Advancing Compassionate AI in Healthcare Through Safety Evaluations of AI-Enabled Medical Devices

Abstract:

Artificial intelligence in medical devices is often introduced with compassionate intent: to extend clinicians' capabilities, improve diagnostic accuracy, and enhance patient care. Yet compassion also requires vigilance. Emerging work on compassionate AI emphasizes that technical ambition must be paired with governance structures that protect patient well-being, promote fairness, and sustain trust. In high-stakes settings like healthcare, this means empirically examining not only what AI promises to do, but how it performs once deployed. Drawing on the Food and Drug Administration (FDA) data on device regulatory approvals and recalls, this study evaluates how AI integration affects post-clearance safety outcomes. We combine a difference-in-differences design with context-tailored matching to compare recall patterns across generational histories of AI-enabled and non-AI devices. Across one- to three-year windows after clearance, AI-enabled devices exhibit a 13.5 to 18.2 percentage-point higher probability of recall. These elevated risks are partly linked to failures in AI clinical functions supporting diagnosis and monitoring, as well as technical functions involving data generation. These results contribute to the broader effort to shape regulatory structures that enable compassionate AI in healthcare. Compassionate AI is not only about designing systems that aim to help; it is also about ensuring that institutions anticipate where harm may arise and intervene early. By identifying device categories where early-stage risks are most acute, our findings highlight the value of more focused oversight during the initial years following clearance. Such targeted scrutiny supports AI systems that advance patient outcomes while preserving room for responsible innovation. In doing so, this work offers empirical grounding for regulatory approaches that align with the principles of compassionate AI –approaches that recognize both the promise of AI-enabled medical devices and the responsibility to ensure they operate safely in practice.

Keywords: Artificial Intelligence, Medical Device, Medical Safety, FDA, 510(k)

1. Introduction

Artificial intelligence (AI) is increasingly embedded in medical devices, reshaping how clinical information is generated, interpreted, and acted upon. Between 2011 and 2023, the number of AI–enabled medical devices (referred simply as "AI devices" in this paper) cleared by the U.S. Food and Drug Administration (FDA) rose from 2 to 223, recording an average annual growth rate of 57.06% (FDA 2025a). These devices employ diverse AI functions, such as automatic diagnosis, image processing, personalized treatment, and robotic assistance, to augment human clinical judgment and decision-making (He et al. 2019, Topol 2019, Benjamens et al. 2020). The rapid adoption of these AI functions in medical devices reflects a broader aspiration toward compassionate AI – devices and systems designed to extend clinicians' capabilities, anticipate patient needs, and ultimately improve patient care and outcomes.

However, compassion in healthcare also requires vigilance, supported by regulatory and monitoring frameworks that foster innovation while mitigating patient-safety risks. As AI systems become more autonomous and technologically complex, concerns have grown about their safety and reliability when introduced into medical devices (Warraich et al. 2024, Lee, Patel, et al. 2025). Although AI promises to reduce adverse outcomes (Lee, Kramer, et al. 2025) – for example, the AI-supported Sepsis ImmunoScore has demonstrated improved accuracy and timeliness in identifying septic patients, saving more lives (Bhargava et al. 2024) – AI devices have also been implicated in serious safety failures. For instance, the recalls of the AI-assisted Medtronic Neurovascular device, linked to 13 injuries and 4 deaths due to improper attachment and potential failures in detecting large-vessel occlusion, underscore the stakes of AI-related errors (FDA 2025d). Such recalls also impose substantial financial burdens on manufacturers (Sedgwick 2023, Christensen 2024).

Evidence also shows that these performance risks can disproportionately affect vulnerable populations; pulse oximeters, for example, have shown higher error rates in Black patients, and AI-based dermatology tools have been found to underdetect skin cancer in people with darker skin (Mundasad 2024). As such, uncertainty persists about how AI integration affects medical device safety and how the FDA's 510(k) regulatory pathway for device clearance should adapt to this evolving technological landscape.

