MIS Quarterly

SPECIAL ISSUE: MANAGING AI

IS AI GROUND TRUTH REALLY TRUE? THE DANGERS OF TRAINING AND EVALUATING AI TOOLS BASED ON EXPERTS' KNOW-WHAT[1]

Sarah Lebovitz
University of Virginia, McIntire School of Commerce,
Charlottesville, VA 22904 U.S.A. {sarah.lebovitz@virginia.edu}

Natalia Levina
New York University, Stern School of Business,
New York, NY 10003 U.S.A. {nlevina@stern.nyu.edu}

Hila Lifshitz-Assaf
New York University, Stern School of Business,
New York, NY 10003 U.S.A. {h@nyu.edu}

Organizational decision-makers need to evaluate AI tools in light of increasing claims that such tools outperform human experts. Yet, measuring the quality of knowledge work is challenging, raising the question of how to evaluate AI performance in such contexts. We investigate this question through a field study of a major U.S. hospital, observing how managers evaluated five different machine-learning (ML) based AI tools. Each tool reported high performance according to standard AI accuracy measures, which were based on ground truth labels provided by qualified experts. Trying these tools out in practice, however, revealed that none of them met expectations. Searching for explanations, managers began confronting the high uncertainty of experts' know-what knowledge captured in ground truth labels used to train and validate ML models. In practice, experts address this uncertainty by drawing on rich know-how practices, which were not incorporated into these ML-based tools. Discovering the disconnect between AI's know-what and experts' know-how enabled managers to better understand the risks and benefits of each tool. This study shows dangers of treating ground truth labels used in ML models objectively when the underlying knowledge is uncertain. We outline implications of our study for developing, training, and evaluating AI for knowledge work.

Keywords: Artificial intelligence, evaluation, uncertainty, new technology, professional knowledge work, innovation, know-how, medical diagnosis, ground truth

Introduction

We are experiencing a significant shift in how knowledge is produced, from focusing on the quality of the human expert to evaluating modern artificial intelligence (AI) technologies. This raises key questions about how we evaluate the performance of human experts as compared to AI technologies. In this study, we focus on specific AI tools that use machine learning (ML) classification methods to draw inferences from training datasets consisting of labeled input–output pairs and classifies new inputs into predefined output classes. One of the challenges with the development of traditional AI tech-

[1]Nicholas Berente, Bin Gu, Jan Recker, and Radhika Santhanam were the accepting senior editors for this paper. Eivor Oborn served as the associate editor.

nologies, such as rule-based expert systems, has been that they often relied on representing experts' knowledge in machine-readable form (Forsythe 1993; Hutchins 1995; Star 1989; Suchman 1987). This was difficult to achieve due to the tacit nature of experts' knowledge (Brown and Duguid 1991, 2001; Kogut and Zander 1992; Orlikowski 2002). Today, however, there is renewed hope among AI creators that they can bypass capturing tacit aspects of experts' knowledge because ML-based AI can implicitly infer how inputs map to outputs. This shift in AI technologies has prompted a burgeoning number of reports that AI can produce higher quality judgments than human experts (e.g., Grady 2019; Moran 2018; Shoham et al. 2018).

Organizational decision-makers are confronting this exploding discourse of the promises of ML-based AI and face decisions about whether and how to incorporate such tools in their organizations (Faraj et al. 2018; Rao and Verweij 2017). Part of assessing the quality of new technologies involves examining the reported "outperformance" claims by examining the technical objects and their implications (Knorr Cetina 1999). While recent research has begun investigating how individual users interact with AI tools in practice (e.g., Christin 2020; Lebovitz 2019; Pachidi et al. 2021), we believe it is critical to understand how managers evaluate AI tools as such evaluations drive their adoption decisions.

To this end, we conducted a field study that focuses on understanding how managers in the field of diagnostic radiology evaluated AI tools for potential organizational adoption. This study is based on the growing area of AI development for medical diagnosis, a field in which experts experience high degrees of uncertainty. Diagnostic radiology, in particular, has been at the cutting-edge of developing AI tools dating back to the 1980s (e.g., Chandrasekaran et al. 1980). Early attempts utilizing rule-based expert systems largely failed, given technical limitations at that time (Oakden-Rayner 2019), but recent technological advances (especially in ML-based tools for image recognition) are spurring renewed interest and investment in diagnostic AI tools. As a result, growing numbers of ML-based AI tools claim to outperform experts and have captured the attention of diagnostic radiology practices worldwide. Our study follows how managers within a major U.S. diagnostic radiology department evaluated five ML-based AI tools for potential adoption and the challenges they encountered. While, in practice, experts tend to rely on rich know-how (accumulated expertise, rooted in situated, social, and tacit practices) to gauge their work quality, managers' evaluation process for these AI tools focused primarily on know-what knowledge outputs (explicit and codified aspects of knowledge) represented in ground truth labels and summary accuracy measures. Although know-what-based measures suggested the AI tools were highly accurate, all five tools performed poorly during internal pilot studies and left managers searching for explanations. Their search ultimately led managers to confront the high uncertainty involved in evaluating human experts' knowledge outputs (know-what) and to recognize that ML-based AI tools did not capture experts' tacit knowledge practices (know-how). Discovering the disconnect between ML-based AI's know-what and human experts' know-how enabled managers to better understand the risks and benefits associated with each AI tool.

Background Literature

Evaluating Professional Knowledge Work

Evaluating knowledge work is at the heart of many professional fields, yet it is highly challenging since it is far from objective. Prior sociology of science and knowledge literature has established how evaluation of knowledge occurs through ongoing processes of contestation and negotiation across and within social groups (e.g., Latour 1987; Pinch and Bijker 1987; Star 1995). Thus, experts in knowledge-intensive fields have been shown to deeply struggle to validate their novel insights, as shown by studies of scientists in biomedical innovation (Dougherty and Dunne 2012; Mengis et al. 2018) or medical professionals (Lebovitz 2019; Menchik 2014): struggling amidst the uncertainty they face about the current state of medical knowledge: "what is not understood about the human body and how it functions is far greater than what is understood" (Northrup 2005, p. 70). Experts often acknowledge the uncertainty about knowledge in their given domain (e.g., Knorr Cetina 1999; Mengis et al. 2018; Rindova and Courtney 2020; Schön 1983), as in the medical literature discussing how many potential medical treatments or tests lack external validation or more reliable measures that can assess the quality of a given medical outcome (Timmermans and Berg 2003).

However, evaluating the quality of knowledge work requires going beyond assessing knowledge outputs, or experts' "know-what," to assessing knowledge practices, or experts' "know-how." For decades, organizational scholars have been investigating the social, tacit, and embodied nature of know-how in knowledge work (Brown and Duguid 1991, 2001; Garud 1997; Hutchins 1995; Kogut and Zander 1992). This literature builds on sociological insights of Ryle (1949) and Polanyi (1958, 1966), who disentangle the explicit aspects of knowledge from the tacit: aspects of knowledge that are socially embedded, learned through experiences, tied to the senses, and cannot be fully articulated. Accordingly, organi-

zational research distinguishes between know-what, or "knowledge as information imply[ing] what something means" (Kogut and Zander 1992, p. 387), and know-how, or "accumulated practical skill or expertise that allows one to do something smoothly and efficiently" (von Hippel 1988, p. 6). Scholars studying knowledge in organizations highlight its situated, social, and enacted nature; they depict know-how as "rooted in action, procedures, routines, commitment, ideals, values, and emotions" (Nonaka and von Krogh 2009, p. 636) and "implicit in our pattern of action and in our feel for the stuff with which we are dealing" (Schön 1983, p. 49).

Experts in a given field are expected to acquire and demonstrate the distinct know-how that internally binds that group (Abbott 1988; Knorr Cetina 1999; Nicolini 2012). Accumulated professional practices and protocols that constituted know-how have been documented to be highly difficult to transfer and master (Leonard-Barton 1995; Szulanski 1996). Through "learning-by-doing whereby knowledge about how to perform a task accumulates with experience over time" (Garud 1997, p. 84), experts come to adopt their field's distinctive know-how, acquiring its unique viewpoint and speaking its language (Brown and Duguid 1991). This can create particular difficulty for evaluating a profession's know-how from the outside, since "the art of one practice tends to be opaque to the practitioners of another" (Schön 1983, p. 271).

When describing the process of professionalization, Abbott (1988, p. 40) shows how experts work to establish and protect their tacit "professional knowledge system" and hold members accountable for their ability to acquire and demonstrate know-how practices defined by that system (e.g., how to structure problems, evoke rules of relevance, apply abstract inference). Focusing on know-how aspects allows professionals to gain legitimacy and ward off potential occupational challenges seeking to over-simplify the rich know-how practices. Bechky (2003) showed how engineers in a cross-disciplinary context were able to "maintain their status as experts" (p. 735) since key aspects of engineers' knowledge (their workmanship, tricks of the trade, and tribal knowledge) were not represented in their central knowledge outputs (e.g., technical blueprints and drawings). Instead of being evaluated based on know-what outputs, their professional value hinged on their ability to demonstrate and enact their know-how by using the outputs within larger practices of problem solving and communication.

Thus, the know-what of knowledge work is difficult to evaluate or even characterize due to the uncertain nature of knowledge in many professional contexts. Instead, more tacit, situated, and social aspects of professionals' knowledge are central to evaluating the quality of the knowledge work. However, today, increasing numbers of AI tools are being designed for and adopted by organizations in knowledge work contexts, raising the important question of how is the quality of these tools being evaluated?

Evaluating the Performance of AI Tools

Generally speaking, evaluating the performance of AI involves assessing its ability to produce the correct output for a given input. The AI tools in this study are image recognition and classification models that use ML-based models to detect and learn patterns between inputs and outputs in pre-labeled data sets in order to assign new inputs to predefined output categories (Bechmann and Bowker 2019; Provost and Fawcett 2001). These tools are designed using neural networks that are trained to discover probabilistic relationships among features of the input and output data and generate a series of relative weights that can be applied to future data inputs. Measuring the quality of this type of AI model involves calculating how often the model's predicted outputs match the label defined as accurate in the data set reserved for model validation (Kohavi and Provost 1998). This calculation is often represented by a metric called the "Area Under the receiver operating Curve," or AUC,[2] and plotted on a two-dimensional graph (see example in Appendix B). The AUC summarizes a model's success and error rates, plotting its relative rate of false negatives (excluding an input from the correct class) and false positives (assigning an input to an incorrect class) (Provost and Fawcett 2013).

These output-based performance measures are central to assessing the performance of ML-based AI tools and are frequently cited to indicate the tool's quality. For example, machine learning competitions are typically judged by comparing submitted models' AUC values[3] and measuring them against the baseline performance of human experts in that domain. Researchers and media outlets often report how AI models today are generating lower error rates than humans, claiming, for instance, that "many consider [image classification] solved—the error rate is incredibly low at around 2%" (Gershgorn 2017). In the context of medical diagnosis, for example, the creators of an AI tool for detecting lung cancer (Ardila et al. 2019) reported their model's AUC of 0.944 and suggested that this highly accurate model is likely to transform patient care.

[2]"AUC" and "ROC" (Receiver Operator Curve) are commonly used interchangeably in practice.

[3]Examples can be found at https://www.kaggle.com/competitions.

These performance measures are then further amplified (and simplified) as they are cited and incorporated into further academic and economic analyses (e.g., Autor 2015; Dhar 2016; Frey and Osborne 2017; Seamans and Furman 2019; Shoham et al. 2018). For instance, researchers predict occupational automation trends by analyzing a prominent dataset (curated by the Electronic Frontier Foundation) that aggregates the highest AUC measures reported for various AI problem domains (e.g., image recognition, speech recognition, language translation) (Felten et al. 2018; Seamans and Furman 2019). Similarly, Frey and Osborne (2017) build on AUC measures published by AI research communities when explaining the assumptions underlying their analysis, including claims that, "today, the problems of navigating a car and decipher handwriting are sufficiently well understood" (p. 259).

Finally, journalists and media outlets cover these economic and technical reports and further incorporate and simplify the message of AI accuracy measures. These stories often concentrate layered technical arguments and research into attention-catching headlines and brief, flashy conclusions. For instance, a recent *New York Times* article (Grady 2019) was titled "AI Took a Test to Detect Lung Cancer. It Got an A" and communicated the tool's accuracy in the first sentence: "Computers were as good or better than doctors at detecting tiny lung cancers on CT scans." *Nature* published an article titled "Rise of Robot Radiologists," which reported the performance of an AI tool as "significantly more accurate at predicting cancer—or the absence of cancer—than practices generally used in clinics" (Reardon 2019, p. S55).

