eczema may be defined differently in the U.S. compared to how it is commonly defined in Thailand. By clearly defining disease and diagnosis requirements during training, organizations can avoid many cultural biases in the labels.

Quality. In addition to leveraging ground truth data in error-seeding exercises to monitor quality performance, other metrics can be employed to monitor quality. "Recall" refers to the measure of sensitivity and completeness, measuring the percentage of total relevant results that are classified by the machine learning algorithm. "Precision" refers to the measure of exactness, measuring the percentage of results that are relevant. Recall and precision rates cannot be maximized simultaneously; however, the F-1 score provides a mean of precision and recall (see Figure 4). This metric is often used for instances when both precision and recall are important to the problem. To measure whether annotators are labeling with a consistent understanding of the defined labels, inter-rater reliability can also be measured. Inter-rater reliability refers to the test-retest method to measure that annotators are labeling with a consistent understanding of the defined labels. These quantifiable metrics for measuring the accuracy of the data labels provide a distinct way to measure the success of the labels and the model.

Calculating the F-1 score to capture a mean of precision and recall

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad \text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \quad \text{F-1 score} = 2\left[\frac{\text{Precision} * \text{Recall}}{\text{Precision + Recall}}\right]$$

Figure 4

AI helping AI

It is essential that organizations leverage best practices and emerging AI tools to effectively accelerate the creation of a labeled medical image repository at speed and scale. Best practices around data preprocessing, ground truth, training, and quality combined with emerging AI trends and technologies reduce the burdens associated with labeling a high volume of complex images. For example, federated learning is one opportunity that combats the need to host highly sensitive medical data in a single location to train an ML model. In federated learning, each healthcare system may train their respective model based on local training data. Then, the models across healthcare systems are merged to create a master AI/ML model, thereby reducing the compliance risk of sharing data across health systems. Leveraging federated learning and other techniques decreases the time spent on the end-to-end data labeling and training process, thereby providing more efficient and scalable diagnostic model development.

Approaches to labeling will vary across teams. The constant should be ensuring that the complexity of the diagnostic question defines the labeler requirements and best practices around labeler selection and model development and training. Careful consideration of these requirements and following best practices are pivotal to accelerating the creation of an accurate and valid repository of labeled medical images.

Endnote

[1] From Drug R&D To Diagnostics: 90+ Artificial Intelligence Startups In Healthcare, CB Insights research brief, September 12, 2019, https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/.

About the authors

Sashi Padarthy

Assistant Vice President, Cognizant Consulting's Healthcare Practice

Sashi Padarthy leads Cognizant's Digital Strategy and Transformation service line. For more than 20 years, Sashi has helped healthcare organizations in the areas of digital strategy, digital transformation, innovation, technology-enabled strategy, new product development, value-based care and operational improvement. He is a Sloan Fellow from the London Business School. Sashi can be reached at Sashi.Padarthy@cognizant.com | www.linkedin.com/in/sashipadarthy/.

Kristin Knudson

Consulting Manager, Cognizant Consulting's Healthcare Practice

Kristin Knudson is a Manager within Cognizant Consulting's Healthcare Practice. She has experience in implementing digital transformation solutions and integrating machine intelligence into care delivery. Kristin holds an MBA in business strategy and a master's degree in biomedical engineering, allowing her to provide a unique perspective on the advantages of leveraging technology in the healthcare industry. Kristin's current interest is exploring applications of machine intelligence to advance digital insights and improve the consumer's overall experience. Kristin can be reached at Kristin.Knudson@cognizant.com | www.linkedin.com/in/kristin-knudson-4266b22a/.

