

DEVELOPMENT OF AN ONLINE TRAINING MODULE
FOR CLINICIANS ON ALGORITHMIC BIAS IN HEALTHCARE

By

Benjamin X. Collins

Capstone Project

Presented to the Department of Medical Informatics & Clinical Epidemiology
and the Oregon Health & Science University

School of Medicine

in partial fulfillment of
the requirements for the degree of

Master of Science

September 2022

School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

Benjamin X. Collins

*“Development of an Online Training Module on Algorithmic Bias in Healthcare for
Clinicians”*

Has been approved

Capstone Advisor

TABLE OF CONTENTS

	Page
Acknowledgements	i
Abstract	ii
Introduction	1
Methods	4
Results	7
Technical Description of the Module	9
Content Description of the Module	12
Discussion	19
Conclusion	23
References	24
Appendix A: List of Reviewed Articles	26
Appendix B: Full List of Concepts	32
Appendix C: Screenshots of Module Slides and Text Transcript	39

ACKNOWLEDGEMENTS

I would like to acknowledge guidance of my capstone mentor Eilis Boudreau, M.D., Ph.D. as well as the feedback on this work from Eran Klein M.D., Ph.D., Amy Forester M.F.A, Lawrence Williams M.S., and Justi Echeles M.A.

ABSTRACT

With the increasing utilization of artificial intelligence/machine learning tools in healthcare, there has been increasing concern that the algorithms used to implement these tools have built-in biases (algorithmic bias). Currently, algorithmic bias is poorly understood by clinicians who have an insufficient opportunity for training on this topic, leading to a risk of harm for patients when AI systems are used to provide clinical care. To improve awareness and understanding among clinicians regarding algorithmic bias, an online interactive module was developed based on a literature review with the content including case studies modeled on real-life AI tools. This module was deployed online while feedback is being received and analyzed in order to improve training in this area.

INTRODUCTION

Artificial intelligence (AI) refers to a branch of computer science for developing digital algorithmic systems capable of complex decisions.¹ Machine learning (ML) is the dominant subset of AI in which algorithms are trained to identify patterns of data from input and then use those patterns to make predictions on new data.¹ Algorithms are a set of step-by-step instructions for computers to perform tasks which can be trained by ML processes to become more accurate.² Bias is a systematic deviation from the truth and the processes that lead to such a deviation.³ Algorithmic bias is an umbrella term for the biases originating from the use of AI as a reflection of data input, methods used, and human judgment. As we increasingly rely on artificially intelligent systems, there is more opportunity for algorithmic bias to impact patient care. For example, Ziad Obermeyer detailed how an algorithm used to select patients that may need extra care functioned with racial bias due to the use of health cost as a proxy for health needs when less money was being spent on Black patients, leading the algorithm to falsely conclude that Black patients were healthier than White patients.⁴ Therefore, clinicians need to be able to recognize the presence of algorithmic bias in AI systems and to determine how to use AI systems to minimize the bias passed into patient care. Nonetheless, clinicians do not currently receive sufficient training on algorithmic bias despite the increasing presence of AI in healthcare. The National Academy of Medicine suggests that training clinicians on the use of AI should be a priority.⁵ Scott, Carter, and Coiera have proposed a “clinician checklist for assessing suitability of machine learning applications in healthcare,” pointing out how, “clinicians need to have some understanding of how algorithms are developed and how to assess their clinical worth.”⁶ However, this lacks a focus on the

risks presented by bias, and using a checklist may not be as effective without a certain base level of knowledge on the topic. Others have brought to attention the importance of AI training in medical education and how this might alter the medical education process.⁷ However, there is little attention on algorithmic bias in the discussion of AI in medical education, the threat that it represents and how to minimize or avoid it. It is imperative for that training to include algorithmic bias which disproportionately harms patients from already marginalized populations.⁸ While clinicians already undergo extensive training, this training continues to be based on traditional curricula that does not incorporate AI approaches, or the more specific issue of algorithmic bias. Nonetheless, the American Medical Association has supported research into how to introduce content on AI into medical education, but there is more work needed on how to best do so. There is anticipation that the digitization of healthcare is supporting a transition from an information age to the age of AI which will require clinicians to gain knowledge on data science and the associated ethical issues and risk for bias.⁹

Although the literature on algorithmic bias in healthcare is extensive, there is little guidance on what clinicians need to know. For example, there is a difference between an AI system which is inaccurate compared to one that is generally accurate but biased. The current literature is more helpful in aiding clinicians to detect the former than the latter. With a curriculum that is already filled, there is little interest in adding more topics despite the clear need for covering this content. In addition, there is a misperception that the benefits of AI need to be proven first in order to justify inclusion in the medical school curriculum.⁹ This is a dangerous perspective as nearly all clinicians will soon be

using AI systems which are redefining the skills required of physicians.^{10,11} It may be the case that more clinicians need to learn about the harms presented by algorithmic bias in order to recognize the importance of addressing the issue, but if it is left out of content for learning then it may not be recognized. To fill this gap in education, supplemental training tools such as online modules may be essential. Because of the lack of guidance on what should be taught, core concepts in this area were extracted from a literature review and then organized into themes to guide the development of a training module for clinicians.

To convey the content of these themes in the module, learning objectives were identified (1) to recognize different forms of algorithmic bias in the use of artificial intelligence for healthcare and the origins of those biases, (2) deconstruct algorithmic bias in the use of artificial intelligence for healthcare by thinking through example cases, and (3) empathize with the importance of recognizing and addressing algorithmic bias in healthcare by exploring the problems resulting from algorithmic bias with attention to the disproportionate impact on already marginalized populations. The objectives of the study include the development of a 2-hour training module to fulfill the above learning objectives, inform training practices for AI in healthcare on what content clinicians need to learn about algorithmic bias, and receive feedback to improve how that training is conducted.

METHODS

To discover the appropriate content to include in an online module for clinicians on algorithmic bias an extensive literature review was performed. PubMed was used to search for articles with content on algorithmic bias in healthcare. This search was limited to articles in the past 5 years in order to ensure relevancy to current technology. Specific concepts within or related to algorithmic bias were extracted from the content of each relevant article. This process continued until reaching a saturation of information as determined by no new concepts in 5 consecutive articles. Concepts were then coded with inductive themes related to algorithmic bias in healthcare such that the themes were derived from the content of the articles. Different concepts were coded into the same theme based on similarity of the concepts. In some cases, concepts were coded into multiple themes. For example, automation bias was extracted as a concept from the content of articles. One article described how, “With automation bias, humans preferentially accept automated/computerized recommendations as a ‘heuristic replacement of vigilant information seeking and processing.’¹⁰ Automation bias is relevant to algorithmic bias in that algorithmic bias can be automated if the user of an AI system does not recognize the potential for bias and accepts the output of the algorithm without thinking about it. As a concept, algorithmic bias suggested two larger themes. First, automation bias is a negative concept which can cause harm to persons or groups. And second, the reason for automation implicitly suggested here is one of humans trusting the output of a computer system. Therefore, automation bias was placed into the themes of harm and of trust. The frequency of concept appearance along with the frequency of theme appearance were analyzed in order to guide selection of the content

for the module. Using frequency as a proxy for the importance of the concept, the top 5 most frequently appearing concepts were selected to be the focus of the module's content.

An early presentation of this content, using 10 concepts, was shown to a group of pre- and post-doctoral fellows at Oregon Health & Science University (OHSU) leading to feedback to focus on a narrower set of concepts. When narrowing the content to 5 concepts, similar concepts were condensed into single concepts when possible, allowing the flow of content to improve with less repetition while avoiding the loss of important content from the module. As content for the module was developed, a set of general recommendations about the use of AI in healthcare and algorithmic bias for clinicians to consider, based on the content of the module, was produced. To provide more interactive content, three cases based on real-life scenarios were created for learners to discuss in an online forum setting. To procure feedback on how to improve training on the use of artificial intelligence in healthcare two anonymous feedback surveys were developed using Qualtrics. One survey was intended to be completed prior to the module and the other following the module in order to pick up on the participant's full experience and how much they learned relative to what they knew prior to the module. The surveys provide both quantitative and qualitative data for analysis. This feedback will be used to develop a set of lessons learned on developing a training module on AI in healthcare for clinicians. The module and the website hosting it were programmed from scratch in HyperText Markup Language (HTML), JavaScript, and Cascading Style Sheets (CSS). The website would ultimately be hosted on GitHub pages at <https://algorithmicbiasmodule.github.io/AI/>. All code is available for the public to view at

<https://github.com/AlgorithmicBiasModule/AI/releases/tag/v1.0>. Initially the module has been publicized locally at OHSU for gaining participants, but it is accessible to anyone with the link. Prior to opening the survey for feedback on the module, the study was approved by the institutional review board at OHSU.

