As the health care industry adopts artificial intelligence, machine learning, and other modeling techniques, it is seeing benefits to both patient outcomes and cost reduction; however, it needs to be cognizant of and ensure proper management of the risks, including bias. Lessons learned from other industries may provide a framework for acknowledging and managing data, machine, and human biases that arise while implementing AI.

As the benefits of artificial intelligence (AI) begin to permeate various areas of health care, the levels of excitement and expectations continue to grow [1, 2]. In fact, over the last few years, there have been several headline-grabbing examples in which AI has augmented our ability to provide better health care through image analysis, robot-assisted surgery, and drug discovery as well as ways to reduce cost through workflow improvements and better claims processing. The initial excitement of augmenting health care workers with more information has been both beneficial and detrimental. While information overload has become a topic of conversation in almost all areas of health care, AI and its champions have worked to consolidate much of the information into a single score that the decision-makers may use to make a decision. While this may work well in reducing the amount of information pushed at a provider, it can also hide nuance and result in different decisions than if all the information was available to the physicians. In addition, AI is also spreading fear as both mistakes and bias are showing up in AI implementations across different industries [3-5]. This has led to both fear in utilizing AI [6] and acknowledgement that we need to do better [7, 8]. Indeed, health care is not immune from bias [9, 10]. Thus, there is a real risk that AI (or any) models based upon historical data will have bias built in. Note bias can generally be defined mathematically; however, fairness has a social context and is often the term utilized in public conversations.

What Health Care Can Learn from Finance

A very well-documented process that requires “fairness” is consumer lending in the financial industry. There was obvious bias against African Americans in the middle of the 20th century that led to legislation requiring consumer and mortgage lending to be “fair” [11] and transparent [12]. Over the decades, this area has seen widespread adoption of algorithmic modeling, forcing the financial industry to develop methods and tools that avoid the pitfalls and biases that may arise in consumer lending. I believe the health care industry can benefit from the adoption of some of these methods and tools.

When modeling an individual to determine whether they are creditworthy, a model goes through many checks and risk evaluations. First, the input data are scrutinized to eliminate all variables that are directly related to areas where fairness must be ensured (aka, the model is blind to these variables). All race, ethnicity, gender, and age variables may not be used (this is very different from other industries, including health care). However, as the industry matured and adopted broader datasets and more complex mathematics of AI, it became clear that even when all known variables related to the protected classes were eliminated from the input, biased outcomes were still possible, and often likely. Thus, the financial regulators and the industry moved to measuring the outcomes of the models and processes. One must not only have unbiased inputs, but also unbiased outputs. So, techniques have been developed to watch and monitor the outcomes, codifying best practices for model validation and monitoring into what is called model risk management (MRM) [13-15].

The health care industry is well behind in its adoption of these concepts. As I have talked with and advised providers and payers in health care, I have noticed a general lack of knowledge about why and how to implement MRM. I have been told of providers adopting models straight from the literature of other providers. When asked about expected benefits and whether the input data was representative, or if the outcomes show any bias, there are only blank stares. I am not implying that the models developed in the literature or at other systems won’t drive benefit, are biased, or are not representative; I am a firm believer that questions of benefits, bias, and appropriateness need to be asked and answered. For health care firms, achieving a sound capability to identify
these issues should come first, quickly followed by a model monitoring process to ensure continued performance. The financial industry has spent years working to address these issues, while others are jumping straight into AI and are learning the lessons the hard way.

**Driving Benefits with Augmented Intelligence**

Human interaction with the machine and “joint” decision making can generally produce a better outcome than either machine or human alone. Many health providers are taking this approach, and I have often heard that the models will recommend or provide a score to a physician, who then uses the model's recommendation to make the final decision. This is a perfect example of augmented intelligence. However, we need to ensure that we include human decisions, and not just the machine’s recommendations, when measuring the outcome of the process, as feeding a fair recommendation into a biased human does not ensure a fair outcome any more than the reverse would. Thus, we need to ensure the process, and not just the model, is as fair as possible. This combined human/model process outcome was often measured in credit decisions where the human credit officers would “override” the model’s recommendations. Generally (but not always) this results in a more accurate outcome, but requires extra diligence for monitoring the overrides themselves [16]. It is the situations in which human intervention does not increase accuracy and fairness that are the most important cases for health care information utilizers to examine.

