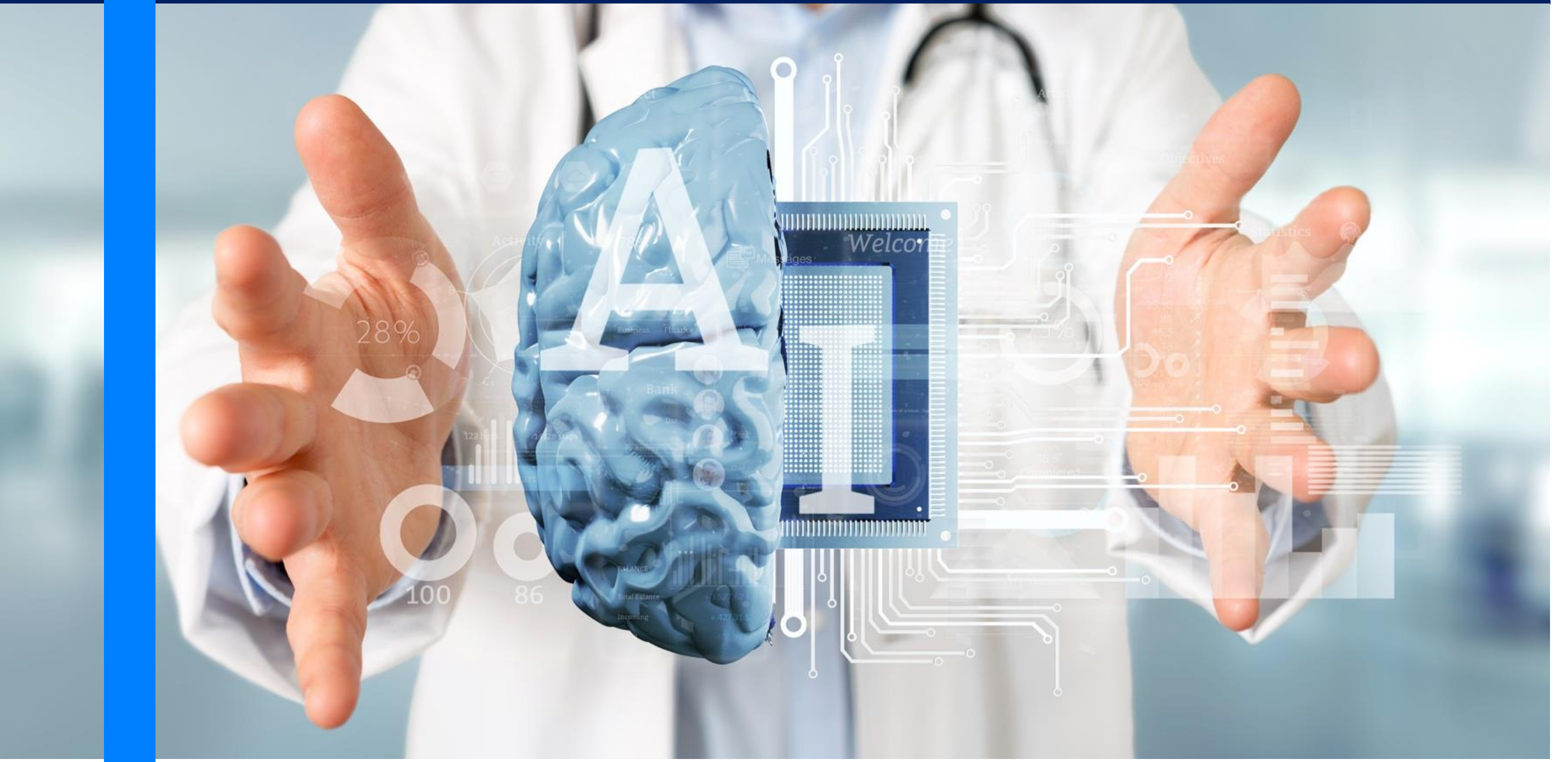
Healthcare & Public Health
Sector Coordinating Councils
PUBLIC PRIVATE PARTNERSHIP

Secure Medtech

Manage Risks

Health Industry Cybersecurity -

Artificial Intelligence & Machine Learning (HIC-AIM)



FEBRUARY 2023

Table of Contents

About the Health Sector Coordinating Council Cybersecurity Working Group

The Healthcare and Public Health Sector Coordinating Council (HSCC) is a coalition of private-sector critical healthcare infrastructure entities organized to partner with and advise gealth sector entities and the government in the identification and mitigation of strategic threats and vulnerabilities facing the sector's ability to deliver services and assets to the public. The HSCC Cybersecurity Working Group (CWG) is a standing working group of the HSCC, composed of more than 370 healthcare and related industry organizations working together to develop strategies to address emerging and ongoing cybersecurity challenges to the health sector.

For more information about your healthcare organization joining the HSCC, please visit https://healthsectorcouncil.org/contact/.

Introduction

Healthcare has continued to evolve from the paper-and-pen world to a digital environment. The opportunities for high-quality, safe and effective care have increased exponentially with this change. Integral to these opportunities is the harnessing of increasing computer power and the revolutionary impact of artificial intelligence (AI) and machine learning (ML). AI/ML could impact every aspect of healthcare, from diagnosis, treatment decisions, predictive analysis, and even administrative functions such as coding and billing.

The promise of AI/ML, however, comes at a price: artificial intelligence systems, especially those dependent on machine learning (ML), can be vulnerable to intentional attacks that involve evasion, data poisoning, model replication, and exploitation of traditional software flaws to deceive, manipulate, compromise, and render them ineffective. Yet too many organizations adopting AI/ML systems are unaware of their vulnerabilities. This potential outcome is the basis of this whitepaper.

It is important for all involved in healthcare technology, including individual providers and administrative and clinical leaders in large healthcare systems to understand the background cyber risks of AI/ML so they are prepared to utilize these advances in the safest way for our communities. In addition, manufacturers that build devices that will use AI/ML will need to understand the downstream impact of flaws or vulnerabilities in their software or security protection.

We envision soon that AI/ML systems may be subject to regulation or cyber security standards or guidelines that healthcare delivery organizations and vendors must be aware of:

- As of this paper's publication, NIST was still developing a Risk Management Framework (RMF) for AI/ML
- In March 2022, Congress introduced the Algorithmic Accountability Act of 2022. The Act requires companies to complete algorithmic impact assessments that provide key details around a given algorithmic system and directs the FTC to create regulations and hire additional staff to enforce them.

What is AI and ML?

The American National Standards Institute (ANSI) defines Artificial Intelligence as:

"(1) A branch of computer science devoted to developing data processing systems that performs functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.

“(2) The capability of a device to perform functions that are normally associated with human intelligence such as reasoning, learning, and self-improvement."

Intelligence, while defined in various ways, is generally the ability to perceive or infer information, to retain this information as knowledge, and to be able to use this knowledge to adapt behaviors to apply to problem sets within a given environment.”

AI is still functionally a toolbox of diverse algorithms and strategies that mimic intelligence in order to solve problems. AI relies on strategies that include formal logic, large databases of information to imitate animal behavior, and model human problem solving.

Machine Learning is a subset of the above attempts, but of all approaches so far, it has been arguably the most successful and has dominated the field. Prominent among machine learning approaches are Deep Learning Neural Networks and Support Vector machines, which are algorithms that seek to model the function of the human brain.

Other Industries Use of AI

AI/ML has gradually become ubiquitous in various forms throughout many industries. There are many lessons from the benefits and risks of AI/ML technology outside of healthcare. In 2009, the potential benefits to medicine from advances in computer science and technology in other industries seemed clear, but according to the report of the President’s Council of Advisers on Science and Technology (PCAST) in 2010, IT and particularly interoperability in US Healthcare were far from the levels of sophistication achieved in many other industries, which include finance, markets, e-commerce, travel planning, navigation, fraud detection, and marketing. Early successful examples of use of AI tools in these industries either (a) included things that a human could do but a machine could do more routinely and, in some cases, better, or (b) provided smoother and more efficient human-machine interaction. Examples include machine learning and advanced data mining, robotics, voice and fingerprint and face recognition for biosecurity, voice control of devices, Internet querying, natural language processing, computer vision and image analysis, and various planning applications.

Actual and Potential uses of AI and ML within Healthcare

In the broadest terms, AI and ML software is most useful when processing large amounts rapidly growing and changing data by helping identify and link suspicious actions.

AI/ML software has both direct and indirect applications in the patient care environment involving minimally invasive surgical procedures, diagnosis verification of radiological imaging, and disaster planning and testing processes.

AI and ML software can also be used for healthcare cybersecurity by analyzing the large volume of enterprise O/S, Application, and Database events in conjunction with network flow analytics to help identify potential insider threats or to better detect suspicious lateral movements within a healthcare network.

What are the general risks for utilizing AI/ML in healthcare that can impact patients?

