

Review

The perils of artificial intelligence in healthcare: Disease diagnosis and treatment

Jung C. Lee

BioMolecular Engineering Program and Department of Physics and Chemistry, Milwaukee School of Engineering,
Milwaukee, Wisconsin, USA.

Received 1 March, 2019; Accepted 15 April, 2019

For the past decade, artificial intelligence (AI) and its related technologies have made remarkable advances in marketing and business solutions based on AI-driven big data analysis of customer queries, and it, when coupled with bioinformatics, seemingly holds out great promise for use in healthcare. In reality, however, AI is still largely a buzzword when it comes to disease diagnosis and treatment. This review addresses the uncertainty of AI applications to disease diagnosis and treatment, not only pinpointing AI's inherent algorithmic problems in dealing with non-patternable stochastic healthcare data, but also revealing the innate fallacy of identifying genetic mutations as a tool for genome-based personalized medicine. Finally, this review concludes by presenting some insights into future AI application in healthcare.

Key words: Artificial intelligence, machine learning, deep learning, bioinformatics, healthcare, genomic medicine, personalized medicine, reference genome, genetic variation.

INTRODUCTION

Artificial intelligence (AI) has been around for decades since its inception at the 1956 workshop in Dartmouth College. However, its technology has been recently hyped with the arrival of machine learning (ML) and deep learning (DL) algorithms whose evolution centered around the artificial neural network (ANN) model to handle complex multi-layered nonlinear data (Schmidhuber, 2014; Bini, 2018). This is in addition to IBM Watson's beating against Jeopardy champions in 2011 and Google AlphaGo's stunning 4-1 victory over a world's best Go player in 2016. Such incredible successes of AI have been largely driven by the integration of Big Data and ML algorithms, rooted in

complex neural networks to process perceptions and make decisions for action in our brain (Rosenblatt, 1958), holding out promise of a revolution in solving all sorts of real-world problems and issues. Armed with image processing, voice recognition, and natural language processing (NLP), today's high-tech companies including Google and Amazon are using AI and its related technologies as a primary growth fuel. These high-tech companies are pouring immense efforts to get machines far smarter in addressing their business challenges, and already starting to shape up our daily lives and the society, both positively and negatively (Yampolskiy and Spellchecker, 2016).

E-mail: lee@msoe.edu, Tel: +414-277-7316, Fax: +414-277-2878.

Author(s) agree that this article remain permanently open access under the terms of the [Creative Commons Attribution License 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

In parallel, there has been a lot of excitement about how AI will disrupt the entire ecosystem in healthcare. AI applications in healthcare have been traditionally concentrated around cancer, neurology and cardiology, mostly through automated medical image analysis to look for specific patterns linked to diseases and disorders (Jiang et al., 2017; Ravi et al., 2017; Mandal et al., 2018). For instance, electronic abnormal mammogram follow-up triggers were reported to flag patient records with delays on mammography with an accuracy of 71% (Murphy et al., 2018). Recently, AI application to other healthcare domains looks equally promising in helping to better streamline and coordinate administrative and clinical processes. This is to serve patients more efficiently and economically in: reducing preventable medical errors associated with robot-assisted surgery and drug dosage determination, improving disease diagnosis and recommending the best treatment options for individual patients, and aiding to develop new medicines (Kalis et al., 2018). These recent AI applications in different healthcare spaces are being primed by such AI companies as PathAI, aiming at error reduction in cancer diagnosis, Freenome, aiming at early cancer detection, BenvolentAI, aiming at providing the right treatments to patients, and Atomwise (established in 2012) aiming at identifying patient characteristics for drug discovery for clinical trials (Daley, 2018). In June 2018, the American Medical Association (AMA) adopted a new policy, Augmented Intelligence in Health Care H-480.940 (American Medical Association, 2018) to promote AI applications in healthcare for benefitting patients, physicians, and the healthcare community. In April 2, 2019, the Food and Drug Administration (FDA) posted a white paper FDA-2019-N-1185, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)". This paper outlines the agency's forceful move to require FDA approval of AI-enabled medical devices prior to commercialization in an effort to ensure their reliable and scalable performance for a wide variety of real-world patients and clinical data (Ross, 2019).

