Journal of the American Medical Informatics Association, 30(7), 2023, 1227–1236
https://doi.org/10.1093/jamia/ocad065
Advance Access Publication Date: 18 April 2023
Research and Applications

Research and Applications

More than algorithms: an analysis of safety events involving ML-enabled medical devices reported to the FDA

David Lyell, Ying Wang, Enrico Coiera, and Farah Magrabi

Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, NSW 2109, Australia

Corresponding Author: David Lyell, PhD, Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, NSW 2109, Australia; david.lyell@mq.edu.au

Received 15 September 2022; Revised 13 March 2023; Editorial Decision 25 March 2023; Accepted 30 March 2023

ABSTRACT

Objective: To examine the real-world safety problems involving machine learning (ML)-enabled medical devices.
Materials and Methods: We analyzed 266 safety events involving approved ML medical devices reported to the US FDA's MAUDE program between 2015 and October 2021. Events were reviewed against an existing framework for safety problems with Health IT to identify whether a reported problem was due to the ML device (device problem) or its use, and key contributors to the problem. Consequences of events were also classified.
Results: Events described hazards with potential to harm (66%), actual harm (16%), consequences for healthcare delivery (9%), near misses that would have led to harm if not for intervention (4%), no harm or consequences (3%), and complaints (2%). While most events involved device problems (93%), use problems (7%) were 4 times more likely to harm (relative risk 4.2; 95% CI 2.5–7). Problems with data input to ML devices were the top contributor to events (82%).
Discussion: Much of what is known about ML safety comes from case studies and the theoretical limitations of ML. We contribute a systematic analysis of ML safety problems captured as part of the FDA's routine postmarket surveillance. Most problems involved devices and concerned the acquisition of data for processing by algorithms. However, problems with the use of devices were more likely to harm.
Conclusions: Safety problems with ML devices involve more than algorithms, highlighting the need for a whole-of-system approach to safe implementation with a special focus on how users interact with devices.

Key words: decision support systems, clinical, machine learning, safety, artificial intelligence, medical devices, decision-making, computer-assisted, clinical decision-making

INTRODUCTION

The goal of artificial intelligence (AI) in healthcare is to *automate* tasks to assist humans,[1] with the aim of improving decision-making, leading to better care delivery and patient outcomes (eg, Ref. [2]). Contemporary AI systems use machine learning (ML) to automate specific sub-tasks; these include detecting or quantifying disease (eg, detecting cancers in screening mammography or cardiac arrhythmias from EKG) and recommending management.[3] ML systems used for the diagnosis, management, or prevention of disease are classed as medical devices and regulated in most nations.[4] Therefore, medical devices represent ML systems usable in clinical practice and by consumers. Their indications for use establish clinician responsibility for key decisions such as diagnosis and management that are informed by ML devices, even if ML provided information is wrong.[3]

As ML devices are implemented into clinical practice, it is crucial to ensure they are safe and deliver expected benefits. Safety research to date has focused on the potential risks of ML and the safety challenges posed by the black box nature of ML algorithms which are not human verifiable. Limitations of ML include susceptibility to biases in training data, and distributional shift over time between training data and the population to which algorithms are applied.[5,6]

These safety problems are theoretically derived from the known pitfalls and limitations of ML and case reports about individual events (eg, Refs [7,8]). With few ML devices approved prior to 2018[3,9] and limited real-world evidence, such theoretical knowledge is an excellent starting point for ensuring safety of these emerging technologies. Additionally, human factors evaluations concerning the potential for bias, skill degradation, human–AI handover, and situational awareness are also needed.[10]

While it is essential to ensure ML algorithms are developed in a safe and robust manner,[11,12] there is an increasing need to focus on safety problems arising from the way ML systems are implemented and how they are used in the real world. For example, an ML system for predicting the onset of sepsis from electronic health record data was found to perform substantially worse in real-world use (AUC, 0.63) than claimed by the manufacturer (AUC, 0.73–0.83).[7] Additionally, sepsis alerts more than doubled in the weeks following the first COVID-19 hospitalizations. Presence of the virus made it difficult for the algorithm to differentiate bacterial sepsis from COVID, thereby limiting the usefulness of alerts.[8]

The aim of this study is to bring greater context and understanding of the real-world safety problems with ML devices through a retrospective analysis of reports about adverse events which are a crucial source of early information on low-frequency safety problems.[13,14] No previous study has systematically collated the real-world safety problems associated with ML systems in healthcare.

Medical device adverse event reporting

The Food and Drug Administration has regulatory responsibility for medical devices in the United States and captures adverse event reports as part of post-market surveillance. Reported events are publicly available via the FDA's Manufacturer and Use Facility Device Experience (or MAUDE). Federal regulations specify mandatory reporting obligations for device manufacturers, importers, and user facilities (the facility using the device, such as hospitals, nursing homes, or outpatient diagnostic facilities), within 30 days of becoming aware of the event.[15] Reportable events occur when Class II or III devices, those devices classified as moderate and high risk, respectively, are suspected of contributing to death or injury, or have malfunctioned in ways which could potentially contribute to death or injury.[15] MAUDE also includes voluntary reports from consumers, caregivers, healthcare professionals, and other concerned individuals submitted via the MedWatch program.[16] Reports from MAUDE have been previously analyzed to examine safety problems with health IT.[17]

In this study, we extend our prior work that examined how ML devices assist clinicians[3] to understand the safety problems with ML in real-world settings. Our goal is to extend what is known about the risks to patients arising from problems with the ML systems themselves, to the way they are implemented and how they are used in the real world by clinicians and consumers.