Two cases where human augmentation through AI is being adopted within the world of health care are radiology and claims processing. In both of these cases, AI improves efficiency and accuracy. In radiology, image AI is able to detect many issues, but difficult/edge cases are forwarded to a radiologist [17]. The models are not only predicting the outcome, but also identifying the specific areas on an image for examination and showing the level of confidence in the prediction. This allows the radiologist to quickly move through routine images and focus on the more difficult
cases. In a very different area of the business of health care, AI models are enhancing claims processing and removing many of the claims that, while not trivial, rely on a larger knowledge base of claims processing employees [18]. Similar to the radiology case, this allows the claims processing employees to focus on the more difficult cases where the model is not certain of the outcome. The vast majority of AI implementations to date in health care have resulted in utilizing resources more effectively and augmenting human decisions. In both of these examples, overall decision effectiveness was improved through the human-machine interactions.

Managing Bias in the Application of AI in Health Care

I tend to think about applications of AI for health care in three main groups: care, claims, and marketing.

Care-based AI is the most obvious case where we want to eliminate bias and treat people fairly within the diverse set of fairness viewpoints. In fact, biases have been recognized and research targeted to resolve the gaps. Whether inspired by the 2004 founding of Go Red for Women by the American Heart Association or more recent work in genetic disease risk-detection algorithms [10], both the public and the research community are more aware of potential biases and are asking the next set of questions. However, I have not yet seen much attention to the potential bias in claims and payments.

Unfortunately, there is a correlation between wealth, ethnicity, and health [19]. Let’s try a hypothetical thought experiment where we want to maximize the efficiency of our claims/payment processing. As we build our model and optimization engine, we would naturally focus on processing the claims most likely to be approved and have the highest payment. This type of model may easily result in biased outcomes, as the claims being paid first may be those of patients with the highest-quality health plans, who are likely to be primarily wealthy and white. Thus, our model will likely
have biases against poor ethnic minorities. Obviously, this is not the intended result, but it is a potential result, as shown in similar analyses of credit modeling [20].

While marketing models have the same potential wealth/payment/ethnicity outcomes, they also have a more deviant way to drive biases into your systems. If your marketing is successful in delivering a high return on investment (ROI), but your patients are not representative of all the diversity in your community, then future models built on your data-sets will not be representative either. In that case, not only did the marketing models drive a disparate impact favoring those who can pay, but there is also a secondary impact on your data such that without careful consideration and processes future modeling and AI will be biased.

Special Case: Social Determinants of Health

The ability to get a ride to a health care facility or to afford to fill a prescription influences the health of individuals. Physicians can recommend to a patient that they eat fresh vegetables and other produce, but if they don’t have access to grocery stores that carry those things, can’t afford to purchase them, or don’t have a stove or microwave to cook in, they are less likely to “comply” with the physicians’ directions. As health care systems move to include data on place, economics, education, and social context in health care data systems (whether collected directly or through partnerships with integrated data systems [IDS] [21]), many of these variables are highly correlated with age, ethnicity, gender, and
other classes of diversity and/or protected groups. How and when these data are utilized in modeling and AI and for what purpose will need to be transparent, actively debated, and only included when societal norms align since the complexity and potential for unintended bias will add additional risk to any implementation. Most large metropolitan areas and many mid-size areas are forming IDS groups to manage just this data that will need to be utilized. Further research with the combination of IDS and health data is required to begin to understand the impact of all the social determinants. For example, utilizing AI to understand how a child’s housing status and the frequency of its change impact education success and health would build a base to test interventions and drive additional preventive services. Measuring the impact of these potential interventions would enable those on the front lines to direct more resources to the best solutions. Working through the necessary practices for implementing AI within the realm of social determinants of health must be done with openness and include a wide diversity of viewpoints such that results are implemented in ways that recognize the benefits while acknowledging the limitations of the human-managed AI systems.

Recommendations

First and foremost is the need for health care to continue to adopt modeling and AI techniques such that the benefits to both individuals and society may be realized. However, care must be taken and lessons learned from other industries. The immediate next step is to adopt processes in which there is active debate on bias, fairness, and benefits during model development. In addition, the ability to continuously monitor a process with embedded AI for performance, potential bias, and fairness linked to societal norms is critical to gaining the trust of not just the patients, but also physicians, nurses, insurance providers, regulators, and all stakeholders of our health care system. NCMJ

Douglas C. Hague, PhD research and industry lead, Data Science Initiative, University of North Carolina at Charlotte, Charlotte, North Carolina

Acknowledgments

D.H. is funded by the Data Science Initiative at the University of North Carolina at Charlotte.