Understanding safety issues of AI use in medical devices is essential not only from a technical standpoint – as discussed in the examples above, where insights can guide improvements in AI performance across clinical tasks – but also from a compassionate perspective. Compassionate AI must support, rather than undermine, human judgment, enabling clinicians to interrogate, contextualize, or override algorithmic outputs when necessary (Gaube et al. 2021, Jussupow et al. 2021, Yin et al. 2025). This aligns with socio-technical perspectives in the information systems literature, which emphasize that system outcomes depend not only on algorithmic accuracy but also on how technologies are embedded in human decision-making processes (Sarker et al. 2019, Rai 2020). In high-stakes clinical environments, compassion requires ensuring that AI systems enhance – not compromise – patient safety.

In response to these concerns, this study (1) empirically investigates how integrating AI into medical devices affects the probability of post-clearance serious recalls[1] and (2) identifies the specific AI clinical functions and technical functions that drive these effects. By examining how AI integration shapes device post-market safety, this study contributes to ongoing conversations about the governance of compassionate AI in healthcare. The empirical findings can provide guidance for a regulatory framework that preserves AI's compassionate promise while strengthening oversight of its clinical and technical components.

Using comprehensive FDA data on device clearance and recalls, we employ a Difference-in-Differences (DiD) design with a context-specific technique that locates appropriate treatment and control groups to analyze the effect of AI integration on medical device safety. At the high level, this approach leverages two comparisons of device recalls: (1) a cross-sectional comparison between the lineage of AI devices and non-AI devices (i.e., devices that do not integrate AI), and (2) a temporal comparison within the lineage of each device (either AI or non-AI).

Specifically, the FDA's list of AI devices (FDA 2025a) provides an authoritative basis for distinguishing AI devices from non-AI devices in our dataset. We obtain the lineage of each device by using the "substantial-equivalence" relationships that underpin the FDA's 510(k) clearance pathway (FDA 2024).

[1] Serious recalls are Class I and II recalls classified by the FDA, which are the situations in which the use of, or exposure to, a recalled device can cause serious patient harm or even death.

This pathway is central to U.S. medical device regulation, as it clears more than 90% of all new devices and 97% of all AI devices in the U.S. (Institute of Medicine 2011, Rathi and Ross 2019, FDA 2025a). Under this pathway, a newly cleared device (referred to as the applicant device) must demonstrate substantial equivalence in intended use (i.e., the purpose for which the device is used), technological characteristics (e.g., design, software/hardware features, compositions, materials), and performance to one or more previously cleared devices already in the market and cite these earlier devices (referred to as predicate devices or predicates) in its 510(k) submission (Basu and Hassenplug 2012, Vo et al. 2020, Zhu et al. 2025). Iteratively linking the most relevant substantially equivalent device (i.e., the *primary predicate*) across successive generations enables the construction of a predicate-history chain for each device. We refer to these citation-based lineages as "predicate chains," which form the basis of our empirical design.

For predicate chains that include AI devices, we define the *focal AI device* as the first AI device to appear chronologically. This device is of primary interest to us, and such predicate chains with an focal AI device constitute our treatment group (hereafter, "AI chains"). In our study design, we restrict attention to the portion of each AI chain up to and including this focal AI device. This restriction ensures clean temporal comparisons and prevents treatment contamination as extending the chains beyond the focal AI device may allow later devices to alternate between AI and non-AI status, confounding the identification of AI's effects.

To define an appropriate control group, we identify, for each focal AI device in the treatment group, a corresponding set of non-AI devices that cite the same *primary predicate* (FDA 2018). We define this set of non-AI devices as *matched focal non-AI devices*. This step yields a group of AI and non-AI devices that share an identical chronological sequence of predecessors up to and including the primary prediate (by definition, all these predecessors are non-AI devices). The predicate chains with the matched focal non-AI devices constitute the "non-AI chains," which serve as the control group.