Output-based accuracy measures clearly play a pivotal role in evaluating the performance of modern AI tools and underscore the need for deeper understanding of how these measures are evoked and interpreted in practice. Today, ML-based AI tools are being developed for growing numbers of knowledge work contexts, such as in criminal justice (Angwin et al. 2016; Christin 2014), human resource departments (Van Den Broek et al. 2020; Weissmann 2018), law enforcement (Waardenburg et al. 2018; Walch 2019), and sales departments (Pachidi et al. 2021). While some organizational research has begun focusing on how individual experts are perceiving and using AI-generated knowledge outputs in their daily work (Knorr Cetina 2016; Lebovitz 2019; Pachidi et al. 2021), we know little about how modern AI tools are being evaluated for potential organizational adoption to begin with. In response to this growing need, this study investigates how managers form evaluations of ML-based AI tools in the context of making medical diagnoses, bringing to light the stifling challenges that arise and the consequences of their evaluation practices for AI adoption.

Research Design and Methods

Research Setting

We conducted an 11-month qualitative study investigating how managers formed evaluations of AI tools across multiple sections of a department of diagnostic radiology at a tertiary hospital in the United States (Urbanside). Diagnostic radiology is a medical specialization whereby highly trained physicians use medical imaging technology (e.g., x-ray, CT-scan, MRI) to provide diagnostic and treatment recommendations to patients and their team of physicians. The AI tools being evaluated at Urbanside were ML-based classification models based on image recognition technologies; they analyzed medical imaging files as inputs and generated outputs in the form of disease classifications or image segmentation files. Because of the regulatory restrictions in the U.S. at the time of the study, the AI tools were not designed to actively learn or dynamically adapt after implementation. Instead, a "frozen" version of a trained model was submitted for regulatory approval, which could then be deployed into clinical settings. Any model tweaks or improvements required additional rounds of regulatory approval and re-implementation. Details of the five tools analyzed in this study are summarized in Table 1.

Methods

Data Sources. We followed a grounded approach to theory development which involved iteratively analyzing data during and throughout the observational period (Charmaz 2014; Glaser and Strauss 1967). The primary data for this study is 11 months of ethnographic observations (Van Maanen 1998) within Urbanside spanning January to November 2019. Observations focused primarily on how Urbanside managers assessed AI tools for potential adoption in their respective sections (chest imaging, breast imaging, pediatric imaging, and neuroradiology). In all, 23 managers were included in our data collection. In this study, managers refer to board-certified diagnostic radiology physicians who were actively shaping the assessments of AI in their subsections. Their roles are similar to manager-professional "hybrids" described in prior organizational research (Croft et al. 2015; McGivern et al. 2015), as they serve dual roles including both managerial and clinical components. The specific manager-professional hybrid roles of this study's managers include department chairs, junior and senior diagnostic radiologists formally leading AI research projects or testing AI vendor tools, as well as medically trained AI specialists working on diagnostic AI projects. Most of the managers in this study performed diagnostic radiology work on a daily or weekly

Table 1. Comparing Five AI Tools Being Evaluated at Urbanside

Tool Name *Urbanside subsection of Radiology Department*	**Nature of the Tool's Diagnostic Task**	**Measure of the Performance of the Diagnosis Task**	**Tool Source**	**Tool Design**
Brain Tumor Segmentation tool *Neuro radiology section*	• Generate segmentation labels for three regions of a brain tumor on MRI (entire tumor, tumor core, and enhancing tumor core) that can be used to calculate precise volumetric measurements. • Context: Physicians pay attention to changes in the size and shape of tumors for making diagnoses and treatment recommendations. Focusing on changes at the edges of each region helps assess tumor development. High variation in brain tumors' appearance, shape, and properties make judging its borders on MRI challenging.	• No absolute measure of performance, little possibility to increase certainty using additional testing. • No single method serves as the agreed-upon standard; multiple methods are utilized depending on the purpose of the exam and produce inconsistent results. • Highly subjective and variable interpretations across different radiologists.	Open source	Cascade of convolutional neural networks that take four separate MRI imaging sequence files as input and generates three segmentation labels. These labels are combined into a single output that is displayed to the user as a visual overlay over the original MRI images.
Bone Age tool *Pediatrics Imaging*	• Classify a child's hand x-ray to a specific numeric value (number of years and month) and classify whether that value is considered "normal" or "abnormal." • Context: Skeletal age, compared to a child's chronological age, is critical for managing growth disorders in children. To assess the stage of a child's growth development, a hand x-ray is compared against a medical atlas containing a series of images to identify the closest match.	• No absolute measure of performance, little possibility to increase certainty using additional testing. • Multiple standards and methods are available; physicians use different medical atlases and analytical approaches within and across hospitals. • Highly subjective and variable interpretations across different radiologists and within the same radiologist at separate times of observation (De Sanctis et al. 2014).	Research group not affiliated with Urbanside	Convolutional neural networks that take a single hand x-ray image as input and first generate a classification of the number of months and number of years and then assign a "normal" or "abnormal" classification. These outputs are presented to the user in an auto-populated report template.
Breast Mammo tool *Breast imaging section*	• Segment abnormal regions on a mammogram image and classify each region to a "malignant" or "benign" output. • Context: Mammography screening is the most popular tool for early breast cancer detection. Typically abnormal regions are identified by analyzing and integrating numerous sources of medical information and inputs. When a region is judged as likely to be malignant, the patient is often recommended for biopsy or additional imaging to make the final diagnosis.	• Highly subjective and variable diagnosis and follow-up recommendations across radiologists (Duijm et al. 2009). Up to 12% of breast cancers in the U.S. are missed during initial mammography. Of the 9-10% of patients recalled for additional imaging, less than half are found to have breast cancer (Lehman et al. 2017). • Professional standards suggest the use of additional imaging modalities (MRI, ultrasound), performing biopsies, or obtaining long-term patient records.	Research group affiliated with Urbanside	Convolutional neural networks input a single mammogram, which is first segmented to identify all lesions present. Each lesion is then classified as "malignant" or "benign." These results are aggregated to an image-level probability of malignancy. Users see an overlay of circles marking "malignant" lesions on the original mammogram image as well as the image's overall "malignancy score."

Tool Name *Urbanside subsection of Radiology Department*	**Nature of the Tool's Diagnostic Task**	**Measure of the Performance of the Diagnosis Task**	**Tool Source**	**Tool Design**
Breast Ultrasound tool *Breast imaging section*	• Classify (physician-marked) lesion on ultrasound image into one of four diagnosis categories (benign, probably benign, suspicious, or probably malignant) with the associated confidence level, and classify the lesion's shape and orientation. • Context: Physicians often cross-validate abnormal regions identified on a mammogram to their appearance on ultrasound, which provides additional information to inform physicians' diagnosis, such as the lesion size, shape, and whether it is solid or fluid-filled.	• Regular disagreement between radiologists on final diagnosis categorization of ultrasound findings; higher concordance reported for lesion size, shape, and orientation (Lazarus et al. 2006). • Professional standards suggest the use of additional imaging modalities (MRI, mammogram), performing biopsies, or obtaining long-term patient outcomes.	Vendor	Patented software design using an ensemble of machine-learning algorithms that takes two regions-of-interest (drawn by the physician user) on an ultrasound image as input and classifies the lesion's diagnostic category, associated confidence level, and shape and orientation.
Chest Triage tool *Chest imaging section*	• Classify chest x-ray into one of 14 disease categories (e.g., pneumonia, pleural effusion, cardiomegaly) and classify whether that outcome is considered "normal" or "abnormal," which is then used to prioritize the work queue. • Context: Chest x-ray is the most common imaging examination globally, as it is critical for the diagnosis and management of many diseases. As physicians analyze cases, the work queue grows longer and longer, and typically utilization a prioritization based on coming from the emergency department, then on a first-in-first-out basis.	• Highly subjective and variable diagnosis interpretations and follow-up recommendations among different radiologists based on the same chest x-ray. • Professional standard suggests the use of additional imaging and modalities (CT scan) or follow-up studies.	Open-source	Convolutional neural networks that take two chest X-ray images (from the front and side views) as input and classifies the disease category and the associated result of "normal" or "abnormal." Users can route cases based on the final labels to prioritize the work queue.

basis (sometimes with a reduced workload) in addition to conducting administrative and AI-related responsibilities. These responsibilities included formally leading and coordinating AI research projects, negotiating with AI vendors, attending and organizing AI conferences and symposia, staying current on published research and regulatory guidelines for diagnostic AI, participating in implementation preparation meetings with hospital IT and infrastructure teams, and so forth.

Five AI tools are the primary analytical focus since their full evaluation occurred while we had field access. In total, 31 AI evaluation meetings were observed and analyzed, wherein managers spent one to two hours presenting cutting-edge AI tool research, debating tool performance at length with data scientists and other stakeholders, discussing internal pilot studies and implementation plans, and so forth. Observations also include managers participating in industry and research conferences, workshops, symposia, and vendor presentations. We conducted 22 semi-structured interviews (Spradley 1979) and numerous informal conversations with managers supplemented the observational data and deepened our understanding of managers' perceptions of the tools throughout the evaluations. Interviewing the same individuals at different time points helped us understand how their perceptions and opinions about the AI tools shifted over time, what was driving those shifts, and what new discoveries resulted. All interviews were conducted in person in administrative offices and were recorded (with informants' permission and consent) and transcribed.

Finally, our data collection included archival analysis of over 150 articles detailing diagnostic AI research across the professional and academic literature. We analyzed how these articles reported dimensions of the AI tools, including the diagnostic context, technical infrastructure, training and validation datasets, and reported accuracy measures. Moreover, while observing evaluation meetings and conferences, particular attention was paid to capturing the technical aspects of the tools and how managers engaged with these aspects in practice. The archival data also included AI tools' regulatory filings as well as materials referenced or distributed at academic events, professional workshops, and vendor presentations.

Data Analysis. The main focus of our analysis was understanding managers' process of evaluating AI tools. In keeping with grounded theory methods, we constantly compared emerging themes and categories across ongoing data collection (Charmaz 2014). Early on, we created detailed accounts of managers' evaluation of each AI tool (based on all authors reading the full set of observation, interview, and archival data multiple times) which chronologically captured the range of practices, interactions, and artifacts that were involved. We identified and examined commonalities and differences among the processes for each of the five main AI tools evaluated. We noted, for instance, how managers in all five evaluation processes were highly focused on a tool's AUC measure and the qualifications of the experts producing ground truth labels. We also noted how these became less prominent in the later phases of the process, as they began discovering the disconnects between these measures and experts' practices. Differences in the evaluation processes were mainly observed in how managers adapted the practices to the unique diagnostic context of each AI tool (for example, a measure called the "DICE" score was used for the brain segmentation task, which operated very similar to the AUC measure but was a more appropriate measure for this task). We frequently zoomed in on the specific nature of managers' practices and zoomed out to see how these practices were subject to institutional and organizational forces (Nicolini 2009). For instance, we noted managers regularly referred to AUC graphs in their meetings and discussions. Zooming out, we read and analyzed the professional radiology literature and discovered that AUC measures have been used for decades as a core professional method of evaluating diagnostic tools (AI or otherwise).

Additional analysis focused on understanding the technical nature and differences among the focal diagnostic scenario of each AI tool. We compared the nature of each tool's predictive task as well as how each tool was reportedly trained and validated. This led us to see the similarity in all five tools' underlying technology, whereby all used image recognition and ML-based classification models. We also observed meaningful differences in each tool's diagnostic task and how managers accounted for the specific risks and benefits associated with each distinct diagnosis context. For example, managers in the chest imaging section had different expectations for the AI tool designed to triage urgent patients than managers in the breast imaging department evaluating tools producing specific diagnosis outputs for physicians to consider. This led us to focus more deeply on how managers assessed the reported claims about AI performance within each diagnostic context and how they weighed the relative risks and benefits of each tool.