RESULTS

In 86 articles reviewed, 886 instances of 141 concepts were revealed in 8 themes. A list of articles reviewed is available in Appendix A.. These articles contained a mean of 10.3 concepts per article. The five most frequent concepts that would be the focus of content for the module: biased data, transparency & explainability, risk of harm, limitations of regulation, and transformation of healthcare. A full list of the concepts in the literature review is provided in Appendix B. The 8 themes are listed in Table 1. Some concepts fit into multiple themes. The order in which the concepts were introduced in the module was based on a logical flow of the content rather than frequency of the concept in the literature review. To improve understanding of the module, basic terms were introduced prior to these concepts. This introductory section included definitions of artificial intelligence, machine learning, algorithm, bias, and algorithmic bias. After each of the concepts were covered in the module, the content was reinforced through the use of two examples of how algorithmic bias can influence healthcare. Recommendations to clinicians on using AI tools and minimizing the risk for algorithmic bias were included at the end of the module. Three case studies with associated questions to be answered in an online discussion forum followed the core module content. The collection of feedback is in process, after which data will be aggregated to assess the performance of the module and determine a set of lessons learned for training clinicians on algorithmic bias.

Theme	Frequency	# Concepts	Description
Background	88	21	Definitions and other general terminology.
Beneficence	106	106	Maximizing the benefits of using AI.
Data	281	39	Related to the collection and use of data to build or train AI.
Function	67	18	Features of how an AI system functions.
Harm	183	24	The risk of harm and types of harm brought about by AI bias.
Healthcare	109	109	Where AI fits in the healthcare system.
Research	48	48	Research on the use of AI.
Trust	201	201	How bias can influence clinician and end-user trust of AI.

Table 1. A list of the eight themes, the number of times they appeared in the review, the number of concepts associated with each theme, and a description of each theme.

The module is presented to participants in 32 slides. It is structured following an agenda of (1) core terminology, (2) main concepts, (3) examples, (4) summary and recommendations, and (5) cases. The following sections provide a technical description of the module and its functionality, then a more detailed description of the content contained within the module and the cases.

TECHNICAL DESCRIPTION OF THE MODULE

Navigation

To facilitate use of the module, the opening screen displays a set of instructions on how to navigate the module (Figure 1), which is slide based and allows the user to control when to move to the next slide or return to a previous slide. There is also an autoplay option to allow the module to continue showing content without requiring user interaction. The module can be paused as needed. These instructions can be viewed by the user as needed. To show the amount of progress through the module there is a slide counter.

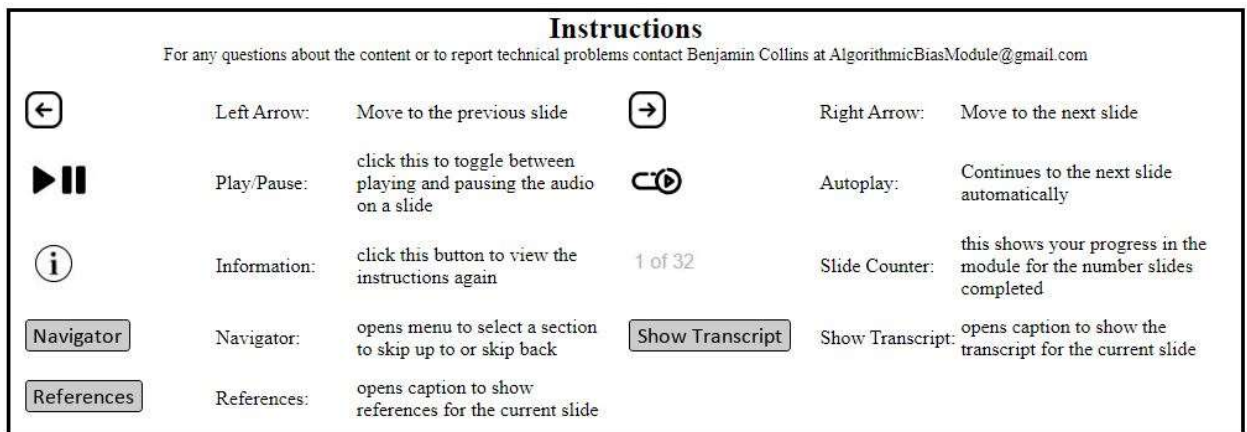


Figure 1. Navigation instructions for viewing the module. This appears prior to initiation of the module and can also be brought back up as needed by the learner.

Transcript

The module automatically plays an mp3 audio for each slide. The content of the audio can be displayed as text in a textbox below the module by clicking on the transcript navigation button. For the cases, a Portable Document Format (PDF) version of the text transcript can be downloaded to view alongside the video.

References

Each slide has a list of references that are associated with the content of that slide. Those references can be viewed in a textbox by clicking on the references button. The references for the cases are listed below each case on the website. There is also a full list of references available as a separate page on the website.

Case Discussions

The case discussions were set apart from the module in order to facilitate construction of a forum that allows for multiple participants. Each case is presented in the format of an mp4 video followed by discussion questions. The discussion is hosted via the website Disqus. Disqus is free for use and does not require login for users to participate allowing learners to remain anonymous. At this time continuing medical education credits are not available, but if that becomes an option in the future, then learners would need to submit with a name to check for completion. Participants are not expected to reply to all available discussion questions but the expectation for full participation is listed for participants in the instructions.

Feedback

Anonymous feedback is received via two Qualtrics surveys. The first survey is a pre-module survey that collects data about the learner's level of clinical experience, area of practice, current knowledge of AI in healthcare, and how important the learner considers the concept of algorithmic bias to be for healthcare. The post-module survey assesses

how much knowledge the learner gained from the module and the learner's thoughts on the quality of the content, including the selection of main concepts.

CONTENT DESCRIPTION OF THE MODULE

The content of the module was based on the literature review, focused on the most frequent concepts present in the review. However, further content was included in order to provide learners with a basis and to describe the content more fully. To view screenshots of the content, see Appendix C.

Introduction (Slides 1-3)

The module content begins with an introduction. The First slide displays the title of the module along with the instructions as shown above in figure 1. This is followed by a slide listing the learning objectives to the participant as listed in the introduction section of this paper. The third slide shows the participant the agenda, as listed in the results section of this paper, on how the content is organized beyond this point.

Core Terminology (Slides 4-9)

The core terminology is derived from background concepts that are essential for understanding algorithmic bias. Although these concepts were not among the most frequent concepts, they are necessary stepping stones for learners to grasp before moving onto the main concepts. Few of the papers identified in the literature review considered definitions or basic concepts. However, as this is intended to be an introductory level module, in order to ensure all individuals taking the module could benefit from the main content, this section was necessary to include.

Term	Definition
Artificial Intelligence	AI describes a branch of computer science for developing digital algorithmic systems capable of complex decisions.
Machine Learning	Machine learning is the dominant branch of AI wherein systems train predictive models by identifying patterns of data from input, and then use those patterns to make predictions in new data.
Algorithm	Algorithms are a set of step-by-step instructions which the computer system relies on to perform tasks modified through the training process of machine learning to be more accurate.
Bias	Bias is a systematic deviation from the truth, and the processes that lead to such a deviation. Bias can come in the form of social biases and statistical biases. Bias contrasts with noise or variation which randomly deviates from the truth. Many errors are a combination of both bias and noise.
Algorithmic Bias	The systematic prejudice found in the development and use of artificial intelligence systems in a reflection of the data input, methods used, and human judgment.

Table 2. Core terminology included in the module with definitions necessary for understanding the main content.

Main Concepts (Slides 10-27)

The content in the main concepts section is based on the frequency of concepts from the literature review. The order of presentation was structured according to a logical flow rather than ordered from most frequent to least frequent. Content from other concepts not among the five most frequent was combined into the content for the main concepts wherever it was logical as a supplement for learning the main concept. An example is the use of the concept of racism & colorism as an example of the risk of harm from

algorithmic bias. Each of the main concepts have dedicated content to outline what the concept is and why it is important for a clinician to understand that concept.