A significant body of technology literatures and articles promotes an assumption across healthcare and other industries that the use of intelligent systems, including ones that use Artificial Intelligence and/or Machine Learning, is will solves problems and automate functions within the healthcare environment. They see results in one area, and associate other potential applications with another area. However, leveraging AI across applications is not simple. These systems are complex, use a multitude of data points, and use algorithms customized for use cases. Transferring to another use case, specifically with aligning with operational processes, is a significant challenge.

Likewise, ethics, including transparency, communication, systems design, and bias, are also challenges. Frameworks for traditional software development and design do not appropriately capture what is needed for designing ethically aligned AI systems. This means that many of the engineers and architects behind them may not understand what is required to design systems, not just technologies that facilitate transparency, communication, bias, and appropriate system usage.

Risks of AI/ML:

- Lack of re-use in AI/ML use cases
- Ethics and Transparency
- Vulnerability management
- Loss of Privacy

In the healthcare space, AI is increasingly used to augment and support the medical decisioning process. As noted in our Concerns below, this use may or may not be subject to FDA regulation in all cases. As Rahman Ladak, Aly Muhammad Ladak, and Imran Ladak indicate in their paper, “The Role of the Physician in Using AI to Enhance Equitable Patient Care”, it is a novel, opaque technology that uses personal information to arrive at decisions that affect patient care. As noted by the Stanford Cyber Policy Research Center, “All modern information systems have vulnerabilities, and AI/ML systems are no different. Indeed, the algorithms and techniques underlying ML systems have weaknesses with no known fixes. The goal, therefore, is risk reduction, not elimination. The urgent effort to build more resilient AI-based systems involves many strategies, both technological and managerial, including sometimes the decision not to deploy AI at all in a highly risky context.”

Beyond biases or vulnerabilities in the AI/ML processes, protecting patient data from re-identification is also a specific threat. Large data sets in aggregate can be used to quickly re-identify patients using numerous data sources to correlate. One of the methods organizations can use to protect patient data is Federated Learning. According to Google Research, it is the use of machine learning to train a high-quality model with training data distributed over a large number of clients. According to Nicola Rieke of Nvidia, it can be used to facilitate collaborative model development without direct data sharing.

One of the other major risks is that AI/ML algorithms will not be exposed to enough data to properly identify specific conditions that will be representative of the environment. The data may not also incorporate samples from patients of different genders, ages, demographics, and environments. Rieke discusses how individual organizations maintain archives of hundreds of thousands of records; however, health data privacy, patient consent, and ethical approval often prevent sharing or training models on multiple sets. This is further complicated by the quality of the data, especially in electronic health records.

An expert determination is one facet of identifying risk. Health systems, according to the HIPAA Security Rule, need appropriate risk analysis completed. However, they may not have effective means of identifying when security incidents occur in AI/ML systems, especially if the system is "closed" or proprietary. While there has been improvement in this area, most organizations have reported that they cannot easily define their inventory of new technologies.

AI/ML Concerns

This paper outlines nine areas of concern and offers suggested approaches to address each one. While the authors have to tried to capture as many relevant points as possible, there have been numerous additional efforts globally to address cyber and non-cyber related concerns of AI/ML. HSCC will continue to provide additional information relevant to industry leadership as they become aware of it.

AIML Concern #1: Organizational Outcomes and Expectation for Performance, Quality, and Precision not clearly articulated

Security Principal Impact Confidentiality, Integrity, Availability	**Healthcare Function Impact** Patient Safety, Continuity of Patient Care, Privacy, Quality	**Operational Impact** Visibility, Control, Resiliency
Business Function Operational, Patient Care	**Stakeholder Function** End User, Administrator, Engineer, Auditor, SecOps, Manager	**Regulatory Guidance** ISO 14155 Clinical Investigations for Medical Devices FDA - Content of Premarket Submissions for Management of Cybersecurity in Medical Devices (Document issued on October 18, 2018.) FDA - Guidance for Industry Cybersecurity for Networked Medical Devices Containing Off the-Shelf (OTS) Software (Document issued on: January 14, 2005) FDA – Post market Management of Cybersecurity in Medical Devices Guidance for Industry and Food and Drug Administration Staff (Document issued on December 28, 2016). ISO 13485 Quality Management Systems ISO14971 updates for AI/ML TIR97 Principles for Medical Device Security - Post Market Risk Management for Device Manufacturers

Description

Much of the success with AI/ML is determined up front, long before the technology is put into production. The organization using AI/ML technology must clearly establish technology objectives for vendors. Specifically, what is the expected performance, precision, margin of error, defined outputs, etc., and how is that tied to healthcare quality in the organization's delivery of services? Failure to clearly define this can lead to outputs that aren't useful, have adverse effects on decision making, and undermine confidence in the technology.

What Might be the Threat Sources?

- Poor quality can lead to a poorly defended system
- More general implementation threats
- May be difficult to identify a compromised system if the quality/performance criteria have not been clearly outlined and tested.
- Hard to identify causes of issues if performance metrics have not been clearly identified and a degradation is identified, such as identifying issue that was caused by adversarial attack or drift over time.
- Data of unknown provenance; who owns and validates this data? "Ground truth" data that are being transmitted, stored, and analyzed should be secured and protected.
- Not keeping up with patches, model, and firmware updates, poor post-market management and oversight.
- End of life and support of various products and/or components in the system.
- Procuring or developing a system that lacks the transparency and/or system controls that adequately support quality monitoring.

What are the Potential Impact Concerns?

- AI output is typically binary and does not provide as much information as a human pathologist and poses a risk for abandonment (precision versus recall).
- Lack of ground truth data with respect to what is being captured in terms of identifying attributes (e.g., age, gender, race, etc.) and the impact it has on the AI/ML outcomes for precision and accuracy.
- De-anonymization or de-identification of data makes it challenging to understand what the technology is doing and can lead to more bias and differences in the patient treatment plan or workflow.
- Explain what the AI doesn't do (limitations of the tool) and potential bias in algorithm.
- Data transitions where the algorithm may not have been trained on is at risk of potential underfitting/overfitting or drift. (Trishan Panch, Heather Mattie & Leo Anthony Celi, 2019)

What to Do?

Data Plan - Acquisition

Clinical Research and the start of the Data Lifecycle

Clinical Research with respect to medical device development and AI have identified some gaps. The data that are currently supported by clinical investigations have not evolved as quickly as the software technology; thereby potentially missing data that would be part of the software's performance safety and effectiveness. *(Trishan Panch, Heather Mattie & Leo Anthony Celi, 2019)* [1].

[1] https://www.nature.com/articles/s41586-021-03430-5

In addition, patient privacy and the 'consent' of information for the clinical trial or investigation can have its own challenges with respect to the protocols needed for ensuring statistical significance safety and efficacy measures. There is a Cybersecurity Privacy Risk Management Framework which includes some criteria with the types of data. (NIST 800-60: Information Type Categorization). Performance requirements to ensure safety and effectiveness should not be impacted by the de-identification or potential privacy concerns and the associated impact on the data used to test the software, as that would impair the ability to accurately determine the performance. *(Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping & James Zou, 2021*[2]

AI/ML Education

- Implementing education and training on AI might be considered for physicians on the interpretation of results (requirements for attaining a license and/or certifications and continuing education credits); education and training should be a shared responsibility between HDOs and AI technology companies and not just a sole burden on the healthcare professional. This type of training can be used as professional credits towards certification. Training enables users to be familiar with the software and AI algorithm to assist with detectability if the algorithm does not perform as intended. This will need to be updated on a regular basis as technology matures.
- Working with trusted parties like RSNA, AMA and AAFP (calls for participants to help evaluate systems, but no certifications or training). *(RSNA Imaging AI certificate)* [3]

ISO13485 Requirement for MDMs – Quality Management System (QMS)

- Training is needed for all project stakeholders that are building, supporting, or servicing the device.
  - System providers may provide the training; this includes the servicers who are also required to take the training that would lead to a certification. The intent is to have a certain criterion that would need to be met in regard to competency, or receive a lower evaluation.
  - Training program requirements (ISO13485); general understanding for all but specific training may be required for data evaluation, measurement, and transitions and/or other roles (qualification and competency).
  - IT Resource constraints and various skill sets within AI/ML domain (data curation/mining/analysis). Defined criteria that are easily measurable for ease of evaluation. *(Haiyan Zhang, Ph.D., Sheri Feinzig, Ph.D., Louise Raisbeck, and Iain McCombe, 2019)*[4]

Data Plan – Data and Requirements Management

FDA pre- and post-market cybersecurity requirement updates

Cybersecurity Premarket – Design

- Design Lifecycle requirements -- Cybersecurity Program Maturity of the MDM should be assessed and measured (ex. scoring like CMMI) high-risk devices to be evaluated against more stringent criteria and clinical significance to be compared against RWE. Some cybersecurity guidance to take into consideration:
  - IEC/TR 80001-2-2:2012 -- Application of risk management for IT-networks incorporating medical devices — Part 2-2: Guidance for the communication of medical device security needs, risks and controls
  - NIST SP 800-53 Rev. 5 -- Security and Privacy Controls for Information Systems and Organizations

[2] Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping & James Zou, evaluating eligibility criteria of oncology trials using real-world data and AI, (2021) https://www.nature.com/articles/s41586-021-03430-5

[3] https://www.rsna.org/ai-certificate

[4] Haiyan Zhang, Ph.D., Sheri Feinzig, Ph.D., Louise Raisbeck, and Iain McCombe, Contributors: Nigel Guenole, Ph.D., Jenny Montalto, Kimberley Messer: The role of AI in mitigating bias to enhance diversity and inclusion, (2019) https://www.ibm.com/downloads/cas/2DZELQ4O

  - ISO/IEC 27002:2013 -- Information technology — Security techniques — Code of practice for information security controls
- Backwards compatibility of old hardware requirement (evaluated on performance against effectiveness); evidence of cybersecure practices
- Cybersecurity trace logging/auditing events for investigation; threat model, data flow diagram, and Security Risk Management file should be included into the design of the product. Example: privacy requirements related to unauthorized access or unauthorized use should be logged.

