AI WORKING TEAMS

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

AFDO RAPS

Bias in Artificial Intelligence

In Healthcare Deliverables

2022

AI WORKING TEAMS

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

Contents

Acknowledgments

This paper was developed under the leadership of the Xavier Health program at Xavier University in partnership with industry professionals, as a planned output from the 2020 Xavier AI Summit.

This paper was developed by the Good Machine Learning Practices Team as a follow-up to the white papers published in 2018, 2019, and 2020 regarding good machine learning practices[1] , explainability[2], and good data quality for AI applications[3]. This paper has been completed by the same team under the new umbrella for the Good Machine Learning Practices Team, the AFDO/RAPS Healthcare Products Collaborative's AI Global Initiative specifically:

- Pat Baird
- Eric Henry
- Jackie Karceski
- Betsy Macht
- Diana Miller
- Rohit Nayak
- Scott Thiel

We'd like to thank everyone who contributed to the creation and the review of this paper – without their work, this paper would not have been possible. Our hope is that this paper provides the foundation for new learnings and best practices in this rapidly evolving field to help deliver the promise and potential of AI.

In February of 2022, the efforts of Xavier Health were assumed by the AFDO/RAPS Healthcare Products Collaborative. Because of the important work done before this transition, the Collaborative has chosen to retain some documents that have Xavier branding and continue to provide them to the communities. If you have questions, please contact Timothy Hsu, Director of Health Technology Initiatives, at thsu@healthcareproducts.org

[1] Baird, P. Nayak, R. et al (2018) Perspectives for Good Practices in Continuously Learning AI Systems in Healthcare, Xavier Health & Xavier University

[2] Baird, P. Nayak, R. et al (2019) Building Explainability and Trust for AI in Healthcare, Xavier Health & Xavier University

[3] Baird, P. Nayak, R. et al (2021) Data Quality for AI in Healthcare

1. Introduction

Background

Over the past few years, Artificial Intelligence (AI), and more specifically, Machine Learning (ML) technology, have experienced rapid adoption in the healthcare space as tools for diagnosis and decision-making. Such tools are intended to address challenges in the healthcare system to both process and put into practice the proliferating medical findings, and also to support delivery on the promise of personalized and precision medicine.

Why is AI for healthcare different? Is there any concern or need to focus on bias differently in this application of AI? The interactions with this committee and the FDA have confirmed a need to explore unintended bias in healthcare AI systems. The aspirational goal of using AI in healthcare products is to enhance the functionality of the product, thus improving the clinical value and user experience. Unintended bias could unintentionally disrupt the ability of AI to deliver on the desired goal.

Overall Goals of this Whitepaper

Unintended bias in AI and AI-driven healthcare applications is an evolving topic that developers, reviewers, and experts are still learning to address effectively and consistently. Bias, as a topic, could benefit from a discussion around standard taxonomy and approaches to identification, and by addressing any identified sources of bias. By opening a dialog and setting standards, unintended bias in AI-enabled systems becomes a visible challenge to consider and manage when defining the parameters needed to collect research data prior to creating an AI algorithm. The goal of this publication is to outline a product developer's framework for bias opportunity detection, assessment, and mitigation of unintended bias. Leveraging established and proven methods currently used for risk analysis in healthcare systems specifically for unintended bias should enable more robust management. It should be noted, that bias can occur at points in the data supply chain and product supply chain. While focused on product developers, when appropriate, this paper will discuss perspectives from regulatory reviewers or other stakeholders of these solutions.

In AI algorithm development, machine learning (ML) provides an opportunity in devising a method for the algorithm to learn. One of the concerns about ML in healthcare is that applications could, unknowingly, be biased against certain patient populations, leading to inequities in healthcare delivery. Reports of inequity lead to questions by patients and caregivers, and they are a barrier to adoption of ML technology in healthcare. Whether it is the perception that bias exists or a reality in the product development, when uncertainty exists in regard to product performance, there will be resistance to adoption of the technology.

This paper is intended to be used by both SaMD and SiMD applications, as the potential sources of unintended bias and bias management techniques apply to both types of applications. This paper is intended to supplement existing standards and good practices in the development of health software such as "ISO/IEC 62304 Medical device software — Software life cycle processes" and "ISO/IEC 82304-1 Health software — Part 1: General requirements for product safety."

Audience and Stakeholders

The intended audience of this paper is broad: It includes developers, implementers, researchers, quality assurance and regulatory affairs professionals, validation personnel, business managers, regulators, and end-users faced with the challenge of assessing the quality and performance of AI-based healthcare applications.

2. Bias Overview

Discussion

Data bias can ripple across the end-to-end process for developing Artificial Intelligence (AI) algorithms used in or as medical devices. If unintended bias is present in the data, there is the potential to adversely influence downstream application of that data in research, clinical analysis, product improvement, and product application (Reddy, Allan, Coghlan, & Cooper). The potential opportunities for the introduction of unintended bias include many aspects of the device development process. Some of these opportunities for bias include, but are not limited to: human error and/or bias in defining the basis for the research needed to develop the medical device, data bias in defining what data is needed to build the basis for the algorithm, population and study bias, lack of transparency, privacy concerns that drive omissions of data or participant-driven errors of omission, programming bias, analysis, and device design bias (Reddy et al.) Depending on the application, the use of a biased system could further compound bias for future releases of the software.

When considering how data used to develop AI-based systems could introduce unintended bias, it is useful to look at some of the different elements where bias could be present. These elements are listed below and explored in more detail in Appendix I: Types of Bias).

1. Initial scope – defining the data needed

   Original data collected from a specific study population could contain a range of biases including race, age, sex, and other demographics. These same elements could contain bias when using secondary data from an existing database or merging information from multiple databases.