Concept	Frequency	Theme	Description and Brief Summary of Content
Biased Data	49	Data	The largest source of algorithmic bias. Biased data is primarily introduced into AI systems by flaws in data collection. An AI system is only as good as the data that was used to train it, so that biased data would result in a biased algorithm.
Explainability & Transparency	32	Trust	Intertwined properties of AI that govern what developers and end-users understand about a system. This concept is related to the black box phenomenon in AI. Explainability and transparency are important for knowing how much trust to place in a system. Commercial developers have an interest in limiting the transparency of systems.
Limits of Regulation	24	Healthcare	Healthcare and governing bodies have a limited system for regulating the use of AI. The FDA has some oversight of AI as medical devices. However, legal responses always lag behind technology development and concerns about algorithmic bias have not been addressed by regulation. Industry has significant influence over AI regulation.
Non-Maleficence	22	Harm	Algorithmic bias presents a risk of harm to patients which clinicians have a duty to minimize or avoid. That harm is often disproportionately distributed to marginalized populations. The origin of that harm can come from anywhere in the lifecycle of an AI system. All new AI tools should be evaluated for safety and bioethicists ideally should be involved in the design process.
Transformation of Healthcare	21	Healthcare	AI tools and the risk for algorithmic bias should be considered in the context of the healthcare system. Healthcare is undergoing a paradigm shift and increasingly making use of big data and AI.

Table 3. List of the 5 most frequent concepts which were included as the focus of the module content, along with the theme they represent, a description of the concepts and a brief summary of the content included in the module for each concept.

Examples (Slides 28-30)

Two practical examples of algorithmic bias were included in the module to help solidify translation of the learned concepts into real use. This would also serve as a way to prepare learners for participating in the case discussions which are constructed around similar real scenarios. The examples are listed and described in table 4.

Example	Description
Melanoma Diagnosis	AI tools have been used to analyze skin lesions for the presence of Melanoma. Data sets used to train these systems have an estimated proportion of 5-10% Black patients. This results in reduced diagnostic accuracy in Black patients. The problem is further complicated in that image capture has been historically affected by racial bias. Digital photography has improved this issue, but measures to correct lighting can also introduce light artifacts or influence color saturation in the image, which could affect the AI output.
Pulse Oximetry	The pulse oximeter measures oxygen saturation of hemoglobin by sending infrared light through the skin. Measurements are known to be affected by skin color such that Black patients are three times more likely to suffer from occult hypoxemia compared with white patients. There is no corrective factor to account for this, but these measurements are still collected as data to be used in AI systems. It is rarely acknowledged that an oximeter that does adjust for skin color was developed in the 1970s but never became widely used.

Table 4. List of two examples included in the module and a description of each example.

Summary and Recommendations (Slide 31 and recommendations page)

In the final slides of the module, there is a brief review of the content. The module then presents a set of recommendations to clinicians on avoiding and minimizing algorithmic bias in healthcare. These recommendations were formed based on the content of the module. In that respect, it focuses on recognizing when bias is present, the risk of harm from that bias, and determining how much to trust the AI system in use. These recommendations are intended to be used as a guide for the learner when participating in the cases, as well as to supplement the thinking process for using AI in real-life.

Recommendations for Clinicians on Avoiding Algorithmic Bias
1. Determine who you can seek guidance from in your organization on using AI systems and be sure to ask questions when they arise. 2. Consider who was represented in the development of the system and their goals. 3. Review how the system was approved and how it is monitored. 4. Check the training data, determine how bias could have influenced the data, and if that bias was adjusted for. 5. If the training data is not available, advocate to your organization and the developers to know more about the system. 6. If the system does not provide a level of certainty, adjust your level of trust in the system output accordingly. 7. Think about the harm that could come from using the AI system and how likely that harm is to occur. 8. Calibrate your trust based on how much you know about the system and the risks associated with using it. 9. Determine how the system fits in the overall care of the patient and the larger healthcare system.

Table 5. Recommendations to clinicians on avoiding algorithmic bias in healthcare, based on the concepts that informed the content of the module.

Cases (Slide 32 and forum)

The cases are similar to the examples in that they are modeled on real-life scenarios.

However, the cases are given more detail in order to inspire discussion among multiple people. In the instructions learners are encouraged to provide thoughtful original answers or reply to another comment while referencing learning points and recommendations from the module. The questions for each case are not structured to necessarily have right or wrong answers but to facilitate discussion and use of the knowledge gained from the module. Learners are not required to answer all of the questions but rather requested to answer at least one question, or provide one reply, for each case. Due to the length of the case, the first case on COVID-19 severity prediction is split into two parts.

Case	Description	Discussion Questions
Covid-19 Severity Prediction Part A	Algorithms were developed to predict the severity of COVID-19 cases in patients in order to guide the distribution of resources. One proprietary algorithm was used without sufficient external validation.	What are some of the pros and cons of implementing the algorithm for predicting COVID-19 severity? How would you interpret results from the algorithm? What else would you want to know before deciding whether to use the algorithm?
Covid-19 Severity Prediction Part B	When a validation study was conducted, the algorithm was discovered to perform with a low sensitivity for predicting adverse events. This case shows the importance of performing external validation studies prior to deploying an algorithm in the care workflow.	Does this study change your initial impression of how to interpret the results of the algorithm and, if so, how? Is this algorithm useful in real clinical practice and why or why not?

Racial Disparities in Knee Osteoarthritis (OA)	When using radiographs to grade the severity of knee OA in patients there is a disparity between the level of pain experienced by Black patients compared to non-Black patients. This disparity is alleviated with the use of algorithms in place of radiologists to assess the level of severity. This study shows that algorithms can be used to address bias when applied to specific problems.	Can you point out and describe some positive aspects of this algorithm? What are some other clinical scenarios where algorithms might be helpful in reducing bias?
Acute Kidney Injury (AKI) Prediction	In order to improve early intervention for AKI, a prediction algorithm was developed using a set of EHR data from the US Department of Veterans Affairs. This algorithm did well overall in predicting AKI within 48 hours in advance. However, it performed worse for female patients. The data used to train the algorithm consisted of only 6% female patients. This case exhibits how biased data results in a biased algorithm.	How would you handle advertising of commercial AI products? What are the pros and cons of using VA data to train AI algorithms?

Table 6. List of cases developed for discussion along with a description of each case and the questions for prompting discussion.

DISCUSSION

From the initial literature review through programming and deployment of the module and continuing into the feedback process, it has been clear that artificial intelligence is an important topic in healthcare, of which algorithmic bias is an essential component for clinicians to be familiar with. However, up to this point in time, there has been little guidance for what clinicians should know although AI is expected to be increasingly used going forward. Part of the reason is that medical education is already filled with content and other important topics are competing for attention. Additionally, the medical education process is slow to adapt. The biggest concern will be that AI becomes widely used in healthcare and the tools introduce biases that are not recognized by clinicians and therefore translate into harm for patients. This is particularly worrisome for marginalized populations who have already been disproportionately harmed by AI in other areas of life such as in policing activities, judicial processes, and housing decisions. Once these tools have been adopted, it becomes difficult to move away from the status quo and those harms could move from being codified in systems to simply being part of life. While regulation can help in some ways, there is a limit to that regulation which also falls behind technology development. Clinicians need to be aware that the tools that have been approved for use may not necessarily be appropriate to use for all patients. But beyond that, they need to know when to trust a system and when to not trust a system. Advocating for AI tools to present a level of certainty as an output as well as for developers to make systems explainable and transparent will help clinicians know how much trust to place in a system. The most significant area to look at for clinicians in respect to the risk for algorithmic bias is if that AI system was trained with biased data.

Clinicians will need access to that data or a summary of that data, and the ability to interpret if that data was biased and to understand what biases could develop in a system over time, even if it was not present initially.

This work suggests that understanding AI and algorithmic bias requires an understanding of where problems originate, how they affect different populations, how to recognize those problems, the kinds of issues are not addressed by regulations, and how all of this is placed within the ever-changing context of providing healthcare. Although implementing AI can require significant technical knowledge, that level of technical know-how is not necessarily a requirement for clinicians in order to understand how to use AI without introducing unfair risk of harm to their patients. Much of the content important for clinicians to recognize and address algorithmic bias is similar to content learned for conducting research, particularly on the use of data. The next step is taking that understanding of data and placing it into the context of the digitization and datafication of healthcare, including the rise of AI and how those AI systems are regulated. This type of learning is well-suited for being presented in a module format since it builds on knowledge that clinicians already have in terms of understanding data.