Data Collection and Training

Cybersecurity Premarket -- Testing

- Verification and Validation activities should be traceable not only to requirements but to third party components or systems such as with a Software Bill of Materials (SBOM).
- Episodic testing, regression, and independent third-party penetration testing should be conducted at a minimum frequency of quarterly, which would coincide with vulnerability monitoring and patch management.
- Variables that added to the training set or might get unexpected results (binary comparisons are easier but typically requires review.

Cybersecurity Premarket - Risk Management

- Acceptance of the technology and collaboration for the use case and expected outputs; clarifications around functionality, data models/validation, translation of the algorithm or expected output for more advanced technologies that would be part of the patient workflow/treatment plan. User trust/over-trust and explain ability is required to assist with the translation. *(Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek, 2020)* [5]
  - Cybersecurity Risk Management -- Data flow system diagram, threat modeling, and attack trees can identify areas that are vulnerable to attack and cybersecurity controls for mitigation.
  - Risk-based approach for software classification of AI software tools; additional cybersecurity documentation required for higher classification of the software (ex. High risk AI/ML must include Data Flow Diagram, Threat Model, Security Risk Assessment, FMEA, and Traceability matrix to V&V including statistical analysis based on intended use and indications for use against the evaluation of training set, etc…).
- Confidence intervals and specific performance expectations should be agreed upon between both the MDM and HDO. Different context of explain ability for various user groups (deep learning algorithm and/or output). (Alex J. DeGrave, Joseph D. Janizek & Su-In Lee, 2021) [6]

Data Cleaning and Evaluation

Cybersecurity Premarket -- Labeling

- Detailed "Instructions For Use (IFU)" on the AI/ML and what was used to train and what the overall residual risks may be
- Evaluation cards for repeatable performance reviews and to document the reported results
- Additional requirements for MDM on labeling requirements for installation/configuration/admin guides while still protecting their IP

[5] Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek, 2020 The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies

[6] Alex J. DeGrave, Joseph D. Janizek & Su-In Lee, AI for radiographic COVID-19 detection selects shortcuts over signal (2021)

- Cybersecurity Supplier Management controls -- SBOM for tracking third-party component hardware/firmware, data and validation this includes third-party Cloud providers
- Labeling requirement (ex. target population, training set characterization, overall residual risks based on indication for use); data sheets that characterize data set, model cards and results after training – parameter and thresholds, algorithm itself for training, purpose, etc.
    - Customized AI/ML Manufacturer Disclosure Statement for Medical Device Security (MDS2) Form -- Addendum for AI/ML in MDS2? [7]

FDA - Submission Requirements

- Regulators may establish metrics to evaluate performance (precision and accuracy); better differentiator of 1% between successes when measuring; tested using data that was not part of the training set (data set split – training, tune the model training/testing, hold out set – evaluation) risk overfitting the model (Cloud Computing for evaluations)
- Regulators should consider establishing the criteria for what types of labeling requirements or additional IFU requirements for safe and effective use and purchasing decisions. *(FDA Virtual Public Workshop – Transparency of Artificial Intelligence/Machine Learning-enabled Medical Devices OCTOBER 14, 2021* [8]
    - When conducting an evaluation on the AI/ML, criteria for success need to be defined. Does the evaluation result in the expected outcome and cover what is expected? What have we not considered? Evaluation criteria that are not strictly performance related such as bias or other factors. Specifics on the populations that it's being tested on (example – Type 1 or Type 2 diabetes); characterization of datasets? Risk is not having the expert or knowledgeable SME as opposed to just having the data.
    - Ensure that the training data that are used for AI/ML pull out 15-20% of the model against real world evidence for comparison. Producing explainable models will assist with the evaluation of the model *(Gunning, (XAI), 2017)* [9]
- Methods of monitoring performance with respect to the varying types of usage of the AI software and continued adoption ensuring that performance metrics and managing user expectations. *Figure 1 – Sankey plot for AI/ML-based medical devices approved in the USA Data from Jan 1, 2015, to March 31, 2020. AI/ML=artificial intelligence and machine learning. NA=not applicable. (Urs J Muehlematter, Paola Daniore, Kerstin N Vokinger, 2015-2020)*[10]

Cybersecurity Post Market

For real-world evidence comparisons, PMCF studies are designed to identify the potential for residual risks of a CE Marked device, and to collect data and gain clarity regarding the long-term clinical performance of the product *(MDCG 2020-7 Post-market clinical follow-up (PMCF) Plan Template A guide for manufacturers and notified bodies, 2020)* [11].

- This approach can be used to confirm the safety and/or clinical performance for a new indication for use or claim that has been CE-certified (approved).

- Post-Market surveillance measurement and monitoring. Post-production monitoring to include cybersecurity proactive monitoring such as threat intelligence and NVD, can assist with gathering the information from real world

[7] https://www.nema.org/standards/view/manufacturer-disclosure-statement-for-medical-device-security

[8] https://www.fda.gov/medical-devices/workshops-conferences-medical-devices/virtual-public-workshop-transparency-artificial-intelligencemachine-learning-enabled-medical-devices

[9] Gunning (2018). Explainable Artificial Intelligence (XAI), DARPA's Explainable Artificial Intelligence (XAI) Program | AI Magazine (aaai.org)

[10] Urs J Muehlematter, Paola Daniore, Kerstin N Vokinger Approval of artificial intelligence and machine learning based medical devices in the USA and Europe (2015–20): a comparative analysis https://www.researchgate.net/publication/348599977_Approval_of_artificial_intelligence_and_machine_learning-based_medical_devices_in_the_USA_and_Europe_2015-20_a_comparative_analysis

[11] https://ec.europa.eu/health/sites/default/files/md_sector/docs/md_mdcg_2020_7_guidance_pmcf_plan_template_en.pdf

evidence. Comparison of measurements and/or metrics over time; trending and assessment/evaluation and investigation.

- Vulnerability Management – Identified threats
    - Vulnerability monitoring or scanning tools to provide real time monitoring; third-party tools that are available that can automatically determine if the target software is source or a compiled binary, then identifies and catalogs all third-party software components, associated licenses, and known vulnerabilities affecting your applications.
    - Supplier components including software/firmware
        - SBOM -- A software bill of materials is a list of all the open source and third-party components present in a codebase. A software BOM also lists the licenses that govern those components, the versions of the components used in the codebase, and their patch status. *(IMDRF, Principles and Practices for Medical Device Cybersecurity, 2020)* [12]
- **System Support, Services and Cybersecurity and Medical Device Complaint Monitoring**
- Cybersecurity signal detection, evaluation and disposition should be part of the formal complaint process to ensure that security issues do not impact performance or have corrupted the data which may result in the AI not performing as intended.
- Cyber forensics – systems that can support of morbidity and mortality conferences; cybersecurity often not included in scope for investigation (system does not support SIC) alongside interconnected systems.
- Patch management – With the number of products that are deployed at an HDO, there are a myriad of patches that are required to be installed. Those include but not limited to software, firmware, anti-virus, O/S, and other third-party of OTS products. In addition, many patches require user interaction and can often disrupt users' productivity. Recommend that AI/ML software patches be designed for seamless patching or silent install. Furthermore, the methods in which these products/components can support and facilitate updates, patching, interoperability, and rollback should be reviewed as part of the criteria of purchasing.
- Some examples are:
    - Encryption/Virtual Private Network (VPN)/Remote Desktop Protocol (RDP)
    - Pre-Qualification of O/S security patches – minimize downtime, benefits (safety and effectiveness)
    - Can't be routinely patched and requires coordination between many departments
    - If (applicable) where software installs/patches requires onsite servicing (are third-party servicers also trained and able to execute on their responsibilities)

Is the change intended to solely strengthen cybersecurity and has been determined to not have any other impact on the software or device?