MATERIALS AND METHODS

We analyzed reports about events involving ML devices that were submitted to MAUDE. *ML devices* were defined as:

1. Class II and III medical devices that were approved by the FDA for use in the United States via the Premarket approval (PMA), Premarket notification (PMN/510k), or De Novo pathways,[18–20] and
2. Utilized ML methods.

We included reports that were submitted to MAUDE between January 01, 2015 and October 22, 2021 as prior reviews indicate most ML devices were approved since 2015.[3,9]

Searching MAUDE for events involving ML medical devices

The methodology for identifying events involving ML devices was challenging. First, the FDA neither reports whether devices utilize ML nor is it possible to search the free text of the FDA approval documents.[9] Second, while published lists of ML devices exist, most do not report any method for confirming ML utilization by the devices identified and therefore cannot be considered gold standard. To overcome these limitations, we searched MAUDE for reports about adverse events involving the ML medical devices that have been identified by previous studies. The results from the MAUDE search were screened to ensure they involved devices utilizing ML and then analyzed.

Methods to compile the list of ML devices from previous studies, search MAUDE and confirm ML utilization are detailed in the following sections.

Strategy to compile ML device search list

We searched Google Scholar by combining the search terms, "AI," "ML," and "FDA" and then used a snowballing approach to identify journal articles and other sources that cataloged AI or ML devices approved by the FDA and were published before August 2021. The search identified 7 journal articles[3,9,21–25] and the American College of Radiology's online database[26] cumulatively reporting 875 ML devices. From this, a search list of 508 unique ML devices was extracted (see Supplementary Appendix SA).

MAUDE search

Using the search list of 508 devices, a systematic search of MAUDE was conducted for the period January 01, 2015 to October 22, 2021. Three search strategies were used to account for variation in event reporting. Two were conducted using the OpenFDA API,[27] searching firstly by approval number, and then by manufacturer and device name where the incident date was after the device approval date, as few event reports were linked to approval numbers. Anticipating greater variation in voluntary reporting from consumers, a third strategy involved a hand search of the MAUDE website by manufacturer and device name[28] for devices intended for use by consumers. For example, some consumer reports identified the involved device by the manufacturer rather than device name. The OpenFDA API search automatically retrieved the first thousand results for individual searches. However, the searches for 5 devices using the manufacturer and device name strategy returned more than 1000 results: 4 devices between 1057 and 3443, and a fifth returning 11 848 results. Review of these 5 devices resulted in their exclusion as none could be confirmed as utilizing ML (as described in the next section), rendering further retrieval unnecessary.

Confirming ML utilization

A significant limitation of the sources used to create the ML device search list was the absence of methods for confirming ML utilization (Table 1, Supplementary Appendix SA). To overcome this limitation, we screened all the devices involved in the adverse events for ML utilization. Such screening was essential to ensure analyzed events were indeed indicative of safety events resulting from the real-world use of ML devices.

Table 1. Generic medical device types associated with 266 events involving 25 unique devices

Generic device	*n* unique devices	*n* events	% of events
Imaging data			
Mammography	1	184	69
Radiotherapy planning	3	13	5
Ultrasound	5	8	3
Computed tomography (CT)	3	6	2
Computer-assisted surgical device	2	2	1
Coronary vascular physiologic simulation software	1	2	1
Computer-assisted detection (CADe)	1	1	<1
Signal data			
Clinical patient monitors	2	28	11
Electrocardiogram (EKG)	4	16	6
Drug dose calculator	2	5	2
Fertility/contraception software	1	1	<1

Accordingly, for each device involved in the event, we reviewed the device's FDA approval documents and manufacturer marketing materials[3,9] for descriptions that the device uses AI, machine or deep learning, or specific ML methods such as neural networks or natural language processing. If neither source described use of ML methods, events involving that device were excluded. For confirming ML use, we accepted manufacturer claims about their products as reliable given that the FDA has general controls which expressly prohibit misbranding, false or misleading advertising.[29] *AI adjacent* devices where the device itself did not contain ML but rather provided data output as an input to another ML-enabled device were also excluded. One example is a continuous glucose monitor where obtained blood glucose readings could be processed by a separate Class I analysis device utilizing ML. As Class I devices are low risk, they do not require FDA approval and are not subject to post-market surveillance.[30]

Data extraction and event classification

We extracted the following FDA-coded variables available in MAUDE using OpenFDA queries: the reportable event type (injury, death, malfunction, or other), date received, and reporter (manufacturer, user facility, importer/distributor, or voluntary). FDA generic device names were used to identify and group devices for analysis. We extracted the narrative text from reports to classify the consequences of events and contributing problems as detailed in the following sections.

Consequences of events

The consequences of events were classified using a standard approach,[17] into:

a. *Harm* (eg, overdose of radiation or irradiating outside the treatment target when delivering radiotherapy);
b. *Near miss events* with potential to harm if not for intervention to prevent it (eg, user recognizes a problem and acts to prevent it);
c. *Hazards* with potential to cause harm (eg, problems that could harm in different circumstances);
d. *Consequences for healthcare delivery* without specific patient harm (eg, needing to reschedule tests);
e. *No consequences for healthcare delivery* (eg, describes problems with a device, but without consequences or healthcare delivery or harm to patients); and
f. *Complaints* which generally describe the users experience but do not indicate harm, hazard, or systemic problems qualifying for the other categories.

Events were classified according to the most serious outcome if multiple outcomes were described. For example, if an event described a near miss and consequences for healthcare delivery, it was classified as a near miss event. Descriptions of harm were summarized from the reports.