The literature review did not necessarily lead to the same content that one might choose to include in the module using other means. However, the frequency of concepts is a viable surrogate for what the scientific and medical communities consider most important and the use of AI systems in healthcare is still recent enough that there is not enough expertise to best direct how to put together content on algorithmic bias for clinicians. The

literature review allows a broad overview of the relevant content and homing in on which of that content is most important. There is a risk for the review to be influenced by publication bias, but this should be mitigated by the large scope of the review. Although the classification of concepts into themes did not play a role in which concepts were selected to be included as the focus of the module, identifying themes aided in ensuring the content presented was sufficiently explained by making reference to concepts from the same theme as needed. . Overall, the process of using frequency of concepts to guide the content of the training module could be expanded to other modes of learning, including the development of a syllabus. There should be some caution with the process of taking the frequency of concept appearance as being reflective of the most important content. In the case of this module, doing so without adding supplemental content would have left the module without the foundational content that was not frequent in the literature. This may be because much of the literature was not targeted toward learners and therefore did not necessarily cover basic concepts. That does not mean, however, that the concepts that were covered in the literature were any less important.

Limitations of the module development include a lack of grant funds to support the development. Free images related to the module content for commercial use are limited. It is also difficult for one individual to program the content from scratch at a level to meet the expectations of end-users. A grant would allow the development of more engaging content and a programmer who has time paid to work on building the module. In regard to the module content, there is currently a limit on the relevance of the cases to all

specialties within medicine. In the future, a module like this would benefit from a wider selection of cases that would be targeted to clinicians and trainees of different specialties.

CONCLUSION

In summary, with the increasing use of artificial intelligence in healthcare, there is a clear need for clinicians to be trained on how to use AI systems in their practice. In this context, the risk of algorithmic bias is an important concept for clinicians to understand for the purpose of avoiding harm to patients. Training on AI, and most concerning, specifically on algorithmic bias, has not become part of the medical education curriculum and receives little attention in post-graduate training. However, waiting until AI has become widespread and the need becomes more apparent is not an option when considering the level of risk and how that risk is unfairly distributed to populations that are already underserved by healthcare. For that reason, alternative forms of training will be an essential component of the e AI will continue to be increasingly used in the healthcare setting, requiring clinicians to be well trained in order to make use of those tools without causing undue harm to their patients, especially for patients from marginalized populations. Although there has been little guidance on what clinicians need to know, and how to train clinicians, this study helps to shed some light on training clinicians on algorithmic bias and in the use of artificial intelligence.

REFERENCES

1. Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. *Healthc Manage Forum*. 2020;33(1):10-18.
<https://doi.org/10.1177/0840470419873123>
2. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: Implications for health systems. *J Glob Health*. 2019;9(22):1-5.
<https://doi.org/10.7189/jogh.09.020318>
3. Brault N, Saxena M. For a critical appraisal of artificial intelligence in healthcare: The problem of bias in mHealth. *J Eval Clin Pract*. 2021;27:513-519.
<https://doi.org/10.1111/jep.13528>
4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.
<https://doi.org/10.1126/science.aax2342>
5. Matheny ME. Artificial intelligence in health care: a report from the National Academy of Medicine. *JAMA*. 2019;323(6):509–10.
<https://doi.org/10.1001/jama.2019.21579>
6. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform*. 2021;28:e100251.
<https://doi.org/10.1136/bmjhci-2020-100251>
7. Kundu S. How will artificial intelligence change medical training? *Commun Med*. 2021;1(8). <https://doi.org/10.1038/s43856-021-00003-5>
8. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019 Dec;9(2):020318.
<https://doi.org/10.7189/jogh.09.020318>

9. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ.* 2019; 5(2):e16048. <https://doi.org/10.2196/16048>
10. Arnold MH. Teasing out artificial intelligence in medicine: An ethical critique of artificial intelligence and machine learning in medicine. *J Bioeth Inq.* 2021; 18(1):121-139. <https://doi.org/10.1007/s11673-020-10080-1>

APPENDIX A. LIST OF REVIEWED ARTICLES

Listed in alphabetical order by first author last name.

Author(s)	Year	Title	Publication
Abràmoff MD et al.	2020	Lessons learned about autonomous AI: Finding a safe, efficacious, and ethical path through the development process	Am J Ophthalmol
Aggarawal R et al.	2021	Patient perceptions on data sharing and applying artificial intelligence to health care data: Cross-sectional survey	J Med Internet Res
Alvarado R and Morar N	2021	Error, reliability and health-related digital autonomy in AI diagnoses of social media analysis	Am J Bioeth
Antes AL et al.	2021	Exploring perceptions of healthcare technologies enabled by artificial intelligence: An online, scenario-based survey	Med Inform Decis Mak
Arnold MH	2021	Teasing out artificial intelligence in medicine: An ethical critique of artificial intelligence and machine learning in medicine	J Bioeth Inq
Asan O et al.	2020	Artificial intelligence and human trust in healthcare: Focus on clinicians	J Med Internet Res
Baird A and Schuller B	2020	Considerations for a more ethical approach to data in AI: On data representation and infrastructure	Front Big Data
Bates DW et al.	2020	Reporting and implementing interventions involving machine learning and artificial intelligence	Ann Intern Med
Brault N and Saxena M	2021	For a critical appraisal of artificial intelligence in healthcare: The problem of bias in mHealth	J Eval Clin Pract
Burlina P et al.	2020	Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases	JAMA Ophthalmol
Butt S et al.	2021	Unintentional consequences of artificial intelligence in dermatology for patients with skin of colour	Clin Exp Dermatol

Byrne MD	2021	Reducing bias in healthcare artificial intelligence	J Perianesth Nurs
Canales et al.	2020	Science without conscience is but the ruin of the soul: The ethics of big data and artificial intelligence in perioperative medicine	Anesth Analg
Cath C	2018	Governing artificial intelligence: ethical, legal and technical opportunities and challenges	Phil Trans R Soc A
Cave S et al.	2021	Using AI ethically to tackle covid-19	BMJ
Challen R et al.	2019	Artificial intelligence, bias and clinical safety	BMJ Qual Saf
Char DS et al.	2020	Identifying ethical considerations for machine learning healthcare applications	Am J Bioeth
Chauhan C and Gullapalli RR	2021	Ethics of AI in Pathology: Current paradigms and emerging issues	Am J Path
Chen M and Decary M	2020	Artificial intelligence in healthcare: An essential guide for health leaders	Healthc Manage Forum
Chin-Yee B and Upshur R	2019	Three problems with big data and artificial intelligence in medicine	Perspect Biol Med
Christou CD, Tsoulfas G	2021	Challenges and opportunities in the application of artificial intelligence in gastroenterology and hepatology	World J Gastroenterol
Clark CR et al.	2021	Health care equity in the use of advanced analytics and artificial intelligence technologies in primary care	J Gen Intern Med
Cockerill RG	2020	Ethics implications of the use of artificial intelligence in violence risk assessment	J Am Acad Psychiatry Law
Cohoon TJ and Bhavnani SP	2020	Toward precision health: applying artificial intelligence analytics to digital health biometric datasets	Per Med
DeCamp M and Lindvall C	2020	Latent bias and the implementation of artificial intelligence in medicine	J Am Med Inform Assoc
Esmailzadeh P	2020	Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives	BMC Medical Inform Decis Mak
Gearhart A et al.	2020	A primer on artificial intelligence for the paediatric cardiologist	Cadiol Young

Ghosh A and Kandasamy D	2020	Interpretable artificial intelligence: Why and When	AJR Am J Roentgenol
Giordano C	2021	Accessing artificial intelligence for clinical decision-making	Front Digit Health
Glauser W	2020	AI in health care: Improving outcomes or threatening equity?	CMAJ
Goodman K et al.	2020	Balancing risks and benefits of artificial intelligence in the health sector	Bull World Health Organ
Guo LN et al.	2021	Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection— A scoping review	J Am Acad Dermatol
Gurupur V and Wan TTH	2020	Inherent bias in artificial intelligence-based decision support systems for healthcare	Medicina (Kaunas)
Hague DC	2019	Benefits, pitfalls, and potential bias in health care AI	N C Med J
Hallak JA and Azar DT	2020	The AI revolution and how to prepare for it	Transl Vis Sci Technol
Howard A and Borenstein J	2018	The ugly truth about ourselves and our robot creations: the problem of bias and social inequity	Sci Eng Ethics
Isaacs D	2020	Artificial intelligence in healthcare	J. Paediatr. Child Health
Johnson SLJ	2019	AI, machine learning, and ethics in health care	J Leg Med
Johnson-Mann CN	2021	Equity and artificial intelligence in surgical care	JAMA Surg
Kelly CJ	2019	Key challenges for delivering clinical impact with artificial intelligence	BMC Med
Kiener M	2021	Artificial intelligence in medicine and the disclosure of risks	AI Soc
Kluge EHW	2020	Artificial intelligence in healthcare: Ethical considerations	Healthc Manage Forum
Krass M et al.	2021	How US law will evaluate artificial intelligence for covid-19	BMJ
Krittanawong C et al.	2017	Artificial intelligence in precision cardiovascular medicine	J Am Coll Cardiol