Uncertainty of AI in disease diagnosis and treatment

Despite such transforming breakthroughs seemingly destined to change the world, AI-powered smart machines still remain not that intelligent in processing non-patternable stochastic healthcare data, structured or unstructured, due predominantly to their inherent algorithmic inability to learn new contexts and adapt to change, making only mediocre advances in disease diagnosis and treatment (Esteva et al., 2017). In addition, AI application to healthcare is limited currently by the quality, bias, consistency, variability, and scale of healthcare data (Warwick et al., 2015; FDA, 2013).

Moreover, the vast majority of healthcare data centered on patients and their diseases are non-discrete, non-patternable, and stochastic in nature (Boddy et al., 2019; JASON, 2017; Wang et al., 2015a), thereby making existing probabilistic statistical analysis useless. Consequently, even coupled with mountains of healthcare big data, AI's critical decision-making around disease diagnosis and treatment is extraordinarily challenging, requiring a new AI algorithm to smartly deal with a multitude of non-patternable stochastic variables or factors associated with each individual disease, epidemic or rare. In this vein, the extensions of AI learning to healthcare solutions for disease diagnosis and treatment is yet to live up to our expectations.

Bias is one of AI's Achilles heels, determining the fate of AI towards the singularity as reflected in the 'garbage in, garbage out' (GIGO) principle for data processing. Depending on the level of quality of and bias in healthcare data, today's ML and DL algorithms, supervised and unsupervised, will inherently learn and replicate the same deeper-seated biases as we have, inevitably and unintendedly failing to make fair decisions (Challen et al., 2019; Canetti et al., 2019). To make it worse, AI is biased by design. The AI's fairness issue is exemplified by the Google's photos app mislabeling two black couple as gorillas and Microsoft's chatbot Tay responding with disruptive and abusive Tweets. In addition, the COMPAS program used for making bail and sentencing decisions in U.S. courts was hugely biased against black defendants, falsely flagging blacks nearly twice as likely to re-offend as whites (Larson et al., 2016). Moreover, AI-based online lenders discriminate against minorities, charging much higher interest rates to minority borrowers compared to white borrowers (Bartlett et al., 2018).

The lack of trust is another key factor that comes up for the success of AI in healthcare. Today's AI and ML algorithms employ a black box just like neural networks in the brain, turning decision-making over to the black box. Thus, not even the most eminent high-profile AI experts really know for sure how AI and ML algorithms internally work to arrive at a final decision (Knight, 2017; Bleicher, 2017). Consequently, the final decision lacks explainability and transparency, not compelling to convince doctors and patients to trust AI's decision-making. In March 2018, an Uber self-driving SUV made a wrong decision, hitting and killing a pedestrian in Tempe, Arizona. According to an MIT Sloan Management Review's research study (Davenport and Bean, 2018), the majority of organizations (82% of those surveyed) had not adopted AI beyond pilot projects, that is, people do not trust AI in healthcare. These together speak of a mortifying reminder of the risk of AI in disease diagnosis and treatment, which must roll out accurate, fair, and trustworthy decisions.

A recent study revealed that AI has not really evolved much since its beginning. It iteratively reuses one form of its existing algorithms nearly every decade, rather than