The focal AI device on an AI chain divides the chain into pre-AI (excluding the focal AI device) and post-AI (including the focal AI device) periods. Similarly, the matched focal non-AI device on a non-AI chain divides the chain into pre-AI (excluding the focal non-AI device) and post-AI (including the focal non-AI device) periods. Because every device in a predicate chain is cleared by the FDA on the basis of

substantial equivalence – ensuring a strict standard that later-cleared products remain functionally and technically very close to earlier generations – those earlier devices in the chain can serve as credible proxies or stand-ins for the focal (or matched) devices at earlier points in time within a DiD framework.

Our recall outcomes are measured for each individual device on the chains based on whether it is recalled within defined post-clearance windows (e.g., one year, two years). In other words, recall outcomes are assessed at the chain-device level (e.g., a focal AI device in an AI chain and the devices on its predicate chain that serve as stand-ins during the pre-AI period) rather than solely at the device level. This assessment captures the fact that devices cleared through the 510(k) pathway can inherit similar technological and regulatory risks from their predicates. Without accounting for these lineage-based risks when comparing recall probabilities between AI and non-AI devices, the effects of AI-related safety risks can be conflated with those inherited from a device's predicate chain. Section 4.1 provides full methodological details.

Our analysis, based on 39,505 applicant devices cleared between 2011 and 2023, shows that integrating AI into medical devices significantly increases the likelihood of recalls shortly after clearance. Relative to the pre-AI period, focal AI devices on AI chains exhibit a 13.5- to 18.2-percentage-point higher recall probability in the post-AI period compared to their matched focal non-AI devices on non-AI chains, when considering one- to three-year recall windows after clearance. The effect is largest within a two-year window and becomes statistically insignificant once the recall window extends beyond three years. These findings remain robust across alternative constructions of predicate chains, different estimation models, the strictness of the procedure for identifying the control group, and alternative adverse-outcome measures.

Further analyses indicate that the elevated recall probability associated with AI devices is partly attributable to issues arising from specific AI functions. In particular, higher risks are linked to AI clinical functions supporting patient assessment (e.g., diagnosis and monitoring) and AI technical functions involved in data generation (e.g., image enhancement and synthetic data creation). We also document heterogeneity across AI function types, AI feature categories, medical specialties, and recall reasons, offering a more granular understanding of where and why AI-related safety risks emerge.

These findings yield several regulatory and managerial implications. Given evidence that AI use may

increase recall probability, stakeholders should remain attentive to the safety risks introduced when incorporating advanced AI capabilities into medical devices. Manufacturers, in particular, should prioritize improving AI-supported diagnostic and monitoring functions, which appear to be key drivers of elevated recall risk. Clinicians, likewise, should integrate AI-based assessments with their own expertise rather than relying solely on algorithmic outputs. Moreover, because our analyses suggest that AI-generated data can contribute to device recalls, manufacturers should invest in improving the quality and reliability of such data (e.g., through more robust "digital twin" representations or richer simulation models), while clinicians should combine AI-generated data with real-world clinical information to support more reliable decision-making. Taken together, the results highlight the need for heightened regulatory attention to AI devices during the first two to three years following market clearance – when recall risks are most pronounced – while also indicating that these risks diminish over time, suggesting that AI devices ultimately achieve safety performance comparable to non-AI devices. These insights can inform the FDA's efforts to refine oversight by prioritizing early-stage monitoring of specific types of AI devices and strengthening evaluations of AI functions involved in clinical assessment and data generation.

This research also makes several theoretical contributions. Most importantly, by linking AI use to human decision-making and medical safety, it contributes to the growing literature in healthcare information systems that calls for a shift from performance-centred to human-centered evaluations of medical AI systems (Gaube et al. 2021, Jussupow et al. 2021, Yin et al. 2025). In doing so, it advances our understanding of how AI can be designed and deployed in ways that are not only intelligent but also safer and more aligned with the fundamental ethos of medical care. Additionally, our finding that AI use may increase the probability of medical device recalls contributes to the governance of compassionate AI use in healthcare by underscoring that careful integration of AI – particularly in its clinical assessment and data-generation roles – is essential for ensuring the safety of medical AI devices.