Example Scenario

- Clinician evaluation of the AI against criteria (ex. reporting thresholds); usability testing was conducted with users by demonstrating screenshots of the software
- AI triaging useful in prioritization and categorization tool however, when using these technologies in more autonomous decision making for patient treatment, is where more of the challenges are being faced in what the algorithm is tracking within its learning based on the expected/unexpected output.
- Additional information and translation of the AI/ML technology of its intended use, indications for use and training validation set should be more understood to assist with purchasing decisions. Further understanding between performance and explain ability should be communicated.

[12] Principles and Practices for Medical Device Cybersecurity (imdrf.org)

- Changes in the number and demographics of patients being seen at hospital. There is a business that is planning on opening a factory in a town that will employ ~13,000 people. How will this be monitored with respect to the technology's precision/accuracy for the potential changes in the population.

AIML Concern #2: Accountability of Outcomes Undefined

Security Principal Impact Confidentiality, Integrity, Availability	**Healthcare Function Impact** Patient Safety, Continuity of Patient Care, Privacy, Quality	**Operational Impact** Visibility, Control, Resiliency
Business Function Operational, Patient Care	**Stakeholder Function** End User, Administrator, Engineer, Auditor, SecOps, Manager	**Regulatory Guidance** Art. 22 GDPR

Description

AI/ML involves 'training' a 'model' to generate certain outcomes. If the model fails to meet expectations, is there accountability for correcting that? There is a higher dependency upon data scientists, AI developers, and AI consumers to communicate and define where roles and responsibilities lie to correct problems. Failure to establish roles and responsibilities can lead to inaction and delays in handling problems and adjustments to the AI/ML model.

What Might be the Threat Sources?

Threat sources are likely to be internal non-malicious actors (AI developers, system users); however there is some risk that unexpected outcomes are a result of malicious actors.

What are the Potential Impact Concerns?

Depending on the use of AI/ML, there is a potential impact to patient safety. If AIML systems are used to evaluate revenue cycle, undefined outcomes could have a financial impact to the organization..

What to Do?

Relevant approaches to address accountability of outcomes undefined in healthcare systems utilizing machine learning modules include:

Explainable Artificial Intelligence (XAI)

In a healthcare environment where decisions from AI have high-risk consequences, it is important to understand why or how the machine learning model concluded what it did. Understanding this model is instrumental in developing public trust. Explainable Artificial Intelligence (XAI), often referred to as white-box machine learning algorithms, presents outcomes so they are understandable to its developers and users.

XAI is a fairly new technology and has been used for things like COVID-19 tracking, cancer diagnosis, and detection of sepsis. *(Shaban-Nejad et al., 2021)* [13] XAI methods and processes are instrumental in assisting with interpretability and providing transparency to users who rely on AI technology. The XAI concept flow is similar to AI in that is uses a training dataset and follows the machine learning process, but it also employs an explainable model and an explanation interface. Vendors offering healthcare AI technology should be strongly encouraged to provide Explainable AI solutions creating transparency and providing a level of accountability.

Baseline and/or Standards

Create a baseline of learning – No matter who is responsible for providing the data source, it is important to establish a baseline of learning. This native baseline allows you to evaluate the output of the problem and quickly identify any variances. These responsibilities should be clearly defined in any contract terms and further define how out-of-bounds conditions are addressed.

NIST Standards - Under Executive Order 13859, federal agencies have been directed to ensure that “technical standards minimize vulnerability to attacks from malicious actors and reflect federal priorities for innovation, public trust and public confidence in systems that use AI technologies” and to “develop international standards to promote and protect those priorities.” Under this initiative, the National Institute of Standards and Technology currently is working to establish a risk management framework for those involved in the design, development and use of AI technology. Organization in both the public and private sector are highly encouraged to assist with NIST’s Request for Information. *(NIST, 2021)* [14]

Failsafe Protocol – Where possible vendors should employ failsafe protocols. Additionally, users of the AI technology can provide oversight, identify, and report any unexpected outcome.

Regulatory Oversight

FDA - In January of 2021 the U.S. Food and Drug Administration released an action plan for AI/ML-based Software as a Medical Device (SAMD). The goal of the project is to develop a regulatory framework for these AI products which is still in the data collection phase. *(Richardson, 2021)* [15]

Other – In addition to work that is being done by US organizations like the FDA and NIST, many other entities are looking at regulatory oversight as a way of establishing accountability for unintended outcomes of AI. The Federal Trade Commission among many US States are currently reviewing risk related to the use of AI.

Example Scenario

Often when diagnosing brain tumors, many MRI tissue contrasts in each modality are collected and compared. This can be a time-consuming task for neuroradiologists. Some AI solutions have been developed to expediate this process and assist with early detection. At times it may be difficult to define one region of the brain from another. This can go out of bounds if you get a lot of bad data. Voxels are three-dimensional pixels, MR generates its images, in plain and out of plain resolution depth of the voxel. If you have a lot of data with varying voxel depth, you could get varying quality of voxel data. Good data and bad data could have a negative effect on AI algorithm.

Good data -- very regular data/images similar brain regions. Small changes can shift regions. Similar voxels in all three plains. Good data. All knowns.

[13] Shaban-Nejad, A., Michalowski, M., Brownstein, J., & Buckeridge, D. (2021). Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, 25(7), 2374–2375. https://ieeexplore-ieee-org.proxy.davenport.edu/stamp/stamp.jsp?tp=&arnumber=9497066

[14] National Institute of Standards and Technology. (2021, August 11). AI Standards: Federal Engagement. NIST. https://www.nist.gov/artificial-intelligence/ai-standards-federal-engagement

[15] Richardson, L. (2021, August 5). How FDA Regulates Artificial Intelligence in Medical Products. The Pew Charitable Trusts. https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2021/08/how-fda-regulates-artificial-intelligence-in-medical-products

Bad data -- cell phone in the room that transmits 750 MHz in range of MR magnet that will interfere and add noise and bad data. Someone sneezes. Device malfunctions.

What do you get if you have unexpected errors or changes to the data? Who is ultimately accountable for these unexpected outcomes?

AIML Concern #3: Transparency for AI/ML may be Missing

Security Principal Impact Integrity, Availability	**Healthcare Function Impact** Patient Safety, Quality, Continuity of Care	**Operational Impact** Control, Safety & Effectiveness, Diminished Trust
Business Function Operational, Patient Care	**Stakeholder Function** Patients / End User, Engineer, SecOps, Clinicians, Privacy	**Regulatory Guidance** FDA Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (Jan 2021)

Description

AI/ML involves 'training' the model to learn the logic/patterns/scopes of the data fed into it. How does one validate the accuracy of the 'training' model? Are the algorithms used to build the intelligence correctly? Can the technology vendor explain in plain language how the model achieves its outputs, and how it ties to healthcare service quality and safety?

Failure to transparently demonstrate the training models can lead to several challenges:

1) *quality problems and inexplicable outcomes*
2) *inappropriate use of an AI model that is not aligned with its design.*

What Might be the Threat Sources?

Determining the accuracy of the 'training' model over the lifecycle is a very complex endeavor. Many AI/ML systems are opaque and without the right controls/expertise the risks of bias introduced into the data sets, or the algorithms could have significant negative consequences.

Bias issues that impact transparency for model assurance fall into 3 general categories:

Coverage Bias – Poorly selected data: Datasets that are not representative of the real population in need of care

Sampling Bias – Incorrect or outdated data that reflect existing/past societal biases (race, ethnicity, socio-economic status).

Confounder Bias – Hidden factors that if unconsidered may influence outcomes. Algorithm may unwittingly rely on factors that imply the solution is working as designed but still produce outcomes that undermine trust in the solution.

How can vendors provide transparency of potential blind spots within their solutions (data and algorithms) in a manner that leads to overall improvement; will the risks of being penalized keep vendors from pursuing transparency?

How does the need for privacy compare/contrast with the need for adequate transparency?

What are the Potential Impact Concerns?

Moving to or adopting AI/ML in clinical practices and workflows that lack transparency in evidence-based decision making or apply unintended biases in clinical operations.

What to Do?

Could be avoidable with better description of the model assumptions and design.

Example Scenario

One of the most widely implemented early warning systems for sepsis in US hospitals is the Epic Sepsis Model (ESM) [16], which is a penalized logistic regression model included as part of Epic's EHR and currently in use at hundreds of hospitals throughout the country. This model was developed and validated by Epic Systems Corporation based on data from 405,000 patient encounters across 3 health systems from 2013 to 2015. However, owing to the proprietary nature of the ESM, only limited information is publicly available about the model's performance, and no independent validations have been published to date, to our knowledge. Research conducted at the University of Michigan in 2021 noted that ESM failed to detect Sepsis in 67% of patients.