We analyzed relationships between themes across the five tools (Golden-Biddle and Locke 2007), noting similarity in patterns. For instance, across all five evaluation processes, managers' focus shifted from one measure to the next in similar phases, from analyzing AUC measures to eventually turning their focus inward to analyze the practices of experts in their field. Analyzing these shifts in managers' focus and practices led us to notice how certain aspects of the tools were scrutinized (or ignored) as the tool evaluation unfolded over time. For example, specific aspects of research articles and vendor materials were central early on, whereas other aspects like the internal pilot study results and internally produced ground truth labels gained prominence later in their evaluations. A key insight from this analysis was observing how, in all five evaluation processes, managers' focus ultimately shifted inward, to scrutinizing the uncertainty and variability of the performance of experts in their field and the lack of reliable measures of quality in many scenarios. We engaged with literatures on the uncertain nature of knowledge work, evaluation of new technologies, and the specific ways AI tools are evaluated. Iterating with our emerging concepts involved in managers' AI evaluation process, we focused even more specifically on literatures related to how knowledge is evaluated in professional contexts (and the importance of both know-what and know-how aspects of knowledge) compared to how it is evaluated in ML-based AI development communities (based on knowledge outputs). As we integrated and adapted concepts of know-how and know-what to the quality measures managers were evaluating, we were able to better conceptualize and theorize their AI evaluation process and its implications.

Findings

Focusing on Reported Claims of the High Quality of AI Tools

Underlying the rise of the new generation of ML-based AI tools were appealing promises of relieving human workloads without compromising quality. Urbanside managers, like

managers in many organizations, were grappling with the competing demands of coping with high work volumes while providing high-quality services: "This is our constant struggle: the 'bottom line' [increasing revenues and decreasing costs] versus the fact that we are here trying our best to do our job for our patients" (Irene). In late 2018, the field of diagnostic radiology experienced an explosion in AI development, marked by rapid increases in vendors, research articles (e.g., Langlotz 2019; Recht and Bryan 2017), educational resources,[4] and media coverage (e.g., Mukherjee 2017; The Economist 2018). Urbanside managers eagerly explored the growing landscape of diagnostic ML-based AI tools and their alluring promises to reduce costs while maintaining (or even increasing) diagnostic quality, as flaunted by one tool's slogan: "providing instant accuracy, anywhere, every time." Urbanside managers combed through academic journal articles, regulatory filings and approvals, and countless media stories and vendors touting impressive new AI tools. They also actively collaborated and partnered with members at the forefront of diagnostic AI development, which deepened their understanding of reports about tools' potential impacts, such as "decreasing the amount of time it takes to do one task" and "improving diagnosis by taking out some of the potential error" (Sadie).

Specifically, five ML-based AI tools stood out to Urbanside managers based on their reported high quality and alluring benefits. The five tools and their associated diagnostic contexts are detailed in Table 1. First, managers in the pediatric section were drawn to the Bone Age tool, which received high praise for its remarkably high accuracy and promises to speed up diagnosis times by eliminating a tedious process from radiologists' workflow. Second, managers in the chest section focused on the Chest Triage tool, based on the highest performing algorithm in a global ML competition and its promises to speed up diagnosis times by classifying urgent cases. Third, the Brain Tumor Segmentation tool, another global algorithmic competition winner, promised to improve diagnosis accuracy by measuring the volume of brain tumors (illustrated in Appendix A): a novel technique without existing standards. Fourth and fifth, developers of the Breast Ultrasound tool (vendor company) and Breast Mammo tool (Urbanside research group) had both received notoriety for their tool's ability to accurately diagnose breast cancer: "the [Breast Mammo tool's] neural network is using [image] *texture* much more than the radiologists are capable of doing" (Chris).

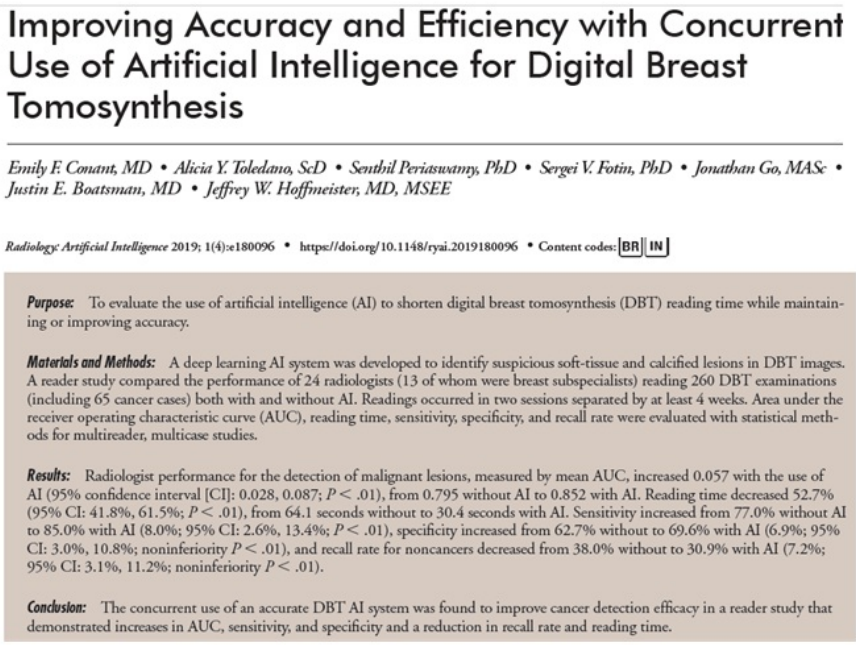
[4]See the American College of Radiology's Data Science Institute at www.acrdsi.org and Radiology Society of North America at www.rsna.org/education/ai-resources-and-training.

Assessing the Reported AUC Accuracy Measures of AI Tools

Assessing the reported accuracy measures of AI tools was critical to forming managers' opinions about the quality of the AI tools: "If there's a well-designed study that proves that the tool's accurate, I'd be comfortable with that" (Miguel). Measures reported in published studies were crucial, and managers specifically focused on reported AUC measures to assess tool quality. The AUC is an aggregated measure ranging from zero to one, that represents the accuracy of the model across different configurations; it summarizes how well the model's predicted outputs match the outputs predefined in the test data set by the developers. Professionals in the field of diagnostic radiology are trained to focus on AUC measures when judging whether *any* technological tool (far before and beyond AI) improves diagnostic accuracy (see example in Appendix B), from assessing the quality of imaging equipment (e.g., x-ray or MRI) to analytical tools built into imaging software (e.g., Tensor Flow Analysis, tomosynthesis). So, Urbanside managers readily focused on AUC measures for diagnostic AI tools and searched for tools "pushing that [AUC] curve into the upper left corner" (Vivian), that is, nearing the optimal score of 1.0. Tools approaching 1.0 were expected to consistently generate near-perfect diagnoses in practice (zero false positive and false negative errors).

AUC measures were reported prominently in the AI materials managers were consulting, including research publications, vendor documentation, regulatory applications, and patent filings. Very often, such materials reported AUC values as the primary evidence of performance, such as in medical journal abstracts that were open for free to the public (see example in Figure 1) and when vendor presentations dramatically revealed AUC values. Typically, a tool's AUC measure was reported *in comparison* to another method, usually expert radiologists or a competing tool, to suggest the tool's ability to improve diagnosis accuracy (see Appendix C).

Thus, managers regarded AUC measures as a short-hand indicator of the quality of an AI tool's outputs and error rates relative to experts', as Cyrus explained: "When we're talking about performance, it comes down to what is the accuracy of the method? It comes down to looking at the AUC number … At what point is the tool better than what we would do on our own?" Managers used the AUC measure to compare the AI outputs to experts' outputs when evaluating the Breast Mammo Tool, "The deep learning model is better than humans in terms of the common metrics used, in terms of AUC" (Chris), and the Bone Age tool, "The tool's performance was shown to be as good as, or slightly *better than*, radiologists' interpretations" (Nadia). Managers also used AUC measures to weigh relative costs of errors within a given



Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis

Emily F. Conant, MD • Alicia Y. Toledano, ScD • Senthil Periaswamy, PhD • Sergei V. Fotin, PhD • Jonathan Go, MASc • Justin E. Boatsman, MD • Jeffrey W. Hoffmeister, MD, MSEE

Radiology: Artificial Intelligence 2019; 1(4):e180096 • https://doi.org/10.1148/ryai.2019180096 • Content codes: BR IN

Purpose: To evaluate the use of artificial intelligence (AI) to shorten digital breast tomosynthesis (DBT) reading time while maintaining or improving accuracy.

Materials and Methods: A deep learning AI system was developed to identify suspicious soft-tissue and calcified lesions in DBT images. A reader study compared the performance of 24 radiologists (13 of whom were breast subspecialists) reading 260 DBT examinations (including 65 cancer cases) both with and without AI. Readings occurred in two sessions separated by at least 4 weeks. Area under the receiver operating characteristic curve (AUC), reading time, sensitivity, specificity, and recall rate were evaluated with statistical methods for multireader, multicase studies.

Results: Radiologist performance for the detection of malignant lesions, measured by mean AUC, increased 0.057 with the use of AI (95% confidence interval [CI]: 0.028, 0.087; $P < .01$), from 0.795 without AI to 0.852 with AI. Reading time decreased 52.7% (95% CI: 41.8%, 61.5%; $P < .01$), from 64.1 seconds without to 30.4 seconds with AI. Sensitivity increased from 77.0% without AI to 85.0% with AI (8.0%; 95% CI: 2.6%, 13.4%; $P < .01$), specificity increased from 62.7% without to 69.6% with AI (6.9%; 95% CI: 3.0%, 10.8%; noninferiority $P < .01$), and recall rate for noncancers decreased from 38.0% without to 30.9% with AI (7.2%; 95% CI: 3.1%, 11.2%; noninferiority $P < .01$).

Conclusion: The concurrent use of an accurate DBT AI system was found to improve cancer detection efficacy in a reader study that demonstrated increases in AUC, sensitivity, and specificity and a reduction in recall rate and reading time.

Figure 1. Details of Accuracy Measures and Select Model Details Included in the First Page Summary of an AI Research Article Published in *Radiology: Artificial Intelligence* (E. F. Conant, A. Y. Toledano, J. W. Hoffmeister, et al., "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis," *Radiology: Artificial Intelligence,* 2019. © Radiological Society of North America)

diagnostic context and grasp the tool's potential risks and benefits: "Would you rather have a model that … found all the abnormal diseases, but had a lot of false positives that the radiologist had to go through? Or one that is particularly specific but might miss something?" (Sadie).

Shifting Focus to the "Ground Truth" Used to Train and Validate ML-Based AI Tools

While assessing a tool's AUC measures, managers also began investigating the source of the "ground truth" created by the AI developers. Here, the term *ground truth* refers to the labels assigned to the data sets used to train a ML model to link new inputs to outputs and to validate its performance. Many diagnostic ML models relied on the diagnosis decisions of licensed radiologists as the ground truth labels, which was considered good practice for AI developers in this field, as illustrated by materials published by the Chest Triage tool creators: "Ground truth is critical in evaluating deep learning models in medical imaging and provide the foundation for clinical relevance when interpreting results in this field—this is why we focus a lot of our effort on considering the *best available ground truth via a panel of medical sub-specialist experts*."

Managers recognized that ground truth labels were integral to measuring ML-based AI performance, as they served as the baseline outputs to compare against the model's outputs. Higher skepticism was expressed towards a model that used diagnosis opinions provided by novices versus seasoned experts. This was illustrated by managers assessing the Bone Age tool, who were frustrated by the lack of visibility into the source of the model's ground truth labels, "If I knew the data was all labeled by pediatric radiologists who had read thousands and thousands of bone ages, I think I would consider [the model] to be probably more vetted, more trustworthy than if it had just been fed college students' reads who were just taught how to read them" (Nadia).

Managers searched for indicators of the caliber of the humans producing the ground truth labels. Indicators like the quantity, qualifications, and years of experience of the individuals labeling the data often appeared in published documentation and the limited summary section of research articles (see Figure 1). This was the case for the Brain Tumor Segmentation tool documentation describing how four physicians followed the same protocol to generate the data labels, which were then refined and approved by a certified neuroradiologist. These details were also uncovered in published regulatory filings, where managers found a detailed table

summarizing who provided ground truth labels for the Breast Ultrasound tool (see Appendix C).

Focusing on Conducting and Assessing Internal Pilot Studies

Satisfied with the reported performance claims for five AI tools, managers turned their focus towards evaluating them within their local work environment. In the case of the Bone Age tool, managers were eager to conduct pilot studies since the tool had previously "only been tested in the hospital where it was developed. How do we know if it will be applicable and function here? … We need to know: does the model generalize to our setting?" (Nadia). Assessing the tool's internal performance was critical before moving towards full implementation, as Savannah from the breast imaging department explained, "If the [Breast Ultrasound tool] study shows positive results, that means we [breast imaging radiologists] are going to be more productive, which is what we want. But we don't want to waste everyone's time before we know for sure."

