CSCI 8980 - Lecture Scribe

Lecture 2: History and Unique Challenges

Scriber: Mert Acar | Monday, January 22, 2024

Various algorithms have been created to support clinicians in diagnostic procedures. These were primarily rule-based systems, originating from empirical observations and clinical studies. A notable example is the Body Mass Index (BMI), devised in the early 19th century by Adolphe Quetelet. BMI employs a straightforward formula involving a person's weight and height to classify their body mass as underweight, normal, overweight, or obese. Another instance is the TNM staging system for cancer, introduced by Pierre Denoix during the 1940s. TNM represents Tumor size, Node involvement, and Metastasis, offering a standardized method for describing how far cancer has spread. A further example is the Patient Health Questionnaire-9 (PHQ-9), a commonly used tool for detecting the presence and severity of depression in both clinical and research environments. Created in the late 1990s as a part of the broader Patient Health Questionnaire, the PHQ-9 is grounded in the diagnostic criteria for major depressive disorder. Although these algorithms were groundbreaking at their launch, they depend significantly on manual calculations and utilize general parameters that overlook individual genetic differences and environmental influences.

The significant shift in medical algorithms came with the adoption of machine learning (ML) technology, which advanced beyond the straightforward, linear calculations of earlier algorithms to embrace pattern recognition, natural language processing, and predictive analytics. A pioneering example from this era is the MYCIN expert system, developed in the 1970s. Edward Shortliffe, a PhD student at Stanford at the time, designed MYCIN to identify bacterial infections like bacteremia and meningitis and to suggest antibiotics, considering factors such as the patient's weight and the possibility of drug resistance. MYCIN used a rule-based system with a dialogue interface, relying on approximately 600 rules to determine the most suitable treatment. A distinctive feature of MYCIN was its capacity to justify its recommendations. When asked why a specific antibiotic was suggested, MYCIN could outline its reasoning, detailing the rules it followed and the rationale behind them, based on the input data. This ability to explain its decision-making process not only fostered trust among healthcare professionals in the system's advice but also offered valuable insights into AI's decision-making, a feature that continues to be important in AI applications today. Despite showcasing the potential of AI in healthcare and its advanced features, MYCIN was never adopted for clinical use.

Another noteworthy system from the 1970s is CASNET, developed at Rutgers University, designed to assist in diagnosing and treating glaucoma. CASNET was a part of a larger initiative focused on understanding and simulating clinical reasoning using AI techniques. In contrast to the rule-based systems prevalent among its peers, CASNET stood out for its application of a network-based method to map out the complex causal relationships inherent in medical knowledge. The foundation of CASNET was laid on causal associational networks, graphically depicting the connections among symptoms, diseases, and treatments. This framework enabled CASNET to accurately represent the intricate and sometimes complicated paths involved in the medical diagnostic and treatment process, particularly in conditions like glaucoma that exhibit considerable variability and uncertainty. Although CASNET introduced an innovative approach and contributed valuable insights into AI's role in medicine, it largely remained a research instrument and did not achieve broad clinical implementation, similar to MYCIN.

Lastly, the INTERNIST-1 system distinguished itself with an extensive knowledge base that encompassed more than 600 diseases and 4,500 disease manifestations, including symptoms, physical signs, lab results, and epidemiological data. This comprehensive database enabled INTERNIST-1 to explore a broad array of potential diagnoses across the full range of internal medicine. Despite its

sophisticated features, INTERNIST-1 encountered obstacles that hindered its clinical application. The system demanded considerable computational power and expert input to maintain and update its vast knowledge base, a challenging task amidst the swift progress in medical science. It employed a detailed scoring system to assess the probability of various diseases based on the patient's information. The difficulties of assimilating such advanced systems into clinical routines, combined with concerns regarding trust, reliability, and ethical issues, curtailed its direct use in patient care.

Initially, machine learning (ML) applications in healthcare faced significant challenges due to the scarcity of meaningful, labeled data, as health records were not extensively digitized, necessitating a considerable amount of manual labor. Additionally, the absence of standardized record-keeping practices posed difficulties, as it prevented different institutions from interfacing or pooling data together, resulting in very limited interoperability. Moreover, similar to the challenges faced by the INTERNIST-1 system, many ML systems required substantial computational resources, which were not readily accessible outside of research environments. This led clinicians and hospital administrators to favor rule-based, relatively simpler systems.

The introduction of computers and a dramatic increase in computing power drastically changed the landscape of medical algorithms. By the late 20th century, the emergence of electronic health records (EHRs) and advancements in data storage and processing abilities facilitated more sophisticated data analyses. The adoption of EHRs in the early 2010s saw a remarkable surge, as depicted in Figure 1. Consequently, algorithms evolved to include a wider range of variables and outcomes, allowing for more accurate risk assessments and diagnostic tools. Statistical software and decision support systems became fundamental to medical research and practice, enabling the analysis of vast datasets to uncover trends and correlations previously unnoticed.