2. Population bias

   When defining the study population, consideration is needed to appropriately reflect the potential user population accurately so that the necessary diversity of the potential user population is reflected in the data collected.

3. Development bias: computational, social scientific, and humanistic

   Lack of diversity among AI developers and researchers could limit perspectives, contexts, and expertise in developing the AI-based application.

4. Reflection bias

   As data is collected and used to monitor device and algorithm performance and/or further develop the system logic through machine learning, data bias can reflect further into the process biasing the model created. This biased data may not be a true reflection of the situational reality.

5. Data bias

   Inaccurate measurement methods, incomplete data collection, non-standardized user reporting, and biased data sources (linking back to population bias), could all lead to data bias.

6. Algorithm bias

   Bias can be introduced based on the data algorithm(s) chosen for use in the device.

7. Intentional bias

   Based upon the intended use and indications for use of the algorithm, there may be bias introduced to ensure focus on the right use and population of individuals and circumstances.

AI-based systems trained on unintentionally biased data will create models that replicate and potentially magnify those biases creating a model that does not accurately reflect the condition being treated and the population being served. This potentially introduces discrimination in the effectiveness of treating the entire patient population and social inequity. Machine learning (ML) can be established in supervised or unsupervised environments. Just as humans make decisions based on knowledge, ML-based systems learn based on the effectiveness of the algorithm(s) that enables that learning. Machine learning decision making within the AI algorithm uses existing data to predict from previous experience within the system logic (Ahmed & Farid, 2018). With a potentially inaccurate AI model, there is then the potential to introduce new error or harm that was not intended or covered in the research and product framework in providing alerts and diagnosis to the user.

Ahmed, F. & Farid, F. (2018). Applying internet of things and machine-learning for personalized healthcare: Issues and challenges. 2018 International Conference on Machine Learning and Data Engineering (iCMLDE). 19-21. doi: 10.1109/iCMLDE.2018.00014

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 0(0), 2019, 1–7. doi: 10.1093/jamia/ocz192

3. AI/ML Bias Mitigation Workflow

General Principles

The means presented in this paper for identifying and addressing unintended bias follow the general pattern laid out for medical device safety risk management in ISO 14971:2019. [1] Although the medical device risk management standard is used as a general workflow model for identifying and mitigating bias, it is worth clarifying that the terms *bias* and *risk* are not synonymous. Reference our definition of *bias,* as compared with the definition of *risk* in section 3.18 of the standard, which is a "combination of the probability of occurrence of harm and the severity of that harm." Our paper in no way implies that bias inevitably leads to harm, although a consideration of biases, with a potential safety impact, is described and linked with safety risk management subject to the ISO 14971:2019 standard. (See Annex B for an overview of the ISO 14971:2019 workflow.)

The proposed workflow below is presented in the spirit of "early and often." The best application of identifying and addressing unintended bias is to begin as early in the product life cycle as possible and to iterate the workflow periodically throughout the life cycle. This ensures the most current use, requirement, design, and implementation information are considered in effectively addressing unintended bias in the system.

Developers should consider incorporating bias mitigation activities (including the activities identified in this paper) into the product design and development plan to ensure consistent and methodical execution during requirements development, design, development, implementation, verification, validation, release, and post-release life cycle phases. Objective evidence of these activities will reside within the Design History File (DHF)/Technical File of the system. Note that a "system" in this model is defined as an integrated composite consisting of one or more of the processes, hardware, software, facilities, and people, which provides a capability to satisfy a stated need or objective (IEC 62304:2006 +AMD1:2015 (3.30))