Importantly, managers had to carefully define what internal measures of quality they would use to assess each tool's performance. In all five cases, managers hand-selected Urbanside experts to produce the ground truth diagnosis labels for each study. For the Bone Age tool, two senior radiologists recorded independent decisions, and the average of the two was recorded as the ground truth for each case. Then, AUC measures were generated to assess how well the AI outputs matched the opinions recorded by the experts. For the Chest Triage tool, the consensus of two chest radiologists' independent assessments was also used as the ground truth. In this case, managers also decided to adjust the label of one disease, *cardiomegaly*, from "abnormal" to "normal," which adjusted the volume of cases ultimately elevated to their prioritized worklist.

Managers debated the specifics of the ground truth labels they would use for the Brain Tumor Segmentation tool: "Is there intra-observer variability? If I do one, and you do one, if he does one, are we happy with that? Or does it then have to go through one person to check them all?" (Vivian). They ultimately decided to compare the tool's outputs to segmentation labels generated by a senior neuroradiologist. Finally, for both the Breast Mammo and Breast Ultrasound tools, managers decided to follow a common standard used in published AI studies: one radiologist's judgment would serve as the ground truth for the majority of cases, complemented by a three-month follow-up assessment or pathology findings when available.

As each study concluded, managers were quite surprised by the internal results, as many results conflicted with the accuracy measures reported by tool developers. For instance, when segmentation results of the Brain Tumor Segmentation tool were projected on the screen against the ground truth label, managers gazed at them in disbelief, "The segmentation is *atrocious*! It takes half the scalp with it. That makes *no sense* … Not only is it highlighting stuff in the brain that is *totally irrelevant*, but it doesn't actually highlight the *important* parts of the tumor!" (Vivian). Desperate to understand these conflicting results, managers searched for explanations, "The results were *far more discrepant* than I expected, and I don't know why" (Anthony).

How the Focus Shifted to Comparing Experts' Process to AI Processes

To investigate the underwhelming internal performance of the AI tools, managers began comparing experts' diagnosis approach to how the ML model determined its outputs. Unpacking the Bone Age tool's poor performance led managers to question whether the ML model was "taught" to use different standards than internal experts' professional training: "Is it possible that our internal standards are slightly different [than those used by the developers]? Or that we've been taught [to generate diagnoses] differently than the algorithm was taught?" (Anthony).

Managers focused on comparing the evidence represented in the data used to train the AI models to the evidence experts used to form diagnosis opinions. Doing so led managers to confront the limitations of models' training data in that only a narrow subset of the relevant diagnosis inputs was captured in the datasets underlying the ML model. In the case of the Breast Ultrasound tool, managers questioned the validity of a model whose outputs were based solely on two cropped ultrasound images, including Savannah remarking, "The software does not take into account certain clinical variables which are so very important," and Lola commenting, "There's a lot of *art of [patient's case] management* that's not accounted for by the tool."

Moreover, managers began scrutinizing *how* ground truth labels were defined and generated, extending their earlier focus on assessing *who* generated the labels. For the Breast Mammo tool, significant discrepancies were discovered between how the labelers generated the ground truth and experts' approach in their daily practice: "The readers were not looking at prior images. It was done intentionally, to keep it, you know, apples-to-apples, a controlled way to do the study. But it's not even close to real practice" (Lola).

Looking at prior images was an essential and fundamental analytical practice for radiologists; not doing so is considered a genuine act of malpractice in the field. Urbanside managers uncovered that experts labeling ground truth cases were only analyzing the current image, driven by developers' desire to draw an "apples-to-apples" comparison between the AI and human outputs. This new realization led managers to confront the limitations of using such measures to evaluate AI tools and increased their skepticism about the tools' reported performance claims.

The Unexpected Consequence of Evaluating AI Performance

Managers continued to dig deeper into how ground truth labels were constructed and ultimately confronted even deeper limitations of using human-generated labels: the deep uncertainty human experts face when generating diagnosis outputs. Managers struggled to reconcile the practice of using expert-generated labels as the ground truth when, in practice, these opinions lacked strong external validation, as explained by Sadie, "All we have for the ground truth is the radiologists … That doesn't necessarily mean that is the right answer, but it's what we have now," and Leslie, "There's no gold standard. The standard we use is the radiologists' read, but is the radiologist right? We don't know!"

Providing medical diagnoses is ambiguous and uncertain knowledge work, plagued with a lack of agreement about what constitutes an absolute or "accurate" opinion. After a radiologist diagnoses a patient, it is difficult to know for certain the accuracy of that diagnosis.[5] In practice, it was common for radiologists to form conflicting conclusions about a particular diagnosis, even when highly trained experts were presented with the same full set of medical information (e.g., De Sanctis et al. 2014; Duijm et al. 2009; Lazarus et al. 2006). Diagnostic errors are a major area of research in this professional field, reporting how such errors impact between 10 and 20 percent of cases, as many as one in five patients overall (Berner and Graber 2008; Bruno et al. 2015). As one Urbanside manager explained solemnly, "We all have misses. Interpretation is hard. It's not necessarily like you weren't looking or paying attention. It's like, you interpreted it, right? And we can be right or wrong when we interpret things" (Leslie).

Acknowledging these limitations, some measures of performance in the diagnostic radiology field are considered more or less reliable for determining the accuracy of a given judgment. Professional standards based on clinical outcomes, such as pathology reports and/or long-term follow-up records, for instance, are considered high quality measures of performance for breast cancer diagnosis because they represent the current best possible evidence to confirm the diagnosis. However, in many diagnostic scenarios (where standards are difficult to obtain in terms of time, cost, technology resources, and patient privacy concerns), the field recommends that one or multiple experts' diagnostic assessments may serve as a measure of diagnosis quality.[6] This was the scenario that managers faced when evaluating both of the AI tools for breast cancer diagnosis, where the professional standard (as well as the expected level of evidence required by most medical journals) would require pathology reports or long-term follow-up records. However, ML research commonly uses experts' diagnoses as the standard, and managers expressed discomfort drawing firm conclusions on that basis, given the high uncertainty and variability of experts' opinions: "The outcomes in breast cancer, by definition, need to be long-term outcomes. So, if some results are false negatives, you need *at least* a year to figure out if they were wrong … How valid are these results if they are not incorporating the long-term outcomes?" (Savannah).

In all five cases of AI tool evaluation, managers faced the crippling limitations of using expert-generated ground truth labels to evaluate ML-based AI tools. When unpacking the poor results of the Brain Tumor Segmentation tool pilot, managers began scrutinizing the ground truth labels they constructed internally: "How good is our standard for segmenting brain tumors? Let's take a look at what the heck was segmented and see how good is it?" (Vivian). Projecting one of the labels on the screen prompted concerned comments about its quality, "In the center, the core, there are some really patchy areas … It's very important we get that [label] right. We can't move forward with anything until we have a gold standard, something to measure the tool against" (Yanis). An extensive debate followed: they described in detail highly varied approaches to how they would have labeled the tumor's regions in practice. Multiple tumor segmenting standards were available in the neuroradiology field, each yielding different outputs. Thus, determining ground truth labels defining

[5]For instance, multiple conclusions may be drawn if a patient never returns to the physician: Was the diagnosis and subsequent treatment accurate and the patient recovered? Was the diagnosis inaccurate, but the patient recovered anyway? Was the diagnosis inaccurate, and the patient worsened, yet never returned to the original physician?

[6]Gathering such assessments are often part of standard medical practices, including routine peer reviews, regular conferencing with other physicians in a patient care team, in addition to the frequent impromptu meetings or phone calls wherein colleagues discuss a current diagnostic question, examine and debating case details together, until agreeing on a unified conclusion.

the brain tumor's edges was subject to deep underlying ambiguity: "According to our standard protocol for this task, that label is considered acceptable. But I think, maybe, that just isn't a good standard, in general" (Vivian).

Similarly, managers sought to understand why the Bone Age tool's outputs frequently diverged from the ground truth labels experts had generated internally. Managers began discussing the high variability between and within even highly trained experts at Urbanside and the field's lack of external methods to validate bone age diagnoses. Confronting the underlying uncertainty about the quality of human experts' knowledge led managers to question their ability to use expert-generated diagnoses to evaluate the accuracy of AI outputs: "I feel like my opinions might be different than other people's opinions. I also don't know how good the intra-observer rate is. And also, the *inter*-observer. I have no idea how good *we* [experts] are. So, is using our opinions a good way to train and test the machine? I don't know" (Nadia).

Outcomes of Evaluating AI Tools

In the end, managers' AI evaluation process, which is summarized in Table 2, resulted in three of the five piloted AI tools moving forward towards deeper exploration in the hospital, while the future for two tools was less clear. For the latter, the Breast Mammo and Breast Ultrasound tools, managers recognized the limitations of using ground truth measures based on experts' diagnoses and struggled to evaluate each tool's performance in a satisfying way: "There were a lot of situations where the model was wrong. There were also a lot of situations where the radiologist was wrong. There was a lot of discordant information" (Stella regarding the Breast Mammo tool). They considered the relative risks of using the tools, from the potential errors it may introduce to the potential increases in costs, as in the case of the Breast Ultrasound tool pilot suggesting slower diagnosis speeds, "For each lesion, you need to draw a little box into two different planes of view, so there's additional time for that extra maneuvering" (Savannah). These additional risks could not be justified since it was not clear what (if any) benefits the tool may generate if implemented.

In contrast, three AI tools were on a path of deeper exploration and potential implementation at Urbanside. Despite lacking a clear assessment of a tool's accuracy (given the underlying uncertainty of expert outputs), managers' AI evaluation process enabled them to have a clearer sense of a tool's relative risks and benefits. For the Bone Age and Brain Tumor Segmentation tools, managers were motivated by the underlying uncertainty and lack of established professional standards for the focal tasks. They viewed implementing these tools as a way to more deeply explore and potentially improve these diagnosis methods, which they viewed as outweighing the risk of the tool generating flawed outputs: "Even if the tool has some degree of error in it, it is still better than [our current method] for measuring the tumors– which is horrible! So almost anything would be an improvement" (Vivian). Furthermore, while the pilots did not analyze diagnosis speeds, managers were hopeful that the tools may increase efficiency, "It makes radiologists' lives easier. It shortens the times for us to measure the tumor" (Alvin).

Finally, managers decided that the benefits of moving forward with the Chest Triage tool outweighed its potential risks: "[The department chair] has already given his stamp of approval. I think as long as it's efficient, there's no questioning it." In this context, managers were notably more tolerant of potential AI errors. They were using the tool to "massively decrease turn-around-times" (Bob) for urgent cases, while still applying experts' full range of practices to make diagnosis decisions: "I'm okay with the model being wrong sometimes [overclassifying cases as urgent], as long as we don't *miss* [truly urgent] cases" (Leslie). Managers were ultimately driven by the promise that the AI tool may greatly improve patients' lives: "Getting to acute patients faster could impact their care and their health and their feelings and avoid possible complications. Isn't that the point of medicine? That's what [the tool] is doing" (Leslie).