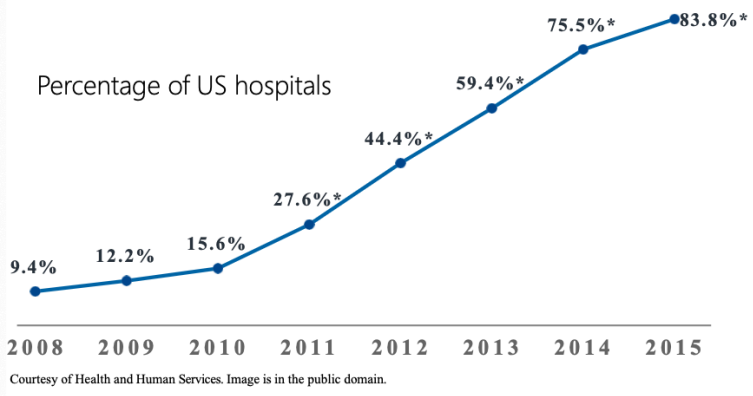

Figure 1 - Adaptation of Electronic Health Records across US hospitals.

The transition to digital records catalyzed the creation of large-scale, open datasets across various medical modalities. For instance, the MIMIC dataset made anonymized data from approximately 40,000 critical care patients available for research. Similarly, the fastMRI project by NYU released an extensive dataset of magnetic resonance imaging scans covering anatomical parts such as the knee, brain, and prostate. As emerging technologies like smartwatches, smartphones, and genomics reports became increasingly integrated into healthcare, the need for standardization surged. This led to the development of standards such as the International Classification of Diseases (ICD-10) diagnostic codes, National Drug Codes (NDC), and the Unified Medical Language System (UMLS).

Advancements in machine learning (ML) and artificial intelligence (AI), driven by improvements in both hardware and software (including GPUs and deep learning frameworks), facilitated the emergence of more sophisticated healthcare algorithms. These advancements made it possible to analyze long-range,

high-dimensional dependencies within datasets—tasks that would be impractical to perform manually. Furthermore, the introduction of self-supervised and unsupervised learning approaches reduced the reliance on labeled medical data, as these methods can uncover meaningful patterns without explicit instruction. Overall, these technological advancements have fostered a more open and democratic environment for research and product development, characterized by the availability of open-source networks, datasets, and computing resources.

A significant driving force behind the progress of machine learning (ML) in healthcare has been the substantial investments made by both private entities and governments. This influx of funding was motivated by the anticipated benefits of automated ML tools in reducing costs, and improving accessibility and maintenance within healthcare systems. The COVID-19 pandemic played a critical role in highlighting the vulnerabilities of relying heavily on manual processes, underscoring the need for more advanced digital healthcare solutions. During the pandemic, the United States witnessed unprecedented levels of digital healthcare funding, with investment levels peaking as the sector combated the crisis. Following the pandemic, spending on digital healthcare returned to its prior upward trend, continuing to grow and evolve.

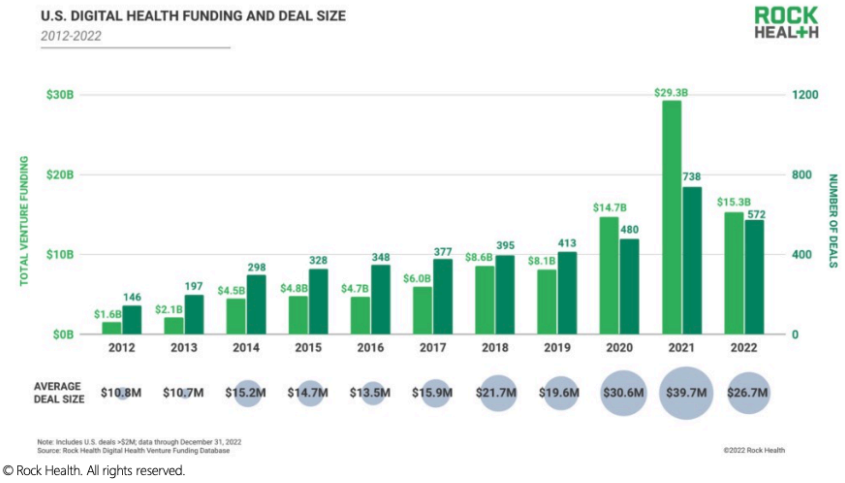

Figure 2 - U.S. digital health funding and deal size depicting the enormous spending during COVID-19 pandemic.