The AI/ML Bias Mitigation Workflow

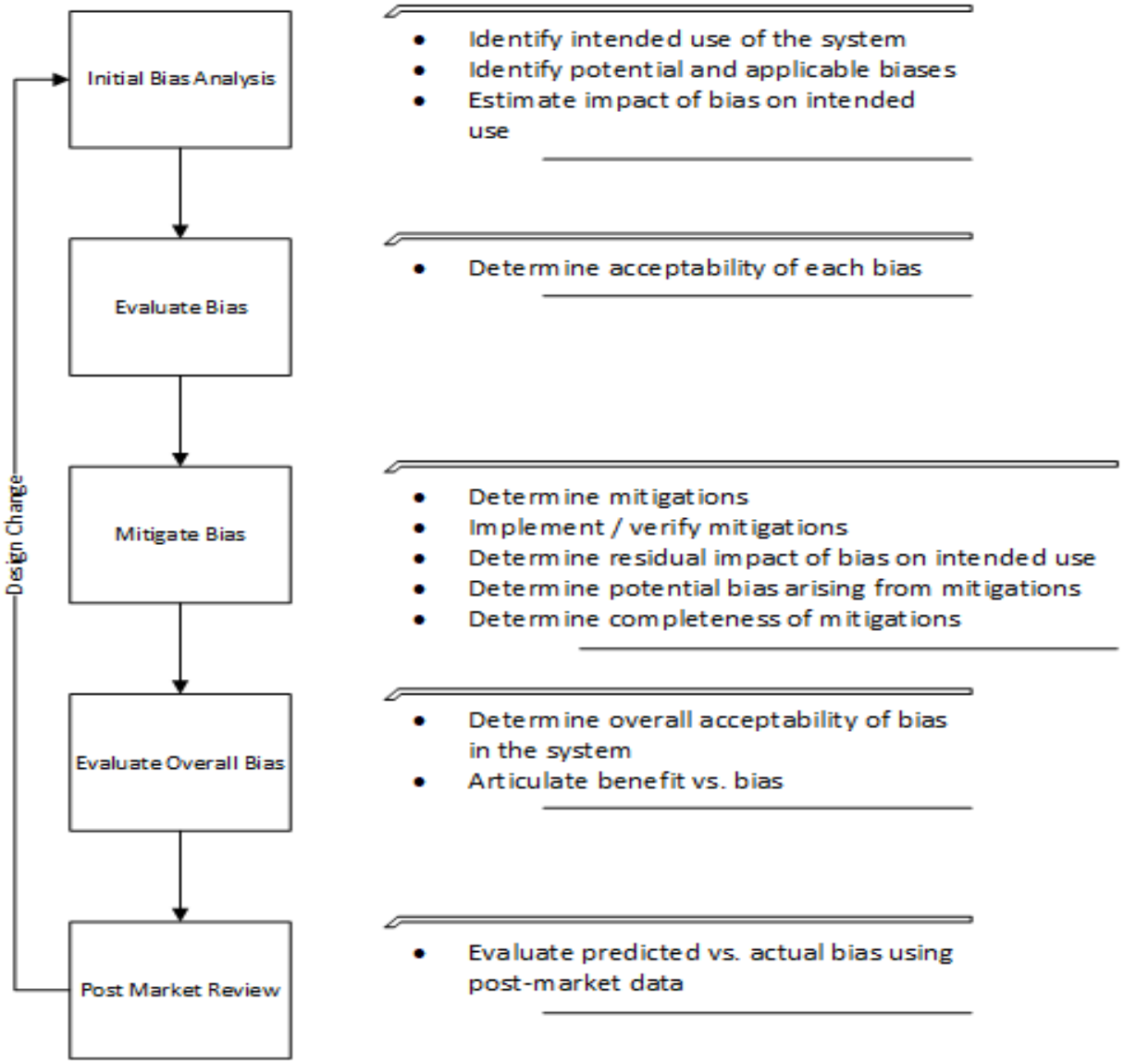

Figure 1

Initial Bias Analysis

Identify the Intended Use of the System

The first step in mitigating bias in AI/ML-based systems is to define the intended use and indications for use of the system against which the impact of unintended bias can be measured. (NOTE: Throughout this section, “intended use” includes both intended use and indications for use.)

The intended use of a medical device is variously identified as the “use for which a product, process, or service is intended according to the specifications, instructions, and information provided by the manufacturer,” and “… should take into account information such as the intended medical indication, patient population, part of the body or type of tissue interacted with, user profile, use environment, and operating principle” while considering “reasonably foreseeable misuse.”[1,2]

In defining “intended use” for 510(k) submissions, the U.S. FDA in 21 CFR §807.92(a)(5) requires, “A statement of the intended use of the device that is the subject of the premarket notification submission, including a general description of the diseases or conditions that the device will diagnose, treat, prevent, cure, or mitigate, including a description, where appropriate, of the patient population for which the device is intended.” For Premarket Approvals, 21 CFR §820.14.20(b)(3) mandates that five elements be part of defining intended use: (1) indications for use, (2) device description, (3) alternative practices and procedures, (4) marketing history, and (5) a summary of studies.

For software systems, a more granular view of intended use and indications for use and an elaboration of the above definitions can be found in a technical report providing guidance for non-device regulated software. The description, however, applies across the spectrum of device and non-device systems. “Specifically, the intended use is meant to describe and explain how the software fits into the overall process that it is automating, what the software does, what one can expect of the software, and how much one can rely on the software to design, produce, and maintain safe medical devices.” This report goes on to detail the three main components of intended use as (1) purpose and intent of the software, (2) software use requirements (e.g., use cases and user requirements), and (3) software requirements. [3]

Using these definitions, the intended use of a system goes beyond a simple statement of intent and includes user and functional levels of system/software requirements. It is worth noting here that the U.S. FDA has proposed a deliverable known as a Software as a Medical Device (SaMD) Pre-Specification (SPS) as a holder of requirements for SaMDit with the purpose of anticipating “modifications to ‘performance’ or ‘inputs,’ or changes related to the ‘intended use’ of AI/ML-based SaMD.” [4] The use of SPSs may be included as an element of intended use, where applicable.

Identify Applicable Biases

Accordingly, a first step in determining the presence of unintended bias could be to identify known foreseeable sources and types of bias associated with the intended use of the system. These elements are listed below and explored in more detail in Appendix I. Where applicable, identify the sequence of events leading from the source of bias to the application of bias during use and foreseeable misuse of the system. Note that there may be multiple combinations of source and type of bias each with multiple potential sequences of events leading to the application of bias during use. Consider providing a separate entry for each bias source/type/sequence

combination noting where bias is intended (e.g., age boundary of the user) and where bias is unintended to enable effective risk analysis.

Estimate the Impact of Bias on Intended Use

The next step would be to evaluate the impact of unintended bias on the intended use of the system. This could be achieved by:

- Providing a subjective narrative for each bias source/type/sequence combination that details the positive, negative, or neutral impact of the application of unintended bias on the intended use of the system.
    - If the bias source/type/sequence combination has an impact on safety (including data privacy), then ensure the combination is considered as part of safety risk analysis.
- Establishing a qualitative scale to categorize each bias source/type/sequence combination incorporating the following elements:
    - Likelihood that the bias source/type/sequence combination will occur during real-world use of the system.
    - Likelihood that there will be impact to the intended use of the system if the bias source/type/sequence combination occurs.
    - Degree (and categorical type) of impact to the intended use of the system if the bias source/type/sequence combination occurs. (Note: A more granular view of this attribute may be provided if there are variable degrees of impact each with a separate likelihood for that impact, given that the combination occurs.
    - Scoring based on the qualitative scale of positive, negative, or neutral impact of the bias source/type/sequence combination to the intended use of the system.
    - Indicator of safety impact or no safety impact (including data privacy) of the bias source/type/sequence combination.
        - If the bias source/type/sequence combination has an impact on safety (including data privacy), then ensure the combination is considered as part of safety risk analysis.