Kukikowski C and Maojo VM	2021	COVID-19 pandemic and artificial intelligence: Challenges of ethical bias and trustworthy reliable reproducibility?	BMJ Health Care Inform
Lennox-Chhugani N et al.	2021	Women's attitudes to the use of AI image readers: A case study from a national breast screening programme	BMJ Health Care Inform
Leslie D	2021	Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?	BMJ
Lindsall CJ et al.	2020	Action-Informed Artificial Intelligence—Matching the Algorithm to the Problem	JAMA
Linthicum KP et al.	2019	Machine learning in suicide science: Applications and ethics	Behav Sci Law
Loftus TJ et al.	2020	Artificial intelligence and surgical decision-making	JAMA Surg
Lopez-Jimenez F et al.	2020	Artificial intelligence in cardiology: Present and future	Mayo Clin Proc
Macrae C	2019	Governing the safety of artificial intelligence in healthcare	BMJ Qual Saf
Matheny ME et al.	2020	Artificial intelligence in health care: A report from the National Academy of Medicine	JAMA
McCoy LG et al.	2020	What do medical students actually need to know about artificial intelligence?	NPJ Digit
McCradden MD et al.	2020	Ethical limitations of algorithmic fairness solutions in health care machine learning	Lancet Digit
McCradden MD et al.	2020	Clinical research underlies ethical integration of healthcare artificial intelligence	Nat Med
Meskó B and Görög M	2020	A short guide for medical professionals in the era of artificial intelligence	NPJ Digit
Moureau JT	2020	Biased intelligence: On the subjectivity of digital objectivity	BMJ Health Care Inform
Morley J and Floridi L	2020	An ethically mindful approach to AI for health care	Lancet
Nabi J	2018	How bioethics can shape artificial intelligence and machine learning	Hastings Cent Rep

Nabi J	2020	Addressing the "wicked" problems in machine learning applications – time for bioethical agility	Am J Bioeth
Nelson CA	2020	Patient perspectives on the use of artificial intelligence for skin cancer screening: A qualitative study	JAMA Dermatol
Nelson GS	2019	Bias in artificial intelligence	N C Med J
Nordling L	2019	Mind the gap	Nature
Norori N et al.	2021	Addressing bias in big data and AI for health care: A call for open science	Patterns (N Y)
Nundy S et al.	2019	Promoting trust between patients and physicians in the era of artificial intelligence	JAMA
O'Reilly-Shah VN et al.	2020	Bias and ethical considerations in machine learning and the automation of perioperative risk assessment	Br J Anaesth
Obermeyer Z and Topol EJ	2021	Artificial intelligence, bias and patients' perspectives	Lancet Digit
Panch T et al.	2019	Artificial intelligence and algorithmic bias: Implications for health systems	J Glob Health
Parikh RB et al.	2019	Addressing bias in artificial intelligence in health care	JAMA
Pham Q et al.	2021	The need for ethnoracial equity in artificial intelligence for diabetes management: Review and Recommendations	J Med Internet Res
Pieszko K et al.	2021	Clinical applications of artificial intelligence in cardiology on the verge of the decade	Cardiol J
Pot M et al.	2021	Not all biases are bad: Equitable and inequitable biases in machine learning and radiology	Insights Imaging
Richardson JP	2021	Patient apprehensions about the use of artificial intelligence in healthcare	NPJ Digit
Rodriguez F et al.	2018	Promise of perils of big data and artificial intelligence in clinical medicine and biomedical research	Circ Res
Röösli E et al.	2021	Bias at warp speed: How AI may contribute to the disparities gap in the time of COVID-19	J Am Med Inform Assoc

Ross P and Spates K	2020	Considering the safety and quality of artificial intelligence in health care	Jt Comm J Qual Patient Saf
Rush B et al.	2018	From big data to artificial intelligence: Harnessing data routinely collected in the process of care	Crit Care Med
Scott I et al.	2021	Clinician checklist for assessing suitability of machine learning applications in healthcare	BMJ Health Care Inform
Straw I and Callison-Burch C	2020	Artificial intelligence in mental health and the biases of language based models	PLoS ONE
Sujan M et al.	2019	Human factors challenges for the safe use of artificial intelligence in patient care	BMJ Health Care Inform
Tat E et al.	2020	Addressing bias: artificial intelligence in cardiovascular medicine	Lancet Digit
Amann J et al.	2020	Explainability for artificial intelligence in healthcare: A multidisciplinary perspective	BMC Med Inform Decis Mak
Vayena E et al.	2018	Machine learning in medicine: Addressing ethical challenges	PLoS Med
Vokinger KN et al.	2021	Mitigating bias in machine learning for medicine	Commun Med (Lond)
Vokinger KN and Gasser U	2021	Regulating AI in medicine in the United States and Europe	Nat Mach Intell

APPENDIX B. FULL LIST OF CONCEPTS

listed in alphabetical order.

Concept	Theme (s)	Number of Articles
Accountability	Trust, Healthcare	19
Accumulation of Bias	Function, Harm	10
Age Bias	Harm	4
AI Assistants	Background	1
AI chasm	Healthcare	2
AI for Addressing Bias	Beneficence, Function	1
AI for Rare Diseases	Healthcare	2
AI Winter	Function	1
AI-Human Partnership	Trust	13
Algorithmic Fairness	Beneficence	19
Anticipatory Design	Beneficence, Function	1
Artificial Neural Network	Background	4
Assessment of AI	Data	18
Assistive vs. Autonomous AI	Background	3
Assumptions	Data, Trust	8

Auditing	Data	6
Automation Bias	Harm, Trust	13
Automation Complacency	Harm, Trust	5
Awareness of AI	Trust	1
Balancing of Values	Function	4
Benchmarking	Data	1
Beneficence	Beneficence	3
Beneficial Bias	Beneficence	5
Biased Data	Data	49
Biased Tools	Harm	2
Bioethical Agility	Beneficence	1
Catalog of Bias	Beneficence	1
Catastrophic Failure	Harm	1
Clinician Participation	Beneficence	3
Cognitive Computing Systems	Background	2
Commercial Bias	Harm	10
Communication of Uncertainty	Function	5

Competency	Function, Trust	3
Confirmation Bias	Harm	1
Conflict of Interest	Research	1
Continuous Learning	Function	3
Data Integrity	Data	23
Data Purpose	Data	1
Debiasing	Beneficence	7
Decentralized AI	Data	1
Decision Trees	Background	1
Deep Learning	Background	7
Defining the Problem	Research	5
Dehumanization of Medicine	Harm, Healthcare	10
Democratization of Information	Beneficence, Trust	2
Deskilling	Healthcare	3
Developer bias	Harm	9
Digital Divide	Harm	2
Disclosure and Consent	Trust, Healthcare	7

Distributional Shift	Function	6
Distributive Justice	Beneficence	8
Dual Valence Problem	Data	1
Edge Cases	Function	2
Education Bias	Harm	1
Ethical Design	Beneficence	15
Ethical Governance	Healthcare	16
Explainability	Trust	32
Fail-Safe	Beneficence	1
Feature Selection	Data, Research	2
Function Bias	Function	1
Gender Bias	Harm	3
General AI	Background	2
Generalizability	Data	8
Geographic and Culture Variance	Data	6
Goldilocks Principle	Beneficence	1
Ground Truth	Research	3

Hidden Stratification	Data	1
Hindsight Bias	Harm	1
Identifying Bias	Beneficence	6
Inclusivity	Beneficence	1
Information Bias	Data	1
Input Bias	Data	6
Interpretability	Trust	13
Justice	Beneficence	3
Latent Bias	Data	4
LGBTQI Discrimination	Harm	2
Limitations of AI	Function	4
Limited Data	Data	10
Limits of Regulation	Data, Healthcare	24
Locked Learning	Function	2
Majority Rule	Harm	1
Medical Robots	Background	1
Medicalization	Data	6

Misclassification	Data	1
Model Bias	Data	1
Model Variance	Data	2
Moral AI	Beneficence	1
Motive	Trust	1
Narrow AI	Background	1
Natural Language Processing	Background	2
Neural Networking	Background	2
Non-Maleficence	Harm	22
Omitted Variable Bias	Data, Research	1
Outcome Noise	Data	1
Out-of-Sample Input	Data	1
Overfitting	Data	8
Patient and Community Engagement	Beneficence	7
Performance Over Time	Function	11
P-Value Hacking	Harm, Research	2
Precision and Accuracy	Research	1