2. Related Work

2.1. AI-enabled Clinical Decision Support

AI-enabled clinical decision support (AI-CDS) tools aims to enhance clinicians' diagnostic, prognostic, and

treatment decisions by learning from large-scale clinical data and embedding predictions or recommendations into clinical workflows. Across the medical literature, a consistent conclusion is that accuracy alone is insufficient; effective AI-CDS must produce actionable outputs, align with clinical objectives, and be supported by implementation and monitoring practices that mitigate safety risks (Obermeyer and Emanuel 2016, Rajkomar et al. 2019). Recent work further underscores the importance of evaluation, regulation, and post-deployment oversight, given concerns about performance drift, subgroup disparities, and unclear liability when AI enters real-world care pathways (Angus et al. 2025).

Information systems research explains why technically strong AI-CDS may still struggle in practice: clinical decisions emerge from human-AI interaction rather than algorithmic performance alone. A robust finding is algorithm aversion – decision makers often discount or reject AI advice even when it is accurate (Turel and Kalhan 2023). This tendency is amplified in healthcare, where clinicians bear ultimate responsibility for patient safety. Yet aversion is not inevitable. Allowing users to adjust AI recommendations increases reliance on imperfect algorithms (Dietvorst et al. 2018), and subsequent studies show that clinicians' responses depend heavily on perceived control, transparency, and role alignment rather than accuracy per se (Berente et al. 2021, Jussupow et al. 2024). These insights help explain why clinicians may override AI-CDS: clinical expertise and professional identity are central to medical work, and systems framed as supportive rather than substitutive are more readily adopted (Hou et al. 2024). More recent research highlights that willingness to delegate to AI is shaped by psychological factors such as loss aversion and framing, reinforcing that trust in AI is behaviorally constructed rather than purely performance-driven (Bockstedt and Buckman 2025).

Taken together, prior work suggests that the success of AI-CDS hinges less on technical accuracy and more on whether systems are designed to collaborate with clinicians and support decision-making for safer patient outcomes. A critical but understudied way to evaluate this collaborative dimension is to compare the safety outcomes of AI-augmented versus non-AI-augmented clinical care. Empirical evidence on this question – particularly for AI-enabled medical devices – remains limited. AI-enabled medical devices are an integral part of compassionate AI, aiming to enhance patient well-being and support safer, more

responsive care. Our study addresses this gap by examining serious recalls of AI versus non-AI medical devices, offering new insight into the safety implications of AI-supported clinical care and decision-making.

2.2. AI Safety in Healthcare

As AI development accelerates, debates over its safety and reliability in healthcare have intensified (Lenharo 2024), yet clear conclusions remain elusive. On one hand, AI has demonstrated substantial promise in clinical tasks such as disease diagnosis – for example, improving sepsis detection accuracy and enhancing patient safety (Bhargava et al. 2024). On the other hand, diagnostic failures can be catastrophic: recent evidence shows that AI systems failed to recognize severely impaired in-hospital patients in 66% of cases (Nong et al. 2025). A broader review of 53 studies similarly finds that AI-powered decision-support tools, such as drug-dosing alerts and clinical alarms, can reduce errors and improve safety when rigorously validated (Choudhury and Asan 2020). Yet the absence of standard benchmarks and consistent reporting continues to hinder widespread adoption, given the potential for safety risks (Ratwani et al. 2024).

The integration of AI into medical devices reflects this same tension. While AI-enabled devices have the potential to improve patient safety, emerging evidence suggests that they may also introduce new risks. For instance, although publicly traded and private manufacturers produce AI devices in roughly equal numbers, the former accounts for more than 98% of recalled AI device units, with recall rates nearly 30 times higher than those of the latter (Lee, Patel, et al. 2025, Lee, Kramer, et al. 2025). Regulatory frameworks have not kept pace: current policies for governing AI devices and reporting safety issues remain incomplete (Benjamens et al. 2020, Warraich et al. 2024, FDA 2025b, Handley et al. 2024).