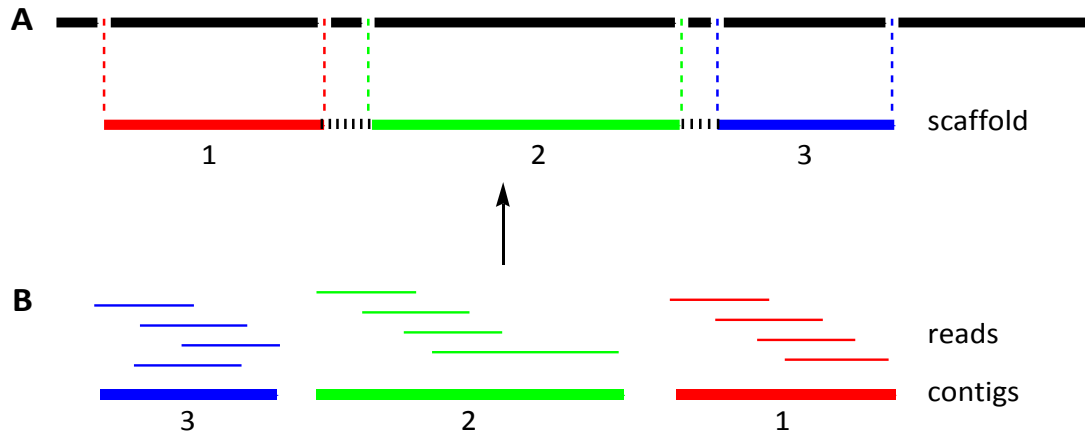


Figure 1. Schematic illustration for genome assembly. (A) A scaffold with gaps is built by aligning each of the three contigs 1 (red), 2 (green), and 3 (blue) against the reference genome (black). The gaps in the scaffold are represented by dotted black lines between the contigs. (B) The three contigs (1-3) in A are constructed, each by aligning a set of short sequence reads through overlapping regions, and they are computationally organized into the scaffold by being positioned against the reference sequence.

its existing algorithms nearly every decade, rather than redesigning a brand-new algorithm (Hao, 2019). This implies that AI decision-making models will not play out well without a next-generation paradigm. Beyond this, healthcare solutions around disease diagnosis and treatment are extremely complicated compared to marketing and business solutions, requiring much harder and frequently life-and-death treatment decisions for patients. Without changing the present AI paradigm, even today's most advanced ML and DL algorithms are doomed to predict exaggerated risks of diseases or simply churn out too many false positives, whereby failing to diagnose disease unambiguously and recommend the right treatment at the right time. For instance, Google Flu Trends (GFT) (Ginsberg et al., 2009), ambitiously designed based on Google search queries to forecast seasonal flu outbreaks two weeks earlier than the Centers for Disease Control and Prevention (CDC), turned out to be an epic failure of AI application to healthcare (Lazer et al., 2014). In July 2018, a STAT medical website's report revealed that IBM Watson Health, the reportedly best AI healthcare system in the world, failed multiple times in recommending safe and accurate treatment options for cancer patients, notwithstanding collaborations with oncologists at the Memorial Sloan Kettering Cancer Institute. This also bodes ill for the future of AI in healthcare, clouding the putatively rosy arrival of the personalized medicine era.

Fallacy of genome-based personalized medicine

Bioinformatics is largely centered on biodata composed of biosequences, biostructures, and their metadata, searching for signature patterns to promote human health

and wellness most commonly by comparatively analyzing homologous biosequences collected from different people. With the putative reference human genome (International Human Genome Sequencing Consortium, 2004), researchers strongly believe that the AI-enabled bioinformatics will efficiently identify such patterns by comparing human genomes, each linked to human disease, launching genomic medicine. The paradigm for genomic medicine is, however, inherently full of fallacies.

Firstly, the putative reference human genome does not intrinsically qualify as a reference, against which a person's genome is compared in search of signature patterns associated with the person's disease (Lee, 2017). It is a composite human genome of which 70% came from just a single donor anonymously named RP11, seriously lacking in diversity (Sherman et al., 2019). As illustrated in Figure 1, it is rather established as a guide for genome sequencing, against which contigs, each constructed from a set of short sequence reads, are aligned and assembled to build often-fragmented scaffolds of individuals and patients (Ekblom and Wolf, 2014). Each individual's genome is unique in both composition and organization (Seo et al., 2016), so that genomic variations found between people do not necessarily support for their relatedness to disease or disorder (Lee, 2017). In this regard, the putative reference human genome should not be a legitimate reference genome to be used to search for genomic variations. In reality, it is not logical to create such a single reference human genome to be universally compared against each individual's genome in search of genomic variations, as individual genomes vary very widely from person to person but uniquely across diverse ethnic groups and populations around the world.

Secondly, the paradigm of genomic medicine is based



Figure 2. Example of alignment of gene sequences. The reference sequence on the top portion (shaded in light pink) of the sequences alignment editor expedites the alignment of homologous gene sequences and helps the identification of genetic variations across sequences. The gap (-) and tilde (~) characters in each sequence are to represent the positions of any insertion/deletion (*aka* indel) and unavailable nucleotide(s), respectively.