Recognizing the potential for innovation, leading technology companies such as Meta, Microsoft, Alphabet, and Apple have announced numerous partnerships with universities and research institutions to introduce AI-powered products into the market. However, developing decision-support products for the healthcare industry, even those intended merely as recommendation tools for clinicians, presents unique challenges. The decisions these products inform are subject to intense scrutiny due to the potentially severe consequences of errors. Moreover, reaching consensus on the fairness and accountability of these algorithms is complex, given the diversity of value systems among different cultures and stakeholders. Additionally, the healthcare sector's reliance on causality poses a challenge for the straightforward application of supervised learning. The main limitation of supervised learning in addressing causality is its focus on identifying correlations rather than establishing causation. While it can indicate a relationship between two variables and predict outcomes based on this relationship, it lacks the capability to discern whether one variable directly causes changes in another. This limitation is inherent in the nature of supervised learning models, which are designed to detect patterns and associations in the training data without grasping the underlying causal mechanisms.

An additional challenge faced by machine learning (ML) applications in healthcare is the issue of imbalanced labels within datasets. When a disease or symptom is rare, positive labels can be vastly outnumbered by negative ones, introducing biases into the model's prediction capabilities. Furthermore,

healthcare data often presents challenges such as varying time intervals between records and missing data fields, which can arise due to concerns over patient anonymity, specific patient conditions, or the practices of physicians. These issues are further exacerbated by strict regulatory standards in the healthcare sector and, at times, a lack of interest from clinical staff, making the deployment of new products even more challenging.

In 2016, Google introduced a groundbreaking deep learning model designed for grading diabetic retinopathy, marking a significant advancement in the application of artificial intelligence (AI) in healthcare. By 2018 and 2019, this model had been deployed across 11 hospitals in Thailand, setting the stage for a real-world evaluation of its efficacy. The findings from this deployment highlighted both the potential and the limitations of AI in healthcare settings. The model faced challenges due to varying lighting conditions, affecting its accuracy in diagnosing the condition. Furthermore, it incorrectly flagged scans as requiring specialist review that nurses deemed obviously normal, demonstrating a disconnect between the model's assessments and clinical judgments. While the AI model did expedite the diagnostic process when its assessments were correct, its inaccuracies led to unnecessary delays in many instances. This example underscores the complexities of integrating AI into healthcare, particularly the need for models to adapt to diverse and real-world clinical environments and to align closely with the practical knowledge and data collecting processes to avoid domain generalization barriers.

Another example is IDx-DR. This system has demonstrated remarkable effectiveness in screening for diabetic retinopathy, referring 100% of patients with at least moderate diabetic retinopathy and more than 96% of patients with diabetic macular edema for further examination. Notably, IDx-DR has also contributed to healthcare efficiency by avoiding 91% of unnecessary specialty visits, indicating its potential to significantly reduce the burden on healthcare systems and patients alike. The development and validation of IDx-DR included training on over 2 million images, a strategy aimed at mitigating bias and ensuring equitable performance across diverse populations, regardless of sex, race, or ethnicity. This extensive training underscores the importance of comprehensive and inclusive datasets in developing AI systems that are reliable and fair, making IDx-DR a pioneering example of how AI can enhance diagnostic processes while being mindful of diversity and inclusivity. A similar domain shift limitation comes from a skin cancer detection model which was trained using 129,450 clinical images with a reported AUC score of 0.9, performing similar to expert dermatologists. However, once the study was validated using an independent set of images curated by experts, the model performance dropped dramatically due to the inherent bias in the training dataset towards light skin tones and common diseases.

About a decade ago, IBM embarked on an ambitious journey with the launch of Watson Health, aiming to transform healthcare through the power of artificial intelligence (AI). The initiative began with a high-profile collaboration with Memorial Sloan Kettering Cancer Center, focusing on training AI to interpret electronic health records (EHR) and provide treatment recommendations. However, by 2022, IBM decided to divest Watson Health, selling it off "for parts" at approximately $1 billion. This sale represented a significant financial setback for IBM, considering the $5 billion spent on acquisitions to establish Watson Health, thus marking a substantial loss. Despite these challenges, one of the notable successes within Watson Health was its partnership with Medtronic to create the Sugar.IQ diabetes assistant. This collaboration combined expertise in diabetes care, AI, and analytics to offer insights into how lifestyle choices, medications, and insulin injections influence diabetes management and the time patients maintain their glucose levels within the target range. The Sugar.IQ app leverages AI and advanced analytics to provide users with a comprehensive view of their glucose levels, offering personalized insights and guidance to aid in daily diabetes management decisions. This innovative system received FDA approval in 2018, showcasing a successful application of AI in enhancing patient care and self-management in diabetes.