Discussion

In this study, we unpack the process of evaluating ML-based AI tools in a context of professional knowledge work (medical diagnosis). Initially, managers focused on specific reports of high AI tool accuracy and assessed AUC measures and expert-generated ground truth labels published by tool creators. Managers chose five tools for further evaluation and conducted pilot studies to assess how well the AI outputs compared to internal experts' outputs. However, the pilots yielded disappointing results and left managers searching for explanations. As managers dug deeper, it brought them to confront the underlying challenge of evaluating the performance of human experts, as many diagnostic scenarios lack strong validation of the diagnosis outcome. As in many knowledge-intensive contexts, experts developed over the years rich know-how practices to form high-quality knowledge outputs. Thus, to evaluate AI outputs, managers began reflecting on the know-how practices that enable internal experts to grapple with uncertainty in their daily work and produce high-quality judgments. As a result, managers came

Table 2. Summarizing the Shifting Focus Across Quality Measures While Evaluating ML-Based AI Tools for Medical Diagnosis

Focus of Evaluation	Brain Tumor Segmentation Tool	Bone Age Tool	Breast Mammo Tool	Chest Triage Tool	Breast Ultrasound Tool
Reported claims of AI tools' high quality	Focusing on promises to increase diagnosis accuracy by adding a precise measurement to physicians' analysis	Focusing on promises to reduce diagnosis time and improve diagnosis accuracy with tool-generated assessment	Focusing on promises to improve diagnosis accuracy and possibly spend less time reading "normal" mammograms	Focusing on promises to reduce diagnosis times and improve outcomes by detecting and prioritizing urgent cases	Focusing on promises to improve diagnosis accuracy and possibly reducing time physicians spend equivocating
AI tools' reported AUC measures	Assessing the reported performance of 0.79, 0.91, 0.84 for matching experts' outputs for respective tumor regions	Assessing the reported accuracy of 0.989 for matching within 12 months of experts' assessment of normal vs. abnormal	Assessing the reported accuracy of 0.93 for classifying images with malignant findings, surpassing radiologists' AUC by 0.11	Assessing the reported accuracy of between 0.90 and 0.97 for 13 disease categories, and 0.85 for one disease category	Assessing the reported accuracy of 0.88, surpassing the mean accuracy of experts' by approximately 0.05
"Ground truth" used to train and validate AI tools	Consisted of manual segmentations drawn by one of four experts following the same protocol, then revised and approved by a board-certified neuroradiologist	Consisted of the average of the assessments of three fellowship-trained pediatric radiologists with nine, eight, and two years of post-fellowship experience	Consisted of biopsy results performed within 120 days of the mammogram	Consisted of diagnoses provided by four radiologists who had four, seven, 25, and 28 years of experience (one was subspecialty trained)	Consisted of pathology or a physician's diagnoses at 1-year follow-up (for cases not biopsied)
Internal measures of quality for pilot studies	Comparing tool outputs to segmentation labels provided by one senior neuroradiologist, then aggregated to generate AUC-like measures	Comparing tool outputs to diagnosis labels of the average of two pediatric radiologists' assessments, then aggregated to generate AUC measures	Comparing tool outputs to diagnosis labels provided by one breast radiologist, then aggregated to generate AUC-like measures	Comparing tool outputs to the consensus of two senior chest radiologists' assessments, then aggregated to generate AUC-like measures	Comparing tool outputs to pathology results (for cases with biopsy), 3-month follow-up of one radiologist (when recommended), or one radiologist's assessment (majority of cases), then aggregated to generate AUC-like measures
Comparing the AI process to experts' process	Comparing the pristine research-grade images used in the limited dataset to the messy and nuanced reality of measuring tumor development	Comparing how the tool was determining a given output to the multiple standards and protocols local experts were trained to use in practice	Comparing how experts produced ground truth labels using practices that did not adhere to their professional standards	Comparing how ground truth labels classified certain diseases as "abnormal" vs. "normal" and deciding to change those labels for their internal use case	Comparing the creators' decision of a limited scope of information to train the tool to the wide array of evidence experts consider in practice
Confronting limits of evaluating their own performance as experts	Confronting whether expert labels were a valid measure of quality when they lacked a single, agreed-upon standard for measuring brain tumor volume	Confronting whether the expert labels were a valid measure of quality when they are highly variable and subjective and based on multiple standards	Confronting the trade-offs of evaluating based on highly variable experts' diagnoses vs. the (highly expensive and often inaccessible) professional standard	Confronting whether experts' diagnoses are a valid quality measure when they are highly variable and subjective	Confronting the trade-offs of evaluating based on highly variable experts' diagnoses vs. the (highly expensive and often inaccessible) professional standard

Table 3. The Quality Measures of Know-What and Know-How Used to Evaluate ML-Based AI Tools

Measure	Description	How Managers Evaluated the Measure
Ground truth measures	• Data labels assigned to every data point in training and validation datasets which form the basis of a ML model's ability to match new inputs to outputs • Constructed by AI creators • Represented prominently in AI literature on ML	• Focusing on assessing experts and their outputs (know-what) that are used to train the model to generate its outputs. • Asking, how well do the labels represent "accurate" outcomes for a given input? • Asking, what is the caliber and qualification of the experts generating labels?
Professional standard of output quality	• The most accurate or best available benchmark for establishing the accuracy of a given knowledge output given current conditions • Set in the context of a professional field • Occasionally represented in AI literature • Standards vary from those that use aggregated opinions of experts to those derived from external validations derived from biopsies and long-term follow-up	• Focusing on assessing experts and their outputs (know-what) • Asking, does the professional field endorse the use of this standard? • Sometimes standards are represented in the AI literature, for example, does the ground truth use the best available standard established by the professional field?
AI accuracy measures	• In the image recognition models in our study, ML-based AI outputs are summarized by AUC measures which compare how well AI outputs match predefined ground truth labels • Represented prominently in AI literature and publicly reported documentation	• Focusing on assessing ML-based AI outputs (know-what) • Asking, how close is the AUC measure to 1.0? A measure of 1.0 suggests the ML-based AI model's know-what perfectly aligns with experts' know-what. • Asking, how does the ML-based AI model's error rates compare to experts' error rates? • Weighing the relative risks and benefits of tradeoffs between false positive and false negative types of errors.
Practically acceptable performance	• The accumulated professional know-how practices that enable experts to reach an adequate level of certainty in a situated problem context • Constituted in daily professional life • Rarely represented in AI literature on ML	• Demonstrating mastery of practices and problem-solving approaches guided by the knowledge system of a given professional field. • Acceptance that one may never achieve full certainty in practice given limits of many knowledge contexts, but that relying on and applying know-how enables acceptable levels of certainty in practice.

to recognize a troubling disconnect between ML-based AI quality measures that were based solely on know-what aspects of knowledge and the rich know-how practices experts rely on in their daily work. These realizations had profound implications for managers' AI evaluations and their assessment of each tool's potential risks and benefits.

A key insight of this study is uncovering the limitations of using know-what-based measures that ignore experts' know-how in evaluating ML-based AI tools for knowledge work. Table 3 summarizes our findings.

Quality measures based on know-what. Ground truth labels used for training and evaluating AI tools, professional standards of diagnostic output quality, and aggregated AI performance accuracy were critical to managers' AI evaluation process but eventually proved problematic due to their emphasis on know-what aspects of knowledge. Regarding ground truth measures, AI creators select ground truth labels that attempt to represent the "accurate" knowledge output for every input in training datasets. In this study, managers initially scrutinized the qualifications of the labelers and treated ground truth labels as taken-for-granted representations of knowledge in their field. Eventually, they recognized how even labels generated by experts limited their evaluations since experts' knowledge outputs were subject to deep underlying uncertainty and ignored know-how aspects of knowledge that were essential to producing knowledge in practice.

The second quality measure based on know-what is a professional standard of diagnostic output quality. These measures are set in the context of a professional field and have strong reputations for being the most accurate or best available benchmark to establish the accuracy of knowledge outputs. In medical diagnosis, for example, a professional standard for the accuracy of a diagnosis is often established using long-term patient outcomes or microscopic evidence. Managers faced the limits of professional standards of diagnostic quality for all five tools in this study. In two cases (bone age and brain tumor segmentation), multiple standards were used unsystematically and yielded varied (yet equally acceptable) diagnosis outputs. For the two breast cancer diagnosis tools,

expert-generated labels were commonly used for evaluation instead of enduring the costs of acquiring the level of evidence defined by the established professional standard.

Finally, aggregated AI performance accuracy measures were the third quality measure based on know-what that influenced and limited managers' AI evaluations. Managers focused specifically on AUC measures, a numeric representation of how well an AI model's outputs matched predefined ground truth labels. Managers viewed AUC measures as objective indicators of the quality of the AI tool. They often failed to appreciate how the AUC measure obscured the uncertain knowledge represented in ground truth labels and how it failed to account for experts' know-how practices.

Quality measure of know-how. This study distinguishes between the previous three quality measures of know-what from a quality measure based on know-how. This measure is associated with experts applying rich professional know-how practices to reach an adequate level of certainty for a situated problem context. In many fields, this involves experts accepting that full certainty may be impossible to achieve given the current state of knowledge in that field and practical costs and constraints. Using a quality measure of know-how in practice means evaluating experts based on their ability to demonstrate the range of tacit practices they have accumulated over time that enable them to form practically acceptable judgments. In our study, we show how managers ultimately confronted the disconnect between experts' know-how-based measures and the know-what-based measures used for evaluating ML-based AI tools.

Theoretical and Practical Implications

AI tools are rapidly emerging in many professional knowledge contexts using expert-provided labels as the ground truth (Mitchell et al. 1990; Smyth et al. 1994), yet the highly uncertain nature of experts' knowledge outputs is going largely unexamined. By identifying how ML-based AI tool evaluation focuses on know-what measures, while professional knowledge work focuses on know-how measures, we offer a number of theoretical implications and future research directions, which are examined in the following section and summarized in Table 4.

This study highlights the vital importance of conducting thorough evaluations of AI tools for contexts of expert knowledge work. Only through a deep process of unpacking the AI tools were managers able to understand AI performance levels and appreciate the potential risks and opportunities of adopting each specific tool. At the same time, recently, innovation in many fields has been accelerated (Lifshitz-Assaf et al. 2021). This became evident in the global pandemic when there was "gold rush" to AI for helping address COVID-19 (Gkeredakis et al. 2021). Our study's findings warn against the rushed adoption of AI tools, particularly for new problems wherein the underlying knowledge is uncertain or immature. The diligence of the evaluation process conducted by managers in our study exposed the gaps between the know-what of ML-based AI tools and experts' know-how. However, in many settings, AI is adopted without such diligent evaluation, and it might take a long time for the resulting damage to be captured. For instance, AI was adopted very rapidly for COVID-19 patient-related care decisions, including treatment choices and patient dismissal decisions. How could these systems be trained properly according to expert's standards? For instance, were ML models trained on data that included tracking of dismissed patients' clinical outcomes? Would there be a way for these ML models to learn if patients were erroneously discharged? Recently, ML researchers themselves started raised such concerns, finding serious flaws with many algorithms developed in the early stages of the pandemic (Roberts et al. 2021). A diligent evaluation of such tools could have surfaced such flaws. As some professional fields begin taking steps to formalize how AI ought to be evaluated (e.g., Mongan et al. 2020), our study speaks directly to what diligent evaluations should encompass.

The Tension Between Evaluating AI and Evaluating Experts

This study illuminates the strong tension between how ML-based AI tools are evaluated (using quality measures of know-what) and how experts evaluate their work (based on know-how). It is critical to examine and understand the limitations of any measures based on know-what aspects of knowledge (e.g., ground truth measures and AUC measures) that ignore the know-how. Prior literature describes how know-what and know-how aspects of knowledge are inherently inseparable (Polanyi 1966; Ryle 1949), in that all knowing emerges from and is rooted in situated practices (Brown and Duguid 1991, 2001; Lave 1988; Orlikowski 2002). Measures using know-what as the sole representation of knowledge are therefore inherently incomplete. In constructing ground truth labels and AUC measures based on know-what knowledge outputs, ML model developers divorce "a view of knowledge as a separate entity, static property, or stable disposition embedded in practice, [from] a view of knowledge as … enacted—every day and over time—in people's practices" (Orlikowski 2002, p. 250). And yet, despite that know-what measures ignore how "knowing is in our action" (Schön 1983, p. 49), they have become a prominent, taken-for-granted means of evaluating ML-based AI tools.