[16] Habib AR, Lin AL, Grant RW. The Epic Sepsis Model Falls Short—The Importance of External Validation. JAMA Intern Med. 2021;181(8):1040–1041. doi:10.1001/jamainternmed.2021.3333

AIML Concern #4: Dubious Quality of Source Data

Security Principal Impact	Healthcare Function Impact	Operational Impact
Integrity	Data Quality, Patient Safety	Financial, Reputational Technical Organization
Business Function	**Stakeholder Function**	**Regulatory Guidance**
Operational, Patient Care	Patients / End User, Engineer, SecOps, Clinicians, Privacy	GDPR, CCPA, HIPAA

Description

Outcomes of AI/ML will only be as good as the quality of data fed into it. Healthcare organizations seeking high precision outputs from the model must give technology vendors assurances the source data are clean, sanitized, and meet certain standards. Also, has the AI/ML model been trained on healthcare data unique to the organization, or on some other samples provided by the vendor? Would those samples satisfy the quality needed by the healthcare organization, or introduce biases that lead to precision shortcomings? Failure to qualify source data may result in inconsistent or inaccurate results.

What Might be the Threat Sources?

Threat sources are likely to be internal non-malicious actors (AI developers, system users); however, there is some risk that unexpected outcomes are a result of malicious actors.

What are the Potential Impact Concerns?

Data bias and poor data hygiene can lead to unintended consequences potentially affecting patient care. If AI/ML systems are used to evaluate revenue cycle, undefined outcomes could have a financial impact to the organization. Some issues that could impact quality of data include:

Ambiguous Data – Often in large databases or data lakes, some errors can creep in even with strict supervision. This situation gets more overwhelming for data streaming at high speed. Column headings can be misleading, formatting can have issues, and spelling errors can go undetected. Such ambiguous data can introduce multiple flaws in reporting and analytics.

Hidden Data – Most organizations typically utilize only a part of their data, while the remaining may be lost in data silos or dumped someplace. For example, customer data available with sales may not get shared with the customer service team, losing an opportunity to create more accurate and complete customer profiles. Hidden data means missing out on discovering opportunities to improve services, design innovative products.

Excessive Data – As the focus is on data-driven analytics and its benefits, too much data does not seem to be a data-quality issue but can be considered as one. Invariably, when looking for data relevant to analytical projects, it's possible to get lost in too much data. Business users, data analysts, and data scientists spend 80% of their time locating the right data and preparing it. Other data-quality issues become more severe with the increasing volume of data, especially with streaming data and large files or database optimize processes.

Duplicate Data – Companies face an onslaught of data from all directions – local databases, cloud data lakes, and streaming data. Additionally, they may have application and system silos. There is bound to be a lot of duplication and overlap in these sources. Duplication of patient contact information, for example, affects patient experience significantly. Duplicate data increases the probability of skewed analytical results. As training data, it can also produce skewed ML models. Rule-based data quality management can help keep a check on duplicate and overlapping records. In addition to these above issues, organizations also struggle with unstructured data, invalid data, downtime of data, redundancy in data, and data transformation errors. *Ankur Gupta (2021, June 21).*

What to Do?

Data Bias -- Some AI systems have been shown to develop an algorithmic bias that compounds existing inequalities such as race, ethnic background, religion, gender, or socioeconomic status. (*Manyika et al., 2019)* [17] For healthcare systems, this may be cause for concern because essentially the AI system may be recommending that the focus of medical care be directed to the wrong demographic.

Customers and vendors responsible for providing the data that is used to train the AI system should ensure that data samples include individuals from diverse backgrounds and perspectives to help minimize AI bias. A general awareness of these algorithmic bias can help healthcare organizations understand and promote diversity, equality and inclusion into the systems that they use[18].

Missing/Incomplete Data – Sometimes when forms or pieces of information are provided, there can be incomplete or missing pieces of data making it difficult to achieve accurate analysis. It is often best to eliminate any data that is missing values.

Inaccurate Data – Accuracy of data plays a critical role in highly regulated industries like healthcare. Inaccurate data does not give a correct real-word picture and the appropriate responses cannot be planned. If customer data is not accurate, personalized customer experiences disappoint, and marketing campaigns underperform. Inaccuracies of data can be traced back to several factors, including human errors, data drift, and data decay. Gartner says that every month around 3% of data gets decayed globally, which is very alarming. Quality of data can degrade over time, and data can lose its integrity during the journey across various systems. Automating data management can help to some extent, but dedicated data quality tools can deliver much better data accuracy.

Typically, data is "sanitized" before training the predictive model; however, it may be difficult to correct all errors potentially impacting that data model or any subsequent models. It is important to establish clear standards for all data that will be used for the model, along with evaluating cleanliness of data and challenging assumptions through each phase of the process. If needed, minimize sample sizes to help ensure data has been sanitized.

Push for Accountability and Oversight -- Additional oversight and support from federal organizations and private entities can also help ensure that bias is kept to a minimum in these systems. Organizations such as the Food and Drug Administration (FDA), the Healthcare Sector Coordinating Council (HSCC), and the National Institute of Health (NIH) are stepping up to provide better oversight and education on the issue. In January of 2021, the FDA announced the Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. The plan calls for the development of a regulatory framework, developer support, and monitoring pilots. *(FDA, 2021)* [19]

Data quality is a big issue and to have strong adoption of AI/ML there requires focus on data quality. One way is to prioritize it in the organizational data strategy and involve and enable all stakeholders to contribute to data quality.

[17] Manyika, J., Silberg, J., & Presten, B. (2019, October 25). What Do We Do About the Biases in AI? Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

[18] Noor, P. (2020). Can we trust AI not to further embed racial bias and prejudice? BMJ : British Medical Journal (Online), 368http://dx.doi.org.proxy.davenport.edu/10.1136/bmj.m363

[19] FDA. (2021, January 12). FDA Releases Artificial Intelligence/Machine Learning Action Plan. U.S. Food and Drug Administration. https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan

Incorporating metadata[20] rules to describe and enrich data in the context of who, what, where, why, when, and how can also help sanitize data to use the organizational data in the right way.

Issues in data quality can be considered as opportunities to address them at the root and prevent future losses and improve business growth.

Data governance council can help strategize and steer the enterprise-wide data governance program to enable data quality and regulatory compliance.

- Define an operational model for data governance
- Develop data governance policies
- Propose data standards, procedures, best practices, tools, and resources

Example Scenario

Data Bias Example -- The AI system is used for early detection of melanoma in patients by evaluating moles. The images used to train an AI system were from patients predominantly with light-colored skin. When a patient with darker skin was scanned, the AI failed to detect the malignant mole therefore delaying proper patient care. *(Noor, 2020)* [21]

Bad data example -- During the middle of the COVID-19 Pandemic, many AI/ML organizations made an attempt to establish early diagnosis of covid and predict patient risk. One such model compared chest x-rays, CT scans and other clinical markers. The data sets were pieced together from various sources and often contained duplicates or bad data. In this case some of the data contained chest scans from children who did not have COVID to help determine what non-COVID cases looked like. As a result, the AI learned to identify children rather than COVID. *(Heaven, 2021)* [22]

[20] Sourced from - Gupta, Ankur (2021, June, 25) The 7 most common data quality issues https://www.collibra.com/blog/the-7-most-common-data-quality-issues

[21] Noor, P. (2020). Can we trust AI not to further embed racial bias and prejudice? BMJ : British Medical Journal (Online), 368http://dx.doi.org.proxy.davenport.edu/10.1136/bmj.m363

[22] Heaven, W. D. (2021, July 30). Hundreds of AI tools have been built to catch covid. None of them helped. MIT Technology Review. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

AIML Concern #5: Absence of Comprehensive Regulatory Oversight

Security Principal Impact Integrity	**Healthcare Function Impact** Patient Safety, Quality, Continuity of Care	**Operational Impact** Safety & Effectiveness
Business Function Regulatory, R&D	**Stakeholder Function** Patients, Healthcare Professionals, Regulators (FDA, OCR, HHS, FTC) - Medical device safety and privacy	**Regulatory Guidance** FDA Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (Jan 2021) • FDA Draft Guidance on Predetermined Change Control Plan

Description

Though there are significant regulatory efforts governing the use of AI/ML in a healthcare environment, the proliferation of commercially available AI/ML products and solutions may impact the ability of stakeholders to manage IT GRC in their environments.

Healthcare AI/ML applications that classify as a medical device will have oversight from FDA, which is tailoring a regulatory framework to these products as evidenced by their AI/ML Action Plan. Examples would include:

- Identifying the risk that a spot in a picture of skin is cancerous
- Suggesting insulin to take based on a picture of the meal that is planned to be consumed
- Analyzing EEG data and monitoring the health of those who have heart conditions and are at risk of complications

However, there are many AI/ML applications in a healthcare context that may not fall into FDA jurisdiction such as the following examples:

- Financial Integrity: Supporting back office or revenue cycle management
- Quality service support: Scanning EHR records to identify potential issues such as unsafe drug interactions or potentially mis-diagnosed conditions
- Wellness products

It should be noted these applications in many cases may be subject to other regulations or guidelines (such as NIST CSF) but may not be included in stakeholder IT GRC efforts.

AI/ML is data hungry; it requires massive amounts of data to properly train.