Contributing problems

Contributing problems described in the reports were classified using an existing framework for classifying safety problems with Health IT,[31] which has demonstrated validity between jurisdictions, including Australia,[31] England,[32] and the United States.[17] Events were first divided into those primarily involving technical problems (device problems) or human factors (use problems). They were then assigned to a single category, focusing on the reporter identified root cause or the most significant precipitating factor leading to the event. We supplemented the framework[32] with 3 new categories to characterize contributing problems that were significant for the analysis of ML device safety and distinct to the existing categories. These were *contraindicated use*, *errors in task execution* when using devices and *algorithm errors* arising from the processing and conversion of input data into outputs. These modifications are described in Supplementary Appendix SB.

Finally, we mapped contributing problems to a model of interaction between user and ML device (Figure 2). We conceptualized the relationship as user and device contributing a unique role to the healthcare task the user seeks to accomplish, with the interaction characterized as inputs and outputs.

Analysis

Two investigators (DL and YW) independently classified consequences and contributing problems described in reports, and then jointly arrived at consensus decisions for each event. Uncertain classifications and disagreements were resolved by consensus involving a third investigator (FM). A clinician (AS) was consulted when required. Descriptive analyses of events were undertaken by consequences and problem type. Subgroup analysis of the events involving patient harm was conducted using the Fisher's exact test at a significance level of $P = .05$.

RESULTS

Events reported in MAUDE

The search identified 266 events involving 25 unique devices (Figure 1). The ML devices identified primarily used image-based data (81%, see Table 1), while the remaining 19% used signal-based data (eg, patient vital signs). Nearly all events fell into 1 of the 3 types subject to mandatory reporting: malfunctions ($n = 238$), injuries ($n = 25$), and deaths ($n = 1$). Two events were coded as other event types, although one of these described a death (see Insulin Dosing Software in Box 1). Manufacturers ($n = 241$) were by far the most common reporters, with user facilities ($n = 11$) rounding out events from mandatory reporters. Fourteen events were voluntarily reported.

Box 1. Deaths involving ML devices

Diagnostic ultrasound

Acute mitral valve insufficiency was not detected on cardiac ultrasound Doppler in a patient who subsequently died. Imaging pre-sets (calibration settings of the device) were made by the user instead of using those provided by the manufacturer. Consequently, signals indicating mitral valve insufficiency were not observed, leading to delayed treatment. The quantifiable effect of the delay on outcome is unknown.

Insulin dosing software

An anonymous voluntary report expressed concerns over the aggressiveness with which an insulin dosing system treated hyperglycemia, with rates of change in blood sugar levels double that of other hospitals. Such rapid changes were described as causing patients to develop metabolic and EKG changes, leading to patients requiring intubation and "several unfortunate outcomes including one patient death." The report did not detail specific events and these events could not be confirmed by the manufacturer.

