The University of Chicago

Artificial Intelligence Techniques in Health Diagnostics: A Systematic Review

Exploring the Current State of AI Diagnostic Tools, Physician Perspectives on AI in Clinical Settings, and Future Implications upon Health Care

Josephine Yau

A thesis submitted in partial fulfillment of the requirements for a
Bachelor of Arts in Public Policy Studies
at the University of Chicago

Paper presented to:
Faculty Advisor, Professor Maria Bautista
Preceptor, Rubina Hundal

April 15, 2024

Abstract

With the development of artificial intelligence (AI) capabilities over the past years, many industries have integrated AI-based tools into their workflow to improve productivity. Healthcare is no exception; AI use in medicine ranges from managing patient data to improving the efficiency of diagnosis. In diagnostics, AI offers an opportunity to improve accuracy and patient outcomes while reducing delays. In this paper, I examine the viability of AI diagnostic technology via an analysis of existing research, including best-diagnosed conditions and common limitations. I then explore interview physicians regarding AI use and their work environment to identify potential gaps between the capabilities of AI and the potential for implementation given attitudes within healthcare. Overall, there is a slight positive correlation between the size of the sample set used to train the algorithms and their performance outcomes. There are also notable differences in both training size and performance based on the body system addressed by a model, which is indicative of shortfalls in research and diagnostic capabilities in certain areas of medicine. These findings signal a need to encourage the construction and use of more comprehensive datasets. This result is supported by the physician accounts, which demonstrated a general interest amongst clinicians in using AI in the future as well as concerns about its abilities and consequences on the healthcare industry in the status quo.

Table of Contents

Introduction

As healthcare tools have developed over time, so has modern medicine's ability to detect and treat a wide range of conditions. The advent of complex computer programs and artificial intelligence, especially, has propelled the development of new technologies due to AI's potential to vastly improve doctor workflows and patient experiences. Artificial Intelligence is an umbrella term describing the category of machine learning and neural network technologies used to process large datasets in order to make deductions and perform cognitive tasks. For example, existing machine learning algorithms are capable of processing large datasets in order to perform complex analyses and create accurate predictive models (Bohr 2020). In the present day, there is an ever-present push for improving efficacy of healthcare through the integration of new technologies due to the rise in medical needs coupled with an international medical worker shortage (Umapathy 2023). This includes both the reduction of healthcare costs and patient turnaround times, as well as the improvement of patient health outcomes.

The belief that AI will eventually become an essential aspect of clinical care delivery and disease diagnosis has also encouraged a wave of studies testing the current capabilities of AI diagnostic applications for various diseases and medical conditions. Many of the existing studies evaluating AI diagnostic models focus on different diseases affecting different areas of the body. As such, they vary greatly in key design aspects such as the AI model used in the experiment and the sample set used to train the model. However, there is a lack of research that evaluates AI performance across medical specializations.

In an effort to standardize the large variance in study models and develop a clearer understanding of the future of AI diagnostics across different medical specialties, I conducted a

meta-analysis of the existing research regarding the performance of AI in detecting various specific conditions. I focused specifically on the age of the study, the condition of interest, the AI model utilized in research, the AI training set characteristics, and finally statistics evaluating the performance and accuracy of each tool. Through this analysis, I determined how testing strength and accuracy has progressed over time, which types of conditions are best detected by AI tools, and what AI model training prerequisites lend to a tool's viability for clinical use. This data provided context into which clinical fields that AI diagnostic tools are best suited to, as well as the most pertinent areas of improvement for fields in which AI applications lag behind in progress. After completing this meta-analysis, I interviewed physicians with varying degrees of AI use and clinical experience to inquire about the current clinical work environment, their experiences and feelings regarding AI technology in medicine, and their perceptions of AI integration in the future. Through these interviews, I sought to identify the most prominent struggles that doctors currently face at work, areas in which AI technology could be of greatest use for improving efficiency, and current systemic and policy barriers to AI diagnostics' expansion into everyday use.

Through my meta-analysis, I identified a slightly positive correlation between sample size and diagnostic performance of the AI model, as well as a decrease in performance variance with greater sample size. I also found that models for traditionally under-addressed conditions had both smaller data sets and reduced performance. In my interviews, I identified several concerns regarding AI tools, including uncertainties regarding lack of data to train algorithms effectively and a change in the medical landscape with increased AI use. As such, policymakers should work to create a legal framework that ensures quality in AI development while also providing medical workers with an increased role and voice in the creation of medical AI tools.

Background

The Current State of the U.S. Health Care Workforce

Emerging from the COVID-19 pandemic, significant gaps have appeared and continue to grow within the health care industry. The health care workforce has taken a large hit, both in numbers and in job satisfaction. Even before the pandemic, analysts projected widespread shortages of primary care physicians across the country ("Primary Care Workforce Predictions"). The onset of COVID-19 placed an increased burden on the health care workforce, as hospital workers not involved in the response were furloughed in the earlier stages of the pandemic (Oster 2021). Primary care practices were also severely affected by insufficient staffing concerns and closures (The Green Center 2020). These primary care gaps left in the wake of the pandemic have yet to be filled despite the increase in patient need, resulting in a significant mental health decline for clinicians (The Green Center 2023). Both during the height of the pandemic and in the present day, the overwhelming work burden resulted in increased rates of health care worker burnout, which is associated with lower-quality care and high worker turnover (Aiken 2023, Rotenstein 2023, Tawfik 2019).

Given the current state of health care employment, hospitals and health systems have attempted to employ solutions to support the health care workforce and bridge staffing gaps. One of those solutions is the assignment of less-complex tasks to less qualified medical staff, so that more qualified medical staff can prioritize the tasks that only they are qualified to complete, a method known as task shifting (Okyere 2017). While task shifting between staff can be beneficial in improving care efficiency, it fails to resolve the worker burnout and mental health challenges and ultimately is not a panacea to the health care workforce crisis (Okyere 2017, Van

Schalkwyk 2020). Other solutions rely on the digitalization of healthcare, such as telemedicine and the adoption of artificial intelligence (Khan 2022).

Limitations of The Current Diagnostic Process

Diagnosis is defined as a "pre-existing set of categories agreed upon by the medical profession to designate a specific condition", as well as the process undertaken to identify this set of categories and ultimately designate a condition to a patient (Jutel 2009, National Academies of Sciences, Engineering, and Medicine 2015). The diagnostic process bears important implications for healthcare stakeholders, including but not limited to patients, physicians, administrators, and health insurance providers.

The process can be best described as a multi-step framework (National Academies of Sciences, Engineering, and Medicine 2015). It begins with a patient developing a health concern. At this point, the patient is the first person to evaluate their symptoms and ultimately decide whether to escalate their concerns to professionals within the health care system. When they decide to seek care, information accumulation begins. During this step, clinical interviews, physical examinations, testing, and consultations with other physicians occur to build a thorough understanding of the patient's health history and current problem. As more information is collected over time, a working diagnosis is eventually developed and refined to match the newest data. This process of gathering, synthesizing, and using information in order to form a diagnosis for the patient is continuous throughout the patient care process. The next steps of the diagnostic process are to communicate the diagnosis to the patient, and to provide treatment to the patient following the appropriate care plan based on the diagnosis. Finally, the final part of the diagnostic framework is to evaluate patient and system outcomes in order to determine areas for improvement in the future.

In practice, the diagnostic process is complex and time-consuming, often involving a larger system of healthcare professionals. Its cyclical nature of constant information collection and reevaluation of diagnoses assists clinicians in providing the optimal care options for their patients but comes at a large cost of resources and time. Diagnostic testing, such as screenings and physical tests, is vital for improving the working diagnosis, as it can identify health conditions within a patient even before symptoms arise (National Academies of Sciences, Engineering, and Medicine 2015). Furthermore, diagnostics also plays a role in the public policy decision-making process, as policymakers often use diagnostic data to inform their decisions regarding resource allocation, research prioritization, and payment policies (Jutel 2009). As such, developing effective screening tests is essential for reducing patient turnaround time and improving patient care, but also for bettering health policy.

The Rise of Artificial Intelligence in Health Care

The potential for artificial intelligence (AI) applications in medicine has been recognized and explored since the 1970s with the development of INTERNIST-1 in 1971 ("AI's Ascendance in Medicine"). While primarily used in an experimental and educational context, INTERNIST-1 was designed to assist physicians within internal medicine in making diagnoses using a ranking algorithm (Miller 2010). This project, which spanned four decades, along with others like it marked the beginning of AI development in the health care industry. In the present day, AI use has expanded within the medical field to include not only diagnostics but also imaging, smart prosthetics, health data, and more (Al'Aref 2018, Shaheen 2021). Even in the realm of diagnostics, artificial intelligence capabilities have expanded outside of internal medicine to include detection and diagnosis of ADHD, among other conditions (Loh 2022). This technological development is corroborated by the influx of AI tools being approved for medical

use. As of the most recent update from the FDA in October 2023, there are 691 approved AI or machine learning (ML)-enabled medical devices, with 108 approved within the past year (FDA 2023). Furthermore, submissions of AI-based medical devices for FDA approval have multiplied in recent years, with an over 30% increase in 2023, and a 39% increase in 2020, from the preceding years (FDA). The widening possibilities of AI within healthcare makes artificial intelligence development a target avenue for addressing current diagnostic and treatment limitations.

Ethical Considerations and Concerns in the Right to Health

As the landscape of healthcare continues to evolve with the use of AI, so do the implications of the right to health as articulated in foundational human rights documents. Fundamentally, government bodies and private businesses are both obligated to ensure that any development and deployment of AI falls within internationally affirmed medical ethics and human rights standards. Articles 12 and 2.2 of the International Covenant on Economic, Social and Cultural Rights (ICESCR) establish states' obligations to make "the highest attainable standard of physical and mental health" available for all without discrimination of any kind (United Nations General Assembly, 1966). This standard vests in signatory nations the responsibility to not only ensure adequate health standards for all, but also strive to optimize health outcomes and safeguard against inequity. Furthermore, states under the ICESCR are obligated to dedicate the maximum resources available, either independently or via international cooperation, to realize the highest attainable standard of health (id., art. 2.1). In the context of digital technologies, this means that states are responsible for ensuring that the delivery of health care via AI does not infringe on people's exercise of their right to access health resources equally and without discrimination.

While governments are the primary duty bearers in ensuring ethical implementation of health technology as part of the human right to health, this does not imply that private parties or businesses have no role in ensuring ethical AI. On the contrary, businesses must respect human rights in the conduct of day-to-day operations. The United Nations' Guiding Principles on Business and Human Rights outlines the bare minimum standard for businesses as a responsibility to "avoid infringing on the human rights of others" and "address adverse human rights impacts with which they are involved" (United Nations OHCHR 2012). This burden on private entities to prevent an infringement upon human rights exists independently of states' obligations to protecting rights. To encourage businesses to follow these principles, nation states are meant to establish ethics-forward policies and adequate oversight. Nevertheless, businesses are expected to follow international human rights standards "regardless of their size, sector, operational context, ownership and structure" (United Nations OHCHR). As such, AI technology developers and health care providers that purchase AI tools for use on patients have a duty to ensure that digital technologies are built and used ethically.