solely on the premise that genetic changes are responsible for diseases, disorders, or medical conditions (Acuna-Hidalgo et al., 2016). The paradigm has prompted to look for disease-causing genetic mutations (most commonly single nucleotide polymorphisms (SNPs)) by comparing gene sequences from patients against their homologous regions in the putative reference human genome, as shown in Figure 2. With a hope to identify genetic variations in patients more efficiently, whole exome sequencing (WES) has been established to sequence only the human exome, or the collection of all known coding regions (or genes) that are translated into peptides, which represents ~1% of a person's full-length genome (Lee, 2017). Nonetheless, the noncoding regions (or non-genes), which are transcribed into RNA but never translated into peptides, are also as functionally and/or structurally equally important as the coding regions (Baker, 2012), embedded with greatly diverse variations including: small or large insertions/deletions (indels), repetitive elements (REs), transposable elements (TEs), and copy number variations (CNVs) (Burger et al., 2011). To our surprise and contrary to our expectations regarding the correlation between mutation and disease, it was uncovered that the peptide-making genes that are essential for proliferation and survival reveal basically no variations across all human populations (Wang et al.,

2015b). More shockingly, the vast majority of the highly variable 3,230 peptide-producing genes among 60,706 individuals from all corners of the world are not linked to any currently known human disease (Lek et al., 2016). The latter two studies further substantiate a notion that genetic variation does not speak of any relatedness to disease-causing phenotypes, but simply of evolutionary changes in people's genomes that have responded differently over time to their past environments and living conditions (Lee, 2017). It is thus clear that without any significant paradigm shift in pattern recognition and genome analysis, today's AI and ML will not warrant any success in identifying medically meaningful signature patterns linked to human diseases and disorders.

DISCUSSION AND CONCLUSIONS

AI and its related technologies have been hugely successful in certain well-defined domains such as game mastering, voice recognition, and language translation. AI-enabled pattern and image recognition algorithms have assisted eminent high-tech companies to make important decisions and necessary adjustments for their marketing and business strategies. In contrast to these well-defined domains, AI-utilized decision-making in

healthcare has made only mediocre advances. Especially, AI in disease diagnosis and treatment (the AI's last resort) faces a number of non-trivial thorny problems to overcome. In particular, most healthcare data around human diseases and their phenotypes involve arrays of non-patternable stochastic variables, so that currently inherently probabilistic AI algorithms will fail to learn about healthcare data, and thus will not be able to make reliable, unambiguous, and transparent treatment decisions for patients. This reminds us of the Feynman trap, which states that something extremely unlikely is 100% likely if it already happened. This also indicates that the future prospect of AI in healthcare is not that promising. The extension of AI to healthcare, disease diagnosis and treatment in particular, is doomed to fail regardless of computing power, without a quantum leap in AI's algorithmic revolution to analyze and process non-patternable stochastic healthcare data. Recently, a team of researchers published a novel scalable deep neural network training model by replacing the conventional fully-connected layers (FCLs) with quadratically fewer sparsely-connected layers (SSLs) without loss of accuracy (Mocanu et al., 2018), opening the doors to a better modeling of the original brain-inspired ANNs where neurons are connected only to handful of other neurons.

Moreover, healthcare data and biodata often lack the quality and fairness required for AI to make right and fair decisions that are explainable to doctors and patients. To make it worse, we do not have quality healthcare data for rare diseases from which several hundred millions of people worldwide are suffering. Furthermore, the majority of bioinformaticians and data scientists – who apply AI to healthcare solutions – lack thorough understanding of real-world healthcare data and biodata. They tend to use oversimplified toy models to solve complex multivariate healthcare problems, thereby throwing out or ignoring many real but seemingly outliers, followed by building healthcare solutions with overly limited reliability and no flexibility. Therefore, they are never capable of dealing with real-world biodata reliably and efficiently. This warrants that they are in dire need of better education and due training opportunities to learn real-world biodata properly in a new type of interdisciplinary setting and employ that learning to design next-generation adaptive AI algorithms, which could make reliable and unambiguous decisions to solve such complex computational problems around healthcare, specifically disease diagnosis and treatment.

CONFLICT OF INTERESTS

The author has not declared any conflict of interest.