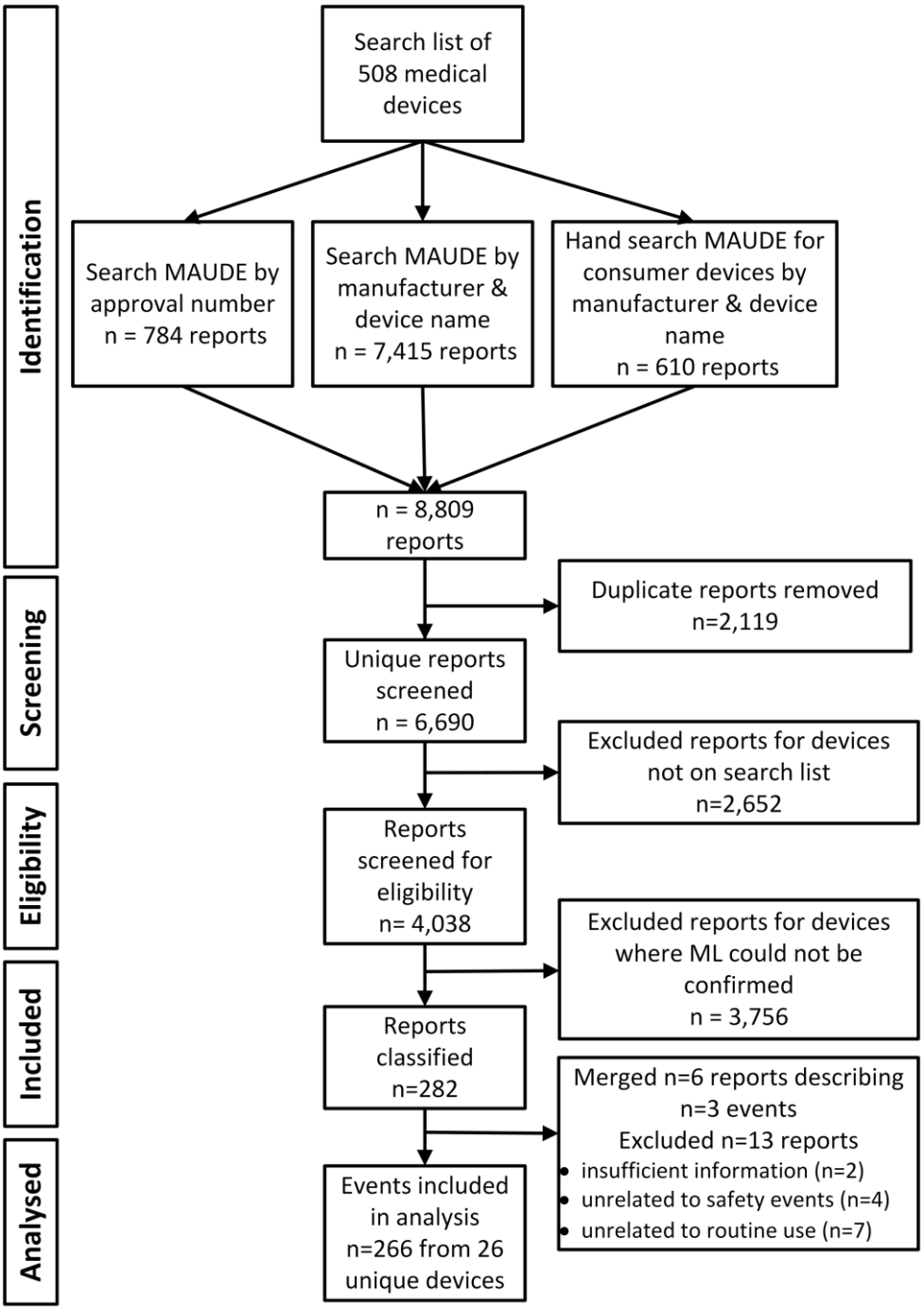

Figure 1. Selection of MAUDE events describing safety problems with ML medical devices.

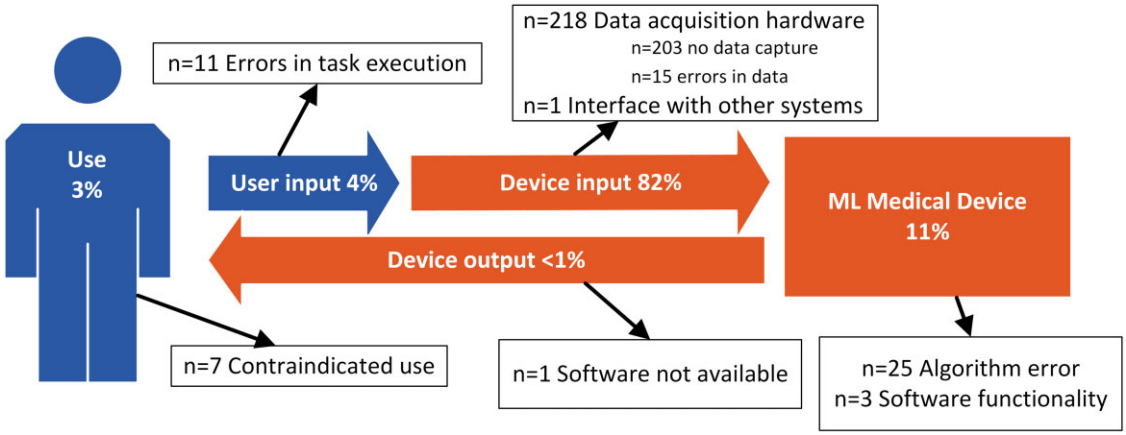

Figure 2. Use problems (Use; User input) and device problems (Device input; ML medical device; Device output) described in events involving ML medical devices.

Table 2. Harm described in 43 events

Harm	Generic device	*n*	Examples
Additional exposure to X-rays	Mammography ($n = 19$), CT ($n = 4$)	23	Scans aborting mid-procedure, cutoff images, or images with artifacts, breast compression plate releasing during image-guided biopsy procedures, required scans to be repeated. In one event X-ray exposure did not stop and had to be manually interrupted.
Radiation treatment delivered to incorrect location	Radiotherapy planning	3	(1) Two users inadvertently altered the planning target volume and radiation was delivered outside the intended treatment target. (2) One patient's position was not correctly calibrated resulting in a 1 cm discrepancy with the radiotherapy plan.
Struck by moving machinery	Mammography	3	Patients were struck by uncommanded movement of scanner c-arms.
Death	Ultrasound ($n = 1$), insulin dosing ($n = 1$)	2	See Box 1
Mispositioned surgical screws	Computer-assisted surgical device	2	Pedicle screws were mispositioned during computer guided surgery. (1) One was attributed to use error possibly due to the trajectory of the pak needle being altered while advancing through soft tissue, causing the navigation system to show a different to actual trajectory. (2) Inaccurate positioning caused the left side of the patients L4 pedicle to be breached, the device was reported to be working within specifications.
Radiation treatment overdose	Radiotherapy planning	2	(1) Caused by the electron calibration curve being incorrectly defined as absolute density resulting in 7 patients receiving higher than intended radiation doses. (2) Manual override caused the patient to be modeled as water leading to a 10% radiation overdose to be delivered.
Movement of machinery during biopsy procedure	Mammography	2	(1) The device lost the target, and the clinician manually moved the needle, resulting in the patient fainting, a 1 cm cut that required suturing and the biopsy not being obtained. (2) The c-arm unexpectedly began to rotate resulting in discomfort, bruising and failure to place the biopsy marker.
Hypoglycemia	Insulin dosing	1	User administered insulin but not the carbohydrate recommended by the device resulting in hypoglycemia.
Device results caused user to delay seeking emergency medical care	EKG	1	See Box 4
Mispositioned biopsy tags	Mammography	1	Incorrectly entered data resulted in biopsy markers being placed in an incorrect location which then had to be surgically removed.
Panic attacked/anxiety from false-positive result	EKG	1	Patient experienced panic attack and anxiety from a false-positive interpretation of atrial fibrillation from a consumer EKG device.
Unplanned pregnancy	Fertility/contraceptive software	1	Consumer became pregnant while using a contraceptive app, the pregnancy was aborted.
Electric shock	Mammography	1	Patient felt electric shock during mammogram but did not require medical treatment.