The Regulatory and Approval Process of Health Care AI Applications

In the United States, medical technologies are regulated and approved for public use by the Food and Drug Administration (FDA). This includes AI and ML-enabled medical devices, which is identified by the FDA as a subcategory of Software as a Medical Device, or SaMD (FDA 2023). The FDA has been reviewing and approving the release of AI/ML-enabled medical devices for almost 30 years, with the first approval for an AI- or ML-based medical tool granted in 1995 (Advanced Medical Technology Association 2024). Like the review process for other medical devices or software, the FDA requires that AI/ML tools undergo specific regulatory pathways and procedures. When assessing the safety and effectiveness of AI/ML algorithms for

medical use, the FDA considers a variety of factors like sample data quality, statistical robustness, and clinical trial results (Advanced Medical Technology Association).

The FDA, Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) jointly outlined their recommendations for AI medical device regulation in the *Good Machine Learning Practice for Medical Device Development: Guiding Principles* (Digital Health Center of Excellence 2021). Meant to promote "safe, effective, and high-quality medical devices", the guidelines focus on themes of multidisciplinary development, widely representative and unbiased data sample sets, thoughtful and secure software engineering practices, ease of user experience, and regular re-evaluation. The full 10 guiding principles published by the FDA are included in **Appendix A**. These standards parallel the principles of "non-discrimination, equality, participation, accountability, reparations and privacy" as laid out by the United Nations Human Rights Council (2023). Through recent publications, international governing bodies have come to a consensus as to the best practices for software developers and companies to employ when constructing AI technologies for medical use.

These standards in turn inform medical device registration procedures such as 510(k) approval process, which is the most common avenue through which medical devices are submitted and evaluated for approval in the United States (FDA 2023). This process can provide a final decision within 90 business days, but other approval methods can take up to eight months (Fenton 2021). The FDA reviews all submitted devices for their overall safety and effectiveness, which includes determining a suitable research dataset diversity "based on the device's intended use and technological characteristics" (FDA). Ultimately, the approval of a medical device entails adherence to all relevant FDA premarket requirements.

Literature Review

Artificial Intelligence as an Avenue for Optimizing Health

The implications of growing artificial intelligence use in medicine are as significant as they are multifaceted. For conditions that are tedious to detect or for which few qualified specialists are capable of an accurate diagnosis, having AI models that can accurately diagnose patients would greatly reduce the workload burden of physicians and the waiting duration for treatment of patients (Loh 2022). Even though many of the AI diagnostic models that currently exist are still in the developmental stages, their levels of success have been unprecedently high. In cases such as with cancer diagnosis and prognosis, AI algorithms have demonstrated accuracy rates higher than the standard statistical analysis (Huang 2020).

As the trend in medicine leans increasingly towards AI technology integration, it becomes critical to anticipate which areas of the healthcare workflow will be supplemented by or completely automated by AI. Examinations of current AI abilities to integrate into operating room (OR) activities, for example, identify multiple ways in which robot algorithms can automate ORs beyond diagnostics. These include, inter alia, independent AI-guided camera manipulation for laparoscopic surgery, automated instrument reconfiguration and motion stabilization during surgery, and automated medical imaging (Kranzfelder 2012). The vast range of complex activities that can be automated with AI offers a promising outlook for the future of medical capabilities. However, with increasing technological complexity comes greater difficulty in understanding how to handle these technologies. As such, additional resources must be put into providing health care workers with appropriate operation assistance programs before these developments can be effectively implemented into common use (Kranzfelder). This requirement

also applies to diagnostic technologies, as AI diagnosis tools are most effective in the hands of a person who understands the required input conditions and can accurately interpret output results.

The gradual automation throughout healthcare represents an intersection of two major goals in healthcare: safety and cost-reduction. On one hand, the introduction of algorithm-supported technologies in healthcare is part of a larger movement towards trauma prevention and less invasive medical procedures. Conventional examinations can result in numerous invasive diagnostic biopsies, especially when traditional exams have low specificity rates (Campanella 2022). This can cause patients to undergo unnecessary procedures that cosmetically or functionally affect hyper visible body areas, such as their face. As such, developing medical interventions that can reduce the physical and emotional trauma a patient must undergo in treatment is of utmost priority for caregivers.

On the other hand, health providers are also looking towards novel AI technologies to facilitate more cost-effective treatments. While the initial process of introducing novel tools into the health care system may incur some significant start-up costs, there are still many budgetary benefits that could emerge from AI implementation. Automation of traditionally time-consuming processes helps to reduce practitioner workloads, freeing up schedules to treat more patients and attend to other priorities. Furthermore, AI applications that have higher success or accuracy rates than human workers can reduce the costs and wasted time associated with human error. Specifically in diagnostics, an accurate initial diagnosis can prevent patients from undergoing misguided interventions, which are oftentimes costly and may even cause harm to a patient's health. With responsible and effective implementation of AI in healthcare, providers can not only foster better health outcomes for patients but also reduce unnecessary expenses in the process.

Limitations of AI in Health Diagnostics

While current research on AI algorithms in clinical diagnosis has been promising, there are still limitations on the capacity of this technology. Many of the tools currently being developed and tested use AI solely to automate the task of diagnosing a condition, such as distinguishing malignant tissue samples from benign samples. However, they currently lack the ability to assist physicians in interpreting images or samples in any greater detail (Elemento 2021). This can become an issue in cases where the images provided are not textbook examples of a conditions, such that a tool that excels with sample images struggles in real-life application and is unable to qualify its diagnoses with probability of accuracy rather than a binary "yes" or "no" response (Pai 2020).

The biggest concern regarding AI applications in healthcare, as well as in general, is the lack of transparency and interpretability (Kiseleva 2022). As artificial intelligence only provides a response to the engineered question and not an explanation as to how it arrived at its conclusion, transparency in every stage of the development of AI is crucial for clinicians to better understand how the tool arrives at its conclusion and whether certain patient factors can render an AI diagnosis unreliable or inaccurate. Understanding the decision-making process of AI tools is especially important for healthcare, as these decisions directly impact human lives. Ensuring transparency of AI is of especially great concern because there is little legal or policy guidance on AI development for commercial or medical use (Kiseleva 2022).

However, transparency in how a model processes images and produces decisions is difficult, if not impossible, to ensure even on a technical level. This is because many AI technologies are designed in a "black box" model. A black box model is one in which the internal structure of the AI technology is invisible to the programmer (Blouin 2023). This means that,

while a person can give an AI tool a data input and receive a decision output in the process of testing or using the application, they cannot see the constructed "neural network" that the AI uses to draw conclusions. An analogy to this would be a teacher tutoring a student in math and quizzing them with questions to ensure that they are able to perform calculations accurately. However, unless the teacher asks the student to explain their exact thought process, the teacher has no way of understanding the student's method of solving math problems. Likewise, an AI tool developer can give their model input data and evaluate their output decisions during the training process to verify that the model can accurately produce the intended outcomes. Unlike the teacher-student example, a programmer is unable to ask their AI application to explain how it "thinks", leaving the internal pathways developed within the AI tool as a mystery.

The current lack of transparency of AI in health care has implications for a physician's ability to trust the technology and be comfortable with using it in their everyday practice. Some of the most prominent factors limiting clinicians' abilities to trust AI systems include a lack of user education, perception bias, and skepticism stemming from insufficient reliability or transparency (Asan 2020). Furthermore, the development of AI also may have an impact on trust within a doctor-patient relationship. Even in times of high medical misinformation and uncertainty such as during the pandemic, many patients retain trust with their primary care physicians (The Green Center 2020). However, the focus on AI development to the extent that they surpass a human clinician's abilities has raised concerns that the authority of physicians in condition diagnosis and interpretation will soon be undermined if a reliable AI tool comes to a different conclusion (Hatherley 2020). Those sounding the alarm in the medical and AI development fields argue that if patients are aware that the decisions made about their health are completely automated, they may not be able to fully trust their physician even if the technology

itself is reliable (Hatherley). These questions in accuracy, ethics, and trust are all forces that currently hamper the ability for health care systems to confidently deploy AI tools in clinical practice. Developers, clinicians, and legislators involved in AI use in health care must address these concerns before AI diagnosis is able to produce the benefits touted by its proponents.

Health Care Perspectives on and Concerns Regarding AI

There are several key groups whose livelihoods are directly impacted by the growing implementation of artificial intelligence in healthcare, namely health care workers, medical students, and the general public seeking care. Health care workers in particular wield significant power in the expansion of healthcare AI technologies, as their understanding of and ability to use new tools successfully impacts the resulting efficacy of technologies in real life environments. In extreme cases of distrust or misinformation, they could refuse to use these tools in providing treatment if there is a consensus against AI use. Accordingly, their sentiments regarding AI are important to determining whether widespread AI use in diagnostics is viable for the near future.

Existing surveys of medical practitioners highlight an enthusiasm for advanced diagnostic technology that can relieve workflow burdens and reduce diagnostic errors. However, they also reveal an elevated level of skepticism regarding algorithm accuracy, legal liability, and the potential replacement of human workers with algorithms. Due to the complex nature of manually diagnosing a condition via visual examination, along with the existing health care worker shortage, many health care workers and general practitioners (GPs) view AI diagnostic tools as helpful clinical implements, especially for particular conditions such as skin cancer that they currently struggle to diagnose (Samaran 2021, Nitiéma 2023). Even so, health care workers harbor a sentiment either erring on the side of caution or altogether negative towards AI for reasons outside of its supplementary abilities. Notably, health care workers have expressed

significant negative sentiments about the potential for AI to replace human workers, AI use in disease screening, and AI's impacts on medical diagnostic procedures in particular (Nitiéma). These findings signal potential organizational roadblocks to further implementation of AI in health care diagnostics specifically.

Status Quo of Policy Regarding AI Applications

Over recent years, artificial intelligence has received increasing attention from lawmakers across the United States, with 2023 seeing the most AI-related laws proposed in state legislatures in a year than ever before (Zhu 2023). Many of these laws function primarily as consumer privacy laws, demonstrating a surge in concern with the impact of data retention and automated profiling. Furthermore, multiple states' legislators and government bodies have expressed concern in the potential impact of AI on healthcare, among other services (Zhu). However, there have been few policies proposed that specifically address data protection within the healthcare space, or that provide specific requirements for tool development that are tailored to AI. Furthermore, the language of legislation that has been introduced is indicative of a general wariness regarding AI deployment and a hesitation to implement AI tools in a broader, public-facing context (Zhu). These factors are indicative of the current gap in information regarding the potential benefits and detriments of AI and an inability as such to confidently encourage or discourage further development.