Table 4. Theoretical Implications of the ML-Based AI Tools Evaluation on Professional Knowledge Work

Focus of Evaluation	Evaluation Based on Know-What or Know-How	Relationship Between Know-What and Know-How	Questions for Future Research
Ground truth measures	Based on know-what	• Know-what knowledge outputs fail to capture the richness of experts' tacit know-how practices yet are taken for granted as "the truth" if produced by trained experts.	• When and how does an introduction of ML tools into a professional field increase the propensity to codify know-what of practice to enable ML tools?
Professional standard of output quality	Based on know-what	• Standards based on know-what may be taken-for-granted as objective and reliable knowledge in a field, but when the underlying knowledge is uncertain, such standards will not be met consistently. • Reliable professional standards of know-what are often costly to obtain at scale.	• How do professional standards of know-what in a field change with the introduction of ML tools? • Does the introduction of ML tools lead to convergence of professional standards of know-what even in areas where know-what knowledge is uncertain and standard creation is less mature? • Do professional standards "degrade" towards those that are easier to obtain at scale? • Will we see the rise of new types of professional standards pertaining to AI-augmented work?
Performanc e accuracy measures (AUC)	Based on know-what	• Aggregating accuracy measures into a single AUC measure obscures the trade-offs made in the construction of the underlying ground truth measures.	• How do AUC measures influence managerial decisions of whether to use ML tools to augment or replace human experts? • What are the risks of comparing human and ML tool performance based on an AUC measure, if ML tools are trained to optimize this measure, while human experts optimize multiple performance outputs?
Practically acceptable performance	Based on know-how	• Know-how is difficult to codify in know-what outputs since know-how is based on tacit, social, situated, and embedded action performed over time. • Know-how measures are the only "measure of quality" available when knowledge is highly uncertain. • As learning unfolds through reflection-in-action, know-how knowledge evolves; know-what knowledge cannot improve without know-how.	• When do tacit and contextual aspects of experts' know-how lead to greater know-what performance? • How can ML tools be developed to take into account more of experts' know-how, and would this impact the performance of ML tools in practice? • Will the adoption of ML tools that replace human experts lead to knowledge stagnation in a professional field due to the erosion of know-how knowledge? • How can "explainable AI" be used to compare machine classification processes with experts' know-how? How can this comparison be used to improve both ML-based AI tool performance and human know-how knowledge?

Currently, ML-based AI creators do not account for these limitations of know-what-based measures when reporting and advertising the performance of tools. Since the days of expert systems, computer science researchers have attempted to capture and codify experts' knowledge (Dreyfus et al. 2000; Forsythe 1993; Simon 1987) and experienced difficulties of faithfully representing tacit know-how in technological form (e.g., Hutchins 1995; Orlikowski 1992; Star 1989; Suchman 1987). Today, however, there is renewed hope that ML-based AI can bypass capturing tacit elements of experts' process by implicitly learning the patterns linking inputs to outputs. This assumes that what is being linked—the ground truth datasets defining the input and output—reliably represent knowledge in that domain. Issues are known to arise when constructing ground truth measures for ML models in contexts where knowledge claims are unreliable (Sheng et al. 2008), such as when outputs are disputed, subject to multiple interpretations, or even unavailable. While some researchers are actively developing benchmarks and methods for improving ground truth acquisition in ML research (Krig 2016; Milan et al. 2013), the majority of ML researchers avoid the problem altogether by selecting domains with less disputed outcomes, such as images of physical objects (e.g., Deng et al. 2009) or audio-signals (Mohamed et al. 2012).

In contrast to AI tool creators ignoring the limitations of know-what, members of professional fields widely acknowledge these limitations in their everyday practice and focus on exercising their know-how. Over decades, experts cultivate rich practices to build practically acceptable levels of certainty in their everyday work (Knorr Cetina 1999). Many professional fields hold their members accountable for their ability to adhere to this "professional knowledge system" (Abbott 1988) rather than enforcing a standard based solely on know-what outcomes. For instance, when radiologists are sued for a diagnostic error, they avoid malpractice charges by showing they adhered to their professional process standards, following all of the best available practices and protocols according to their know-how. Professionals must trust that their know-how will help them avoid negative outcomes, or else the uncertainty they face in their know-what may be paralyzing.

Our study also relates to the growing discourse and body of research focusing on "explainable AI" (e.g., Barredo Arrieta et al. 2020; Bauer et al. 2021; Fernández-Loría et al. 2020; Guidotti et al. 2018). Managers in this study were actively searching for ways to better understand the "black box" of the AI tools during their evaluations. Indeed, one of the key drivers behind the movement towards explainable AI research is the eagerness of managers performing AI evaluation to understand how the tools were developed and what processes they use to derive outputs. Unpacking such processes can help managers compare AI processes with experts' know-how. Such comparison may create an opportunity for learning, where it can reveal if a crucial part of know-how is not captured by the machine; vice-versa, it may enable human experts to learn from the AI processes. Moreover, if explainable AI for ML-based tools reveals inconsistencies in how machines classify similar input data (e.g., using different parts of the image every time a classification is done), this may prompt managers to question the reliability of such AI tools. Explainable AI, however, would not be able to address critical challenges that we have identified in this study associated with the limitations of the ground truth labels and the lack of uniform professional standards that characterize many areas of professional knowledge.

The Risky Consequences of Treating Constructed Quality Measures Objectively

Our study highlights the constructed nature of the measures commonly used to evaluate ML-based AI tools and the consequences of treating them as objective means of judging knowledge. Ground truth labels and AUC measures were often presented in quantified forms, which are known to increase the appearance of objectivity while hiding the situated, embodied, and equivocal nature of the underlying knowledge being represented (Bechmann and Bowker 2019; Espeland and Stevens 2008; Pentland 1993). However, despite the popularity of the term "ground truth," prior scholars have argued that such measures are far from objective or neutral, but are socially constructed and subject to ongoing debate and contestation (Bowker and Star 2000; Gitelman 2013; Latour 1987; Timmermans and Berg 2003). In particular, prior literature examining the constructed nature of categories and labels in (digital and nondigital) classification systems has emphasized the high degree of subjectivity, abstraction, and influence underlying decisions to form and define labels (Bechky 2021; Bowker and Star 2000). Recently, research has begun investigating the influential role of ML-based AI tool creators in constructing ground truth labels and arbitrating the "right" knowledge that AI tools should generate (Bechmann and Bowker 2019; Pasquale 2015). Further research is needed, however, to better understand how constructed ground truth labels shape the performance of ML-based AI tools and impact their adoption and use in organizational contexts.

Our study illuminates how ground truth measures were treated objectively, despite their constructed nature. Prior literature describes how knowledge that is removed from situated work processes begins to take on a more objective and static quality (Berger and Luckmann 1966; Latour 1987) that may then be represented in and associated with technological artifacts (Pentland 1995). Part of the process of forming judgments about new technologies has been shown to involve individuals questioning the measures, artifacts, and quantifications meant to represent knowledge in a given field (Anthony 2018; Espeland and Stevens 2008; Pentland 1993). Attending to these issues involves actively unfolding and scrutinizing the technology and its related artifacts, "unraveling of the features of physical and technical objects, of their details, composition, hidden sequences, and behavioral implications" (Knorr Cetina 1999, pp. 71-72).

In our study, the spell of objectivity surrounding AI measures that are based on know-what was broken when managers began comparing these measures to experts' rich professional know-how and recognizing the serious disconnect. Focusing on a specific ground truth label (previously assumed to represent the "accurate" diagnosis), they recontextualized it within its situated problem scenario and analyzed the diagnosis as an expert would in their daily practice. After discussing and debating the patient's specific condition, potential imaging nuances, and conflicting evidence of disease progressions, they concluded that multiple diagnoses were equally likely and recording any one as the "ground truth" was unacceptable. In other words, as managers drew on experts' rich and multifaceted know-how practices, they reconsidered the

validity of ground truth labels based solely on experts' know-what knowledge outputs.

Understanding the constructed nature of AI performance claims is not only useful for managers evaluating AI tools; it is also critical for anyone engaging with or referencing quality measures of know-what for academic or policy-related purposes. Our study points to the risks of taking-for-granted measures based on know-what and ignoring their limitations when incorporating them into aggregated studies and performance claims. Take, for instance, ImageNet models' high AUC measures (Gershgorn 2017; Parloff 2016) which are commonly cited as clear evidence of AI's strengths and abilities to support analyses of employment and automation trends (e.g., Dhar 2016; Frey and Osborne 2017; Seamans and Furman 2019). Our findings would urge for explicitness about the constructed nature of the measures underlying ImageNet performance claims and the limitations of being built upon popular datasets using crowdsourced image labeling as the ground truth. Explicitness about possible risks and limitations is critical, since, as this study showed, these claims often serve as the driving force behind the adoption of highly consequential AI tools at growing scales.

Relying on AI Outputs May Severely Limit Learning

Finally, if ML-based AI tools are indeed implemented at scale in knowledge-intensive contexts, the dilemmas that surfaced in this study would be further exacerbated and lead to major consequences. Namely, organizational and professional learning processes may disappear if know-what-based AI outputs dominate decision-making contexts and erode experts' tacit know-how. Once AI tools are perceived as taken-for-granted and objective (Berger and Luckmann 1966; Latour 1987), and experts rely on seemingly accurate AI outputs as a welcome reprieve from their uncertainty, fundamental organizational changes may follow. Moving forward, AI tools and their trusted outputs may influence the social and technological ensemble that generates the very data on which the tool is trained (Faraj et al. 2018; Orlikowski and Scott 2014; Pachidi et al. 2021). For instance, in a recent study, Pachidi et al. (2021) find that sales professionals responded to a new tool by continuing to work according to their expert know-how and used the tool in symbolic ways only. However, their perfunctory use generated new data which further trained and legitimized the predictive model, and, ironically, resulted in the entire sales staff being laid off.

In the future, a "new know-how," which is augmented and influenced by AI outputs, may eventually represent the sole remaining source of knowledge in an organization and stunt the possibility for learning (in addition to improving the tool). Organizational researchers have theorized about the dynamic way of knowing transforms through humans "interacting, discovering 'truth,' justifying observations, defining problems, and solving them," which fuels how "knowledge alternates between tacit knowledge that may give rise to new explicit knowledge and vice versa … Tacit and explicit knowledge mutually enhance each other towards increasing the capacity to act" (Nonaka and von Krogh 2009, p. 638). This process is likely to dissolve if AI tools' explicit outputs overshadow the tacit aspects of experts' knowledge processes and know-how (Feldman 2004). If, however, both humans and AI tools operate in parallel, both retain the potential to learn and evolve, and there is the important potential to continue observing and scrutinizing the performance of both over time. However, operating in parallel is highly difficult and expensive in practice, especially as organizational leaders urge AI tool adoption based on promised efficiency gains. Based on this study, we suggest two possible ways of addressing these concerns. In areas where scientific knowledge is highly uncertain, human experts must remain the final arbiter of decision-making in practice. For fields with more established knowledge claims, based on our study, we recommend that AI tools should be trained and validated on quality measures that more closely resemble know-how and experts' practically acceptable performance.

Ever since the Turing Test, the performance of computers has been measured in head-to-head comparisons with the performance of humans. Today, the sentiment has not changed, and comparisons are routinely drawn between AI tools and human experts in growing numbers of domains. Moreover, today AI tools are being developed for increasingly more critical individual, organizational, and societal decisions. Such decisions, however, are often embedded in contexts where human experts routinely experience uncertainty-producing judgments and, as a result, rely on their know-how to address limitations of their know-what. In such contexts, organizational actors should be cautious in developing and adopting AI tools that are based on human experts' know-what knowledge. Such tools may not only produce poor decisions that are consequential but may also limit our ability to learn how to improve such decisions in the future.

Acknowledgments

We would like to thank the anonymous reviewers and the special issue editors for their instrumental comments and insights throughout the review process. We appreciate the helpful feedback provided by Foster Provost, Beth Bechky, and Susan Scott as well as researchers at the ICIS 2020 AI in Practice PDW and in the Work in the Age of Intelligent Machines (WAIM) community. Finally, we are deeply grateful for and inspired by the individuals at "Urbanside" who generously allowed us to study their daily work.

References

Abbott, A. 1988. *The System of Professions: An Essay on the Division of Expert Labor*, Chicago: University of Chicago Press.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. "Machine Bias," *ProPublica* (https:// www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Anthony, C. 2018. "To Question or Accept? How Status Differences Influence Responses to New Epistemic Technologies in Knowledge Work," *Academy of Management Review* (43:4), pp. 661-679.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., and Shetty, S. 2019. "End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography," *Nature Medicine* (25:6), pp. 954-961.

Autor, D. H. 2015. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *The Journal of Economic Perspectives* (29:3), pp. 3-30.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82-115.

Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. 2021. "Expl(AI)n It to Me—Explainable AI and Information Systems Research," *Business & Information Systems Engineering* (63:2), pp. 79-82.

Bechky, B. 2003. "Object Lessons: Workplace Artifacts as Representations of Occupational Jurisdiction," *American Journal of Sociology* (109:3), pp. 720-752.

Bechky, B. A. 2021. *Blood, Powder, and Residue: How Crime Labs Translate Evidence into Proof*, Princeton, NJ: Princeton University Press.

Bechmann, A., and Bowker, G. C. 2019. "Unsupervised by Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media," *Big Data & Society* (6:1).

Berger, P., and Luckmann, T. 1966. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*, New York: Anchor Books.