It is a complex process to introduce specific AI regulations to address unique aspects of some AI devices such as adaptability.

What Might be the Threat Sources?

Threat sources include:

- Misuse of personal information
- Exposure of personal information
- Incorrect semantic interpretation of shared information results in data integrity (mislabeled information)
- Incomplete and poor data sets
- Retraining of AI systems in a targeted manner by adversaries
- Manipulating data to fool algorithms
- Training set attacks
- Adversarial training
- Inference attacks

What are the Potential Impact Concerns?

While the FDA has worked aggressively to implement new regulations concerning AI/ML in clinical settings, the increasingly prevalence of AI/ML may mean that algorithmic processes may be embedded in IT systems that the healthcare organization may not even be aware of.

What to Do?

These AI/ML services as described are the devices for which FDA is proposing a novel regulatory structure for as described in the 2019 Discussion Paper "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback".

Example Scenario

A vendor provides a revenue cycle management tool that aids in patient billing and collection. The vendor has chosen to implement AI/ML processes that inappropriately identifies socio-economic categories of patients due to algorithmic bias that results in disproportionate collection efforts in certain demographics than others, which draws public scrutiny and investigation without the HDO being aware of the use of AI/ML to bias RCM processes.

AIML Concern #6: Lack of Business Leader Knowledge

Security Principal Impact Integrity	**Healthcare Function Impact** Patient Safety, Quality, Continuity of Care	**Operational Impact** Safety & Effectiveness
Business Function Regulatory, R&D	**Stakeholder Function** Patients, Healthcare Professionals, Regulators (FDA, OCR, HHS, FTC) - Medical device safety and privacy	**Regulatory Guidance** FDA Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (Jan 2021) • FDA Draft Guidance on Predetermined Change Control Plan

Description

Business leaders may not have a good understanding of how AI/ML works, or its impact within their organization. Understanding what healthcare assets use AI/ML is important as well -- AI/ML may function behind-the-scenes as engine for a technology product that isn't explicitly purposed as such. Business leaders need to know bounds that AI/ML functions in, and the due diligence required to ensure quality AI/ML outputs. Failure to have educated decision-makers in the healthcare delivery sector can lead to trust factors, over-confidence, over/under dependency.

What Might be the Threat Sources?

Given that this concern deals with the vulnerability of exploiting lack of business leader knowledge, the threat sources are the same as those identified in other AIML concerns. Essentially, the lack of understanding of both the application of AI/ML in a system as well as how it works means that a business leader lacks the necessary information to complete a risk assessment. This is presuming the AI/ML source is provided legitimately from a known vendor. The issues are further complicated if the vendor source is attempting to social engineer access or provide a product with malicious intent.

What are the Potential Impact Concerns?

Education is considered a fundamental tenet to a sound cybersecurity program. AI/ML poses a new complexity to system operations that may result in clinical decisions or recommendations beyond the involvement of the clinician. Lack of understanding and awareness of AI/ML by business leaders may result in acquiring AI/ML based products that are misconfigured, easily exploitable, or maliciously designed to impact operations, steal sensitive data, or disrupt the "human in the loop" in clinical decision-making, which may result in adverse medical events that impact patient safety, affect regulatory compliance, or impact the business reputation of the facility or system.

What to Do?

As with other cybersecurity education and awareness programs, executive education is critical facet of organizational awareness of cybersecurity risks in AI/ML. Organizations should provide training in the foundations of AI/ML, how AI/ML is used in clinical applications, an overview of cybersecurity risks of AI/ML, and how to make procurement and acquisition decisions for AI/ML based products.

Example Scenario

In a real-world example, EHR provider Practice Fusion paid $145 million fine in 2020 for implementing changes to its Clinical Decision Support (CDS) system, adjusting the algorithms at the request of a pharmaceutical company that would provide real-time prompts to physicians during the encounter with the net result being to increase the prescriptions of opioids for a broader range of clinical encounters. This algorithm was specifically developed at a point when opioid prescriptions were beginning to decline because of increased scrutiny and limits by physicians to prescribe opioids.

While the intent of this modification was to increase pharmaceutical sales, it increased the risk of opioid misuse and creating additional instances of Substance Use Disorders (SUDs).

AIML Concern #7: Unintended Consequences – Change Management

Security Principal Impact Integrity	**Healthcare Function Impact** Quality, Patient Safety	**Operational Impact** Resiliency
Business Function Operational	**Stakeholder Function** DevOps, Engineer	**Regulatory Guidance**

Description

Activities such as algorithm adjustments, model re-training with new datasets, baseline software patching, or changes to the operational data feeding the AI model can impact outcomes and produce 'model drift'. This is in addition to normal service availability. Are there mechanisms to validate that the behavior of the model is as it was before the change? Failure to have procedures in place that qualify performance and quality of model outputs after a change may result in substandard or faulty healthcare outcomes.

What to Do?

Relevant approaches to address unintended consequences (change management) in healthcare systems utilizing machine learning include:

- Model development and validation checklists to improve the transparent reporting of prediction modelling studies in medical settings and to assess the risk of bias for non-randomized studies. These checklists provide a framework for creating well-defined baselines that can facilitate the implementation of procedures that qualify performance and quality of model outputs after a change.
- Retrain an AI/ML model periodically to avoid model drift resulting in false negatives and false positives due to changes in data distribution. Pipeline triggers can be used for continuous training of models on availability of new training data to offset model performance degradation.
- Identifying outliers or anomalies within the expected dataset when developing the model. These outliers should be either filtered or alerted upon during both training and operation of the ML model. **An MLOps pipeline used to proactively predict unintended usages could be deployed providing guard rails of what not to process when retraining system.**
- **Secure any input into an AI/ML to maintain the integrity of the model**. Ideally, the data should also be signed and verified prior to any training process. Storing and versioning input data and the generated artifacts (e.g., data transformations) of a machine-learning pipeline allows for better organization and tracking of model and dataset lineage across pipeline runs.
- Continuous monitoring of production AI/ML systems includes defining system logging strategies, establishing continuous evaluation metrics, and identifying an appropriate model-retraining policy. Most models cannot explain how they arrive at a decision. Teaching exceptions is difficult. Event can happen before anomaly is identified. A side-channel monitoring system like Security Information & Event Management (SIEM) or custom system could be used to test expected results versus production outcomes. By looking at out-of-band data points, you may be able to identify events that look normal but show suspicious properties like ransomware, bitcoin mining or intellectual

property extraction (I.e., COVID-19 testing or vaccine). The operational performance of the model should be considered to avoid any potential Denial of Service (DoS) vectors associated with data pollution.

Example Use Case

1.) A systematic review of diagnostic deep learning algorithms for medical imaging in 10 randomized trial registrations and 81 non-randomized studies taken from various sources (Med-line, Embase, Cochrane Central Register of Trials, and World Health Organization trial registry from 2010 to June 2019) was performed by expert clinicians.

2.) Predictions from a pediatric intensive care unit (PICU) in-hospital mortality risk model were displayed using model-agnostic, instance-level explanations based on feature influence, as determined by Shapley values. Focus group sessions solicited critical care provider feedback on the proposed displays, which were then revised accordingly.

AIML Concern #8: Adversarial Data Input Poisoning

Security Principal Impact	Healthcare Function Impact	Operational Impact
Integrity, Availability	Patient Safety, Quality, Continuity of Care	Control, Resiliency
Business Function	**Stakeholder Function**	**Regulatory Guidance**
Operational	SecOps, Engineer	*Computer Fraud and Abuse Act of 1984 (Kumar, et. al, 2018)*[23]

Description

Data Input Poisoning is the introduction of "bad" data into the training set, affecting the model's output. AI/ML requires lots of data to build identifiable logic/patterns/scopes for reporting or decision-making. Protecting the sources of data input is critical. Adversarial "threats" or "actors" with access to input could introduce data sets that pollute the training information used to instruct the AIML model. Failure to secure and validate the AIML inputs, and corresponding data, can contaminate the AIML model with outputs that are inaccurate or well out- of-scope.

What Might be the Threat Sources?

Threat sources are likely to be a malicious individual or group of individuals that include:

1. Hostile nation-state actors seeking to disrupt an AIML process as part of an infrastructure attack
2. Rogue competitive forces seeking to undermine a competitor's solution, or facility
3. Malicious insiders trying to undermine the use of an AIML platform for personal profit

Potential targets and datasets in a healthcare environment may include (but are not limited to): data analytic systems (healthcare, financial, or research), IoT devices, network and security monitoring systems, and facilities control systems.

Methods that actors may employ in an AIML Data Poisoning attack:

- Training phase attacks (poisoning) - Attacks on data collection, processing, and training include data injection, modification, and corruption of logic.
- Testing phase attack (evasion) - Attacks on the AIML model (e.g. HopsSkipJump, Fast Gradient Method, Carlini & Wagner, Crafting Decision Tree, and Zeroth Order Optimization), or its results. *(Newaz, et. al, 2020)*[24]
- Integrity and Availability - Device vulnerability exploitation (e.g. zero-day known and unpatched or unknown and unpatched), potentially damaging data collection and integrity.