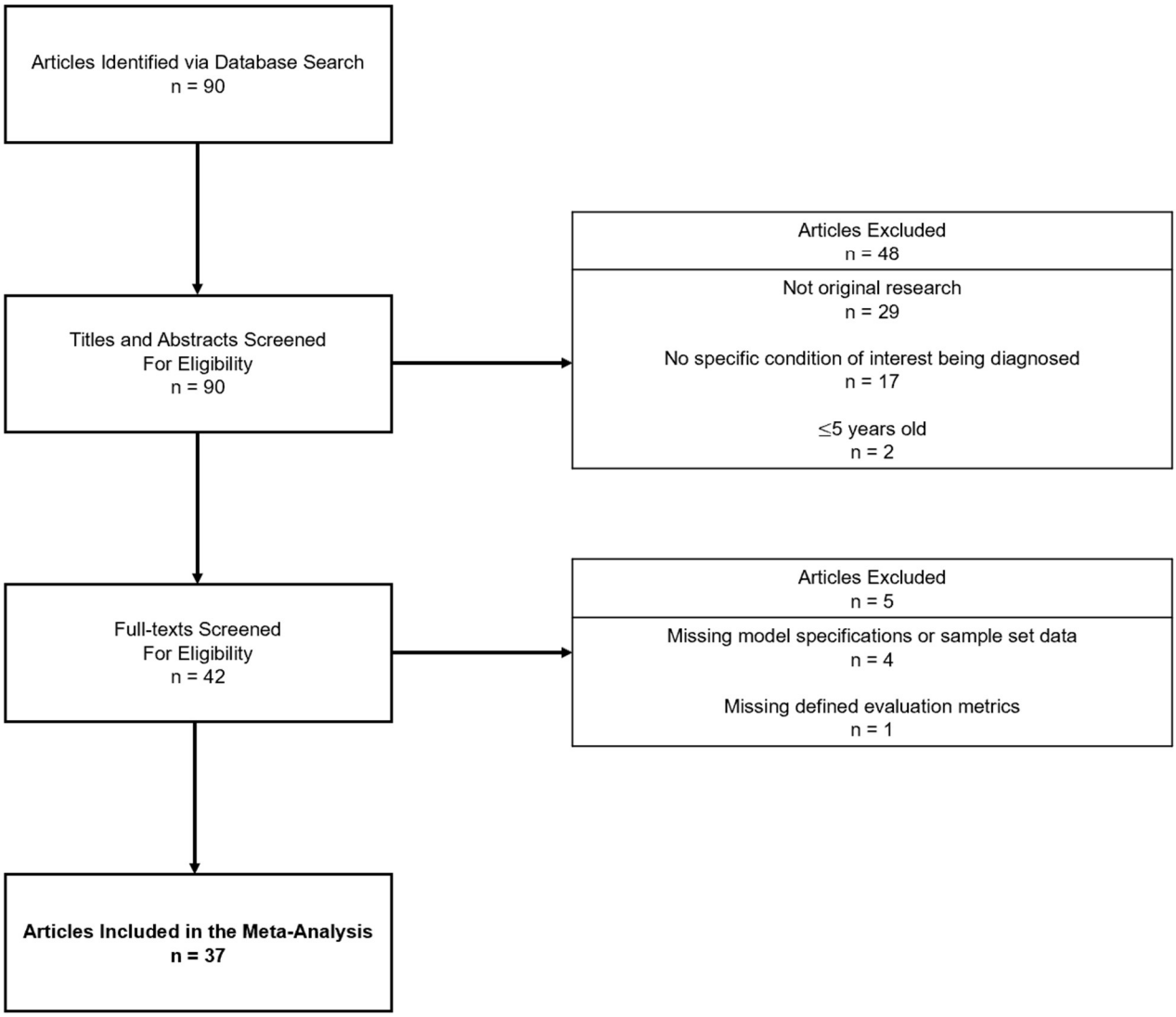
Methodology & Data Collection

My study consists of a quantitative meta-analysis component and a qualitative interview component. Each of these aspects has an individual methodology that was followed in the process of collecting, processing, and analyzing the relevant data.

Quantitative Data Methods

For the meta-analysis, I utilized published articles listed in the ScienceDirect and PubMed databases as the data for analysis. Publications were collected as datapoints by searching on the aforementioned databases with relevant keywords including “artificial intelligence”, “machine learning”, “diagnosis”, “deep learning”, and “automated detection” and evaluating their title for topicality. A standard set of criteria was employed to evaluate the suitability of each publication for the purpose of this analysis. Those which did not conform to all criteria were rejected and removed from the final dataset (**Figure 1**). I included articles concerning original research, utilizing an artificial intelligence model, with a focus on automated detection and diagnosis of a specific medical condition and that were published within the last 5 years. I excluded papers that were not original research such as reviews and opinion papers, that lacked a specified condition of interest, that did not clearly describe the AI model or the model’s training sample size and process, and that did not have defined metrics for evaluating detection capabilities.

Figure 1. A flow chart demonstrating the data collection process.



In total, 90 articles were originally collected based on the key term search and precursory title examination. Out of those, a total of 48 articles were excluded from data analysis through a review of their title and abstract based on the established criteria, with 2 articles being older than 5 years, 8 articles lacking original research, and 3 articles lacking a specific condition of interest. Another 5 articles were excluded through a screening of their full texts due to either lacking clear AI model specifications or sample set data (n= 4), or for missing defined performance evaluation

metrics (n= 1). Across the remaining 37 articles included in the analysis, a total of 64 separate models or target conditions were identified. As such, each of the individual models or target conditions were separated in data analysis and treated as unique datapoints. This consideration was made because while multiple individual models or conditions of interest may be included in the same overall study, each entails a separate diagnostic process.

The independent variable in this analysis is the set of study characteristics, and the dependent variable is the accuracy outcomes. These variables were captured via examination of the selected published studies. After identifying the necessary data for the independent variables for each study, I then compiled this information through Microsoft Excel (**Figure 2**). The prevalence of each condition was determined through statistics regarding the most up-to-date global findings.

Figure 2. A sample data collection table demonstrating the data extracted in the meta-analysis.

Study	Year Conducted	Condition of Interest	Affected Organ	Affected Body System	Prevalence of Condition
Bhattacharjee, 2022	2020	Tooth Decay (Dental Caries/ Cavities)	Tooth	Digestive System	46.2% (primary teeth); 53.8% (permanent teeth)
Bonnevie ED, et al., 2023	2023	Ulcerative Colaritis	Colon	Digestive System	

AI Model Type(s)	Single Class vs Multiclass Mod	Sample Size	Method of Data Collection	Performance	Notes/Key
12-layer **CNN** via PyTorch (Deep Learning), ResNet-18 & ResNet-27 (pre-trained image classifiers), two-stage curriculum learning	single	Web-searched: 314 (129 +, 185 -) (251 training, 63 testing) ImageNet1k: >1.2 million Field-collected: 192 (136 +, 56 -) (157 training, 35 testing)	Web-searched (dental blogs, dental presentations, journals); Two-stage curriculum learning via ResNet-18 and ResNet-27 (ImageNet1k); Field-collected (consenting human participants photographed with a sterilized intraoral camera)	ResNet-27 with two-stage curriculum learning: Accuracy = 82.8% Sensitivity = 1.0	CNN = Convolutional neural network
VGG-16 (**CNN**) trained with HALO AI (v2.3, v3.5)	single	N= 19 (9 training, 10 testing)	Self-collected (H&E stained UC specimens; stained and imaged human coloretal pinch biopsies assessed by 3 independent observers)	Nuclear Detection: R^2= 0.9844	VGG = visual geometry group

In collecting performance outcomes, the most common metrics of performance provided were sensitivity, specificity, AUC, and accuracy. In the case that none of these metrics were identified or derivable from the information provided by the study, detection rate was collected instead. For articles in which more than one AI model is developed and tested, the characteristics and outcomes of each model are reported separately.

After the initial data collection, I reorganized this data to isolate information that I planned to include in my regression analyses. Target data included the data subset sample sizes, the individual performance metrics, body system affected, and AI model specifications (**Figure 3**).

Figure 3. A sample data collection table demonstrating the data reorganized to highlight target information for data analysis.

Name	Yea	Condition	Organ	System	Mode	Clas	SizeTota	SizeTrain	SizeVali	SizeTes
Bhattacharjee, 2022	2022	Tooth Decay (Dental Caries/ Cavities)	Tooth	Digestive System	CNN	single	506	408	0	98
Bonnevie ED, et al., 2023	2023	Ulcerative Colaritis	Colon	Digestive System	CNN	single	19	9		10
Sun H, et al., 2023	2023		Knee	Skeletal System	CNN	single	3784	1638		2146

Method	Accuracy	Precisio	NPV	AUC	Sens	Spec	F1	Other
Web-searched (dental blogs, dental presentations, journals); Two-stage curriculum learning via ResNet-18 and ResNet-27 (ImageNet1k); Field-collected (consenting human participants photographed with a sterilized intraoral camera)	0.828				1			
Self-collected (H&E stained UC specimens; stained and imaged human coloretal pinch biopsies assessed by 3 independent observers)								R2= 0.9844
Retroactively-collected hospital data (Picture Archiving and Communication System (PACS) at Shanghai Changzheng Hospital, AP radiographs); Field-collected (knee plain AP radiographs from adult hospital patients);		0.922			0.955			

Statistical analysis was performed in R Studio and Microsoft Excel. In my evaluation of the data collected, I estimated the relationship between the sample set size & the performance outcomes of the AI diagnostic models, which I hypothesized to be a positive correlation (i.e., a larger sample size would be correlated with superior performance). Building upon analysis of condition-specific variables, I evaluated the relationship between the affected body system and performance, following my hypothesis that certain areas of medicine are more suitable for implementing AI diagnostics in the present-day than others. Furthermore, I calculated the relationship between the year that the study was conducted and performance to evaluate my hypothesis that more recently developed models would be more likely to exhibit superior performance compared to older models. Finally, I categorized each of the studies by the type of artificial intelligence model used (i.e. Deep Learning framework) to determine which models were most frequently used, as well as to calculate the correlation between the type of model used and performance. The correlation calculations involving the performance outcomes were

performed through a multiple regression analysis to determine the strength of each independent predictor variable's effect upon the performance outcome of a model.

Qualitative Data Processing and Analysis

I conducted 6 semi-structured interviews with physicians to determine where doctors struggle most in efficiency of workflow as well as to better illustrate the viability of AI implementation in the workplace. The purpose of the qualitative interview portion of my paper is to determine physician perceptions of their current work environment, current attitudes within the healthcare industry towards the use of artificial intelligence in clinical diagnostics, and the greatest perceived barriers to implementation. Specifically, my interviews focus on 1) the current clinical environment and the greatest struggles that doctors currently face at work, 2) areas in which technology could be of greatest use for improving efficiency, and 3) current perceptions of AI in healthcare and in diagnostics especially, and 4) suggestions for future advancements in diagnostic AI use.

I conducted interviews with currently employed physicians, with a distinction drawn between those employed in large practices (defined as those with 50 physicians or more) or medium-sized practices (10-50 physicians) and those working in small practices (10 physicians or less). Medium- and large-sized practices tend to more often be hospitals and clinics that are a part of a larger healthcare system, while small practices tend to be private clinics. This factor is notable as physicians employed within hospitals and other large health systems are more likely to have experience with advanced technology in screening and diagnosis compared to physicians employed in smaller private practices and are also likely more familiar with many of the conditions targeted within the data. I also differentiated between different size practices because

the difference in workplace environment may impact one's opinions on AI technology. In my interview process, I first obtained verbal consent from each subject for interviewing and transcription before commencing the interview. I aimed to recruit interview subjects from varying sizes of clinical workplaces and different areas of specialization in order to examine the effects of different work experiences on perceptions of AI in healthcare. I recruited interview subjects regardless of prior experience with AI, as I wanted to cover the perspectives both of those who have worked with AI technology in the past and those who have not. Ultimately, 3 of the physicians worked in large practices, 2 worked in medium-sized practices, and 1 worked in a small practice. Of all the physicians, 3 were general practitioners or internal medicine generalists, and 3 were specialized surgeons or physicians. I recruited physicians from multiple locations and healthcare systems: the interviewees were each recruited from unique clinics and hospitals either in California (n=4, or 66%) or Texas (n=2, or 33%). Due to the diverse locations and conditions represented in this study, there is great variety in the circumstances and needs of the workplaces represented. This is reflected in my interview as another point of comparison, which I address through my questions tailored to personal and community attitudes regarding the use of AI in healthcare.