Evaluate Bias

Determine Acceptability of Each Bias

Based on the impact estimation above, determine whether each unintended bias source/type/sequence combination is acceptable (requiring no further mitigation) or unacceptable (requiring mitigations be identified and implemented). Consider one of the following methods for this determination:

- If a subjective narrative was used for impact estimation, provide a further narrative rationalizing whether the unintended bias source/type/sequence combination is acceptable or unacceptable.
- If a qualitative or quantitative scale was used for impact estimation, use a series of decision matrices to make the acceptability determination. For example:

- Determine the likelihood of a unintended bias source/type/sequence combination having an impact on the intended use of the system by correlating the likelihood that a bias source/type/sequence combination will occur with the likelihood it will have an impact on intended use if it occurs.
- Determine the acceptability of the impact to the intended use of the system by correlating the likelihood that a unintended bias source/type/sequence combination will have an impact on the intended use of the system with the degree of impact. (Note: If the positive/neutral/negative indicator is “positive” or “neutral,” the impact could be considered acceptable).

Mitigate Unintended Bias

Determine Mitigations

For the unintended bias source/type/sequence combinations considered unacceptable, determine (and document in the bias analysis) mitigations (e.g., changes to requirements, architecture, design, data, code/algorithms, and/or algorithm training) necessary to bring bias source/type/sequence combinations to an acceptable state. The goal of determining appropriate mitigations is to reduce the likelihood that the bias source/type/sequence combination will occur, reduce the likelihood the bias source/type/sequence combination will have a negative impact if it occurs, and/or reduce the level of impact of the bias source/type/sequence combination.

Implement/Verify Mitigations

Where mitigations have been identified, implement them in the system using controlled design change and configuration management, where applicable.

Verify implemented mitigations through both static and dynamic means (e.g., requirement reviews, technical reviews, code reviews, static code analysis, unit testing, integration testing, functional system testing).

Determine Residual Impact of Bias on Intended Use

Using the methods defined above to estimate and evaluate bias, determine the residual impact of unintended bias to the intended use of the system after bias mitigation mechanisms have been implemented and verified.

Determine Potential Bias Arising from Mitigations

Perform an estimation and evaluation of potential new or changed unintended bias source/type/sequence combinations that may arise from the implementation of bias mitigations. Where applicable, identify, implement, and verify mitigations on these new or changed bias source/type/sequence combinations.

Determine Completeness of Mitigations

Review all bias mitigation activities to ensure that the impacts from all identified bias source/type/sequence combinations have been considered and that all bias mitigation activities are complete.

Evaluate Overall Bias

Determine Overall Acceptability of Bias in the Systems

Taking into consideration the residual impacts of all bias source/type/sequence combinations, determine whether the overall residual impact of bias source/type/sequence combinations is acceptable or unacceptable. If the overall impact is considered unacceptable, consider identifying and implementing further bias mitigations, making changes to intended use, identifying and documenting an acceptable benefit vs. bias assessment (refer to next paragraph), or reconsidering whether to release the system in its current configuration.

Articulate Benefit vs. Bias

For unintended bias source/type/sequence combinations that remain unacceptable, gather and review data to determine if the benefits of the intended use of the system (e.g., technical, clinical, economic) outweigh the residual impact for this specific combination. Consider also making an overall benefit vs. bias determination at the system level. If the benefits do not outweigh the residual impact, consider further system design changes (including bias mitigations and/or changes to intended use), or reconsidering whether to release the system in its current configuration.

Disclose Residual Bias

Disclose known residual bias source/type/sequence combinations including their predicted impact on the intended use of the system to relevant internal and external stakeholders (e.g., company management, customers, users, regulatory authorities). This may be done via a formal report, release notes, instructions for use, or other applicable communication mechanisms.

In support of transparency, clear disclosure of intended bias needs to be disclosed as well.

Post-Market Review

Evaluate Predicted vs. Actual Bias Using Post-Market Data

It is important that the understanding and mitigation of unintended biases that may be introduced by the system remain current and are addressed in a timely manner. For example, “Drift” (sometimes called “Temporal Bias”) arises from changes that occur over a period of time. These may be changes in clinical practice, change in patient demographics, changes to the healthcare landscape (e.g., COVID-19), etc. A product that was developed and trained just

a few years ago may have a change in performance over time, even if the product itself has not changed. To that end, there is a need to continuously collect, review, and adjudicate post-release information related to the use of the system in its intended environment and of other similar systems with similar intended uses (i.e., real-world data).

Effective post-market review of bias in a system may be executed in the following manner:

1. Information Collection

Collect information from a wide variety of trustworthy sources through a series of real-time and periodic activities, as it relates to the use of the system or to similar systems with similar intended uses. Sources of such information may include:

- Continuous post-release testing activities
- Complaints or other forums for user feedback (including surveys)
- Service reports
- Published adverse events
- Post-market studies
- Technical, scientific, and clinical peer-reviewed literature
- Adjudicated media sources
- Independent data sources (e.g., user facility forums or data sharing platforms, vendor data sharing platforms, industry groups, government information sharing systems, clinical data registries)

2. Information Review

Triage incoming information, addressing safety-critical items in a time frame commensurate to the risk posed and in compliance with legal/regulatory requirements, and review remaining collected information at a predetermined frequency with the following objectives:

- Identify necessary changes to existing residual impact assessments of bias source/ type/sequence combinations (i.e., impact level, likelihood of occurrence, likelihood of impact).
- Identify new bias source/type/sequence combinations.
- Identify new uses or foreseeable misuses of the system that may drive new or changed bias source/type/sequence combinations.
- Identify changes to the stated benefits of the system.
- Identify changes in what may be considered positive, negative, or neutral bias.
- Identify changes in the criteria for determining the acceptability of bias at both a bias source/type/sequence combination level and a system level.
- Identify new bias mitigations for existing bias source/type/sequence combinations.