To support more effective regulation, researchers have called for improving the quality of FDA adverse event data, developing post-market regulatory schemes tailored to AI devices, and implementing unbiased performance monitoring and periodic audits (Everhart and Stern 2022, Babic et al. 2025). Despite growing attention to medical device safety, empirical evidence examining recalls specifically from the perspective of AI integration remains limited. One exception is Lee, Kramer, et al. (2025), which examines recalls among 112 AI devices and finds that recalls are generally rare but occur early after clearance and

predominantly involve devices lacking clinical validation. While this research provides preliminary insights, it has three limitations. First, the study's scope is narrow, examining only 112 AI devices without comparing them to a control group of matched non-AI devices. Consequently, the study does not determine whether observed recall patterns are specific to AI integration or reflect broader patterns in medical device safety. Second, the analysis is primarily descriptive rather than causal, documenting patterns without establishing whether AI integration itself drives recall risk or whether recalls stem from other device characteristics correlated with AI adoption. Third, the study does not examine which specific AI functions contribute to safety issues, limiting actionable insights for regulators and manufacturers.

Our research addresses these key limitations. We compare AI devices to non-AI devices, employing a rigorous difference-in-differences design to establish causal effects, and decompose effects by specific AI clinical and technical functions. This approach enables us to move beyond description to causal inference about AI's impact on device safety and to identify the specific AI capabilities that elevate recall risk.

3. Data and Study Sample

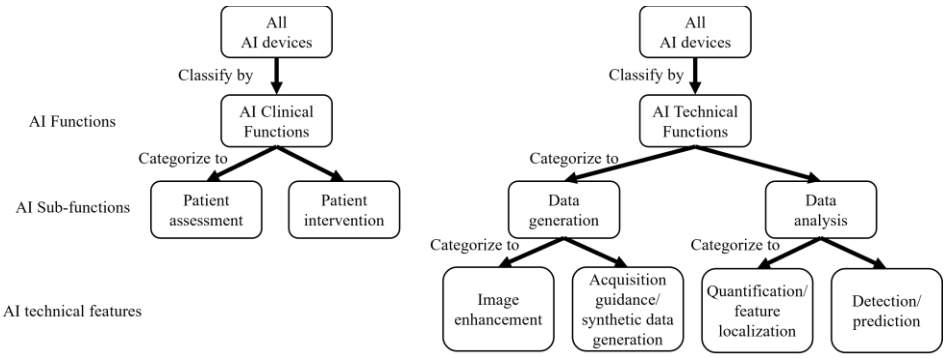
Our study leverages six publicly available datasets, curated using custom data mining techniques. The first dataset comprises 510(k) summary submission documents, submitted by manufacturers to the FDA for clearance of each device, from 2003 to 2023 (FDA 2023b). These documents provide the predicate information for each newly cleared device, allowing us to construct predicate chains and match devices based on citation links – both essential for identifying the causal effect of AI use on device safety (details in Sections 4.1.1 and 4.1.3).

The second dataset is the FDA 510(k) device clearance records over the same period (2003 – 2023), which contain basic device information, including device clearance date, name, manufacturer, product code, and medical specialty (FDA 2023c). This data provide control variables for the estimation models.

The third dataset is a list of AI devices cleared by the FDA from 2011 to 2023 (FDA 2025a). This list allows us to identify AI devices using the FDA as the authoritative source, reducing identification errors. The starting year for this list is 2011 because it marks the year when the number of AI devices began to increase, from fewer than 2 devices per year (Warraich et al. 2024).