Berner, E. S., and Graber, M. L. 2008. "Overconfidence as a Cause of Diagnostic Error in Medicine," *The American Journal of Medicine* (121:5, Supplement), pp. S2-S23.

Bowker, G. C., and Star, S. L. 2000. *Sorting Things Out: Classification and Its Consequences*, Cambridge, MA: MIT Press.

Brown, J. S., and Duguid, P. 1991. "Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation," *Organization Science* (2:1), pp. 40-57.

Brown, J. S., and Duguid, P. 2001. "Knowledge and Organization: A Social-Practice Perspective," *Organization Science* (12:2), pp. 198-213.

Bruno, M. A., Walker, E. A., and Abujudeh, H. H. 2015. "Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction," *RadioGraphics* (35:6), pp. 1668-1676.

Chandrasekaran, B., Mittal, S., and Smith, J. 1980. "RADEX–Towards a Computer-Based Radiology Consultant," in *Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal (eds.), Amsterdam: North Holland, pp. 463-474.

Charmaz, K. 2014. *Constructing Grounded Theory*, Thousand Oaks, CA: SAGE Publications.

Christin, A. 2014. "When it Comes to Chasing Clicks, Journalists Say One Thing but Feel Pressure to Do Another," *Nieman Lab*. (http://www.niemanlab.org/2014/08/when-it-comes-to-chasing-clicks-journalists-say-one-thing-but-feel-pressure-to-do-another/, accessed October 18, 2018).

Christin, A. 2020. "The Ethnographer and the Algorithm: Beyond the Black Box," *Theory and Society* (49:5), pp. 897-918.

Conant, E. F., Toledano, A. Y., Periaswamy, S., Fotin, S. V., Go, J., Boatsman, J. E., and Hoffmeister, J. W. 2019. "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis," *Radiology Artificial Intelligence* (1:4) (DOI: 10.1148/ryai.2019180096).

Croft, C., Currie, G., and Lockett, A. 2015. "Broken 'Two-Way Windows'? An Exploration of Professional Hybrids," *Public Administration* (93:2), pp. 380-394.

De Sanctis, V., Di Maio, S., Soliman, A. T., Raiola, G., Elalaily, R., and Millimaggi, G. 2014. "Hand X-Ray in Pediatric Endocrinology: Skeletal Age Assessment and Beyond," *Indian Journal of Endocrinology and Metabolism* (18:Suppl 1), pp. S63-S71.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li F. F. 2009. "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA: IEEE Computer Society Press, pp. 248-255.

Dhar, V. 2016. "When to Trust Robots with Decisions, and When Not To," *Harvard Business Review*.

Dougherty, D., and Dunne, D. D. 2012. "Digital Science and Knowledge Boundaries in Complex Innovation," *Organization Science* (22:5), pp. 1467-1484.

Dreyfus, H., Dreyfus, S. E., and Athanasiou, T. 2000. *Mind Over Machine*, New York: Simon and Schuster.

Duijm, L. E. M., Louwman, M. W. J., Groenewoud, J. H., van de Poll-Franse, L. V., Fracheboud, J., and Coebergh, J. W. 2009. "Inter-Observer Variability in Mammography Screening and Effect of Type and Number of Readers on Screening Outcome," *British Journal of Cancer* (100:6), pp. 901-907.

Espeland, W. N., and Stevens, M. L. 2008. "A Sociology of Quantification," *European Journal of Sociology/Archives Européennes de Sociologie* (49:3), pp. 401-436.

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), pp. 62-70.

Feldman, S. P. 2004. "The Culture of Objectivity: Quantification, Uncertainty, and the Evaluation of Risk at NASA," *Human Relations* (57:6), pp. 691-718.

Felten, E. W., Raj, M., and Seamans, R. 2018. "A Method to Link Advances in Artificial Intelligence to Occupational Abilities," *AEA Papers and Proceedings* (108), pp. 54-57.

Fernández-Loría, C., Provost, F., and Han, X. 2020. "Explaining Data-Driven Decisions Made by AI Systems: The Counterfactual Approach," *ArXiv:2001.07417* (http://arxiv.org/abs/2001.07417).

Forsythe, D. E. 1993. "Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence," *Social Studies of Science* (23:3), pp. 445-477.

Frey, C. B., and Osborne, M. A. 2017. "The Future of Employment: How Susceptible Are Jobs to Computerisation?," *Technological Forecasting and Social Change* (114), pp. 254-280.

Garud, R. 1997. "On the Distinction between Know-How, Know-What, and Know-Why," *Advances in Strategic Management* (14), pp. 81-102.

Gershgorn, D. 2017. "The Data That Transformed AI Research—and Possibly the World," *Quartz*. (https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/).

Gitelman, L. 2013. *"Raw Data" Is an Oxymoron*, Cambridge, MA: MIT Press.

Gkeredakis, M., Lifshitz-Assaf, H., and Barrett, M. 2021. "Crisis as Opportunity, Disruption and Exposure: Exploring Emergent Responses to Crisis through Digital Technology," *Information and Organization* (31:1).

Glaser, B., and Strauss, A. 1967. *Discovering Grounded Theory*, Chicago: Aldine Publishing Company.

Golden-Biddle, K., and Locke, K. 2007. *Composing Qualitative Research*, Thousand Oaks, CA: SAGE Publications.

Grady, D. 2019. "A.I. Took a Test to Detect Lung Cancer. It Got an A," *The New York Times*, Health, May 20 (https://www.nytimes.com/2019/05/20/health/cancer-artificial-intelligence-ct-scans.html).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2018. "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* (51:5), Article 93.

Hutchins, E. 1995. *Cognition in the Wild*, Cambridge, MA: MIT Press.

Knorr Cetina, K. 1999. *Epistemic Cultures. How the Sciences Make Knowledge*, Cambridge, MA: Harvard University Press.

Knorr Cetina, K. 2016. "What If the Screens Went Black? The Coming of Software Agents," in *Beyond Interpretivism? New Encounters with Technology and Organizations*, L. Introna, D. Kavanah, S. Kelly, W. Orlikowski, and S. Scott (eds.), Berlin: Springer.

Kogut, B., and Zander, U. 1992. "Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology," *Organization Science* (3:3), pp. 383-397.

Kohavi, R., and Provost, F. 1998. "Glossary of Terms," *Machine Learning* (30), pp. 271-274.

Krig, S. 2016. "Ground Truth Data, Content, Metrics, and Analysis," Chapter 7 in *Computer Vision Metrics: Survey, Taxonomy, and Analysis*, Cham: Springer International Publishing, pp. 247-271.

Langlotz, C. P. 2019. "Will Artificial Intelligence Replace Radiologists?," *Radiology: Artificial Intelligence* (1:3), p. e190058.

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, MA: Harvard University Press.

Lave, J. 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*, Cambridge, UK: Cambridge University Press.

Lazarus, E., Mainiero, M. B., Schepps, B., Koelliker, S. L., and Livingston, L. S. 2006. "BI-RADS Lexicon for US and Mammography: Interobserver Variability and Positive Predictive Value," *Radiology* (239:2), pp. 385-391.

Lebovitz, S. 2019. "Diagnostic Doubt and Artificial Intelligence: An Inductive Field Study of Radiology Work," in *Proceedings of the 40th International Conference on Information Systems*, Munich.

Lehman, C. D., Arao, R. F., Sprague, R. B., Lee, J. M., Buist, D. S. M., Kerlikowske, K., Henderson, L. M., Onega, T., Tosteson, A. N. A., Rauscher, G. H., and Miglioretti, D. L. 2017. "National Performance Benchmarks for Modern Screening Digital Mammography: Update from Breast Cancer Surveillance Consortium," *Radiology* (283:1), pp. 49-58.

Leonard-Barton, D. 1995. *Wellsprings of Knowledge: Building and Sustaining the Sources of Innovation*, Boston, MA: Harvard Business School Press.

Lifshitz-Assaf, H., Lebovitz, S., and Zalmanson, L. 2021. "Minimal and Adaptive Coordination: How Hackathons' Projects Accelerate Innovation Without Killing It," *Academy of Management Journal* (64:3), pp. 684-715.

McGivern, G., Currie, G., Ferlie, E., Fitzgerald, L., and Waring, J. 2015. "Hybrid Manager-Professionals' Identity Work: The Maintenance and Hybridization of Medical Professionalism in Managerial Contexts," *Public Administration* (93:2), pp. 412-432.

Menchik, D. A. 2014. "Decisions about Knowledge in Medical Practice: The Effect of Temporal Features of a Task," *American Journal of Sociology* (120:3), pp. 701-749.

Mengis, J., Nicolini, D., and Swan, J. 2018. "Integrating Knowledge in the Face of Epistemic Uncertainty: Dialogically Drawing Distinctions," *Management Learning* (49:5), pp. 595-612.

Milan, A., Schindler, K., and Roth, S. 2013. "Challenges of Ground Truth Evaluation of Multi-Target Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 735-742.

Mitchell, T. M., Mabadevan, S., and Steinberg, L. I. 1990. "LEAP: A Learning Apprentice for VLSI Design," in *Machine Learning*, Y. Kodratoff and R. S. Michalski (eds.), San Francisco: Morgan Kaufmann, pp. 271-289.

Mohamed, A., Dahl, G. E., and Hinton, G. 2012. "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing* (20:1), pp. 14-22.

Mongan, J., Moy, L., and Kahn, C. E. 2020. "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers," *Radiology: Artificial Intelligence* (2:2), p. e200029.

Moran, G. 2018. "This Artificial Intelligence Won't Take Your Job, it Will Help You Do it Better," *Fast Company*. (https://www.fastcompany.com/90253977/this-artificial-intelligence-wont-take-your-job-it-will-help-you-do-it-better).

Mukherjee, S. 2017. "A.I. Versus M.D.," *The New Yorker*, Annals of Medicine (https://www.newyorker.com/ magazine/2017/04/03/ai-versus-md).

Nicolini, D. 2009. "Zooming In and Out: Studying Practices by Switching Theoretical Lenses and Trailing Connections," *Organization Studies* (30:12), pp. 1391-1418.

Nicolini, D. 2012. *Practice Theory, Work, and Organization: An Introduction*, Oxford, UK: Oxford University Press.

Nonaka, I., and von Krogh, G. 2009. "Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory," *Organization Science* (20:3), pp. 635-652.

Northrup, J. 2005. "The Pharmaceutical Sector," in *The Business of Healthcare Innovation*, L. R. Burns (ed.), Cambridge, UK: Cambridge University Press, pp. 27-102.

Oakden-Rayner, L. 2019. "The Rebirth of CAD: How Is Modern AI Different from the CAD We Know?," *Radiology: Artificial Intelligence* (1:3).

Orlikowski, W. J. 1992. "The Duality of Technology: Rethinking the Concept of Technology in Organizations," *Organization Science* (3:3), pp. 398-427.

Orlikowski, W. J. 2002. "Knowing in Practice: Enacting a Collective Capability in Distributed Organizing," *Organization Science* (13:3), pp. 249-273.

Orlikowski, W., and Scott, S. 2014. "What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector," *Organization Science* (25:3), pp. 868-891.

Pachidi, S., Berends, H., Faraj, S., and Huysman, M. 2021. "Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing," *Organization Science* (32:1), pp. 18-41.

Parloff, R. 2016. "From 2016: Why Deep Learning Is Suddenly Changing Your Life," *Fortune*, September 28 (https://fortune.com/longform/ai-artificial-intelligence-deep-machine-learning/).

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information* (Reprint ed.), Cambridge, MA: Harvard University Press.

Pentland, B. T. 1993. "Getting Comfortable with the Numbers: Auditing and the Micro-Production of Macro-Order," *Accounting, Organizations and Society* (18:7), pp. 605-620.

Pentland, B. T. 1995. "Information Systems and Organizational Learning: The Social Epistemology of Organizational Knowledge Systems," *Accounting, Management and Information Technologies* (5:1), pp. 1-21.

Pinch, T., and Bijker, W. 1987. "The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other," in *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, T. P. Hughes, W. Bijker, and T. Pinch (eds.), Cambridge, MA: MIT Press, pp. 17-50.

Polanyi, M. 1958. *Personal Knowledge: Towards a Post-Critical Philosophy*, Chicago: University of Chicago Press.

Polanyi, M. 1966. *The Tacit Dimension*, Chicago: University of Chicago Press.

Provost, F., and Fawcett, T. 2001. "Robust Classification for Imprecise Environments," *Machine Learning* (42:3), pp. 203-231.