Prediction Drift	Function	1
Pre-Existing Bias	Harm, Data	32
Privilege Bias	Harm	2
Quantifying Bias	Data	1
Racism and Colorism	Harm	34
Reflexivity Statements	Beneficence, Research	1
Reinforcement Learning	Background	7
Relational Justice	Beneficence	6
Reliability	Trust	6
Reward Hacking	Function, Harm	1
Saliency Maps	Beneficence	3
Selection Bias	Data, Research	5
Self-fulfilling Bias	Harm	1
Site-Specific Bias	Data	3
Social Bias	Harm	2
Sociotechnical	Beneficence, Function	6
Spectrum Bias	Data	1

Standardization	Data	14
Statistical Bias	Data	3
Study Quality	Research	1
Study Reproducibility	Research	2
Subgroup Analysis	Data, Research	6
Supervised Learning	Background	9
Support Vector Machines	Background	1
Synthesized Data	Data	2
Technological Imperative	Healthcare	1
Tracing AI Developing	Beneficence	3
Transformation of Healthcare	Healthcare	21
Transparency	Trust	27
Trust in AI	Trust	31
Underfitting	Data	3
Underspecification	Data	1
Unsupervised Learning	Background	7
Use Context	Function, Healthcare	3

Validation of AI	Data, Trust Research	19
Veracity	Trust, Research	1
What is AI	Background	17
What is Algorithmic Bias	Background	3
What is an Algorithm	Background	2
What is Bias	Background	4
What is Machine Learning	Background	13

APPENDIX C. SCREENSHOTS OF MODULE SLIDES AND TEXT TRANSCRIPT

Slide One

Navigator

References

An Introduction to Algorithmic Bias in Health Care for Clinicians



1 of 32

Transcript

i

autoplay: off

▶|| paused

→

In taking this module, you will learn to recognize different forms of algorithmic bias in the use of artificial intelligence for health care and the origins of those biases, recalling them for multiple choice questions; To deconstruct algorithmic bias in the use of artificial intelligence for health care by discussing example cases; And to empathize with the importance of recognizing and addressing algorithmic bias in health care by exploring the problems resulting from algorithmic bias and the disproportionate impact on already marginalized populations.

Slide Two

Navigator

References

Objectives

In this module you will learn to...

- (1) Recognize different forms of algorithmic bias in the use of artificial intelligence for health care
- (2) Deconstruct real-life examples of algorithmic bias
- (3) Empathize with the importance of recognizing and addressing algorithmic bias in health care

2 of 32

Transcript

i

autoplay: off

▶|| playing

←→

In this module, you will learn, first, to recognize different forms of algorithmic bias in the use of artificial intelligence for health care and the origins of those biases; second, to deconstruct algorithmic bias in the use of artificial intelligence for health care; and third, to empathize with the importance of recognizing and addressing algorithmic bias in health care by exploring the problems resulting from algorithmic bias and the disproportionate impact on already marginalized populations.

Slide Three

Navigator

References

Agenda

- 1 Core Terminology
- 2 Main Concepts
- 3 Examples
- 4 Summary and Recommendations
- 5 Cases



3 of 32

Transcript



autoplay: off



playing



We will begin the module by reviewing some core terminology essential for understanding the content. Then we will delve into the main concepts of algorithmic bias, followed by looking at some examples. You will then receive a brief summary of the module and a set of general recommendations on how to approach using AI systems in health care to mitigate bias. After viewing the module you can then move on to cases where you will apply the concepts learned here.

Slide Four

Navigator

References

1. Core Terminology

- 1 Artificial Intelligence
- 2 Machine Learning
- 3 Algorithms
- 4 Bias
- 5 Algorithmic Bias

4 of 32

Transcript



autoplay: off



playing



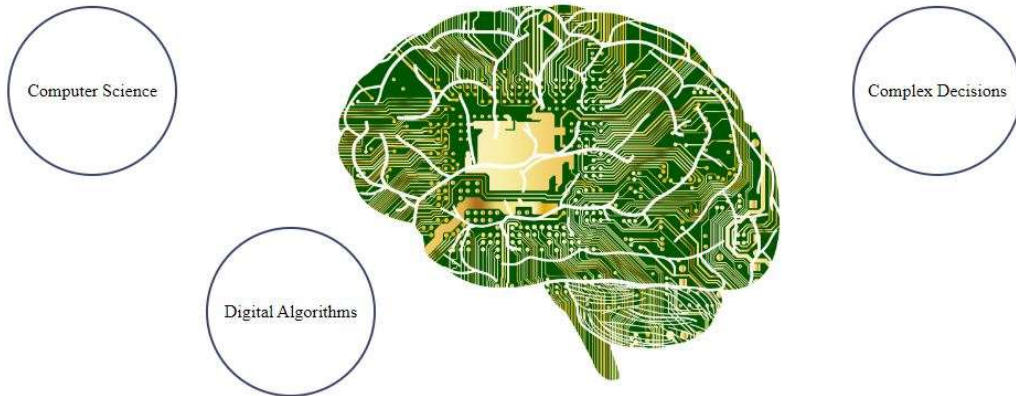
Lets begin with the core terminology you will need to understand algorithmic bias in health care. There are five key terms: Artificial Intelligence, Machine Learning, Algorithms, Bias, and Algorithmic Bias. This will provide a starting point, but you can learn some other important terms in the reading listed on the overview page.

Slide Five

Navigator

References

What is Artificial Intelligence?



5 of 32

Transcript



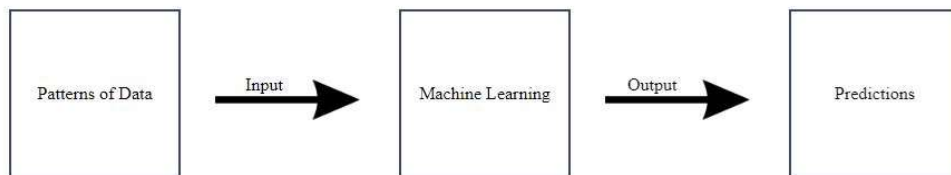
First, what is artificial intelligence? AI describes a branch of computer science for developing digital algorithmic systems capable of complex decisions

Slide Six

Navigator

References

What is Machine Learning?



6 of 32

Transcript



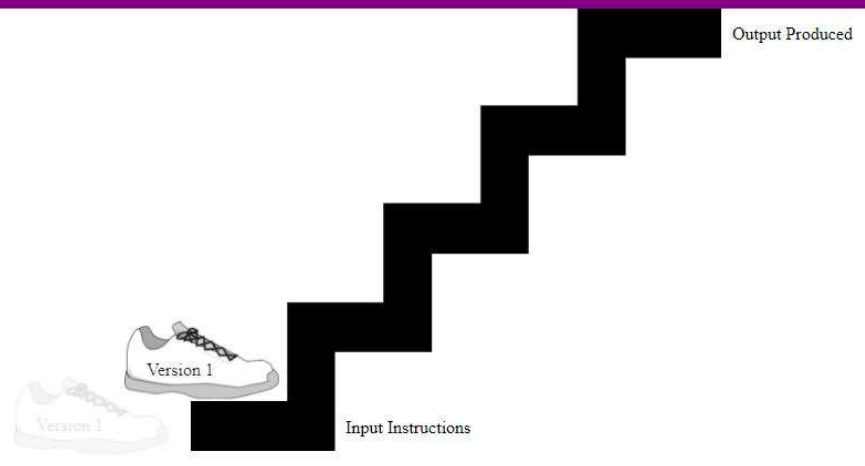
Next, what is machine learning? Machine learning is the dominant branch of AI wherein systems train predictive models by identifying patterns of data from input, and then use those patterns to make predictions in new data.

Slide Seven

Navigator

References

What is an Algorithm?