The fourth dataset is a taxonomy of AI use in medical devices developed by Singh et al. (2025). In this work, the authors introduce a taxonomy that categorizes all FDA-cleared AI devices as of the end of 2024 into several subcategories based on AI clinical function (clinical assessment, such as diagnosis or monitoring, versus clinical intervention, such as surgery or treatment guidance) and AI technical function (AI used for data generation, such as image enhancement or synthetic data simulation, vs. data analysis, such as triage or prediction). Figure 1 illustrates how this taxonomy categorizes AI devices by their AI functions and subfunctions, and Table 1 provides illustrative examples of devices with different combinations of AI subfunctions.[2] Such a taxonomy from medical practitioners allows us to decompose AI use in medical devices by clinical and technical functions, thereby enabling a more nuanced understanding of the effects of AI use on medical device safety and the underlying functionality-level mechanisms.

Figure 1. Summary of the Taxonomy for AI Use in Medical Devices (Singh et al. 2025)



The fifth dataset is the device recall records from 2003 to 2024 (FDA 2023a), which provides our main outcome variable. Considering the 2024 data for the outcome variable provides at least a one-year recall window for devices cleared in 2023 (details are provided in Section 4.1.4).

The final dataset is the Manufacturer and User Facility Device Experience (MAUDE) history, which records numeric counts of user-generated complaints of device adverse events (e.g., deaths, injuries, malfunctions) between 2003 and 2024 (FDA 2023d). This dataset provides alternative outcome variables (i.e., user-reported adverse events) in addition to recalls to support the robustness of the main results.

[2] To guarantee enough sample size for the categories of different AI technical features, we have combined similar features to a single category, including combining “acquisition guidance” and “synthetic data generation”, and cominbing “triage”, “detection”, “diagnoisis”, and “predictive” (calling it “detection/prediction”).

Table 1. Illustrative Examples of AI-Enabled Medical Devices by Clinical and Technical Function

Device ID	Clear year	Device name	Manuf-acturer name	Medical specialty	AI clinical function	AI technical function	Description of Compassionate AI Use
K230365	2023	Sonio Detect	Sonio	Radiology	Patient Assessment	Data Generation and Analysis	The device uses AI in supporting image analysis and interpretation, assisting clinicians in fetal ultrasound exams and improving patient care.
K232699	2023	Low Ejection Fraction AI-ECG Algorithm	Anumana, Inc.	Cardiov-ascular	Patient Assessment	Analysis	The device uses AI in aiding the screening of patients at risk of heart failure, supporting early detection and management of serious conditions to reduce patient suffering.
K211379	2021	syngo.via RT Image Suite	Siemens Medical Solutions USA Inc.	Radiology	Patient Intervention	Data Generation	The device uses AI in enhancing the reliability of radiologic image data and treatment planning, supporting more effective radiation therapy and reducing patient harm.
K191370	2019	DreaMed Advisor Pro	DreaMed Diabetes Ltd	Clinical Chemistry	Patient Intervention	Data Analysis	The device uses AI in supporting diabetes management by analyzing individualized patient data to generate tailored, clinically relevant recommendations for a patient.

Our study sample contains 39,505 unique applicant devices cleared between 2011 and 2023 (796 AI devices vs. 38,709 non-AI devices; the overall fraction of AI devices is 2.01%). Table 2 provides the summary statistics of the number of AI vs. non-AI devices by medical specialty.

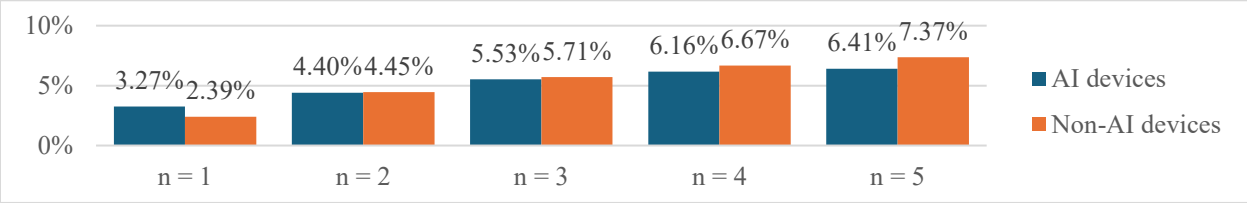
Table 2. Distribution of AI vs. Non-AI Devices by Medical Specialty