3. Action

   If the review of information dictates some action be taken (e.g., new or changed bias source/type/sequence combinations, estimates, or evaluations), return to the top of the bias mitigation workflow incorporating the organization's design change process.

4. Conclusion

Currently, medical devices that use AI-enabled algorithms utilize machine learning (ML) as a mechanism to "learn" during the algorithm development process. Being deligient about the identification and mitigation of unintended biases is important as a guard against bias impacting certain patient populations and resulting inequities in healthcare delivery. Unless bias is circumvented, resulting reports of inequity may lower trust in the output of an AI-enabled medical device and create a barrier to adoption of ML technology in healthcare.

Unintended bias in AI-enabled healthcare applications (including medical devices) will continue to be an evolving topic that developers, regulators, and experts must continue to address. This paper leverages existing risk management and related processes and proposes a framework for bias opportunity detection, assessment, and mitigation of unintended bias.

.

Citations

[1]	ISO/TC210 (Application of Risk Management to Medical Devices Working Group), "ANSI/AAMI/ISO 14971:2019," Association for the Advancement of Medical Instrumentation, 2019.
[2]	International Medical Device Regulators Forum, "Software as a Medical Device (SaMD): Key Definitions," IMDRF, 2013.
[3]	AAMI Medical Device Software Working Group, "AAMI/ISO TIR80002-2:2017 (Medical device software - Part 2: Validation of software for medical device quality systems," Association for the Advancement of Medical Instrumentation, 2017.
[4]	U.S. Food and Drug Administration, "Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD))," U.S. FDA, 2019.
[5]	IEC/TC56 and ISO/TC262 (Risk Management), "IEC 31010:2019," International Electrotechnical Commission, 2019.

Appendix I: Types of Bias

Discussion

In developing this paper, the authors performed a literature review to develop a list of different types of bias as a reference tool. It should be noted that there are two broad categories of bias. Data bias can occur within the data supply chain when the data used for training the device algorithm does not reflect the population or when incomplete data is used to train the AI model (Sunarti et al., 2020). Cognitive bias is when variations in decisions and judgments form a pattern that may not align with rational decision-making processes and could introduce errors (Azzopardi, 2021).

Please note this table was developed with references to 24027 when it was still in draft, but information sourced was removed from the published version of 24027. The information referenced is pertinent and of significant value to this whitepaper and so it is referred to as 24027 (draft).

Table 1: Data Bias and Cognitive Bias Types

Table 1A Data Bias Type	Description	Source
Aggregation Bias	Arises during model construction, when distinct populations are inappropriately combined. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.	SC42 24027 (draft)
Algorithmic Bias	Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm.	Survey of Bias & Fairness: Mehrabi et al., 2019 Fu et al., 2020
Algorithmic Focus Bias	Deliberate non-use of certain information (e.g.,. anonymizing data wipes out relevant information).)	Survey of Bias & Fairness
Confounding Variables	Some factors influence dependent variables, some factors influence independent variables, and confounding variables impact both, which leads to a relationship between independent and dependent variables, which is false.	SC42 24027 (draft)
Content Production Bias	Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contentcontents generated by users [99]. An example of this type of bias can be seen in [97] where the differences in use of language across different gender and age groups is discussed. The differences in use of language can also be seen across and within countries and populations.	Survey of Bias & Fairness
Data Aggregation	Aggregating data sets that have different distributions can result in bias in the system.	SC42 24027 (draft)

Table 1A Data Bias Type	Description	Source
Data Labelling Bias	The chosen labels and labelinglabelling process can introduce bias by not properly representing the variety of data being modeledmodelled. This can also be a result of bias in the person performing the labeling.	SC42 24027 (draft)
Data Processing	The data processing activities may introduce bias – which values to include or exclude may be based on rules that had cognitive bias in them.	Srinivasan & de Boer, (2020)
Deployment Bias	Occurs after model deployment, when a system is used or interpreted in inappropriate ways.	SC42 24027 (draft)
Distributed Training	Collecting data from different sources may lead to variations between one collection point and another, which can introduce bias.	SC42 24027 (draft)
Emergent Bias	Emergent bias happens as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually sometimesome time after the completion of design [46]. This type of bias is more likely to be observed in user interfaces, since interfaces tend to reflect the capacities, characteristics, and habits of prospective users by design [46]. This type of bias can itself be divided into more subtypes, as discussed in detail in [46].	Survey of Bias & Fairness
Evaluation Bias	Occurs during model iteration and evaluation. It can arise when the testing or external benchmark populations do not equally represent the various parts of the use population. Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.	SC42 24027 (draft)
Gender Bias	An unintended but systematic neglect of either women or men;, stereotyped preconceptions about the health, behavior, experiences, needs, wishes, etc.,. of men and women; or and neglect of gender issues relevant to the topic of interest.	Hamberg, 2008
Historical Bias	Arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.	SC42 24027 (draft)