The 6 interviews were conducted between March 2024 and April 2024. Interviewees were recruited through email or via phone number, and all interviews were conducted either as a video call over Zoom or over a mobile phone call. The interviews were semi-structured, meaning that the interviews were structured using the same set of questions from a predetermined interview guide, but varying probes were used with individuals in order to extend upon their responses as needed. The contact information of interviewees was procured from online contact information databases posted on workplace websites, as well as by word of mouth from mutual

indirect connections. However, I did not have any direct personal connections with any of the interviewees prior to facilitating the interview.

Interview audios were recorded with the permission of the interviewees via Zoom software if interviewed through Zoom, or Windows Sound Recorder (version 11.2312.5.0) if interviewed over the phone. The resulting audio files were transcribed using the Otter.ai software. After the automatic transcription process was completed by Otter.ai, transcripts were manually reviewed for accuracy before being uploaded to a shared cloud-based folder accessible only to researchers directly involved in the study. During the manual review process, transcripts were also edited to omit the name of the interviewee and other identifying information, instead substituting this information with pseudonyms to ensure interviewee anonymity. If interviewees did not provide permission to be recorded, then notes were taken on Microsoft Word throughout the duration of the interview on the responses provided. These notes were reviewed after the interview for accuracy. Similarly to the transcripts, the notes were edited to omit and substitute any personal identifying information for anonymity.

Quantitative Findings

Data Set Characteristics

While some data characteristics such as sample size were guaranteed for every point by my data collection requirements, not every study included the same measures of performance. **Table 1** below displays a summary of the data set used to conduct my meta-analysis with details regarding the individual statistics collected. The most used performance measurements were

Accuracy (n=41) and Sensitivity (n=44). As such, I focused my analysis of performance on these two statistics in relation to the tested independent variables.

Table 1: Data Set Summary

Statistic	N	Mean	St. Dev.	Min	Max
Year	64	2,021.328	1.594	2,018	2,024
Total Size	64	24,130.450	88,594.060	19	703,970
Training Size	64	21,706.090	87,107.070	9	695,030
Validation Size	35	2,421.486	3,871.364	0	17,919
Testing Size	59	1,805.305	6,938.627	10	52,870
Accuracy	41	0.863	0.137	0.228	0.995
Precision	28	0.869	0.132	0.400	1.000
NPV	7	0.898	0.101	0.697	0.990
Sensitivity	44	0.899	0.091	0.554	1.000
Specificity	34	0.912	0.139	0.430	1.000
F1	17	1.150	0.445	0.840	1.940

The characteristics of the studies found through my search bring to light specific trends in AI algorithm development for health diagnostics. Out of the 64 models included in my analysis, the majority of them were tailored to diagnose conditions related to the digestive system (n=14), or the cardiovascular (n=13) and circulatory systems (n=10). Another commonly addressed body system is the nervous system (n=12).

This was notable as many of the conditions associated with these body systems are internationally recognized as global health concerns, including tooth decay and cardiovascular abnormalities. However, there was a lack of studies concerning body systems such as the reproductive system (n=2). Even in more frequently addressed body systems, there was a lack of research regarding traditionally stigmatized diseases, such as colon cancer. These statistics mirror existing trends in which many of the most prevalent or deadliest diseases are particularly underfunded and understudied due to existing stigma attached to these conditions (Samuelson

2019). One example of this is ovarian cancer, a condition which is significantly poorly funded and overlooked yet extremely prevalent and aggressive in its progression (Samuelson). While there were two studies regarding reproductive health within the data sample, they were focused on prostate cancer and breast cancer; both conditions typically receive greater public attention and adequate funding (Samuelson). Overall, it is noteworthy that current AI development for the healthcare industry follows existing health research trends. This could be a result of the lack of existing research of these conditions, and thus a lack of images or datasets needed to create effective diagnostic algorithms.

The models observed in my data demonstrated similar structures but varied greatly in development sample sizes. The vast majority of the models observed were Convolutional Neural Networks (CNNs) (n=52), which is a type of deep learning (DL) AI model. This is likely because CNNs are designed to be highly compatible with image data. Much of the data currently examined manually by doctors to diagnose health abnormalities exists as images taken during screenings or biopsies. This makes developing AI tools adept at interpreting image data of great priority for improving medical efficiency. Other types of models used were Deep Neural Networks (DNNs, n=6), Natural Language Processing (NLP, n=4), and Machine Learning (ML, n=1). Deep learning models, such as CNNs and DNNs, typically have high performance compared to other machine learning models, and thus are more reliable for diagnostic applications. As such, it is not surprising to find them as model of choice for health diagnostics.

A major consideration in the development of AI-enabled tools is the sample size of data used in model construction. This is especially true for deep learning models, which typically have complex structures consisting of many parameters (Lin 2021). The total sample sizes represented in the models analyzed were mostly over 1000 datapoints, and just under half of the

models (n=29, or roughly 45%) used a minimum of 5000 datapoints in their total sample size. However, deep learning models require a lot more sample size data and parameters than the average AI algorithm to make it effective (Lin). In order to facilitate the development of AI for a wider range of conditions, it is imperative that sufficient image databanks are created for innovators to use in algorithm training.

Statistical Analysis

At the beginning of the study, I set out to test three main hypotheses. First, I hypothesized that there would be a positive correlation between sample size of the dataset used to build the model and the precision metrics used in testing. Alongside this, I hypothesized that there was a positive correlation between the year a study was conducted and diagnostic performance. Finally, I hypothesized that there would be notable differences in average sample sizes when stratified by the body system affected in diagnosis, contributing to models addressing some body systems performing significantly better on average than those targeting other systems. I formed this prediction due to an understanding that some conditions are over- or under-funded in proportion to their disease burden, which causes disparities in disease treatment and care outcomes. (Gross 1999, Samuelson 2019).

Relationship between Sample Size and Performance

Before beginning my full data analysis, I first determined that a linear model would be the most beneficial regression with which to analyze the relationship between sample size and performance. To this end, I modeled this relationship first as a simple linear regression, and then as a logarithmic regression. In both cases, I looked for the F-statistic and the p-value as indicators of statistical significance and best model fit.

Table 2: Comparing Model Fit

	Dependent variable:	
	Accuracy	
	Linear	Logarithmic
Total Size	−0.00000**	
	(0.00000)	
Total Size		−0.019
		(0.013)
Constant	0.897***	1.025***
	(0.024)	(0.106)
Observations	41	41
R^2	0.148	0.058
Adjusted R^2	0.126	0.034
Residual Std. Error (df = 39)	0.128	0.135
F Statistic (df = 1; 39)	6.756**	2.422
Note:	*p<0.1; **p<0.05; ***p<0.01	

As seen in **Table 2**, I compared the simple linear model (1) to the logarithmic model (2) for the relationship between total sample size and accuracy. The F Statistic for the linear model is larger and has a p-value demonstrating at least 95% confidence interval significance. In contrast, the logarithmic model has a smaller F Statistic, and has a p-value larger than 0.1. This demonstrated that a linear regression would be the most beneficial in modeling the target relationship. From this outcome, I chose to model the relationship between the sample size and performance outcomes as a linear relationship.

With the linear model established, I first designed my simple regression models as

$$(Accuracy = \beta_0 + \beta_1 * \text{TotalSize} + \epsilon) \text{ and } (Sens = \beta_0 + \beta_1 * \text{TotalSize} + \epsilon)$$

so that I could observe the raw relationship between the total sample size and performance, as measured by accuracy and by sensitivity. I then expanded my regression to control for potential confounding variables. Specifically, I controlled for model class (e.g., whether the model was designed to recognize a single state or multiple states) and the publication year of the study. I structured the Class variable as a binary input such that multi-class models were identified as 1 and single-class models were identified as 0.

One consideration that was relevant in my analysis was that even if a study is utilizing a large sample size of data to train the model, the data itself may be of poor quality. This can occur when the image data is self-collected by patients or taken using lower-quality imaging equipment such as a cell phone. As such, I reviewed my dataset to determine that the method of data collection for each model was of medical or scientific quality, utilizing standard equipment. Through this review, 2 entries were removed due to utilizing consumer-sourced images or inferior photography equipment. The dataset without the two removed entries was utilized for all other calculations performed during my analysis. My full regression results are summarized in **Tables 3** and **4** below.

Table 3: Accuracy Results - Total Size

	Dependent variable:			
	Accuracy			
	(1)	(2)	(3)	(4)
Total Size	1.972e-07	2.241e-07	1.003e-06	1.014e-06
	(7.279e-07)	(7.455e-07)	(1.412e-06)	(1.434e-06)
Class		0.007		0.010
		(0.028)		(0.035)
Year)2019			-0.092	-0.093
			(0.159)	(0.162)
Year)2020			-0.055	-0.055
			(0.060)	(0.061)
Year)2021			0.025	0.019
			(0.042)	(0.046)
Year)2022			0.073	0.072
			(0.047)	(0.047)
Year)2023			0.026	0.018
			(0.046)	(0.053)
Year)2024			0.050	0.050
			(0.056)	(0.057)
Constant	0.885***	0.883***	0.851***	0.851***
	(0.015)	(0.018)	(0.035)	(0.035)
Observations	39	39	39	39
R^2	0.002	0.004	0.219	0.221
Adjusted R^2	-0.025	-0.052	0.043	0.014
Residual Std. Error	0.080 (df = 37)	0.081 (df = 36)	0.077 (df = 31)	0.078 (df = 30)
F Statistic	0.073 (df = 1; 37)	0.066 (df = 2; 36)	1.243 (df = 7; 31)	1.067 (df = 8; 30)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Sensitivity Results - Total Size

	Dependent variable:			
	Sens			
	(1)	(2)	(3)	(4)
Total Size	1.079e-07	1.121e-07	1.792e-07	1.274e-07
	(1.319e-07)	(1.365e-07)	(2.133e-07)	(2.233e-07)
Class		-0.004		0.028
		(0.029)		(0.037)
Year)2019			-0.040	-0.034
			(0.124)	(0.125)
Year)2020			0.015	-0.002
			(0.072)	(0.076)
Year)2021			0.018	0.014
			(0.074)	(0.074)
Year)2022			0.076	0.074
			(0.073)	(0.073)
Year)2023			-0.019	-0.038
			(0.070)	(0.074)
Year)2024			0.002	0.002
			(0.084)	(0.084)
Constant	0.896***	0.897***	0.881***	0.881***
	(0.014)	(0.018)	(0.065)	(0.065)
Observations	44	44	44	44
R^2	0.016	0.016	0.154	0.168
Adjusted R^2	-0.008	-0.032	-0.010	-0.022
Residual Std. Error	0.092 (df = 42)	0.093 (df = 41)	0.092 (df = 36)	0.092 (df = 35)
F Statistic	0.669 (df = 1; 42)	0.337 (df = 2; 41)	0.938 (df = 7; 36)	0.882 (df = 8; 35)
Note:				*p<0.1; **p<0.05; ***p<0.01