Table 3. Harm by use or device problems

Problem type	Patient harm				Total	
	Yes		No			
	n	%	*n*	%	*n*	%
Use problems	10	56	8	44	18	7
Device problems	33	13	215	87	248	93

Consequences of events

Of the 266 events we analyzed, most (66%, $n = 175$) described hazards which in different circumstances could lead to harm. These included the risk of contact with machinery due to uncommanded movement of scanner c-arms (the C-shaped arm that houses the X-ray generator/emitter on one side and the detector on the other. The arm is movable allowing flexibility in achieving desired positioning for imaging) and shattered radiation shields. Others related to inaccurate measurements or results that could lead to misdiagnosis or patient misunderstanding of device results that could then influence decision-making. Harm was the second most observed consequence described in 16% of events ($n = 43$), summarized in Table 2.

Twenty-four events (9%) were device failures without harm or consequences for healthcare delivery, while 8 events (3%) described consequences for care delivery, such as having to reschedule scans due to nonfunctional equipment. There were 11 near miss events (4%) where users intercepted potentially harmful problems, including recognizing and preventing calculated overdose of insulin or radiation from being administered. Five events (2%) were classified as complaints, 4 of which described discomfort or skin irritation from EKG electrodes, and 1 expressed that a contraindication preventing their use of a device should have been more prominently declared in the product labeling.

Safety problems contributing to events

Our classification of the problems with ML devices and their use is summarized in Figure 2.

While use problems only contributed to 7% ($n = 18$) of all events, a disproportionally high percentage (56%; Table 3) were associated with patient harm, compared to device problems (13%; Fisher's exact test, $P < .001$), with the relative risk of use problems leading to harm being 4.2 (95% CI 2.5–7).

In the following sections, we give a brief overview of the categories of use and device problems (Figure 2).

ML medical device

Eleven percent of events ($n = 28$) involved problems located within the medical device. Most of these were associated with algorithm errors where incorrect or inaccurate results of data processing were the primary contributor ($n = 25$). These involved a wide variety of problems including devices with inaccurate fractional flow reserve derived from CT (FFRCT) values (see Box 2); problems with image enhancement; incorrect positioning for pedicle screws; inaccurate measurements of bladder urine volume, and problems with radiotherapy treatment plans; being unable to classify cardiac rhythms or incorrect measuring of heart rate from EKG; inaccurate measures of cardiac index or cardiac output calculated by patient monitors; calculations of higher than expected insulin doses, and incorrect prediction of ovulation by a contraceptive app. Other problems related to software functionality including contours in radiotherapy plans not saving, a software bug causing total daily insulin to be misreported and insulin dosing software pursuing blood glucose reductions >200 mg/di/hr, which were deemed to be too aggressive by the reporter (Box 1).

Device input

Eighty-two percent of events ($n = 218$) involved problems with data acquisition by the device and can be subdivided by whether the problem manifests in the failure to capture data (no data) or erroneous data. No data capture accounted for the majority ($n = 203$), of which most involved various mechanical problems and failures such as uncommanded or unexpected movement of c-arms or compression plates, detachment of or broken device components, electrical arcing, overheating, burning, or shocks. Other device failures included failure to power on, scans terminating mid procedure, devices freezing during operation, error messages, or other failures preventing use. Most of these events related to one model of a mammography device ($n = 184$).

Another group of safety events occurred when data was acquired but contained errors or contamination ($n = 15$). These included the presence of artifacts in images, portions of images being cutoff, or known lesions or administered contrast barely visible in scans. Reports about such events commonly described corrective actions to resolve problems rather than identifying a specific cause, such as device calibration returning the device to expected operation, or checking device components, settings, general maintenance, replacement of worn components, lubrication of machinery, and system reboot resulting in the device functioning as expected. Three devices involved patient monitors affecting SpO_2 readings, one indicated the SpO_2 board assembled was to be replaced without identifying the specific problem or if that was the cause. The other 2 reported devices under investigation for suspected higher than expected SpO_2 readings.

User input to device

Four percent of events ($n = 11$) involved errors in task execution or use of devices. These comprised use of incorrect settings (see Diagnostic ultrasound in Box 1), problems with user calibrations or patient positioning during procedures (Box 3). Several reports involved radiotherapy planning devices. One described skin burns attributed to the physician mistakenly adding a "bubble" outside the tumor, another with no patient impact was attributed to the target area being moved before the treatment plan was approved (Box 3).

Box 2. Example of an event demonstrating an algorithm error in the measurement of fractional flow reserve derived from CT (FFRCT)

A device provided false-negative fractional flow reserve derived from CT (FFRCT) result, indicating no clinically significant deterioration (0.83 and 0.92, respectively), for a patient following an acute myocardial infarction involving left anterior descending (LAD) and diagonal arteries. A second analysis returned FFRCT values of <0.50 and 0.74. The patient underwent additional testing and an emergency percutaneous coronary intervention including invasive coronary angiography. The report did not describe whether the false-negative impacted patient care.

Box 3. Examples where user input to ML devices contributed to events

Patient mispositioned for radiotherapy treatment

A patient was over-irradiated due to incorrect calibration of the position of the patient's jaw resulting in 1 cm discrepancy with the radiotherapy plan. Patient consequences were not reported.

Radiotherapy target moved prior to approval

An initial radiotherapy treatment was administered to an incorrect location. It appeared the user, working from home during the COVID-19 pandemic, was working on a laptop with a small screen. They moved the target area by accidentally clicking on an unintended software function. Subsequent review and approval of the radiotherapy plan failed to detect the incorrect target location. No adverse consequences for the patient were reported.

Box 4. Example of contraindicated use of an ML medical device contributing to events

A patient suffering a myocardial infarction reported delaying medical care after receiving an automatic interpretation of "normal sinus rhythm" from an over-the-counter consumer EKG device. The device detects arrhythmias but is not indicated or capable of detecting heart attack. The consumer commented "I was having a[n] actual heart attack. 100% blockage of the LAD [left anterior descending artery] and it said nothing was wrong. I delayed going [to seek emergency care] because of this and probably suffered more damage."