Table 1A Data Bias Type	Description	Source
Interpretation Error	Due to the layers of complexity in deep neural networks used in some AI, there can be errors in the interpretation of data [BJM]. Source 7 added to Sources tab.	?
Linking Bias	Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users [99].	Survey of Bias & Fairness
Longitudinal Data Fallacy.	Observational studies often treat cross-sectional data as if it were longitudinal, which may create biases due to Simpson's paradox. As an example, analysis of bulk Reddit data [9] revealed that comment length decreased over time on average. However, bulk data represented a cross-sectional snapshot of the population, which in reality contained different cohorts who joined Reddit in different years. When data was disaggregated by cohorts, the comment length within each cohort was found to increase over time.	Survey of Bias & Fairness
Measurement Bias	Arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance.	SC42 24027 (draft)
Missing Features and Labels	Data may be incomplete, and the missing data may not be random. There may be underlying reasons for missing information, and this can lead to bias.	SC42 24027 (draft)
Non-normality	An assumption is made that there is a normal distribution, but there is not.	SC42 24027 (draft)
Non-representative Sample	{24027 does have text for this, but it seems repetitive with sampling bias in general.}	SC42 24027 (draft)

Table 1A Data Bias Type	Description	Source
Omitted Variable Bias	Omitted variable bias occurs when one or more important variables are left out of the model. An example for this case would be when someone designs a model to predict, with relatively high accuracy, the annual percentage rate at which customers will stop subscribing to a service, but soon observes that the majority of users are canceling their subscription without receiving any warning from the designed model. Now imagine that the reason for canceling the subscriptions is the appearance of a new strong competitor in the market offeringwhich offers the same solution, but for half the price. The appearance of the competitor was something that the model was not ready for; therefore, it is considered to be an omitted variable.	Survey of Bias & Fairness
Other Sources - Noise	Noise in the data can affect product performance.	SC42 24027 (draft)
Other Sources - Outliers	Extreme data values that are low probability, if captured in the data set, can lead to an over-representation of the probability of that occurring again.	SC42 24027 (draft)
Popularity Bias	Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation — —for example, by fake reviews or social bots [96]. As an instance, this type of bias can be seen in search engines [61, 96] or recommendation systems where popular objects would be presented more to the public. But this presentation may not be a result of good quality; instead, it may be due to other biased factors.	Survey of Bias & Fairness
Population Bias	Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population. An example of this type of bias can arise from different user demographics on different social platforms, such as women being more likely to use Pinterest, Facebook, Instagram, while men arebeing more active in online forums like Reddit or Twitter.	Survey of Bias & Fairness
Representation Bias	Arises while defining and sampling a development population. It occurs when the development population underrepresentsunder-represents, and	SC42 24027 (draft)

Table 1A Data Bias Type	Description	Source
	subsequently fails to generalize well, for some part of the use population.	
Sampling Bias	Sampling bias arises due to non-random sampling of subgroups. As a consequence of sampling bias, the trends estimated for one population may not generalize to data collected from a new population.	Survey of Bias & Fairness
Selection Bias - Coverage Bias	The data is chosen in a way that is not representative of the target population. The sample population does not match the target population (e.g.,. not all user groups are sufficiently represented.)	SC42 24027 (draft)
Selection Bias - - Non-response Bias	The data is chosen in a way that is not representative of the target population. Many times, data collection is on a voluntary basis, however, there are some populations that often will not participate, resulting in an underrepresentationunder-representation of those populations.	SC42 24027 (draft)
Selection Bias - Sampling Bias	The data is chosen in a way that is not representative of the target population. The data being collected isn't collected randomly, and therefore doesn't represent the population.	SC42 24027 (draft)
Self-Selection Bias	Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves. An example of this type of bias can be observed in situations where survey takers decide that they can appropriately participate in a study themselves. For instance, in a survey study about smart or successful students, some less successful students might think that they are successful andto take the survey — — which would then bias the outcome of the analysis. In fact, the chances of this situation happening areis high, as the more successful students probably would not spend time filling out surveys that would increase the risk of self-selection bias.	Survey of Bias & Fairness
Simpson's Paradox	Pooled data produces a mean that masks biases in subsets of the data. Attributes about the entire population, on average, areis different than the attributesattribute for a specific subset.	SC42 24027 (draft)

Table 1A Data Bias Type	Description	Source
Temporal Bias	Temporal bias arises from differences in populations and behaviors over time. An example can be observed in Twitter where people talking about a particular topic start using a hashtag at some point to capture attention, then continue the discussion about the event without using the hashtag [99, 120].	Survey of Bias & Fairness
Training Data Bias	Model deviates from actual population statistics because training data is biased in some way. An example would be self-driving cars learning regional norms.	Survey of Bias & Fairness
Transfer Context Bias	The context of use is different from what it was trained on. Example of self-driving car on wrong side of the road, example of research hospital patients vs. rural clinic.	Survey of Bias & Fairness

Table 1B **Cognitive Bias Type**	**Description**	**Source**
Automation Bias	Sometimes people trust software systems when they shouldn't; this lack of critical thinking can impact product performance. Errors of automation bias tend to occur when decision-making is dependent on computers or other automated aids and the human is in an observatory role but able to make decisions. Examples of automation bias range from urgent matters like flying a plane on automatic pilot to such mundane matters as the use of spell-checking programs. As we work to automate decisions to reduce human error there is the potential to create automation bias.	SC42 24027 (draft)
Behavioral Bias	Behavioral bias arises from different user behavior across platforms, contexts, or different datasets [99]. An example of this type of bias can be observed in [88], where authors show how differences in emoji representations among platforms can result in different reactions and behavior from people and sometimes even leadleading to communication errors.	Survey of Bias & Fairness