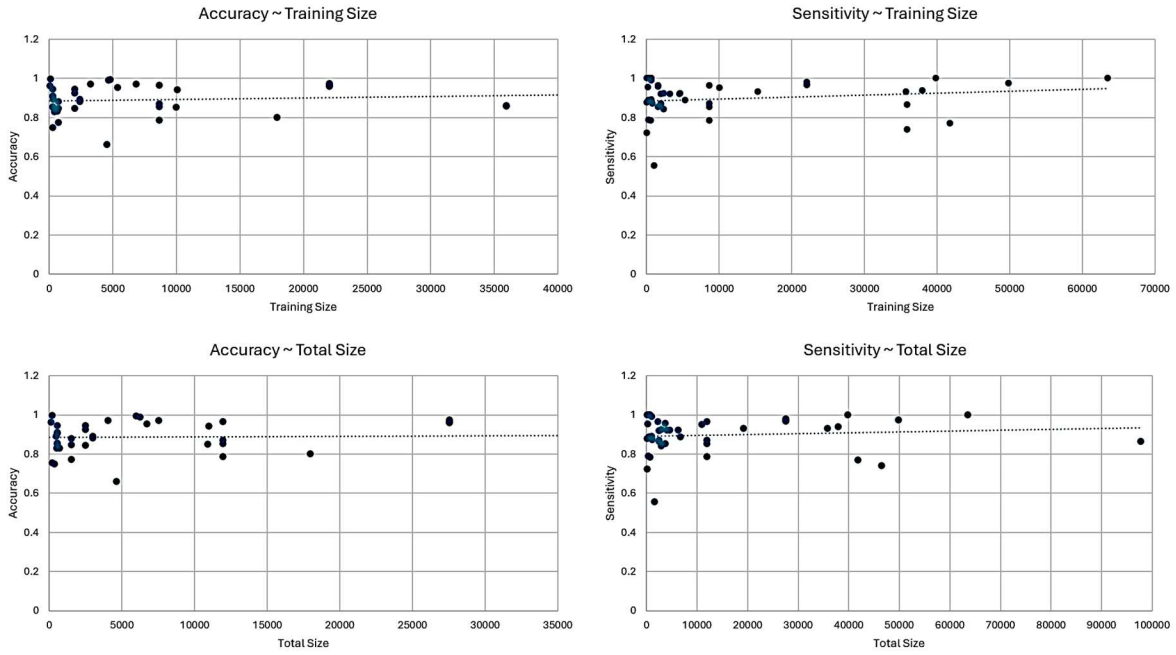
The total sample size has statistically zero effect upon either accuracy or sensitivity. Models designed to identify multiple conditions or states (e.g. “classes”) tend to be slightly more accurate than single-class models. However, multi-class models tend to provide either a very small decrease in sensitivity (without controlling for year) or a slight increase in sensitivity (when controlling for year). As multi-class models are typically trained with several sets of data

in order to distinguish between each target condition, these results are consistent with the intuition that more complex and detailed training contributes to improved model performance.

When controlling for the year of study publication, models from 2019 tend to have slightly decreased accuracy and sensitivity, and models from 2020 tend to have slightly better performance than those from 2019, but decreased accuracy overall. In contrast, models from 2021- 2024 all tend to have slightly improved accuracy and sensitivity. This indicates that, while there is not a statistically significant linear trend of improvement of models from year to year, there is a slight trend of improved model performance from 2019 to 2024.

Given the negligible slightly positive effect of total sample size on the performance of AI models, I wanted to measure the effect of specific types of sample size on performance. I speculated that the training sample size would have a stronger correlation with diagnostic performance, as this subsection of data directly contributes to the formation of the AI model. I first plotted scatterplots of relationship between training size and performance, as well as total size and performance for comparison.

Figure 4. Scatterplots of training size and total size in relationship with performance



From this preliminary examination, there is a slight positive relationship between training size and performance, as well as with total size and performance. The relationship is stronger when performance is measured in terms of sensitivity compared to accuracy. Additionally, the trend demonstrated is stronger between training size and performance than between total size and performance.

To quantify the relationship between training size and performance and confirm the stronger effect of training size on performance compared to total size, I repeated the regression analysis performed previously, but instead for the relationship between the training sample size and performance. The full results are included in **Appendix B**. The relationship between training sample size and performance, while demonstrating a slightly stronger positive effect than in the total size analysis, is still negligible and statistically close to zero. I interpret this result as being partially due to my dataset size, which may not be large enough to establish statistical

significance. Even so, the slightly larger positive effect of training size on performance over total size suggests that the size of the training set specifically may be more influential on the performance of an AI model than the overall amount of data used in a study.

Relationship between Target Body System and Performance

My second hypothesis for my analysis regarded the relationship between the body system of interest in a model and its performance. To do this, I first examined the training size statistics in relation to each body system, as displayed in **Table 5**.

Table 5. Summary statistics of the training set sizes for each targeted body system, with the greatest and least values of each statistic emphasized.

	Training Size				
System	Count	Min	Mean	Median	Max
Cardiovascular	**13**	281	26319.8	**35970.0**	63528
Circulatory	10	**4699**	**82256.1**	15354.0	**695030**
Digestive	14	**9**	1484.1	754.0	5392
Integumentary	**1**	267	**267**	**267**	**267**
Nervous	2	100	1456.83	754	6863
Reproductive	5	698	2349	2349	4000
Respiratory	12	90	6653.4	4559	15371
Skeletal	2	1638	1638	1638	1638
Urinary	2	694	5392.5	5392.5	10091

A notable finding from this data set is the large range of training set sizes included within the models in my dataset. Models targeting the circulatory, cardiovascular, and respiratory systems had the largest training sizes on average. The circulatory system was especially noteworthy for having the largest maximum and minimum training sizes across the dataset. On the other end, models addressing the digestive and integumentary systems had the smallest training sizes on average, and also represented the smallest values within each statistic identified.

This disparity in training sizes by body system aligns with existing trends in medicine in which certain conditions or body systems receive much less funding proportional to their health burden than others (Gross 1999, Samuelson 2019).

After reviewing body system data characteristics, I performed a regression for the relationship between training size and performance and included a system fixed effect to control for differences between models addressing different body systems. I structured my analysis as

$$Performance = \beta_0 + \beta_1 * TrainingSize + Class + \sum \beta_{System} * I_{System} + \epsilon$$

such that any effects related to a target body system could be accounted for in analysis. The full regression results can be found in **Appendix C**. For both performance measures, I compared the system-controlled regression (9) to the prior training size performance regression controlled for class and year (8).

This analysis yielded mixed findings. Some body systems demonstrated positive correlations with both accuracy and sensitivity (circulatory, respiratory, and urinary). These also generally showed the strongest positive relationships with performance amongst the body systems measured. Other systems, however, had mixed performance results such that one measure is positive, and the other is negative or negligible. One example is of nervous system models, which displayed a notable positive relationship with accuracy (12.8% average increase), but only had a statistically negligible slight negative impact upon sensitivity (-4.8% average decrease). While these findings are not statistically significant due to a smaller data set size, they suggest that certain body systems are easier to diagnose with AI models. This could be for a variety of reasons, including the aforementioned trend of condition- and organ-related disparities in medical attention and funding within healthcare.

Qualitative Findings

During my interviews with clinicians, I encountered diverse perspectives and considerations that each individual doctor had developed due to their unique workplaces and specializations. However, there were many points of agreement amongst the interviewees. These points both corroborate and conflict with my prior findings in different areas while providing essential nuance and perspective regarding the health care work environment.

General Sentiments

Overall, the sentiment of those interviewed towards AI development can be described as positive. All of the doctors expressed interest in using AI in the future if it were available for their needs, as well as hope for improvements in future that would enhance their ability to improve patient outcomes. When asked to rate their comfortability with future use of diagnostic AI tools in medical practice from 1-10, with 10 being most comfortable, all of the doctors provided a rating between 5 to 10. Those on the lower end of this range cited the current "range of pros and cons" as well as being "unsure" how AI would be applied in their work as reasons for their ratings. Those on the higher end, while generally comfortable with AI health technologies, did not consider AI implementation as a full replacement of human involvement but rather a supplement for doctors' typical tasks. While those with lower comfortability rankings shared the belief that AI would function at most as a supplemental tool for clinicians, they expressed concern that hospital management may attempt to cut costs or substitute unfilled specialist positions with AI diagnostic tools, leaving medical staff to make up for the labor and skill gap. A couple of the doctors admitted to being comfortable with using AI technology at work despite having "no idea how it's programmed", relying on software developers and government

regulatory bodies to ensure that AI tools are designed well and are safe for use. However, even with the overall enthusiastic attitude towards AI in health care, interviewees cautioned against widespread application or overreliance on AI. Their biggest concerns are with current technological shortcomings and AI's consequences on medical liability, clinician employment, and workplace dynamics.

The interviewees all had varying experience with or knowledge of AI in the workplace, ranging from having very little understanding of AI to being intimately involved with medical technology development and implementation. Likewise, some doctors reported more experience with AI-enabled medical software in their daily work than others. One clinician, "Dr. T", described how her team uses AI-based diagnostic tools to help identify abnormalities in their patients' polysomnographies, which is then reviewed by a human technician to verify the results before any course of treatment is taken. She noted that AI software was helpful in her practice, as it allowed her to begin her reviews with a general understanding of her patient's sleep patterns and focused her efforts on correcting and clarifying the AI-generated flags rather than manually reviewing the full study without context. Other doctors, however, had little to no experience with AI used in diagnosis. Instead, the two AI-based medical applications that had been used by almost every interviewee were a comprehensive patient health record and billing system and an automatic notes dictation software.

Clinician Struggles in the Workplace

During the interviews, physicians agreed that the greatest struggles they face at work are staffing issues, management disconnect, and feeling overburdened with increasing responsibilities. These problems were mentioned by interviewees regardless of their workplace size or department of work. Their descriptions of their workplace environment provide a

valuable and nuance context within which to evaluate the usefulness of AI diagnostic tools in practice.

Staffing Issues

While staffing shortages are a longstanding issue that small clinics and large hospital systems alike face, clinicians' concerns with staff go beyond just a lack of employees. Rather, some also pointed to high turnover rates in medical staff and nurse roles as a compounding concern. One doctor, "Dr. U", lamented that high nurse turnover rates have resulted in more new or inexperienced nurses who are less likely to be familiar with typical courses of action or to be proactive in monitoring and caring for patients. The clinician is then obligated to keep a closer eye on patients and provide guidance to nurses in ways they normally would not have to, resulting in frustration and fatigue. Staffing shortages and high turnover together constitute a major struggle for physicians.

Management Disconnect

Another concern voiced by the interviewees is the perceived disconnect between hospital management and medical staff. Being familiar with multiple large hospital groups, Dr. U explained that there are "metrics that Medicare and Medicaid ding hospitals for" under their hospital quality performance rating, including the length of hospital stay per patient. A longer than normal stay is typically penalized by the government and reflects poorly on the hospital. As such, hospital management is incentivized to limit unnecessarily long patient stays. However, Dr. U found that this initiative also causes administration to pressure doctors to discharge patients early, causing doctors to have to spend additional time appealing to management on behalf of their still-recovering patients. The aforementioned staff shortages and high turnover rates also

impact hospital case managers, who are in charge of moving patients to other secondary care facilities as needed to reduce the length of hospital stay and allow for intake of new patients. Like with nurses, an influx of inexperienced or novice case managers has resulted in increased frustration for doctors who may lose track of their patients or be prevented from taking on new patients due to case mismanagement. Another interviewee voiced that, because of these issues, "management doesn't seem to care about doctors or patients" in the eyes of medical staff, a sentiment which creates distance and distrust between administrators and clinicians.

The disconnect between administrators and technicians in health care stems from the different motivations that affect each role. One of the interviewees, "Dr. V", operates in both a medical and administrative role in their practice. As an administrator, Dr. V is motivated to provide the best quality care to the maximum number of patients possible, at the best margins. As such, management pushes doctors for accurate and thorough medical records and billing. On the other hand, his goals as a clinician are to treat less patients per day but optimize the quality of care with adequate time to perform tasks well. In this role, his primary focus is on his patients' health and his clinical tasks, and the required billing work becomes more of a tedious afterthought. While management would prefer for health records to consist of numerous discrete data fields to ensure that all potential conditions are accounted for in billing, doctors find this structure unwieldy when trying to diagnose patients and prefer to take quicker, less structured notes. These clashing goals and priorities result in misunderstandings and distrust between hospital management and medical workers.