Provost, F., and Fawcett, T. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, Sebastopol, CA: O'Reilly Media.

Rao, A. S., and Verweij, G. 2017. "Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise?," Report, September 8, PricewaterhouseCoopers Australia

Reardon, S. 2019. "Rise of Robot Radiologists," *Nature* (576:7787), pp. S54-S58.

Recht, M., and Bryan, R. N. 2017. "Artificial Intelligence: Threat or Boon to Radiologists?," *Journal of the American College of Radiology* (14:11), pp. 1476-1480.

Rindova, V., and Courtney, H. 2020. "To Shape or Adapt: Knowledge Problems, Epistemologies, and Strategic Postures under Knightian Uncertainty," *Academy of Management Review* (45:4), pp. 787-807.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., and Schönlieb, C.-B. 2021. "Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans," *Nature Machine Intelligence* (3:3), pp. 199-217.

Ryle, G. 1949. *The Concept of Mind*, London: Hutcheson.

Schön, D. A. 1983. *The Reflective Practitioner: How Professionals Think in Action*, New York: Basic Books.

Seamans, R., and Furman, J. 2019. "AI and the Economy," *Innovation Policy and the Economy* (19:1), pp. 161-191.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. 2008. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, pp. 614-622.

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., Lyons, T., Etchemendy, J., Grosz, B., and Bauer, Z. 2018. "The AI Index 2018 Annual Report," Stanford, CA: AI Index Steering Committee, Human-Centered AI Initiative, Stanford University.

Simon, H. A. 1987. "Making Management Decisions: The Role of Intuition and Emotion," *The Academy of Management Executive (1987-1989)* (1:1), pp. 57-64.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. 1994. "Inferring Ground Truth from Subjective Labelling of Venus Images," *Advances in Neural Information Processing Systems*, pp. 1085-1092.

Spradley, J. 1979. *The Ethnographic Interview*, New York: Holt, Rinehart and Winston.

Star, S. L. 1989. "The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving," Chapter 2 in *Distributed Artificial Intelligence*, L. Gasser and M. N. Huhns (eds.), San Francisco: Morgan Kaufmann, pp. 37-54.

Star, S. L. 1995. *Ecologies of Knowledge: Work and Politics in Science and Technology*, Albany, NY: SUNY Press.

Suchman, L. 1987. *Plans and Situated Actions*, Cambridge, UK: University of Cambridge Press.

Szulanski, G. 1996. "Exploring Internal Stickiness: Impediments to the Transfer of Best Practice within the Firm," *Strategic Management Journal* (17:S2), pp. 27-43.

The Economist. 2018. "AI, Radiology and the Future of Work," *The Economist*, June 8 (https://www.economist.com/leaders/2018/06/07/ai-radiology-and-the-future-of-work).

Timmermans, S., and Berg, M. 2003. *The Gold Standard: The Challenge of Evidence-Based Medicine*, Philadelphia: Temple University Press.

Van Den Broek, E., Sergeeva, A., and Huysman, M. 2020. "Managing Data-Driven Development: An Ethnography of Developing Machine Learning for Recruitment," *Academy of Management Proceedings* (2020:1), p. 17689.

Van Maanen, J. 1998. *Qualitative Studies of Organizations*, Thousand Oaks, CA: SAGE Publications.

von Hippel, E. 1988. *The Sources of Innovation*, Oxford, UK: Oxford University Press.

Waardenburg, L., Sergeeva, A., and Huysman, M. 2018. "Hotspots and Blind Spots," in *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology*, U. Schultze, M. Aanestad, M. Mähring, C. Østerlund, and K. Riemer (eds.), New York: Springer, pp. 96-109.

Walch, K. 2019. "The Growth of AI Adoption in Law Enforcement," *Forbes*, July 26 (https://www.forbes.com/sites/cognitiveworld/2019/07/26/the-growth-of-ai-adoption-in-law-enforcement/).

Weissmann, J. 2018. "Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women," *Slate*, October 10 (https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html; accessed July 28, 2019).

About the Authors

Sarah Lebovitz received her Ph.D. from New York University's Stern School of Business and is currently an assistant professor at the University of Virginia's McIntire School of Commerce. Her main research interests are in understanding how emerging technologies are adopted in organizations and how they impact professionals and their knowledge work practices. Currently, based on a one-year qualitative field study, her research examines how machine-learning-based AI tools are evaluated and used to make consequential medical diagnosis decisions. Her recent work also investigates hackathons to understand how innovation processes are being accelerated and transformed by new technologies such as 3D printing and open-source platforms.

Natalia Levina received her Ph.D. in Information Technology from MIT's Sloan School of Management and is a Toyota Motors Corp Term Professor at New York University's Stern School of Business. She also holds a Research Environment Professor part-time position at the Warwick Business School. Her main research interest is in understanding how people span organizational, professional, cultural, and other boundaries while developing and using new technologies. Currently, her research explores the evaluation and adoption of AI in medicine, diverse modes of open innovation, theories of smart contracts, epistemic underpinnings of entrepreneurship and innovation, and firm-community interaction in crowdsourcing. She served in a number of editorial positions at *Information Systems Research, MIS Quarterly, Organization Science,* and *Information & Organization*. She co-founded and chaired the AIS Special Interest Group on Grounded Theory Method.

Hila Lifshitz-Assaf is an associate professor at New York University's Stern School of Business. She is also a faculty associate at Harvard University's Lab for Innovation Science. Her research focuses on developing an in-depth empirical and theoretical understanding of the micro-foundations of scientific and technological innovation and knowledge creation processes in the digital age. She investigates new forms of organizing for innovation such as crowdsourcing, open source, Wikipedia, hackathons, and artificial intelligence. She earned a doctorate from Harvard Business School. Her dissertation study was an in-depth 3-year longitudinal field study of NASA's experimentation with open innovation online platforms and communities, resulting in a scientific breakthrough. Her work received the prestigious INSPIRE grant from the National Science Foundation and multiple best paper awards.

Appendix A

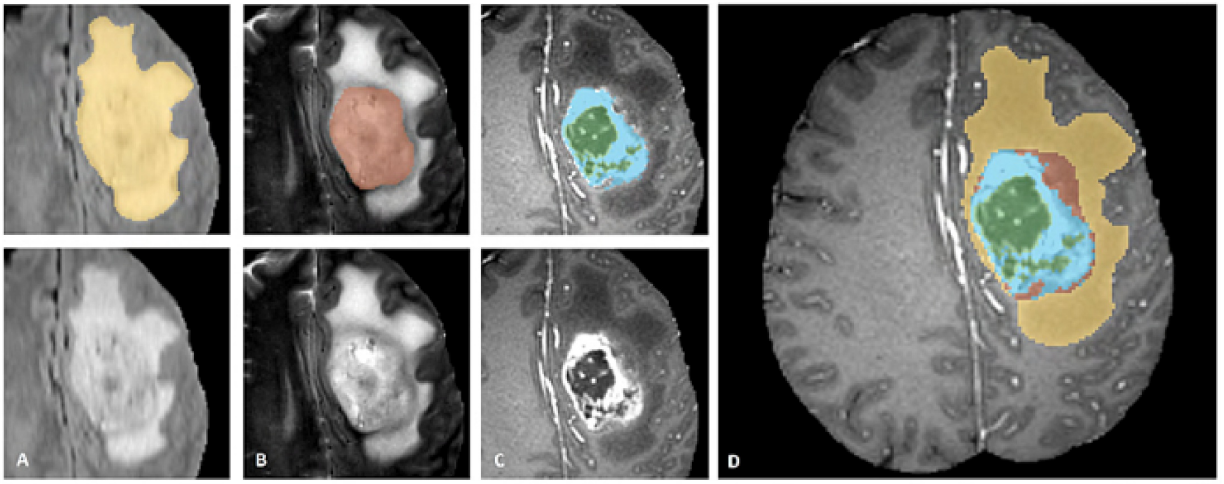


Fig.1: Glioma sub-regions. Shown are image patches with the tumor sub-regions that are annotated in the different modalities (top left) and the final labels for the whole dataset (right). The image patches show from left to right: the whole tumor (yellow) visible in T2-FLAIR (Fig.A), the tumor core (red) visible in T2 (Fig.B), the enhancing tumor structures (light blue) visible in T1Gd, surrounding the cystic/necrotic components of the core (green) (Fig. C). The segmentations are combined to generate the final labels of the tumor sub-regions (Fig.D): edema (yellow), non-enhancing solid core (red), necrotic/cystic core (green), enhancing core (blue). (Figure taken from the BraTS IEEE TMI paper.)

Figure A1. Featured on the Brain Tumor Segmentation Tool Website (The tool promises to segment three tumor regions (small images, left), which are combined in the final output (far right image).)

Appendix B

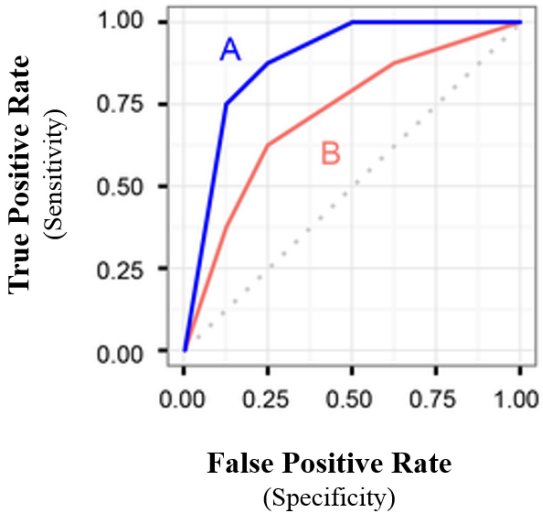


Figure B1. AUC Graph Depicting the Superior Performance of Tool A (e.g., 3D imaging for breast cancer screening) over tool B (e.g., 2D imaging)

Appendix C

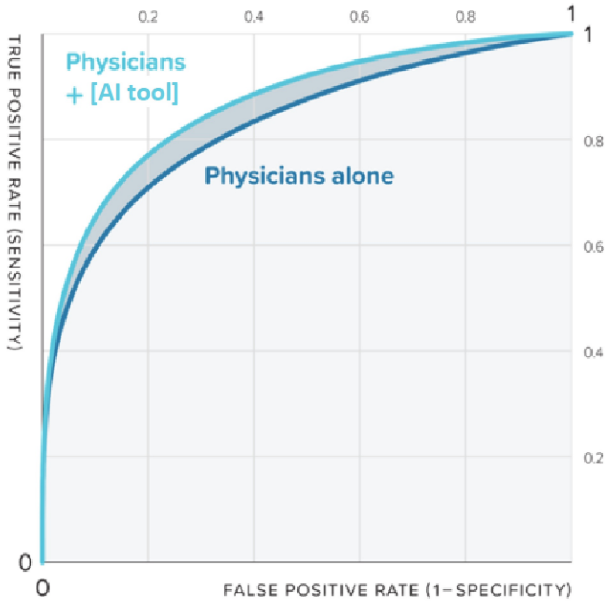


Figure C1. AUC Graph Displayed on Breast Ultrasound Tool Vendor's Website

Reader ID	Board Certification/Specialty	Breast Fellowship Trained and/or Dedicated Breast Imager	Years of Experience - Mammography and/or Breast Ultrasound	Academic Institution Affiliation (Yes/No)	MQSA Qualified Interpreting Physician
1	Diagnostic Radiology	No	13 Years	No	Yes
2	Diagnostic Radiology	No	4 Years	No	No
3	Diagnostic Radiology	Yes	7 Years	Yes	Yes
4	Breast Surgeon	No	0 Years	No	No
5	OB/GYN	No	20 Years	No	No
6	Diagnostic Radiology	No	13 Years	Yes	No
7	Diagnostic Radiology	No	3 Years	Yes	No
8	OB/GYN	No	0 Years	No	No
9	Diagnostic Radiology	Yes	15 Years	No	Yes
10	Diagnostic Radiology	No	13 years	No	No
11	Diagnostic Radiology	Yes	30 Years	No	Yes
12	Diagnostic Radiology	Yes	10 Years	Yes	Yes
13	Diagnostic Radiology	No	0 Years	No	No
14	Interventional Radiology	No	4 Years	No	No
15	Breast Surgeon	No	25 Years	Yes	No

Figure C2. FDA Filing for Breast Ultrasound Tool Showing Details of Physicians Involved in the Tool's Validation Study