Table 1B Cognitive Bias Type	Description	Source
Cause-Effect Bias	Cause-effect bias can happen as a result of the fallacy that correlation implies causation. An example of this type of bias can be observed in a situation where a data analyst in a company wants to analyze how successful a new loyalty program is. The analyst sees that customers who signed up for the loyalty program are spending more money in the company's e-commerce store than those who did not. It willis going to be problematic if the analyst immediately jumps to the conclusion that the loyalty program is successful, since it might be the case that only more committed or loyal customers, who maymight have planned to spend more money anyway, are interested in the loyalty program in the first place. This type of bias can have serious consequences due to its nature and the roles it can play in sensitive decision-making policies.	Survey of Bias & Fairness
Confirmation Bias	This is a type of implicit bias where data could be collected or labeledlabelled in a way that confirms assumptions that the human has. It is the inclination to interpret information in a way that aligns with an individual'sindividuals beliefs.	SC42 24027 (draft) Matey et al., 2021
Experimenter's Bias	The humanhumans involved makes an assumption based on their experience, but this assumption is not true for all circumstances. The user continues to train the model until it agrees with the user's point of view.	SC42 24027 (draft)
Funding Bias	Funding bias arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study. As an example, this manifests when employees of a company report biased results in their data and statistics in order to keep the funding agencies or other parties satisfied.	Survey of Bias & FairnessFairness
Group Attribution Bias	People sometimes assume that what is true for one is true for everything in that group, ignoring differences between individuals and/or cultures. In addition, groups are biased toward attributing their success to factors that are internal to their group.	,SC42 24027 (draft) Yanbo, 2020 Goncalo & Duguid (2008)

Table 1B Cognitive Bias Type	Description	Source
Implicit Bias	Implicit bias reflects an individual's true thinking but not a shared opinion. The humans involved make an assumption based on their experience, but this assumption is not true for all circumstances.	SC42 24027 (draft) Bullard, 2020 Lin et al., 2020
In-Group Bias	The user favors characteristics of his or her groups, family, etc.	SC42 24027 (draft)
Interpretation Bias	Misinterpretation of outputs by user.	Survey of Bias & FairnessFairness
Latent Bias	Just as latent errors are generally described as errors "waiting to happen,"" in complex systems, latent biases are biases waiting to happen.	DeCamp & Lindvall, 2020
Observer Bias	Observer bias happens when researchers subconsciously project their expectations onto the research. Inaccuracy in observer feedback can produce inaccurate data influencing results in an unintended direction. This type of bias can happen when researchers (unintentionally) influence participants (during interviews and surveys) or when they are selective in identifying participants or statistics that will favor their research.	Matey et al,l, 2021
Out-Group Homogeneity Bias	The users assumes that things outside of a group are similar with each other.	SC42 24027 (draft)
Presentation Bias.	Presentation bias is a result of how information is presented [8]. For example, on the Web, users can only click on content that they see, so the seen content gets clicks, while everything else gets no click. And it could be the case that the user does not see all the information on the Web [8].	Survey of Bias & FairnessFairness
Ranking Bias.	The idea that top-ranked results are the most relevant and important will result in attraction of more clicks than others. This bias affects search engines [8] and crowdsourcing applications [78].	Survey of Bias & FairnessFairness

Table 1B Cognitive Bias Type	Description	Source
Social Bias	Social bias happens when other people's actions or content coming from them affect our judgment. [8]. An example of this type of bias can be a case where we want to rate or review an item with a low score, but when influenced by other high ratings, we change our scoring thinking that perhaps we are being too harsh [8, 125].	Survey of Bias & FairnessFairness
Societal Bias	Bias that is shared by many individuals – societal beliefs are the drivers of this type of bias.	SC42 24027 (draft)
What You See Is All There Is Bias	User looks for information that confirms their beliefs.	SC42 24027 (draft)

Appendix II: Review of Risk Management Process

This paper proposes to use a modified version of the risk management process flow found in the ISO/IEC 14971:2019 “Application of risk management to medical devices” standard. Some readers of this paper might not be familiar with that standard; the purpose of this annex is to provide a brief overview of that standard. It should be noted that there is also a companion document, ISO/IEC 24971:2020 “Guidance on the Application of ISO 14971” which provides additional insight and suggestions to ensure good risk management practices.

Risk Management spans the entire product development life cycle – starting with the planning phase, all the way through product launch and post-market support. Figure 2 shows the major process steps as outlined in ISO/IEC 14971.

The following is a brief summary of the major process steps in risk management. Other requirements such as management responsibilities, the need for competent personnel, documenting deliverables in a risk management file, etc., will not be covered in this annex.