Potential conflicts of interest. D.H. has an active sole-proprietorship where he advises corporations, including health care organizations, on how to manage their models and AI. These services are provided through International Institute for Analytics and Reference Point. He is also an advisory board member of Duco Corporation.

References

1. Davenport TH, McNeil D. Analytics in Healthcare and the Life Sciences: Strategies, Implementation Methods, and Best Practices. Upper Saddle River, NJ: Pearson FT Press; 2014.
2. Burke J. Health Analytics: Gaining the Insights to Transform Health Care. Hoboken, NJ: John Wiley & Sons; 2013
3. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Published October 9, 2018. Accessed April 19, 2019.
4. Vincent J. Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day. The Verge. https://www.theverge.com/2016/3/24/11297050/tae-microsoft-chatbot-racist. Published March 24, 2016. Accessed April 19, 2019.
5. Buolamwini JA. Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers [dissertation]. Cambridge: Massachusetts Institute of Technology; 2017.
6. Maher K. Without Humans, A.I. Can Wreak Havoc. The New York Times. https://www.nytimes.com/2019/03/12/opinion/artificial-intelligence-wikipedia.html. Published March 12, 2019. Accessed April 19, 2019.
7. Sharma K. Can We Keep Our Biases from Creeping into AI? Harvard Business Review. https://hbr.org/2018/02/can-we-keep-our-biases-from-creeping-into-ai. Published February 9, 2018. Accessed April 19, 2019.
8. Chowdhury R, Mulani N. Auditing Algorithms for Bias. Harvard Business Review. https://hbr.org/2018/10/auditing-algorithms-for-bias. Published October 24, 2018. Accessed April 19, 2019.
9. Tello M. Racism and Discrimination in Health Care: Providers and Patients. Harvard Medical School website. https://www.health.harvard.edu/blog/racism-discrimination-health-care-providers/patients-20170161015. Published January 16, 2017. Accessed April 19, 2019.
10. Kim M, Patel K, Teng A, Berens A, Lachance J. Genetic disease risks can be misestimated across global populations. Genome Biol. 2018;19(1):179.
11. Equal Credit Opportunity Act, 15 USC §1691 (2011).
12. Federal Trade Commission. Fair Credit Reporting Act: United States Code Title 15 Section 1681. Washington, DC: FTC; 2018. https://www.ftc.gov/system/files/fgpo2016/04-fair-credit-reporting-act-0918.pdf. Accessed April 19, 2019.
13. Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. Federal Reserve website. https://www.federalreserve.gov/speechesreleases/sr1107.htm. Updated April 4, 2011. Accessed April 19, 2019.
14. Babel B, Buehler K, Pivonka A, Richardson B, Waldron D. Derisking machine learning and artificial intelligence. McKinsey & Company website. https://www.mckinsey.com/business-functions/artificial-innovation/our-insights/derisking-machine-learning-and-artificial-intelligence. Published February 2019. Accessed April 19, 2019.
15. Regan S, Klein L, Jacobs M, Kazmi, S. Model Behavior. Nothing Artificial. New York, NY: Accenture Consulting; 2017. https://www.accenture.com/us-en/_acncms/pdf-70/Accenture-Emerging-Trends-Validation-Machine-Learning-AI-Models.pdf. Published 2017. Accessed April 19, 2019.
16. Tasche D. Bounds for rating override rates. Journal of Credit Risk. 2012;8(4):3-29.
17. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284(2):574-582.
18. Hehner D, Körs B, Martin M, Uhrmann-Klingen E, Waldron J. Artificial intelligence in health insurance: Smart claims management with self-learning software. McKinsey & Company website. https://www.mckinsey.com/business-functions/healthcare-systems-and-services/our-insights/artificial-intelligence-in-health-insurance-smart-claims-management-with-self-learning-software. Published September 2019. Accessed May 28, 2019.
19. Narayanan A. Tutorial: 21 fairness definitions and their politics. Oral presentation at: ACM Conference on Fairness, Accountability, and Transparency; February 23, 2018; New York, NY. https://www.youtube.com/watch?v=5jXLuYdnvjk. Accessed May 4, 2019.
20. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Paper presented at: NIPS’16 Proceedings of the 30th International Conference on Neural Information Processing Systems; December 5-10, 2016; Barcelona, Spain. https://dl.acm.org/citation.cfm?id=3157469. Accessed April 19, 2019.
21. University of North Carolina at Charlotte Urban Institute, Division of Academic Affairs. Institute and the Institute for Social Capital. UNC Charlotte website. https://ui.uncc.edu/programs/isc. Accessed May 14, 2019.