Input Instructions

Output Produced

Version 1

Version 1

7 of 32

Transcript

autoplay:off

playing

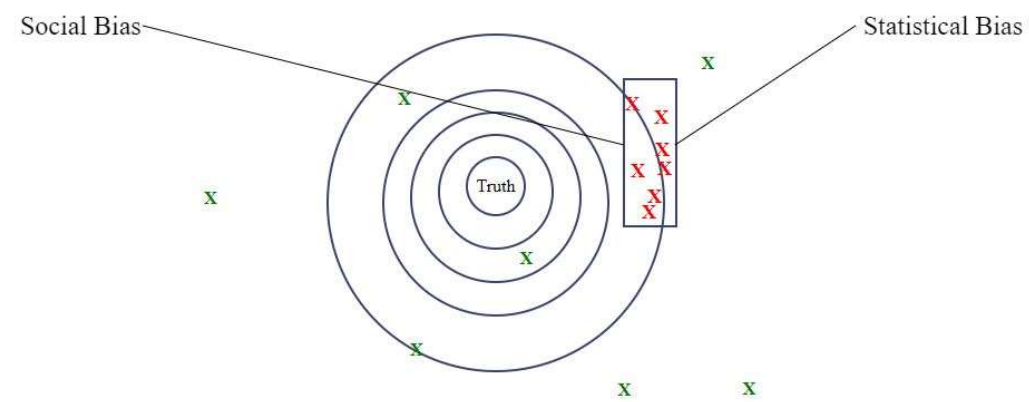
Algorithms are a set of step-by-step instructions which the computer system relies on to perform tasks modified through the training process of machine learning to be more accurate.

Slide Eight

Navigator

References

What is Bias?



Social Bias

Truth

Statistical Bias

8 of 32

Transcript

autoplay:off

playing

Bias is a systematic deviation from the truth, and the processes that lead to such a deviation. Bias can be come in the form of social biases and statistical biases. Bias contrasts with noise or variation which randomly deviates from the truth. Many errors are a combination of both bias and noise.

What is Algorithmic Bias?



Data Input



Methods



Human Judgment

9 of 32

Transcript



autoplay:off



playing



And now, what is algorithmic bias? This is the systematic prejudice found in the development and use of artificial intelligence systems in a reflection of the data input, methods used, and human judgment.

2. Main Concepts

- 1 Biased Data
- 2 Transparency and Explainability
- 3 Regulation of AI in Health Care
- 4 Risk of Harm From Algorithmic Bias
- 5 Transformation of Health Care

10 of 32

Transcript



autoplay:off



playing



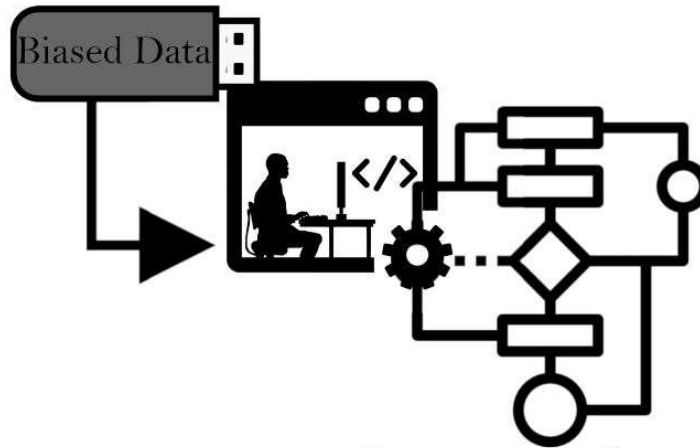
Now that we have the core terms defined, we can move to some of the main concepts in algorithmic bias in health care. These concepts are derived from an extensive literature review and discussions with experts on AI in health care. We will cover 5 main concepts: biased data, transparency & explainability, regulation of AI in health care, risk of harm, and the transformation of health care. What is covered here is just the tip of the iceberg, but representative of essential knowledge for clinicians who will be interacting with AI tools in providing patient care.

Slide Eleven

Navigator

References

Concept #1 - Biased Data



11 of 32

Transcript



The first concept we will explore is biased data which is the largest source of bias in AI algorithms.

Slide Twelve

Navigator

References

Where does biased data come from?

Flaws in Data Collection



Preexisting Social Biases



System Drift



Feedback Loops



12 of 32

Transcript



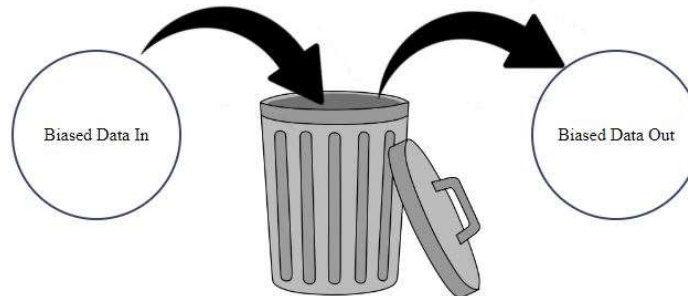
Click each of the four areas to learn about different major sources of biased data in AI systems.

Slide Thirteen

Navigator

References

Garbage in - garbage out



13 of 32 Transcript



The importance of biased data is in how this data influences the development of the AI algorithm, in a problem often referred to as garbage in, garbage out. In other words, AI systems are only as good as the data on which they are trained.

Slide Fourteen

Navigator

References

Preexisting bias is captured in data



Codification of preexisting biases



Risks for unintended consequences



Amplification of preexisting biases

14 of 32 Transcript



Another way in which data is biased is when preexisting social biases are captured in the data and then codified into the algorithm. In any AI tool, there are risks for unintended consequences when users are not aware of the social biases being captured in the data. And furthermore, those biases can become amplified by the system. Click each line to hear examples of how each biased process could occur.

Slide Fifteen

Navigator

References

Avoiding biased data

Diverse



Accessible



Standardized



New



Accurate



Complete



15 of 32

Transcript



autoplay:off



playing



To avoid using biased data, it is essential for AI algorithms to rely on diverse data sources which are accessible and standardized. Due to the poor integrity of data historically, the data used for an algorithm should be new whenever possible, in addition to being accurate and complete.

Slide Sixteen

Navigator

References

Concept #2 - Transparency and Explainability



"Any sufficiently advanced technology is indistinguishable from magic." -Arthur C. Clarke

16 of 32

Transcript



autoplay:off



playing



The next concept, transparency and explainability, is important for understanding the role of using AI systems in health care. When advanced technology is not well understood, as Sci-Fi writer Arthur C. Clarke suggests, it is indistinguishable from magic.

Slide Seventeen

Navigator

References

The Black Box



17 of 32

Transcript



Transparency and Explainability are best described in the frame of black boxes in AI. The black box refers to a situation where an AI algorithm has learned to provide an output but no one is able to explain how the system derived the output. This can be problematic in terms of our inability to know if that output is correct and fair. This is especially an issue in health care where the ability to explain is essential to maintaining trust in the health care system.

Slide Eighteen

Navigator

References

Expanding Transparency and Explainability



Hindered by commercialization



There is a burden on clinicians to learn more

18 of 32

Transcript



The ability for users of an AI system to understand how the output was achieved is often purposefully hindered by commercial entities, where developers want to keep proprietary tools unexamined. This means that tools in use can provide incorrect or unfair output without the user being able to determine that is the case. Clearly, this can be an issue when AI tools are used to provide clinical care. From a different perspective, it is important to consider that any AI systems with some level of transparency and explainability will increase the burden on clinicians to learn more with the expectation to be educated users of those systems.

Slide Nineteen

Navigator

References

(Mis)trust in AI

Auditing

Certainty

Medium-Level of Trust

At a medium level of trust, you may have some idea about how well the system works for a test population, but you may not know the characteristics of that population and so it is unclear whether you should apply the system to you individual patient. At this point, you would want to seek out more information about the system and how it was trained. A system that does not report to the user its level of certainty can be a reason to hold off on having a high level of trust in a system. Additionally, even if a system initially performs well, not being able to audit the system should suggest not having a high level of trust as the level of system performance may not remain high over time.

0.0

0.5

1.0

19 of 32

Transcript

autoplay:off

playing

Underlying the concept of explainability and transparency is the inclination to trust, or not trust, the use of AI in health care. As clinicians, there is a need to determine how much trust to place in an AI system. Part of knowing how much trust to place in the system is in the transparency of those systems and the ability to audit how well the algorithm is performing. Those systems should also regularly report to users the level of certainty with which it is providing an output rather than as a simple binary output. Based on what is known about a system, clinicians should calibrate their level of trust in the system, with the level of trust also dependent on the context in which the system is used. Click the left and right arrows next to the calibration tool appearing on the

Slide Twenty

Navigator

References

Concept #3 - Regulation of AI in Health Care

20 of 32

Transcript

autoplay:off

paused

To effectively use AI tools, clinicians should have some understanding of the regulation for those tools, and for understanding bias, specifically the limits of regulation.