REFERENCES

Acuna-Hidalgo R, Veltman JA, Hoischen A (2016). New insights into the

- generation and role of de novo mutations in health and disease. *Genome biology* 17(1):241.
- American Medical Association (AMA) (2018). Augmented intelligence in health care H-480.940. https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMA_Doc%2FHOD.xml-H-480.940.xml
- Baker M (2012). The changes that count. *Nature* 482:257-262.
- Bartlett R, Morse L, Stanton R, Wallace N (2018). Consumer-lending discrimination in the era of FinTech. <https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf>.
- Bini SA (2018). Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care? *Journal of Arthroplasty* 33: 2358-2361.
- Bleicher A (2017). Demystifying the black box that is AI. *Scientific American*, August 9. <https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/>
- Boddy A, Hurst W, Mackay M, El Rhalibi A, Baker T, Montañez CA (2019). An Investigation into Healthcare-Data Patterns. *Future Internet* 11(2):30.
- Burger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R (2011). The genomic complexity of primary human prostate cancer. *Nature* 470(7333):214.
- Canetti R, Cohen A, Dikkala N, Ramnarayan G, Scheffler S, Smith A (2019). From Soft Classifiers to Hard Decisions: How fair can we be?. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* pp. 309-318.
- Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality and Safety* 28(3):231-237.
- Daley S (2018). Surgical robots, new medicines and better care: 17 examples of AI in healthcare. December 5, <https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare/>. Daley
- Davenport TH, Bean R (2018). The problem with AI pilots. July 26. <https://sloanreview.mit.edu/article/the-problem-with-ai-pilots/>.
- Eklom R, Wolf JBW (2014). A filed guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* 7:1026-1042.
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115-118.
- FDA (2013). Guidance for industry: electronic source data in clinical investigation. <https://www.fda.gov/downloads/drugs/guidance/ucm328691.pdf>.
- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009). Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012-1014.
- Hao K (2019). We analyzed 16,625 papers to figure out where AI is headed next. *MIT Technology Review*, January 25.
- International Human Genome Sequencing Consortium (IHGSC) (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945.
- JASON (2017). Artificial intelligence for health and health care. https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf.
- Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*. 2(4):230-243.
- Kalis B, Collier M, Fu R (2018). 10 Promising AI applications in Health Care. *Harvard Business Review*, May 10.
- Knight W (2017). The dark secret at the heart of AI. *MIT Technology Review*, April 11.
- Larson J, Mattu S, Kirchner L (2016). Machine Bias. *ProPublica*, May 23.
- Lazer D, Kennedy R, King G, Vespignani A (2009). The parable of Google Flu: Traps in big data analysis. *Science* 343:1203-1205.
- Lee JC (2017). Are genes and their mutations responsible for disease? *International Journal of Structural and Computational Biology* 1(1):1-5.
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T,

- O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.
- Mandal S, Greenblatt AB, An J (2018). Imaging intelligence: AI is transforming medical imaging across the imaging spectrum. *IEEE Pulse* 9(5):16-24.
- Mocanu DC, Mocanu E, Stone P, Nguyen PH, Gibescu M, Liotta A (2018). Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature communications* 9(1):2383.
- Murphy DR, Meyer AN, Vaghani V, Russo E, Sittig DF, Wei L, Wu L, Singh H (2018). Electronic triggers to identify delays in follow-up of mammography: Harnessing the power of big data in health care. *Journal of American College Radiology* 15:287-295.
- Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ (2017). Deep learning for healthcare informatics. *IEEE Journal of Biomedical and Health Informatics* 21(1):4-21.
- Rosenblatt F (1958). The perception: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65:386-408.
- Ross C (2019). FDA developing new rules for artificial intelligence in medicine. April 2, <https://www.statnews.com/2019/04/02/fda-new-rules-for-artificial-intelligence-in-medicine/>.
- Schmidhuber J (2014). Deep learning in neural networks: An overview. *Neural networks* 61:85-117.
- Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J, Kuk J (2016). *De novo* assembly and phasing of a Korean human genome. *Nature* 538:243-247.
- Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, Levin AM (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics* 51(1):30-35.
- Simonite T (2014). IBM aims to make medical expertise a commodity. *MIT Technology Review*, July 21.
- Wang JY, Ho HY, Chen JD, Chai S, Tai CJ, Chen YF (2015a). Attitudes toward inter-hospital electronic patient record exchange: Discrepancies among physicians, medical record staff, and patients. *BMC Health Services Research* 15(1):264.
- Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM (2015b). Identification and characterization of essential genes in the human genome. *Science* 350:1096-1101.
- Warwick W, Johnson S, Bond J, Fletcher G, Kanellakis PA (2015). A framework to assess healthcare data quality. *The European Journal of Social and Behavioral Sciences* 13(2):1730-1735.
- Yampolskiy RV, Spellchecker MS (2016). Artificial intelligence safety and cybersecurity: a timeline of AI failures. *arXiv:1610.07997*.