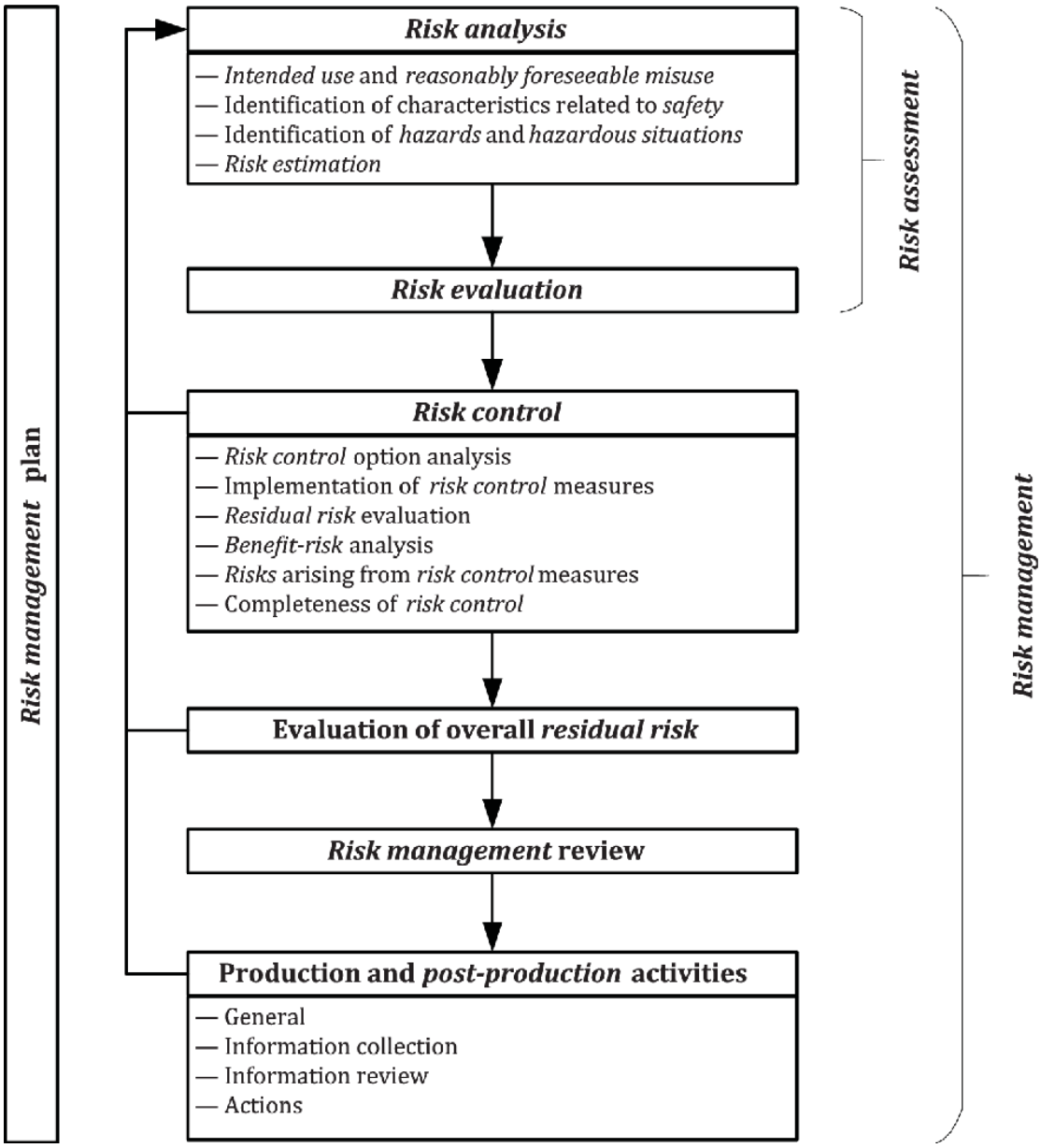

Figure 2

Risk management plan

There should be a risk management plan that includes the scope of risk management activities, assignment of responsibilities, requirements for review of the activities, risk acceptability criteria, method to evaluate overall residual risk, activities of the

implementation and effectiveness of the risk control measures, and activities to collect and review post-production information. For systems that continue to learn over time, the plan should include activities that support the algorithm change activities.

Risk Analysis

Intended use and reasonably foreseeable misuse: To be able to evaluate the risks associated with a product, it is necessary to know what the product is intended to do. Additionally, medical devices are often misused, so the risk management process requires the development team to consider "reasonably foreseeable misuse."

Identification of characteristics related to safety: Certain elements of a product design can influence the safety of a medical device. For instance, a medical device that is designed to operate using A/C power has different safety concerns than a battery-powered device or a purely mechanical device.

Identification of hazards and hazardous situations: To ensure a thorough analysis, the 14971 standard does not immediately ask the question, "How can this hurt people?" Instead, it breaks the process down into multiple steps.

14971 defines harm as "injury or damage to the health of people, or damage to property or the environment," defines hazard as a "potential source of harm," and defines a hazardous situation as "circumstance in which people, property, or the environment is/are exposed to one or more hazards."

For example, just because a device operates on A/C (hazard) does not mean the patient will get an electric shock (harm). Even if there was line voltage appearing on electrodes (a hazardous situation), this doesn't always mean the patient will get an electric shock. (It might be likely, but it is not always the case.)

Examples of hazards include electrical hazards, biological hazards, chemical hazards, etc. Hazardous situations would depend on the intended use of the product. For example, an infusion pump may have a hazardous situation if it delivers too much medication, and a different hazardous situation if it delivers too little.

Risk estimation: This step provides an estimation of the risks associated with the product, including both the probability and severity of the harm.

Risk evaluation: The manufacturer should evaluate the risks to determine if the risk is acceptable or not, and implement risk controls to reduce the risk, where appropriate. Annex C of ISO/TR 24971 [7] provides additional guidance for risk acceptability and risk evaluation.

Risk control: After analyzing different hazards, hazardous situations, and harm, developers will likely need to implement some sort of risk control to reduce that particular risk. For example, high electrical power may warrant double-insulation as a risk

control. The risk controls that are put into place will need to be tested to ensure they are effective, and ensure that they are not introducing new risks.

Evaluation of overall residual risk: Not every risk can be completely eliminated, so it's important to understand any "residual risk" after the risk control measures are implemented and their effectiveness verified. A product may have controlled individual risks to an acceptable level, but the overall product might not be acceptable.

Risk management review: Before product launch, there needs to be a review to ensure that all of the process steps were followed and the benefits of the products outweigh the risks.

Production and post-production activities: Things change over time – design or supplier changes, bug fixes, creative users that use the product in unexpected ways, etc. Therefore, monitoring the performance of medical devices over time is an important aspect of ensuring the continued safety of the product.