[23] Kumar, R. S. S., O'Brien, D. R., Albert, K., & Vilojen, S. (2018). Law and Adversarial Machine Learning. arXiv preprint arXiv:1810.10731. https://arxiv.org/abs/1810.10731

[24] Newaz, A. K. M., Haque, N. I., Sikder, A. K., Rahman, M. A., & Uluagac, A. S. (2020). Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems. arXiv preprint arXiv:2010.03671. https://arxiv.org/pdf/2010.03671.pdf

Success of an attack will depend on an actor's access to the AIML model or data structure, and may include (1) untargeted attacks, which seek to alter the output of the AIML classifiers, (2) targeted attacks, where actors seek to alter the input data in order to change the output, and (3) targeted device attack: altering the function of a singular device to affect a direct or indirect harm on the patient or the patient's data. *(Newaz, et. al, 2020)*[25]

Threat actors may or may not have prior knowledge of the targeted machine-learning algorithm or dataset, although knowledge of the algorithm increases attack efficacy. Adversarial capabilities depend on an adversary's knowledge of:

- Data distribution
- Smart Health System Architecture
- Output labels
- ML Model used *(Newaz, et. al, 2020)*[26]

What are the Potential Impact Concerns?

Potential technical harms from data poisoning involve attacks on data integrity, including input poisoning, output misclassification, and results tampering. Examples of technical harm include:

- Direct or indirect harm to patient data -- erroneous detections (type I – false positives, and type II – false negatives) and treatment(s). This can lead therefore to harm to patients themselves
- Service availability -- denial of the ability of the AIML solution to function or be accessible, including performance degradation
- Cascading failures of AI driven systems that are chained together *(HC3, 2020)*[27]
- Damage to organizational/patient trust in AIML outputs or the algorithmic and/or analytical integrity of the AIML processes

What to Do?

Properly sanitize and review inputs during both the training and inference process of any machine learning application. Identify outliers or anomalies within the expected data set when developing the model. Outliers should be either filtered or alerted upon during training and operation of the ML model using the following approaches:

- **Retrain model periodically to avoid concept drift resulting in false negatives and/or positives as "normal" changes over time and after the initial training period**. A smart attacker can trick an anomaly detector.
- **Deploy an additional AI/ML pipeline to proactively predict unintended usages**. This provides guard rails when retraining the system and helps sanitize the data set. This helps ensure the model knows what NOT to process during retraining.
- **Consider using a side-channel monitoring system like a Security Information & Event Management (SIEM) or custom system to test expected results versus production outcomes.** By looking at out-of-band data points, you may be able to identify events that look normal but show suspicious properties like ransomware, bitcoin mining or intellectual property extraction (I.e., COVID-19 testing or vaccine), Most models cannot explain how they arrive at a decision. Teaching exceptions is difficult. Events can happen before anomalies are identified
- **Maintain the integrity and performance of the model with security**. Input into an AI/ML model should be secured to maintain the integrity of the model. Ideally, the data should also be signed and verified prior to any training process.

[25] Newaz, A. K. M., Haque, N. I., Sikder, A. K., Rahman, M. A., & Uluagac, A. S. (2020). Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems. arXiv preprint arXiv:2010.03671. https://arxiv.org/pdf/2010.03671.pdf

[26] Newaz, A. K. M., Haque, N. I., Sikder, A. K., Rahman, M. A., & Uluagac, A. S. (2020). Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems. arXiv preprint arXiv:2010.03671. https://arxiv.org/pdf/2010.03671.pdf

[27] HC3 (2020). AI Application and Security Implications in the Healthcare Industry. https://www.hhs.gov/sites/default/files/ai-application-and-security-implications-in-healthcare-industry.pdf

Moreover, the operational performance of the model should be considered to avoid any potential Denial of Service (DoS) vectors associated with data pollution.

Example Scenario

While training a medical imaging model to identify tumors, an attacker pollutes the data intended to render a negative cancer diagnosis with positive cancer imagery. This results in the model rendering a negative cancer diagnosis when presented with imagery containing tumors.

AIML Concern #9: Inversion, Inference and Model Extraction Attacks

Security Principal Impact Confidentiality	**Healthcare Function Impact** Privacy	**Operational Impact** Visibility
Business Function Operational	**Stakeholder Function** Administrator, Engineer, Auditor, SecOps, Manager, Executive	**Regulatory Guidance** *Health Insurance Portability and Accountability Act (HIPAA)* *Health Information Technology for Economic and Clinical Health (HITECH) Act* *State Breach Laws* *FTC Section 5 – Unfair and Deceptive Trade Practices*

Description

An attacker may use several existing adversarial confidentiality attacks to obtain information about an AI/ML model or training set. Four examples of these are:

1. *Model Extraction Attack – Gathering information from a model to reconstruct a substitute model that behaves very much like the original.*
2. *Membership Inference Attack – Determining whether an input (potentially an individual) was part of a training set for the model.*
3. *Model Inversion Attack -- Inverting the model to gain additional information on sensitive features about a subject known to be in the dataset. (Sensitive here means features the model's creator does not want to be revealed to the consumer of the model's output.)*
4. *Property Inference Attack -- Extracting data set properties not explicitly encoded as features in the model.*

What Might be the Threat Sources?

Threat sources are likely external actors who have access to a model's interfaces as a regular consumer or an insider with partial knowledge of the model architecture, hyper-parameters, training setup, or complete access to the model itself. Specific potential threat sources include:

1. Hostile nation-state actors seeking to reconstruct the model, determine whether an individual was part of the training set, or gain knowledge about individuals in the training dataset
2. Rogue actors seeking to steal a model or gain personal information to conduct cybercrime or other misuses
3. Malicious insiders trying to exploit the use of an AI/ML platform for personal gain
4. Competitors seeking proprietary information from which they may benefit

The type and methods of attack available to an attacker vary based on the amount of information the attacker can gain on the model and training set, the learning architecture, and ML Algorithm used[28]. There are black-box, white-box, and gray-box confidentiality attacks. Black-box attacks typically assume the attacker does not know the model, parameters, architecture, or training data. In contrast, white-box attackers have complete access to model parameters or loss gradients during training. Gray-box attacks are a scenario when an attacker has some knowledge but not to the extent of a white-box attack[29]. Attacks on centralized learning architectures (model and training data collocated during the training phase) often include the development and use of shadow models or a meta-model[30]. Attacks on distributed learning architectures typically require that the attacker access the central server or is a participant in the distributed architecture[31]. Most known attacks (83.3%) target neural networks with decision trees (11.9%) second[32].

These attacks should be concerning for healthcare entities, notably when the model's creator trained it using personal information generally and ePHI specifically[33]. If a threat actor exploits a vulnerable model, a breach may occur. For example, by simply submitting a few hundred similar variations of subject(s) input data to the model, outputs of the AI/ML model could reveal to an attacker, individual identities of members of the training set or disclose their personal protected health information.

What are the Potential Impact Concerns?

The first concern associated with adversarial attacks on confidentiality is the potential leak of an individuals' personal information. An attacker may use knowledge of the subject's identity or information about the subject to cause physical, mental, or financial damage to that individual.

The next concern is the impact on the organization owning or controlling the AI/ML model. The organization owning or governing the use of the model could suffer economic damages resulting from legal action taken either by the impacted individual/s or a state or federal agency, lost revenue from reputational damage, and investment in an AI/ML solution that the organization must now re-engineer or stop using. In the case of model extraction, the organization may lose a competitive advantage provided by the model.

Data leakage, in the form of the identity of a member of a training dataset or information about a member of the training dataset, is a potential violation of HIPAA if the creator or host of the model is a covered entity or business associate. No current determination or guidance suggests that an AI/ML model is ePHI in and of itself and therefore subject to HIPAA; however, this concept has been explored in the context of GDPR[34]. Nevertheless, training an AI/ML model using ePHI likely pulls that AI/ML model's data pipeline and the AI/ML application within the scope of HIPAA regulations. If so, a

[28] Maria Rigaki and Sebastian Garcia. 2020. A Survey of Privacy Attacks in Machine Learning. ArXiv:2007.07646v2[cs.CR].
[29] Maria Rigaki and Sebastian Garcia. 2020. A Survey of Privacy Attacks in Machine Learning. ArXiv:2007.07646v2[cs.CR] , 6-7.
[30] Maria Rigaki and Sebastian Garcia. 2020. A Survey of Privacy Attacks in Machine Learning. ArXiv:2007.07646v2[cs.CR] , 12.
[31] Maria Rigaki and Sebastian Garcia. 2020. A Survey of Privacy Attacks in Machine Learning. ArXiv:2007.07646v2[cs.CR] , 16-17.
[32] Maria Rigaki and Sebastian Garcia. 2020. A Survey of Privacy Attacks in Machine Learning. ArXiv:2007.07646v2[cs.CR] , 18.
[33] S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," 2018 IEEE 31st Computer Security Foundations Symposium (CSF), 2018, pp. 268-282, doi: 10.1109/CSF.2018.00027
[34] Veale M, Binns R, Edwards L. 2018 Algorithms that remember: model inversion attacks and data protection law. Phil.Trans. R. Soc. A 376: 20180083. http://dx.doi.org/10.1098/rsta.2018.0083

failure to effectively manage the risk of unauthorized disclosure of the ePHI used to train the model is likely to be viewed by the Office for Civil Rights as a HIPAA violation, assuming the organization is subject to HIPAA.