Even so, Dr. V finds the idea of AI implementation to be less contentious than some may expect, saying that "the general public thinks [using AI tools to augment care] is a point of misalignment [between administration and clinicians, but that's not necessarily true". He

explained that while there is a fear of losing jobs amongst some doctors, most believe that the existing physician shortage will prevent many hospitals from considering mass layoffs as a cost-cutting measure. Doctors are also primarily motivated by their ability to treat patients and improve people's health conditions, which many of the interviewees cited as one of their favorite aspects of their work. Furthermore, depending on the type of practice they work in, clinicians' salaries can be based on the number of patients they see. As such, this provides an additional incentive for doctors to increase their productivity in diagnosis and treatment. As both administrators and doctors perceive more advanced tools as increasing productivity, the implementation of AI/ML-enabled medical devices is often supported by both parties. While Dr. V's unique position allowed him to articulate the factors influencing both management and clinicians, his observations align with the positive sentiments expressed by other interviewees towards using AI in their medical practices in the future.

Increasing Burdens and Responsibilities

A third concern that was articulated by many of the interviewees was feeling overburdened by an increase in responsibilities. This was often viewed as a result of the first two issues as well as post-pandemic health trends. One interviewee, "Dr. W", who works in a small practice, noted that despite his best treatment efforts, his "ability to help is impacted by the patient's external situation". Dr. U also noted that the patients she sees "now come in very sick" and in worse conditions than they were before. These observations support nationwide trends of increased patient acuity after the COVID-19 pandemic (Requarth 2022, American Hospital Association 2022). This increased health burden, alongside the redistribution of responsibilities due to staffing shortages and high turnover rates, has caused clinicians like Dr. U to feel that "[hospitals are] asking too much of doctors these days".

While these shifts have resulted in a more stressful healthcare work environment, the most time-consuming tasks have remained the same. All of the interviewees mentioned documentation and paperwork as the most tedious tasks within their workdays. One of the doctors described the importance of documentation as “if you didn’t document it, it didn’t happen”, but also felt that the hours spent taking and reviewing notes “kind of takes away from patient care”. On average, the interviewees reported spending at least half of their workday on documentation or paperwork-related tasks, even with the use of speech recognition notetaking software. These permanent tasks, alongside the increase of new workplace pressures after the pandemic, contribute to feelings of burnout. The overall sentiments expressed by the doctors of overburdening and burnout corroborate the existing literature of rising post-pandemic stress within the healthcare industry, as well as the resulting rise in healthcare worker turnover rates (Aiken 2023, Rotenstein 2023, Tawfik 2019).

Perceived Benefits of AI

Through the interviews, the clinicians affirmed a shared belief that artificial intelligence diagnostics had the potential to enhance care in certain areas of health. When probed, they provided multiple potential benefits that AI could provide in their daily work. These included timesaving, improved monitoring of patients, expanding treatment options, and enhancing medical research.

One of the most mentioned benefits of AI diagnostics was its potential to save time, allowing doctors to treat more patients in a day. Clinicians such as Dr. T often rely on AI diagnostic tools to identify general patient health trends and to highlight discrepancies or abnormalities that require further examination. Without the heightened computing power of these applications, they must take more time to review the available data and synthesize a potential

diagnosis and treatment plan. As Dr. X noted, this increased efficiency can help to reduce the heightened levels of burnout that many doctors currently face.

Another benefit identified by the interviewees was the ability to track and receive prompt updates on changes in a patient's condition in real time. Typically, these are tasks performed at regular intervals by nurses and other medical staff. However, with current staffing shortages, there are longer periods of time in which a patient may be unattended. This can result in seemingly stable patients becoming unresponsive or experiencing sudden health crises without help for longer periods of time. With AI supplementation, health systems can make up for staffing concerns and ensure constant care so that doctors can be immediately responsive to patient needs.

Besides improving care efficiency, interviewees mentioned AI's capacity to augment clinicians' course of treatment by providing better care suggestions. Dr. V provided examples of AI tools that could suggest up-to-date treatment options that are better suited as a course of action than the standard care plan given patient characteristics, or that can predict odds of remission so that doctors can be more informed on a patient's condition. Alternatively, AI diagnostic tools can be used to quickly identify rare or specific conditions that doctors may be less familiar with, or that they encounter less often in the course of their work. This could then streamline the care process by pinpointing elusive diseases that may otherwise be difficult to diagnose promptly without the aid of a specialist.

Some of the interviewees who were more involved or familiar with medical research within their institutions also mentioned that AI diagnostic tools are good for conducting and synthesizing research for use in care. One reason for this is because, as explained by Dr. V, a lot of clinical trials are quite small and thus are not significant or reliable on their own. However, AI

can compile datapoints from multiple trials and produce more significant results. These outcomes are then more diverse and widely applicable to patients that clinicians may encounter, thus increasing the span of medical knowledge that a doctor can reference when diagnosing a patient.

Perceived Detriments, Concerns, and Barriers to Implementation

While the interviewees were generally in favor of AI implementation in the future, they voiced concerns as to the potential pitfalls of existing AI tools as well as the negative consequences that AI implementation could have on medical workplaces and physicians. The scope of their worries ranged from individual consequences upon patients and physicians to a fundamental shift in the healthcare landscape. Overall, many physicians felt that AI in its current state could not be implemented widely in a diagnostic context before their larger looming questions and issues surrounding this technology are resolved.

A concern that many doctors had was the negative impact of AI on clinicians. Employment worries varied amongst the participants. While some doctors were adamant that AI would never fully replace human doctors in their practice, others were concerned that hospital systems would use AI to “cut corners” and delay the hiring of needed accredited doctors, or lay off doctors in specializations that already utilize diagnostic technology more frequently such as radiology and pathology. However, the interviewees who were more optimistic of future employment trends were unconvinced of the severity of these staffing cuts, pointing towards the current lack of skilled radiologists across the country that can operate AI tools effectively as evidence that their jobs would not be put at risk. Separate of employment concerns, some interviewees voiced concerns that having greater AI use in hospitals, while immensely useful for ensuring proper patient care, could result in hostile work dynamics and feelings of distrust. One doctor referenced a local health system that came under fire by nursing unions for adopting AI-

assisted camera feeds to measure how often patients are turned. While the hospital employed this technology to ensure that patients were receiving quality care, nurses contended that the camera feeds were an invasion of privacy and a sign of institutional distrust of nurse competence. The union also argued that the technology could be used to monitor nurses' activities, and to identify and fire those with suboptimal performance or hand hygiene rates. While AI diagnostic technologies have the potential to enhance care and reduce work burdens, their differing effects on different roles can create tensions between medical colleagues and lead to a more negative workplace culture.

Another barrier to implementation identified by interviewees was the lack of clear liability guidelines or legal protections for doctors interested in using new AI diagnostic tools with their patients. When doctors make a mistake, they can be subject to a malpractice lawsuit or other legal action. However, due to the relative recent development of AI-based diagnostic technology, physicians are unsure as to where the legal burden of a faulty AI model lies, and if their employers are able to offer legal protections when implementing new AI tools. Dr. V cautioned that "pioneer physicians" who choose to be one of the first to adopt a new technology are typically the person held liable if the technology fails or is faulty, not the companies that developed the AI tool. This in turn makes implementation of new medical devices scary, especially in the case of AI where doctors are unable to understand how the algorithms process data.

The interviewees' perceived detrimental effects of AI also extended to patients' health. A major point of concern was the technology's potential to reduce the quality of care provided under certain conditions. For example, Dr. X warned that AI tools implemented in the current workplace may be used without proper supervision or skilled validation of outputs due to a lack

of specialists. Another clinician, “Dr. Y”, hesitated to recommend AI for widespread use, likening AI diagnosis to a “cookie cutter” solution inappropriate for an extremely diverse patient population. Yet another issue for doctors was the possibility of reducing the human aspect of healthcare. Dr. U explained that one of the most important aspects of receiving care is the human interaction between patients and the staff treating them. She observed that her patients often seem lonely or scared, and as such seek out company and comfort through conversations with the medical staff treating them. However, this aspect of care is not quantifiable, and as such not considered as part of traditional measures of care performance or outcomes. If doctors are pushed to interact less often with patients due to the advent of AI diagnostic tools for efficiency, then this could undermine the patient-provider relationship in healthcare.

Finally, interviewees also expressed concerns with the models being developed in the status quo. Dr. X noted that the biggest objection to newly implemented tools amongst medical staff in his practice is that the tool design is not user-friendly. Without medical practitioners involved in the development of medical devices and software, software engineers may design solutions that look optimal to a developer but are difficult or tedious to use in the medical context. Another concern that resonates outside of the healthcare industry is in regards to data privacy. As Dr. X pointed out, HIPAA provides strict standards for medical data privacy within the United States. However, the development and use of AI-based medical tools entails placing sensitive user data into the hands of private companies that may not be as motivated as hospitals to protect patient privacy, or to be fully transparent about how they design their models. These considerations make clinicians less willing to use AI diagnostic tools in the present day.

While data privacy in itself is an issue, there is also a limit to the complexity and types of models that can be developed due to gaps in publicly available medical image databases. As

mentioned by multiple interviewees, medical scans and images are often considered to be not just liable to HIPAA laws but also valuable intellectual property. As such, many private hospitals, educational institutions, and private imaging companies who generate these image datasets often keep their databases private, only making them accessible to those within their network under strict conditions. This, in turn, stunts AI development, limiting both the performance and applicability of an AI tool within a large population. The limited access to diverse training data may also cause algorithmic bias to develop within the AI model, especially when attempting to use the tool on a demographic that is underrepresented in the training data. This lack of large public datasets results in AI tools with low positive predictive values (PPV), which doctors like Dr. W are wary of using in their daily work. While emerging studies seem promising, doctors are not yet convinced that AI diagnostic models are fully trained and equipped for widespread implementation.

Discussion

There are several points of alignment and contrast between the quantitative findings deduced from the meta-analysis, the clinician statements given in the interviews.

Several of the studies collected for the meta-analysis noted either a significantly improved or equal level of performance compared to raw physician diagnosis (Nishi 2020). The articles that studied AI-augmented physician performance alongside raw AI performance also found that physicians were able to make slightly more accurate diagnoses while using AI tools (Nijiati 2021). The interviews support this finding, as several clinicians perceived AI tools to be useful in diagnosing uncommon conditions and suggesting optimal treatment plans based on patient data. The statistical confirmation of AI efficacy in assisting clinicians in their everyday

work furthers the hopes of physicians and healthcare providers that AI diagnostic tools can be beneficial in streamlining healthcare workflows and would improve patient outcomes.

The wide range of opinions regarding employment outcomes after AI implementation provides context to the ongoing debate as to the role that AI should play in health care. In areas with acute shortages of skilled clinicians, AI-enabled technology is seen as an opportunity to bridge the gap between patient health burden and provider care capacity (Qin ZZ 2021). However, the physicians interviewed overwhelmingly believed that high performance AI will be seen by management as an opportunity to reduce salary costs by forgoing the acquisition of medical specialists. Studies included in the meta-analysis focused heavily on AI performance capabilities and their ability to augment clinician diagnosis, but failed to examine the effects of AI in the hands of non-specialized practitioners. As such, this nuance provides reason for doctors to oppose implementation of AI diagnostic tools, and is a point of conflict that must be resolved before AI diagnostic applications can be successfully used within health systems.