Another consumer consulted the device but sought emergency care despite the results: "I used this device 2 minutes prior to calling ambulance and device said everything was fine with my heart. Suffered major heart attack that morning."

Common to these examples is a mistaken expectation that the device should detect heart attack. The manufacturer noted that the device labeling specifies it is not intended to detect heart attack, adding in one report that it is possible for an EKG taken during "heart attack to still be normal sinus rhythm." Event reports show how consumers equated a device result of normal sinus rhythm as "nothing was wrong," "show[ed] up as normal," "normal EKG," and "device said everything was fine with my heart."

Another described an incorrectly calculated radiotherapy dose attributed to the use of a non validated couch, resulting in differences in densities between the actual couch and that modeled by the device.

Use of ML medical devices

Three percent of events ($n = 7$) were attributed to the contraindicated use of devices. Six involved a device intended for use by consumers to acquire Lead-I and Lead-II electrocardiograms, detecting normal sinus rhythm, and several arrhythmias. Five of these events involved consumers who received an interpretation of a normal sinus rhythm while suffering a heart attack (myocardial infarction). The manufacturer attributed these events to use errors, referring to the device's indications for use and the fact that it did not detect infarctions. One consumer delayed seeking care (see Box 4), 2 others did not, while the remaining 2 were described as may have delayed and unknown. The sixth event was a voluntary consumer complaint about product labeling after purchasing a device only to discover it was contraindicated due to their pacemaker. In the final event, the indicated carbohydrate treatment plan for insulin dosing software was not followed resulting in the patient experiencing a hypoglycemic event.

Device output

One event (<1%) described a problem with device output where the device would freeze while viewing images and stop responding to user input (software not available).

DISCUSSION

Main findings

The strength of this study lies in the contribution of the first systemic analysis of safety events arising from real-world use. To date, what is known about the safety problems with ML medical devices has been theoretically derived[5,6] and based on case reports about specific events (eg, Refs [7,8]). All of these focus primarily on algorithmic problems, where, for a specific input, machine learned algorithms provide an output that is wrong.

Consistent with case studies and the theoretical risks reported in the prior literature (eg, Refs [5–8]), we observed problems involving algorithmic errors. However, by examining all problems associated with ML devices, we found that safety problems are much more than algorithms. We identified problems in all stages of ML device use (Figure 2), most of which involved the data processed by ML devices; however, problems with the way ML devices were used and for what purpose were proportionally higher in the events involving harm.

Most events involved problems with data acquisition. These included 15 events where poor-quality data were acquired, including known lesions not visible or the presence of artifacts in scans, which could in turn impact the validity of any algorithmic outputs based on it. Likewise, erroneous input from users contributed to incorrect results. These problems commonly concerned errors in data input, selected settings, or calibrations, or use that differed from that expected by the device.

The contraindicated use of devices by consumers demonstrates the potential for increased risk of harm when ML devices are consulted for making healthcare decisions but the results are misunderstood. Five events involving a consumer facing EKG capable of detecting arrhythmias, described a failure to detect myocardial infarction, a condition the device was neither indicated nor capable of detecting. Yet, consumer comments suggested they had considered device results when deciding whether to seek emergency medical care (see Box 4).

While contraindicated use mostly affected consumers, clinicians were not immune. Clinician use of an insulin dosing device resulted

in hypoglycemia when the recommended dose was administered without the recommended carbohydrates. The importance of adhering to the indications for use was highlighted in a recent FDA letter to healthcare providers about the *Intended Use of Imaging Software for Intracranial Large Vessel Occlusion* (LVO).[33] The letter outlines the FDA's concern regarding evidence "providers may not be aware of the intended use of these devices," and that contraindicated use risks "misdiagnosis resulting in patient injury or death."[33] While our MAUDE search list included known ML devices for detecting LVO, we did not observe any related events.

For consumers, ML devices are a double-edged sword, providing access to capabilities they do not possess, such as detecting cardiac arrhythmias from EKG. However, consumers simultaneously lack the expertise to fully understand what the results communicate, for example, that normal sinus rhythm does not exclude infarction, and device limitations, such as Lead-I EKGs being incapable of detecting infarctions. So, while device labeling indicates the device does not detect heart attack, consumers may not fully appreciate the difference between arrhythmias and infarctions. Likewise, it is easy to see how these nuances may not be at the forefront of someone's mind while experiencing concerning cardiac symptoms and faced with the decisions of whether they should be sufficiently worried to call an ambulance. Use of ML devices by nonexperts, including consumers, carries greater risks, particularly the risks associated with improper use or for improper purposes. However, the FDA's letter[33] indicates clinicians with domain expertise are also not immune to contraindicated use.

Implications

There are 3 broad implications for safe implementation and use of ML devices which are equally applicable for users, manufacturers, researchers, regulators, and policy makers.

A whole-of-system perspective

Firstly, there is the importance of focusing on the entire system of use. Our analysis reveals problems in all stages of ML device use (Figure 2), including the quality of data processed by devices, how devices are used and what they are used for. Algorithms also contributed to events, indicating that research into the pitfalls of ML is indeed well-placed.[5,6] However, as ML is increasingly implemented into real-world practice, the focus needs to shift from algorithms to the overall system, and how data capture hardware, software, and users can affect data quality.

Use problems demonstrate the need for a whole-of-system perspective. Manufacturers show awareness of whole-of-system factors, such as the risk of consumers relying on arrhythmia detection results in deciding whether to seek emergency treatment. However, their responses to events in attributing the cause to use error appears to dismiss the contribution of the device, "the device likely had no malfunction, did not cause or contribute to the [event], and the incident was a result of user error," "internal testing show there is no issue with the [device] function… inaccurate scans may be attributed to training or improper scanning of the patient," and "event was not caused by a malfunction of the [device]… The system and software functioned as designed. This was a use error related event."