Jacqueline Zelener

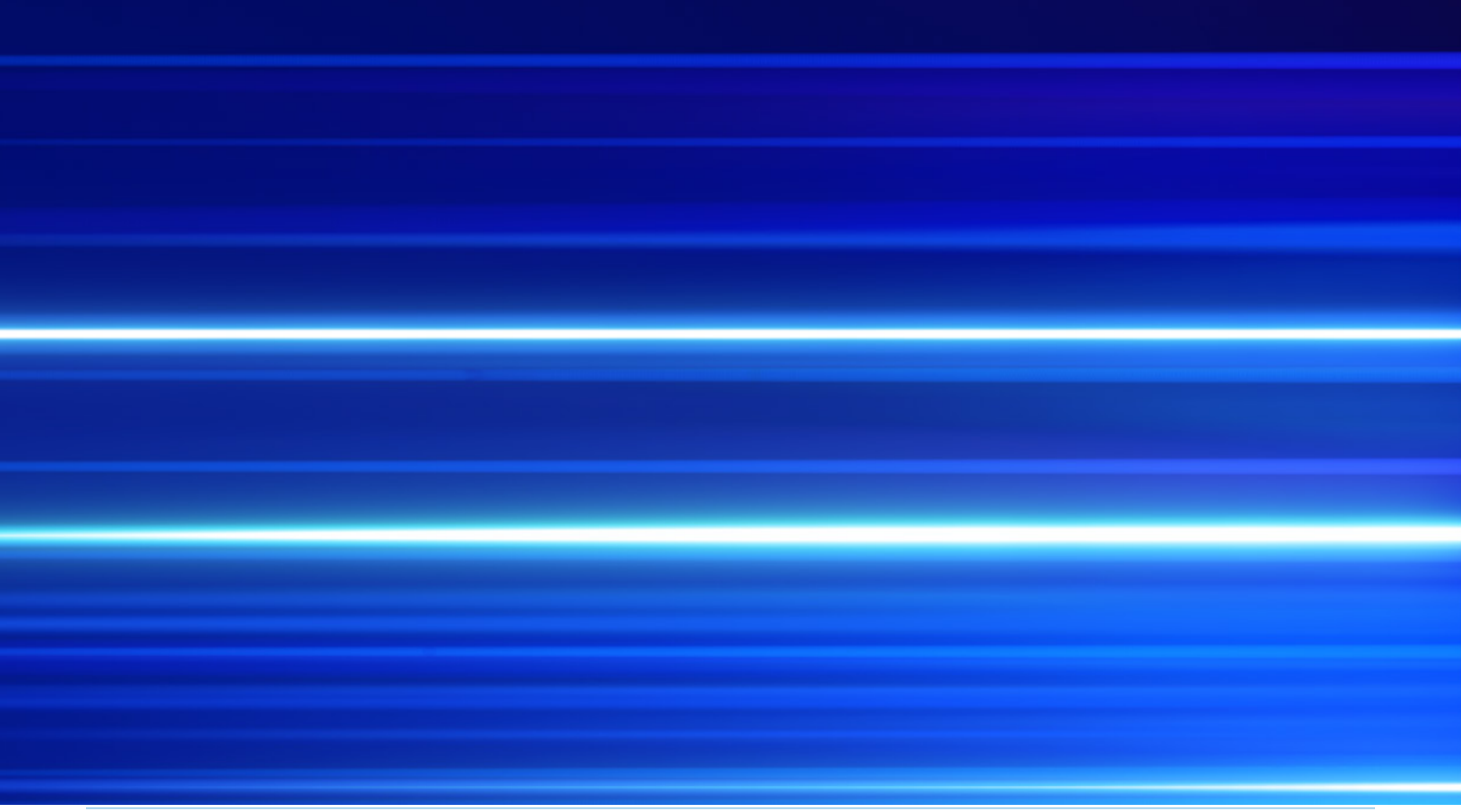
Senior Consultant, Cognizant Consulting's Healthcare Practice

Jacqueline Zelener is a Senior Consultant in Cognizant Consulting's Healthcare Practice. She has valuable experience in payer operational processes, process optimization, business readiness, and medical management. Jacqueline holds a master in health administration and has worked in operations and clinical research, which provides her with a dynamic outlook on the intersection of business outcomes, technology, and healthcare. Jacqueline can be reached at Jacqueline.Zelener@cognizant.com | www.linkedin.com/in/jacquelinezelener/.

Mary Helen Turnage

Senior Consultant, Cognizant Consulting's Healthcare Practice

Mary Helen Turnage is a Senior Consultant within Cognizant Consulting's Healthcare Practice. She has experience in payer operational readiness as well as process optimization. Mary Helen has an MBA in strategy and finance and has worked in clinical recruitment and strategic execution. She is especially interested in how technology is changing the workforce landscape in healthcare. Mary Helen can be reached at Mary.Turnage@cognizant.com | www.linkedin.com/in/maryhelenturnage/.



About Cognizant Healthcare

Cognizant's Healthcare business unit works with healthcare organizations to provide collaborative, innovative solutions that address the industry's most pressing IT and business challenges — from rethinking new business models, to optimizing operations and enabling technology innovation. As a global leader in healthcare, our industry-specific services and solutions support leading payers, providers and pharmacy benefit managers worldwide. For more information, visit www.cognizant.com/healthcare.

About Cognizant Digital Business

We help clients build digital businesses and innovate products that create new value — by using sensing, insights, software and experience to deliver on what customers demand in the digital age. Through IoT we connect the digital and physical worlds to make smart, efficient and safe products, operations and enterprises. Leveraging data, analytics and AI we drive intelligent decisions and anticipate where markets and customers are going next. Then we use those insights, combining design and software to deliver the experiences that consumers expect of their brands. Learn more about how we're engineering the modern enterprise at www.cognizant.com/digitalbusiness.

About Cognizant

Cognizant (Nasdaq-100: CTSH) is one of the world's leading professional services companies, transforming clients' business, operating and technology models for the digital era. Our unique industry-based, consultative approach helps clients envision, build and run more innovative and efficient businesses. Headquartered in the U.S., Cognizant is ranked 194 on the Fortune 500 and is consistently listed among the most admired companies in the world. Learn how Cognizant helps clients lead with digital at www.cognizant.com or follow us @Cognizant.

Cognizant

World Headquarters

500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277

European Headquarters

1 Kingdom Street
Paddington Central
London W2 6BD England
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102

India Operations Headquarters

#5/535 Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060

APAC Headquarters

1 Changi Business Park Crescent,
Plaza 8@CBP # 07-04/05/06,
Tower A, Singapore 486025
Phone: + 65 6812 4051
Fax: + 65 6324 4051

Codex 5914.2