The variance in experience with diagnostic AI tools in the workplace across interview participants indicates that some specialties have been more open to AI in certain areas of their work, while other practices may still be resistant to or considered too complex for introducing an AI tool. This aligns with the belief held by several interviewees that some professions may be more adversely affected by AI implementation, especially if hospital administrators and management believe that AI can more competently replace human examination for a specific area of medicine. This belief is supported by the meta-analysis findings regarding body system performance disparities, as models addressing body systems with greater available image documentation also displayed notable increases in average performance. On the other hand, the AI tools that are most widely used in health care right now tend to be for administrative tasks,

specifically billing and documentation. One reason for this commonality could be due to demand, as doctors find these tasks to be the most tedious and time-consuming in their workday. Another factor is the prioritization of these tasks by management, who are motivated in their roles to charge accurately for medical procedures and to keep track of any metrics related to clinical activities and performance. As such, this widespread adoption of administrative medical software could be indicative of internal stakeholder goals influencing the types of AI tools that hospitals are willing to invest in for use in the health workplace.

Policy Recommendations & Conclusion

The expanding role of artificial intelligence in health care calls for increased government involvement and policy implementation in order to guide technological development in a direction that is beneficial for patients and practitioners alike. From the existing data, AI diagnostic technologies have proven themselves to be effective if developed appropriately. However, the current unclear legal landscape and lack of clear programming standards tailored for medical AI development prevents medical stakeholders from readily adopting existing diagnostic tools.

Government bodies, as part of their commitment to medical ethics and human rights, should tighten and clarify regulatory standards for AI/ML medical device development. First, they should require a minimum size for the training sample set that can ensure a specific level of performance from emergent AI tools. As evidenced in the meta-analysis, there is a slight positive effect of training set size on model performance. This is also supported by the existing literature—studies that have tested identically-designed CNN models with the same total sample size but different ratios of training to testing have found that accuracy increases with larger training sets (Yotsu et al. 2023). Ultimately, a robust training set size is crucial to ensure that

models can perform at high levels of accuracy. Beyond the size of the training set, states should also ensure that the data sets used to train AI technologies meet adequate diversity criteria to ensure that an algorithm is generally applicable to an entire population regardless of race, gender, or other factors such that inadvertent discrimination in health outcomes is avoided. While governments should create specific legislation and accountability measures to confirm that AI development businesses comply with these standards, businesses should regardless consider their commitment to using large and diverse datasets as part of their duty to respect the right to health. As noted earlier, the FDA performs tests on the external validity of each model that is submitted for approval. Beyond just noting the extent to which an AI tool can be generally applied to a population and using this criterion privately to make approval decisions, the FDA and other similar regulatory institutions should publicly release guidance for constructing diversely-applicable and usable AI/ML-enabled medical devices. These guidelines should also include a requirement of model review or certification by medical practitioners in order to ensure that the tool is designed appropriately for the healthcare workplace. This measure is to ensure that concerns of usability like those raised by interviewees can be addressed before a tool is made publicly available. The incorporation of healthcare workers into the design process can provide clinicians with an increased understanding of the technology being offered to them as well as improved trust in the performance of the algorithm, thus helping to overcome the institutional hurdles faced in introducing AI tools in healthcare. In this way, software companies are made aware of the level of data diversity and multidisciplinary collaboration that is expected of them and can strive to maintain these standards in their work.

Another concern that governments and AI developers should work to address is the need to develop larger public image datasets, with a prioritization of currently underfunded and

underdiagnosed conditions. The growth of AI in healthcare is often touted as an opportunity to make up for existing care discrepancies. However, my analysis shows that existing research efforts have yet to bridge these gaps. Many body systems that are typically under-addressed continue to be overlooked in the development of medical imaging databases, which then limits the performance and capabilities of AI models meant to address under-researched or rare conditions. Even if an AI developer is able to source a large amount of non-professional images, the lack of professional images available does have a significant effect on the potential performance of the tool. In the case of one tool studied in my meta-analysis, a model trained on 58,000 smartphone images taken by consumers performed at an extremely low accuracy (Zaar et al. 2020). Cases such as this provide greater impetus for public institutions to create databases with not only a high quantity, but also a high quality of images. Even though it is important not to attention away from conditions that are already topical, it is critical to prioritize the development of diagnostic technologies for the conditions that are undertreated so that AI applications can most effectively make up for existing gaps in health care. To this end, governments should work to incentivize the development of AI diagnostic tools targeting currently under-addressed diseases as well as the construction of more comprehensive image datasets of these conditions. One way in which governments can do this is by creating greater incentives for private institutions to share and pool imaging data to counter existing corporate and intellectual property interests. Alternatively, public initiatives can be developed between government agencies and public educational institutions to collect images. These measures are necessary not only for developers to be able to design effective algorithms, but also for medical professionals and the general public to gain confidence in the usability of these applications.

Ultimately, the increased volume of AI/ML-based medical devices that have been developed over the past years parallels the growth of AI-related legislation across the nation (FDA 2023, Zhu 2023). However, the lack of policy specifically addressing healthcare-related concerns in an industry where a faulty algorithm can constitute high risk for patient health leaves a gap in security and comfortability for the doctors and patients most directly affected by diagnostic AI use. Increased tailored and clear regulation, expanded multidisciplinary development efforts, and greater data availability are all steps that governing bodies must take to ensure that AI diagnostic applications are able to augment the provision of healthcare with minimal negative consequences to the stakeholders involved.

References

American Hospital Association. “Pandemic-Driven Deferred Care Has Led to Increased Patient Acuity in America’s Hospitals.” *American Hospital Association*, Aug. 2022, www.aha.org/guidesreports/2022-08-15-pandemic-driven-deferred-care-has-led-increased-patient-acuity-americas.

“AI’s Ascendance in Medicine: A Timeline.” *Cedars Sinai*, 20 Apr. 2023, www.cedars-sinai.org/discoveries/ai-ascendance-in-medicine.html#:~:text=Scientists%20began%20laying%20the%20groundwork,the%20empowering%20technology%20has%20proliferated.

Advanced Medical Technology Association. “Artificial Intelligence in Medical Technology Myths vs. Facts.” *AdvaMed*, 21 Feb. 2024, www.advamed.org/wp-content/uploads/2024/02/AI-Myths-vs.-Facts.pdf.

Aiken, Linda H., et al. “Physician and nurse well-being and preferred interventions to address burnout in hospital practice.” *JAMA Health Forum*, vol. 4, no. 7, 7 July 2023, https://doi.org/10.1001/jamahealthforum.2023.1809.

Al’Aref, Subhi J, et al. “Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging.” *European Heart Journal*, vol. 40, no. 24, 2018, pp. 1975–1986, https://doi.org/10.1093/eurheartj/ehy404.

Asan, Onur, et al. “Artificial Intelligence and human trust in healthcare: Focus on clinicians.” *Journal of Medical Internet Research*, vol. 22, no. 6, 19 June 2020, https://doi.org/10.2196/15154.

Blouin, Lou. “AI’s Mysterious ‘Black Box’ Problem, Explained.” *Dearborn News*, University of Michigan-Dearborn, 6 Mar. 2023, umdearborn.edu/news/ais-mysterious-black-box-problem-explained.

Bohr, Adam, and Kaveh Memarzadeh. “The rise of Artificial Intelligence in healthcare applications.” *Artificial Intelligence in Healthcare*, 2020, pp. 25–60, https://doi.org/10.1016/b978-0-12-818438-7.00002-2.

Digital Health Center of Excellence. “Good Machine Learning Practice for Medical Device Development.” *U.S. Food and Drug Administration*, FDA, 27 Oct. 2021, www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

Elemento, Olivier, et al. “Artificial Intelligence in cancer research, diagnosis and therapy.” *Nature Reviews Cancer*, vol. 21, no. 12, 17 Sept. 2021, pp. 747–752, https://doi.org/10.1038/s41568-021-00399-1.

FDA. “Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices.” *U.S. Food and Drug Administration*, 6 Dec. 2023, www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

Fenton, Robert. “How Long Does the FDA Medical Device Approval Process Take? [Timeline].” *Qualio*, Qualio QMS for Life Sciences, 27 July 2021, www.qualio.com/blog/fda-medical-device-approval-process#:~:text=The%20FDA%20approval%20process%20can,is%20not%20a%20fast%20process.

Gross, Cary P., et al. “The relation between funding by the National Institutes of Health and the burden of disease.” *New England Journal of Medicine*, vol. 340, no. 24, 17 June 1999, pp. 1881–1887, https://doi.org/10.1056/nejm199906173402406.

Hatherley, Joshua James. “Limits of trust in medical AI.” *Journal of Medical Ethics*, vol. 46, no. 7, 27 Mar. 2020, pp. 478–481, https://doi.org/10.1136/medethics-2019-105935.

Huang, Shigao, et al. “Artificial Intelligence in cancer diagnosis and prognosis: Opportunities and challenges.” *Cancer Letters*, vol. 471, 28 Feb. 2020, pp. 61–71, https://doi.org/10.1016/j.canlet.2019.12.007.

Lin, Lin et al. “Combining collective and artificial intelligence for global health diseases diagnosis using crowdsourced annotated medical images.” Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference vol. 2021 (2021): 3344-3348. doi:10.1109/EMBC46164.2021.9630868

Loh, Hui Wen, et al. “Automated detection of ADHD: Current trends and future perspective.” *Computers in Biology and Medicine*, vol. 146, 2022, p. 105525, https://doi.org/10.1016/j.compbiomed.2022.105525.

Khan, Nagina, et al. “Post-COVID-19: Can Digital Solutions lead to a more equitable global healthcare workforce?” *BJPsych International*, vol. 20, no. 1, 7 Apr. 2022, pp. 18–23, https://doi.org/10.1192/bji.2022.12.

Kiseleva, Anastasiya, et al. “Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations.” *Frontiers in Artificial Intelligence*, vol. 5, 30 May 2022, https://doi.org/10.3389/frai.2022.879603.

Kranzfelder, Michael, et al. “Toward increased autonomy in the surgical OR: Needs, requests, and expectations.” *Surgical Endoscopy*, vol. 27, no. 5, 13 Dec. 2012, pp. 1681–1688, https://doi.org/10.1007/s00464-012-2656-y.

Miller, R. A. “A history of the internist-1 and Quick Medical Reference (QMR) computer-assisted diagnosis projects, with lessons learned.” *Yearbook of Medical Informatics*, vol. 19, no. 01, 2010, pp. 121–136, https://doi.org/10.1055/s-0038-1638702.

National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. 29 Dec. 2015, Washington, DC: The National Academies Press, https://www.ncbi.nlm.nih.gov/books/NBK338596/

Nitiéma, Pascal. “Artificial Intelligence in Medicine: Text Mining of Health Care Workers' Opinions.” Journal of medical Internet research vol. 25 e41138. 27 Jan. 2023, doi:10.2196/41138

Okyere, Eunice, et al. “Is task-shifting a solution to the health workers’ shortage in northern Ghana?” *PLOS ONE*, vol. 12, no. 3, 2017, https://doi.org/10.1371/journal.pone.0174631.

Oster, Natalia V., et al. “COVID-19’s Effect on the Employment Status of Health Care Workers.” *Center for Health Workforce Studies*, University of Washington, May 2021, familymedicine.uw.edu/chws/wp-content/uploads/sites/5/2021/05/Health_Employ_Status_PB_May_26_2021.pdf.

Pai, Raghav K et al. “A review of current advancements and limitations of artificial intelligence in genitourinary cancers.” *American journal of clinical and experimental urology* vol. 8,5 152-162. 15 Oct. 2020, https://pubmed.ncbi.nlm.nih.gov/33235893/.