Slide Twenty-One

Navigator

References

Limits to the regulation of AI systems



- FDA recognizes AI systems as a medical device
- Legal response lags behind technology development
- FDA has not directly addressed concerns about algorithmic bias
- Uncertainty about liability in the use of AI
- Industry can exert significant influence over AI regulation

21 of 32 Transcript



The legal response to new technology always lags behind the development. At the time of developing this module, the FDA has not addressed concerns regarding algorithmic bias in the tools used for health care. There also remains significant uncertainty about liability in the development and use of AI. And, as with other areas of health care, industry can exert a significant influence over the regulation of AI. In Europe, there has been ongoing debate on a right to explanation. In thinking about the level of explainability in AI systems in health care, it is important to keep in mind how this can also limit the effectiveness of AI systems.

Slide Twenty-Two

Navigator

References

Concept #4 - Risk of harm from algorithmic bias



22 of 32 Transcript



Medical professionals have a duty of nonmaleficence, to do no harm or more specifically weigh the risk of harm against the benefits. The clinical use of AI algorithms with bias presents a clear risk of harm that clinicians must be able to fairly weigh against their benefits.

Slide Twenty-Three

Navigator

References

Potential Harm



23 of 32

Transcript

i

autoplay:off

playing

⏮ ⏭

In general, the use of AI tools lends to a high risk of bias, and that bias represents potential harm. Where bias is involved, it also means that the risk of harm is disproportionately distributed to marginalized populations. And this harm can arise anywhere in the lifecycle of an AI system from the development of the system in deciding what to make an algorithm for and why, to the collection and use of biased data that informs the algorithm, to where and how the algorithm is deployed, as well as how it is evaluated.

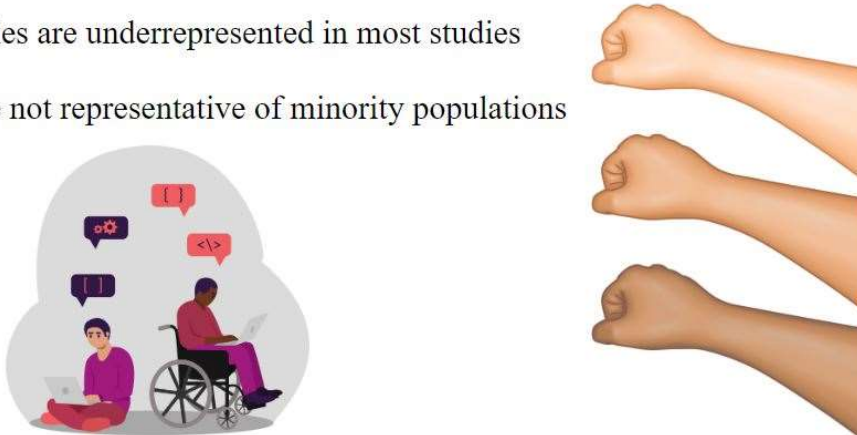
Slide Twenty-Four

Navigator

References

Racism and Colorism

- Racial minorities are underrepresented in most studies
- Developers are not representative of minority populations



24 of 32

Transcript

i

autoplay:off

playing

⏮ ⏭

The most commonly referenced form of harm from algorithmic bias is racism and colorism. This goes back to the issue with biased data where racial minorities are underrepresented in most studies. But it goes back even further in that the developers of AI algorithms are often not representative of minority populations. This influences the decisions made in the development process, especially where social factors are involved and may not receive enough attention.

Slide Twenty-Five

Navigator

References

Avoiding Harm

- All new clinical tools should be evaluated for safety
- Bioethicists should be involved in the design of AI systems
- Patients expect their clinicians to ensure systems are safe



25 of 32 Transcript



To avoid harm in using AI systems in health care, clinical tools should be evaluated fully for safety and effectiveness in an independent validation process. During the development and evaluation process bioethicists should be involved. And clinicians need to be aware of how effective and safe these tools are, including in specific populations so they can determine which circumstances and patients are suitable for using the tool and how to interpret the results.

Slide Twenty-Six

Navigator

References

Concept #5 - Transformation of health care



26 of 32 Transcript



The final concept, transformation of health care, places the risk of algorithmic bias and the use of AI tools into context of how the health care system is changing, and how the role of clinicians is changing alongside that.

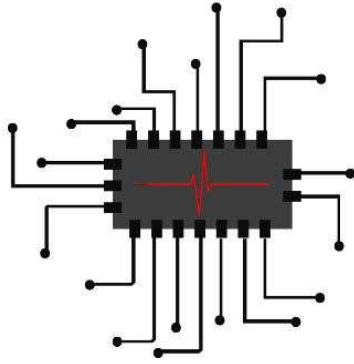
Slide Twenty-Seven

Navigator

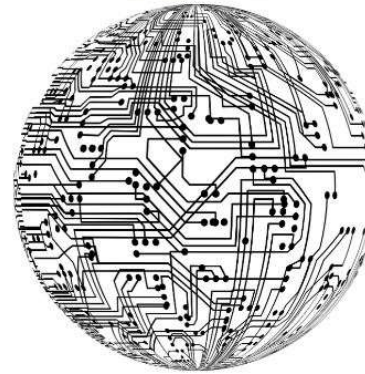
References

Datafication and Digitalization of Health Care

750 quadrillion bytes of data daily



Paradigm Shift



27 of 32 Transcript



autoplay:off



playing



There is a clear paradigm shift occurring with the datafication and digitalization of health care, impacting what clinicians are expected to know and how they are expected to interact with patients. Each day, 750 quadrillion bytes of patient data is produced which is far more than what human clinicians can work with, making AI algorithms increasingly necessary in the health care work flow.

Slide Twenty-Eight

Navigator

References

3. Examples



28 of 32 Transcript



autoplay:off



playing



Now we will go through a couple of examples, keeping in mind the terminology and concepts just covered.

Slide Twenty-Nine

Navigator

References

Melanoma Diagnosis



29 of 32

Transcript



The analysis of skin lesions by AI to determine the presence of melanoma has become the classic example of algorithmic bias in health care. AI systems for the diagnosis of skin lesions are often trained with images of these lesions on white patients, while data sets have an estimated proportion of Black patients of 5-10%. This leads to reduced diagnostic accuracy in Black patients compared to the purported level of the system. In the case of melanoma, this can be very consequential in that the estimated 5 year survival rate is 70% for Black patients compared to 94% for white patients. The problem here goes deeper than inclusion, as the process of image capture has historically been affected by racial bias. Digital photography has brought about advancements that

Slide Thirty

Navigator

References

Data Collection with Pulse Oximeter



30 of 32

Transcript



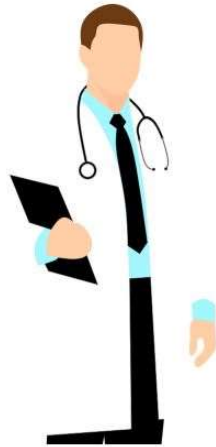
The pulse oximeter measures oxygen saturation of hemoglobin by sending infrared light through the skin. Measurements of the pulse oximeter are known to be affected by the patient's skin color, as the device systematically overestimates oxygen saturation levels in nonwhite patients. As a result, Black patients are three times more likely to suffer from an occult hypoxemia compared with white patients. Despite the difference in quality of measurement from the devices base on skin color, there is no corrective factor to account for this and these measurements are collected as data to be used in AI systems. Furthermore, it is rarely acknowledged that an oximeter that does adjust for skin color was developed in the 1970s but never became widely used.

Slide Thirty-One

Navigator

References

Summary and Recommendations



31 of 32

Transcript



In summary, algorithmic bias is the systematic prejudice found in the development and use of artificial intelligence systems in a reflection of the data input, methods used, and human judgment. The source of that bias is prominently from the use of biased data, the lack of transparency and explainability, and poses a risk in health care due to inadequate regulation. The harms of algorithmic bias disproportionately affect marginalized populations. Nonetheless, the use of AI technology is increasing along with the amount of data in health care, and so clinicians need to recognize the presence of bias, understand the limitations of tools and regulation, and know when to trust the AI output and when not to trust it. To help put things together, here is a set of general

Slide Thirty-Two

Navigator

References

4. Cases



click to navigate to cases

32 of 32

Transcript



Three cases with four videos based on real-life scenarios have been prepared for you to further explore algorithmic bias in health care. Each case is followed by a small set of questions. To receive credit for completing the module, you will need to participate by answering at least one question for each of the case videos. There is not a right or wrong answer to the questions. The goal is to provide thoughtful answers that use the concepts learned here in the module. To go to the cases, you can either click in the menu or click the image presented here of 3 casebooks.