Encouragingly, 2 reports from a radiotherapy planning device manufacturer, described proactive measures, "the device functions as designed, but usability improvements will be considered to prevent similar use errors in the future." However, the absence of such remarks in MAUDE reports does not preclude similar actions by other manufacturers. A whole-of-system approach requires an expansion of the boundary from use of devices as specified in the instructions to building in resilience for real-world use, such as greater tolerances or safety measures to prevent user mistakes as well as poor quality or unexpected data.

Importance of the human-ML medical device interaction

Secondly, there is the importance of the human–device interaction and how those interactions inform decisions and actions. Despite only comprising 7% of events, use problems were 4 times more likely to result in harm compared to device problems. This finding is consistent with previous research on safety events involving health IT where human factors issues were proportionally higher in the events involving patient harm.[32]

The relationship between user and device is established in the approved indications for use. In our previous analysis of clinician-facing ML devices most devices providing diagnostic or treatment recommendations were *assistive*,[3] requiring users to confirm or approve ML recommendations and be responsible for outcomes. These devices are characterized by indicated caveats, such as, "The clinician retains the ultimate responsibility for making the pertinent diagnosis based on their standard practices" and "Should not be used in lieu of full patient evaluation or solely relied on to make or confirm a diagnosis" (eg, Ref. [3]). Other devices functioned more autonomously providing clinicians with either information they may incorporate into their decision-making or make triage decisions to expedite reading of time sensitive cases. However, these devices do not replace clinician review and screening, rather they aim to increase accessibility where patients with positive findings are referred to consultant physicians for assessment and management.[3]

Safe implementation of ML medical devices into IT infrastructure

Thirdly, there is the need for safe implementation that considers how ML devices integrate into IT infrastructure. ML devices, especially those as a software device, are reliant on data and integration with IT infrastructure and other systems. For example, one event described problems importing CT scans from the radiological information system into a radiotherapy planning system.

ML adjacent medical devices

Finally, we noted the presence of many *ML adjacent devices* in the MAUDE search results. These devices do not contain ML, but instead acquire or manage data that can be processed by separate ML applications, which are not necessarily classed as medical devices. For example, CT scanners and PACS (picture archiving and communication system) are ML adjacent as they provide data which can be processed by ML devices detecting LVO. The most common example in the MAUDE search results were blood glucose meters. However, their regulatory approvals show no evidence of ML utilization, instead, glucose readings obtained could be uploaded to separate ML applications for analysis. Having already highlighted the importance of data to safety, these ML adjacent devices form part of the context, workflows, and infrastructure needed to ensure safe implementation and use.

Limitations

The safety events we analyzed are limited to the events reported to MAUDE, most of which come from mandatory reporters and likely favor events meeting regulatory thresholds for reporting. Reports are also likely to favor more common devices. Likewise, there may

be a selection bias whereby reporters favor highly salient events, while others may be under reported.[34] Accordingly, FDA reports may not be representative of all safety problems that can occur with ML devices.

Moreover, event reports identify problems from the user perspective but are not able to identify precisely where in the chain of algorithms problems arose. Reports were authored by those with an interest in the device and reflect the expertise of the reporter and are written from their perspective. Most reports were submitted by device manufacturers, while others were authored by users. Furthermore, reports only provide a "snapshot" of the events, we were limited to analyzing them as reported and it was not possible to independently verify their accuracy or determine the root cause of events. Nevertheless, post-market surveillance and reporting provides the most comprehensive dataset of adverse events and is one that is relied on by regulatory agencies, such as the FDA for identifying new aspects of known safety problems, and new, unforeseen problems for the first time.

Lack of gold standard reporting on ML utilization by medical devices

Another limitation for research sampling ML devices is the absence of definitive reporting by the FDA on whether devices utilize ML.[3,9] The FDA's own published list of ML medical devices is not based on information captured by the pre-market review process, but rather was compiled from public facing FDA data and external sources (Supplementary Appendix SA), most of which were included in our search list.[35] We overcame this limitation by screening devices for ML utilization, which resulted in the exclusion of a sizable number of devices identified in the literature as ML-enabled. This provides confidence that the events analyzed are indeed indicative of ML-enabled devices.

Despite efforts to catalog medical devices using ML,[9,23,24,26] an accounting for exactly what ML does within these devices is notably absent. Medical devices are more than a *singular* algorithm, instead, they are end-to-end products comprising the multiple hardware and software components needed by devices to perform tasks. Devices comprise multiple algorithms, each performing a specific function and utilizing appropriate methods, only a small portion of which are ML. The problem is compounded by the lack of specificity in manufacturer and regulator reporting. The FFRCT manufacturer offers more details than most, describing, "advanced deep learning (AI) methodology to precisely extract coronary anatomy." That ML output is used by "physics-based computational models [to] assess pressure and flow changes in coronary arteries." Output of the physics computation model are used to calculate FFRCT. Here the function of ML is limited to the extraction of coronary anatomy.

CONCLUSION

The promise of ML for healthcare is held back by the challenges of implementation. Unresolved concerns around safety and limited implementation in real-world practice have meant research has been largely based on theory and case studies focusing on ML algorithms. Our analysis of safety events reported to the FDA provides the first systematic examination of the problems and consequences associated with the use of ML medical devices for diagnosis, treatment, and prevention of disease in the real world.