“Primary Care Workforce Projections.” *Bureau of Health Workforce*, Health Resources and Services Administration, 21 Sept. 2021, bhw.hrsa.gov/data-research/projecting-health-workforce-supply-demand/primary-health.

Requarth, Tim. “‘It Just Seems like My Patients Are Sicker.’” *The Atlantic*, Atlantic Media Company, 1 Sept. 2022, www.theatlantic.com/health/archive/2022/08/america-health-premature-death-disability-post-pandemic/671276/.

Rotenstein, Lisa S., et al. “The association of work overload with Burnout and intent to leave the job across the healthcare workforce during COVID-19.” *Journal of General Internal Medicine*, vol. 38, no. 8, 23 Mar. 2023, pp. 1920–1927, https://doi.org/10.1007/s11606-023-08153-z.

Samaran, Romain et al. “Interest in artificial intelligence for the diagnosis of non-melanoma skin cancer: a survey among French general practitioners.” European journal of dermatology : EJD vol. 31,4 (2021): 457-462. doi:10.1684/ejd.2021.4090

Samuelson, Kristin. “Many of the deadliest cancers receive the least amount of research funding.” *Northwestern Now*, 18 July 2019, https://news.northwestern.edu/stories/2019/07/disparities-cancer-research-funding/

Shaheen, Mohammed Yousef. "Applications of artificial intelligence (AI) in Healthcare: A Review." *ScienceOpen Preprints*, 25 Sept. 2021, https://doi.org/10.14293/s2199-1006.1.sor-.ppvry8k.v1.

Tawfik, Daniel S., et al. "Evidence relating health care provider burnout and quality of care." *Annals of Internal Medicine*, vol. 171, no. 8, 2019, p. 555, https://doi.org/10.7326/m19-1152.

The Green Center. "Quick Covid-19 Primary Care Survey Series 24 Fielded December 11-15, 2020." *The Larry A. Green Center*, 17 Dec. 2020, static1.squarespace.com/static/5d7ff8184cf0e01e4566cb02/t/5fde274bd85ca26442f7bf17/1608394572823/C19+Series+24+National+Executive+Summary.pdf.

The Green Center. "Quick Covid-19 Primary Care Survey Series 37 Fielded March 13-19, 2023." *The Larry A. Green Center*, static1.squarespace.com/static/5d7ff8184cf0e01e4566cb02/t/5fde274bd85ca26442f7bf17/1608394572823/C19+Series+24+National+Executive+Summary.pdf.

Umapathy, Vidhya Rekha, et al. "Perspective of artificial intelligence in disease diagnosis: A review of current and future endeavours in the medical field." *Cureus*, vol. 15, no. 9, 21 Sept. 2023, https://doi.org/10.7759/cureus.45684.

United Nations General Assembly. "International Covenant on Economic, Social, and Cultural Rights." *Treaty Series, 993*, *3,* 1966. New York.

United Nations Human Rights Council. "Digital innovation, technologies, and the right to health." *Report by Special Rapporteur on the right of everyone to the enjoyment of the highest attainable standard of physical and mental health*, 21 April 2023, UN Doc A/HRC/53/65.

United Nations Office of the High Commissioner for Human Rights (OHCHR). *Guiding Principles on Business and Human Rights Implementing the United Nations "Protect, Respect and Remedy" Framework*. United Nations, 2011.

Van Schalkwyk, May CI, et al. "The best person (or machine) for the job: Rethinking task shifting in Healthcare." *Health Policy*, vol. 124, no. 12, 2020, pp. 1379–1386, https://doi.org/10.1016/j.healthpol.2020.08.008.

Yotsu, Rie R., et al. "Deep learning for AI-based diagnosis of skin-related neglected tropical diseases: A pilot study." *PLOS Neglected Tropical Diseases*, vol. 17, no. 8, 14 Aug. 2023, https://doi.org/10.1371/journal.pntd.0011230.

Zhu, Katrina. "The State of State AI Laws: 2023." *EPIC*, 3 Aug. 2023, epic.org/the-state-of-state-ai-laws-2023/.

Meta-Analysis Articles

Adedinsewo D, et al., 2020

Ahn J M, et al., 2018

Al-Sarem M, et al., 2022

Attia ZI, et al., 2019

Bhattacharjee, 2022

Bonnevie ED, et al., 2023

Campanella G, et al., 2022

Celik B, et al., 2023

Christ M, et al., 2024

Feng S, et al., 2022

Foersch S, et al., 2021

George-Jones NA, et al., 2021

Golovanevsky M, et al., 2022

Howard JP, et al., 2021

Jain DK, et al., 2022

Jameela T, et al., 2022

Jaugey A, et al., 2023

Kimura K, et al., 2019

Kono M, et al., 2021

Lee JH, et al., 2018

Li M, et al., 2023

Lin L, et al., 2021

Marginean F, et al., 2021

Mascarenhas Saraiva M, et al., 2021

Nijiati M, et al., 2021

Nishi T, et al., 2020

Ribiero AH, et al., 2020

Serte S & Demirel H, 2021

Sharma A, et al., 2023

Sun H, et al., 2023

Tang MCS, et al, 2021

Wu L, et al., 2023

Wu L, et al., 2023

Yang J, et al., 2020

Yoo JW, et al., 2022

Yousefi B, et al.,, 2020

Zaar O, et al., 2020

Zhu J, et al., 2023

Programs and Software

Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer.

Interviews

Dr. T*. “Interview”, 28 March 2024.

Dr. U*. “Interview”, 3 April 2024.

Dr. V*. “Interview”, 5 April 2024.

Dr. W*. “Interview”, 6 April 2024.

Dr. X*. “Interview”, 6 April 2024.

Dr. Y*. “Interview”, 8 April 2024.

*Pseudonym used for anonymity of participants.

Appendix

Appendix A

Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle	Good Software Engineering and Security Practices Are Implemented
Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population	Training Data Sets Are Independent of Test Sets
Selected Reference Datasets Are Based Upon Best Available Methods	Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
Focus Is Placed on the Performance of the Human-AI Team	Testing Demonstrates Device Performance During Clinically Relevant Conditions
Users Are Provided Clear, Essential Information	Deployed Models Are Monitored for Performance and Re-training Risks are Managed

Source: Digital Health Center of Excellence 2021

Appendix B

Table 5: Accuracy Results - Training Size

	Dependent variable:			
	Accuracy			
	(5)	(6)	(7)	(8)
Training Size	8.555e-07	9.155e-07	9.953e-07	1.013e-06
	(1.340e-06)	(1.371e-06)	(1.753e-06)	(1.781e-06)
Class		0.008		0.010
		(0.028)		(0.035)
Year)2019			-0.031	-0.031
			(0.104)	(0.106)
Year)2020			-0.051	-0.052
			(0.060)	(0.061)
Year)2021			0.025	0.019
			(0.042)	(0.047)
Year)2022			0.075	0.074
			(0.047)	(0.048)
Year)2023			0.027	0.020
			(0.046)	(0.053)
Year)2024			0.050	0.050
			(0.056)	(0.057)
Constant	0.881***	0.878***	0.852***	0.852***
	(0.016)	(0.019)	(0.035)	(0.035)
Observations	39	39	39	39
R^2	0.011	0.013	0.215	0.217
Adjusted R^2	-0.016	-0.041	0.037	0.008
Residual Std. Error	0.079 (df = 37)	0.080 (df = 36)	0.077 (df = 31)	0.078 (df = 30)
F Statistic	0.408 (df = 1; 37)	0.244 (df = 2; 36)	1.210 (df = 7; 31)	1.039 (df = 8; 30)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Sensitivity Results - Training Size

	Dependent variable:			
	Sens			
	(5)	(6)	(7)	(8)
Training Size	1.159e-07	1.214e-07	1.707e-07	1.228e-07
	(1.340e-07)	(1.392e-07)	(1.948e-07)	(2.063e-07)
Class		-0.005		0.027
		(0.029)		(0.037)
Year)2019			-0.031	-0.027
			(0.116)	(0.117)
Year)2020			0.015	-0.001
			(0.072)	(0.076)
Year)2021			0.018	0.014
			(0.073)	(0.074)
Year)2022			0.077	0.074
			(0.073)	(0.073)
Year)2023			-0.019	-0.038
			(0.070)	(0.074)
Year)2024			0.002	0.002
			(0.084)	(0.084)
Constant	0.896***	0.898***	0.881***	0.881***
	(0.014)	(0.018)	(0.065)	(0.065)
Observations	44	44	44	44
R^2	0.018	0.018	0.155	0.169
Adjusted R^2	-0.006	-0.030	-0.009	-0.022
Residual Std. Error	0.092 (df = 42)	0.093 (df = 41)	0.092 (df = 36)	0.092 (df = 35)
F Statistic	0.749 (df = 1; 42)	0.381 (df = 2; 41)	0.946 (df = 7; 36)	0.887 (df = 8; 35)
Note:				*p<0.1; **p<0.05; ***p<0.01

Appendix C

Table 7: Accuracy - Body Systems

	Dependent variable:	
	Accuracy	
	(8)	(9)
Training Size	1.013e-06	9.882e-07
	(1.781e-06)	(2.149e-06)
Class	0.010	0.022
	(0.035)	(0.046)
Year)2019	-0.031	
	(0.106)	
Year)2020	-0.052	
	(0.061)	
Year)2021	0.019	
	(0.047)	
Year)2022	0.074	
	(0.048)	
Year)2023	0.020	
	(0.053)	
Year)2024	0.050	
	(0.057)	
System)Cardiovascular System		0.097
		(0.090)
System)Circulatory System		0.141
		(0.091)
System)Digestive System		0.118
		(0.109)
System)Nervous System		0.128
		(0.112)
System)Reproductive System		-0.011
		(0.133)
System)Respiratory System		0.091
		(0.107)
System)Urinary System		0.172
		(0.128)
Constant	0.852***	0.759***
	(0.035)	(0.111)
Observations	39	39
R^2	0.217	0.195
Adjusted R^2	0.008	-0.055
Residual Std. Error	0.078 (df = 30)	0.081 (df = 29)
F Statistic	1.039 (df = 8; 30)	0.780 (df = 9; 29)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 8: Sensitivity - Body Systems

	Dependent variable:	
	Sens	
	(8)	(9)
Training Size	1.228e-07	9.179e-08
	(2.063e-07)	(1.493e-07)
Class	0.027	-0.020
	(0.037)	(0.036)
Year)2019	-0.027	
	(0.117)	
Year)2020	-0.001	
	(0.076)	
Year)2021	0.014	
	(0.074)	
Year)2022	0.074	
	(0.073)	
Year)2023	-0.038	
	(0.074)	
Year)2024	0.002	
	(0.084)	
System)Circulatory System		0.014
		(0.046)
System)Digestive System		-0.036
		(0.046)
System)Nervous System		-0.048
		(0.054)
System)Reproductive System		0.107
		(0.101)
System)Respiratory System		0.035
		(0.061)
System)Skeletal System		-0.009
		(0.080)
System)Urinary System		0.037
		(0.104)
Constant	0.881***	0.912***
	(0.065)	(0.043)
Observations	44	44
R^2	0.169	0.146
Adjusted R^2	-0.022	-0.081
Residual Std. Error	0.092 (df = 35)	0.095 (df = 34)
F Statistic	0.887 (df = 8; 35)	0.643 (df = 9; 34)
Note:		*p<0.1; **p<0.05; ***p<0.01