Additionally, for sensitive information not governed by HIPAA, it will likely be covered by a wide variety of state, federal, and international laws that govern the use and disclosure of medical or other sensitive information. For example, a health application with AI/ML functionality containing or trained with sensitive health information and offered by an organization not subject to HIPAA may fall under FTC regulation and enforcement.

What to Do?

Defenses against *model extraction* include:

- Detection of anomalies in model queries.[35]

Defenses against *model inversion* include:

- Set all loss gradients below a certain threshold to zero.[36]
- Adding random noise.[37]

Defenses against membership inference include:

- Differential Privacy (DP) introduces random noise into the process to make it harder for an adversary to gain information on an individual within the training set. DP also provides a quantitative measure of privacy. There is a trade-off that the data scientist must make between privacy and model accuracy.
- Regularization – reducing overfitting and increasing model generalization.
- Lowering the precision of the prediction vector or adding noise.

Defenses against property inference include:

- Model compression might be effective in neural networks.[38]

Example Scenarios

Model Extraction Attack: By examining the inputs and outputs, an attacker can replicate the model so that the "stolen" model behaves as the target model behaves. Having this information is almost the same as having the original model itself. The attacker can use this shadow model to conduct any of the other attacks or launch other schemes to profit from the value of the data or use of the model.[39]

Membership Inference Attack: An attacker uses a membership inference attack to identify an individual (along with sensitive information) in the dataset used to train the model.[40] For example, a hospital system uses patient clinical records to train a model to predict which treatment is most likely to succeed given a patient's clinical symptoms or genetic makeup. A membership inference attack would "turn the machine learning against itself" by training an attack model to reveal the details about the training dataset. The attacker can accomplish this by examining the inputs (making

[35] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. 2019. PRADA: protecting against DNN model stealing attacks. In 2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, Stockholm, Sweden, 512–527.

[36] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep Leakage from Gradients. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). CurranAssociates, Inc., Vancouver, Canada, 14747–14756.

[37] Titcombe, Tom, Hall, Adam James, Papadopoulos, Pavlos, Romanini Daniele. 2021 Practical Defences Against Model Inversion Attacks for Split Neural Networks, Workshop on Distributed and Private Machine Learning (DPML)

[38] Tianhao Wang. 2019. Property Inference Attacks on Neural Networks Using Dimension Reduction Representations. https://scholar.harvard.edu/files/tianhaowang/files/pia_19.pdf. 17-18

[39] Andrew J. Lohn. 2020. Hacking AI: A primer for Policymakers on Machine Learning Cybersecurity.

[40] Maoqiang Wu, Xinyue Zhang, Jiahao Ding, Hien Nguyen, Rong Yu,, Miao Pan, Stephen T. Wong. 2020. Evaluation of Inference Attack Models for Deep Learning on Medical Data. arXiv:2011.00177 [cs.LG]

queries) and examining the outputs. He is re-engineering what is in the black box that is the machine learning algorithm. [41] In doing so, he learns that a targeted individual was part of the training set.

Model Inversion Attack: An attacker uses a model inversion attack on a model trained to predict hospital readmissions. By starting with known data about a population and iteratively making small, virtually undetectable adjustments to that data, the attacker can eventually reveal individual names and attributes. The attack results in the attacker gaining information about a particular individual in the training set, medical condition. The attacker then exploits this information to harm the individual.

Property Inference Attack: Since training deep neural networks is expensive in computational power and data requirements; a hospital chooses to become a part of a consortium of hospitals that shares predictive models. In this way, the hospitals can create more precise models. An attacker deploying a property inference attack infers information about the training set used to create the models. This approach could lead through inference to identifiable information about the members of the training datasets even though they are not attributes contained within the training data itself.[42]

##

Additional References

Habib AR, Lin AL, Grant RW. The Epic Sepsis Model Falls Short—The Importance of External Validation. *JAMA Intern Med*. 2021;181(8):1040–1041. doi:10.1001/jamainternmed.2021.3333

Burt, A. (2021, May 3). New AI Regulations Are Coming. Is Your Organization Ready? Harvard Business Review. https://hbr.org/2021/04/new-ai-regulations-are-coming-is-your-organization-ready

[41] Reza Shokri, Marco Stronati, Congheng Song, and Vitaly Shmatikov. 2017 Membership Inference Attacks Against Machine Learning Models | IEEE Conference Publication | IEEE Xplore

[42] Tianhao Wang. 2019. Property Inference Attacks on Neural Networks Using Dimension Reduction Representations. https://scholar.harvard.edu/files/tianhaowang/files/pia_19.pdf.

National Institute of Standards and Technology. (2021, August 11). AI Standards: Federal Engagement. NIST. https://www.nist.gov/artificial-intelligence/ai-standards-federal-engagement

Richardson, L. (2021, August 5). How FDA Regulates Artificial Intelligence in Medical Products. The Pew Charitable Trusts. https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2021/08/how-fda-regulates-artificial-intelligence-in-medical-products

Shaban-Nejad, A., Michalowski, M., Brownstein, J., & Buckeridge, D. (2021). Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, 25(7), 2374–2375. https://ieeexplore-ieee-org.proxy.davenport.edu/stamp/stamp.jsp?tp=&arnumber=9497066

Heaven, W. D. (2021, July 30). Hundreds of AI tools have been built to catch covid. None of them helped. MIT Technology Review. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

FDA. (2021, January 12). FDA Releases Artificial Intelligence/Machine Learning Action Plan. U.S. Food and Drug Administration. https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan

Manyika, J., Silberg, J., & Presten, B. (2019, October 25). What Do We Do About the Biases in AI? Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Noor, P. (2020). Can we trust AI not to further embed racial bias and prejudice? BMJ : British Medical Journal (Online), 368http://dx.doi.org.proxy.davenport.edu/10.1136/bmj.m363

Sourced from - Gupta, Ankur (2021, June, 25) The 7 most common data quality issues

https://www.collibra.com/blog/the-7-most-common-data-quality-issues

Thomas C Redman (2017, November 27) Seizing Opportunity in Data Quality

https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

https://www.medicaleconomics.com/view/why-artificial-intelligence-in-health-care-needs-regulation

Barda et al. A qualitative research framework for the design of user-centered displays of explanations for machine learning predictions in healthcare. BMC Med Inform DecisMak 20, 257, (2020).

Nagendran et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368, (2020)

Mirsky, Y., & Mahler, T., & Shelef, I., & Elovici, Y. (2019). CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. 28th USENIX Security Symposium, 473,474. https://www.usenix.org/conference/usenixsecurity19/presentation/mirsky

Veale M, Binns R, Edwards L. 2018 Algorithms that remember: model inversion attacks and data protection law. Phil.Trans. R. Soc. A 376: 20180083. http://dx.doi.org/10.1098/rsta.2018.0083

Acknowledgements

The Health Sector Coordinating Council Cybersecurity Working Group (CWG) thanks the following individuals who served on the CWG Emerging Technology Task Group and contributed to the development of this white paper. The CWG also recognizes the many subject matter experts who provided their insightful perspectives about the complex concepts discussed in this paper.

HSCC Cybersecurity Working Group Emerging Technology Task Group Members

Preethi Amurthur, Philips

Lee Barrett, EHNAC

Robert Bastani, HHS ASPR

Penny Chase, MITRE

Regina Farmer, McKesson

Ed Gaudet, Censinet

Darrell Hall, HHS

Michael Holt, Virta Labs

Dr. Mark Jarrett, Northwell Health

Catherine Lowe, MedSec

Jon Moore, Clearwater

Nimi Ocholi, Medtronic

Bill Proffer, Leidos

Chris Reed, Medtronic

Barry Robson, Ingine

Shawn Savadkohi, (formerly) San Mateo County Health

Julie Sisk, First Health Advisory

Jim St. Clair, Linux Foundation Public Health

Mac Stevens, Spok

Christine Sublett, Sublett Consulting

Kenneth Wilder, ClearData

Subject Matter Expert Contributors

Troy Adams, HHS

Dr. Sven Cattell, AI Village

Fotios Chantzis, OpenAI

Matthew Diamond, FDA

Hugo Espiritu, Johns Hopkins University Applied Physics Laboratory

James Harbinson

Aaron Heath, Syneos Health

Dr. Arvind Rao, University of Michigan

Dr. Flo Reeder, MITRE

Barton Rhodes, Lacework