ML and algorithm failures indeed contributed to safety events, however, the safety problems observed involve much more than algorithmic errors. While issues with data acquisition was the most frequently observed problem, problems with the way in which devices were used and for what purpose were more likely to result in patient harm. Ensuring safety of ML devices requires a shift of focus from ML algorithms to the whole-of-system, the human–ML device interaction, and safe implementation into workflows and with IT infrastructure.

FUNDING

NHMRC Centre for Research Excellence (CRE) in Digital Health (APP1134919) and Macquarie University.

AUTHOR CONTRIBUTIONS

DL and YW conceived this study, and designed and conducted the analysis with advice and input from FM and EC. DL and YW screened ML devices for inclusion, performed data extraction and classification, in consultation with and review by FM. Disagreements were resolved by consensus with DL, YW, and FM. DL drafted the article with input from all authors. All authors provided revisions for intellectual content. All authors have approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of Andre Nguyen who undertook background research tasks to support this study; Matt Millett who assisted in identifying devices used by consumers from the search list; and Anindya Susanto (AS) who provided clinical expertise.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

All data relevant to the analysis are reported in the article.

REFERENCES

1. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Hum Factors* 1997; 39 (2): 230–53.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25 (1): 44–56.
3. Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. *BMJ Health Care Inform* 2021; 28 (1): e100301.
4. Federal Food, Drug, and Cosmetic Act, Section 201(h), 21 U.S.C. § 321(h) (2022) United States of America. https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title21-section321&num=0&edition=prelim.
5. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. *ArXiv* 2016; arXiv:1606.06565, preprint: not peer reviewed. doi: 10.48550/arXiv.1606.06565.

6. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019; 28 (3): 231–7.
7. Wong A, Otles E, Donnelly JP, *et al.* External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021; 181 (8): 1065–70.
8. Wong A, Cao J, Lyons PG, *et al.* Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. *JAMA Netw Open* 2021; 4 (11): e2135286.
9. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020; 3 (1): 118.
10. Sujan M, Furniss D, Grundy K, *et al.* Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* 2019; 26 (1): e100081.
11. Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it. *Acad Radiol* 2020; 27 (1): 132–5.
12. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022; 4 (5): e384–97.
13. Runciman WB, Edmonds MJ, Pradhan M. Setting priorities for patient safety. *Qual Saf Health Care* 2002; 11 (3): 224–9.
14. Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. *J Am Med Inform Assoc* 2017; 24 (2): 246–50.
15. Medical Device Reporting, 21 C.F.R. § 803, 2014. https://www.ecfr.gov/current/title-21/chapter-I/subchapter-H/part-803. Accessed April 2023.
16. U.S. Food & Drug Administration. MedWatch: The FDA Safety Information and Adverse Event Reporting Program. 2022. https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program. Accessed April 2023.
17. Magrabi F, Ong M-S, Runciman W, Coiera E. Using FDA reports to inform a classification for health information technology safety problems. *J Am Med Inform Assoc* 2012; 19 (1): 45–53.
18. U.S. Food and Drug Administration. Medical Devices Database Search: Device Classification Under Section 513(f)(2)(De Novo). https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/denovo.cfm. Accessed April 2023.
19. U.S. Food and Drug Administration. Medical Devices Database Search: 510(k) Premarket Notification. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm. Accessed April 2023.
20. U.S. Food and Drug Administration. Medical Device Database Search: Premarket Approval (PMA). https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm. Accessed April 2023.
21. Hamamoto R, Suvarna K, Yamada M, *et al.* Application of artificial intelligence technology in oncology: towards the establishment of precision medicine. *Cancers* 2020; 12 (12): 3532.
22. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. *Cancer Cell* 2021; 39 (7): 916–27.
23. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. *Lancet Digit Health* 2021; 3 (3): e195–203.
24. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021; 27 (4): 582–4.
25. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021; 31 (6): 3797–804.
26. American College of Radiology DSI. AI Central. 2021. https://aicentral.acrdsi.org/. Accessed August 2021.
27. U.S. Food & Drug Administration. Device API Endpoints Adverse Events Overview. n.d. https://open.fda.gov/apis/device/event/. Accessed April 2023.
28. U.S. Food and Drug Administration. MAUDE—Manufacturer and User Facility Device Experience. 2021. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm. Accessed April 2023.
29. U.S. Food & Drug Administration. General Controls for Medical Devices. 2018. https://www.fda.gov/medical-devices/regulatory-controls/general-controls-medical-devices. Accessed April 2023.
30. U.S. Food & Drug Administration. Class I and Class II Device Exemptions. 2022. https://www.fda.gov/medical-devices/classify-your-medical-device/class-i-and-class-ii-device-exemptions. Accessed April 2023.
31. Magrabi F, Ong M-S, Runciman W, Coiera E. An analysis of computer-related patient safety incidents to inform the development of a classification. *J Am Med Inform Assoc* 2010; 17 (6): 663–70.
32. Magrabi F, Baker M, Sinha I, *et al.* Clinical safety of England's national programme for IT: A retrospective analysis of all reported safety events 2005 to 2011. *Int J Med Inform* 2015; 84 (3): 198–206.
33. U.S. Food & Drug Administration. Intended Use of Imaging Software for Intracranial Large Vessel Occlusion—Letter to Health Care Providers. 2022. https://www.fda.gov/medical-devices/letters-health-care-providers/intended-use-imaging-software-intracranial-large-vessel-occlusion-letter-health-care-providers. Accessed April 2023.
34. Runciman WB, Kluger MT, Morris RW, Paix AD, Watterson LM, Webb RK. Crisis management during anaesthesia: the development of an anaesthetic crisis management manual. *Qual Saf Health Care* 2005; 14 (3): e1.
35. U.S. Food & Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed February 2022.