Medical specialty	All devices	AI devices	Non-AI devices	AI adoption rate
Radiology	5058	627	4431	12.40%
Hematology	222	11	211	4.95%
Pathology	85	2	83	2.35%
Cardiovascular	4435	80	4355	1.80%
Neurology	1782	24	1758	1.35%
Ophthalmic	826	8	818	0.97%
Gastroenterology & Urology	2031	11	2020	0.54%
Clinical Chemistry	1353	7	1346	0.52%
Microbiology	911	4	907	0.44%
Anesthesiology	1399	6	1393	0.43%
Ear, Nose, & Throat	528	2	526	0.38%
Physical Medicine	741	1	740	0.13%
General & Plastic Surgery	4405	5	4400	0.11%
Obstetrics/Gynecology	974	1	973	0.10%
General Hospital	3997	3	3994	0.08%
Orthopedic	7160	3	7157	0.04%
Dental	3008	1	3007	0.03%
Immunology	332	0	332	0%
Clinical Toxicology	258	0	258	0%
Total	**39505**	**796**	**38709**	**2.01%**

Notes: AI adoption rate is calculated as (AI-enabled devices / All devices) × 100%

We explore and compare the recall rates of AI devices with those of non-AI devices across different recall windows after device clearance. As shown in Figure 2, the recall rate differences across the five recall windows are not statistically significant ($p = 0.88$, t-test assuming two-tailed equal-variance). However, the raw data show that the recall rates of AI devices are slightly higher than those of non-AI devices when considering shorter recall windows following market approval. This observation indicates that AI may yield different effects on device recalls for different recall windows, and further motivates the need for an empirical analysis. For example, the Siemens SOMATOM computed tomography (CT) scanner systems were recalled within one year after clearance because specific installed CT systems with AI-driven software could cause workflow interruptions[3]. To more accurately estimate the effect to inform better regulatory policies, we need to rely on rigorous econometric models, as discussed in the next section.

Figure 2. Devices Recall Rates within Different Numbers of Years After Device Clearance Year



Notes: *n* represents the number of years considered for the recall window after the device clearance year.

4. Empirical Methodology and Results

Estimating the effect of AI integration on medical device recall probabilities requires careful attention to potential endogeneity. A first concern arises from selection bias: manufacturers may selectively incorporate AI into devices – particularly software-driven or lower-risk products – where the expected benefits of AI (e.g., performance gains or cost advantages) are greater, because of AI's reliance on software and the lower cost of failure (He et al. 2019, Benjamens et al. 2020). Similarly, many mechanical, implantable, or disposable devices (e.g., surgical scissors, hip implants, catheters) are inherently incompatible with AI, rendering direct comparisons inappropriate. A second concern is that AI adoption may be systematically associated with certain medical specialties or intended uses, making it difficult to disentangle the effect of AI from specialty-specific recall patterns. In our sample of 796 AI-enabled devices, 88.82% fall within

[3] Recall report: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfres/res.cfm?id=189946

radiology (78.77%) or cardiovascular (10.05%) categories (Table 1), underscoring the specialty concentration. To mitigate these endogeneity concerns, we employ a Difference-in-Differences (DiD) design combined with context-specific matching to identify appropriate treatment and control groups.

4.1. Difference-in-Differences Setup

Our analyses focus on devices cleared through the 510(k) pathway, which relies on demonstrating applicant devices' substantial equivalence to previously cleared predicate devices – typically those with similar intended uses and technological characteristics (FDA 2024). This process creates chains of predicate citation histories (referred to as predicate chains) that trace the evolution of devices across successive generations. Given the "substantial equivalence" relationship, on each predicate chain, devices at earlier points serve as proxies (or stand-ins) for earlier generations of devices that appear later on the chain.

Our DiD framework exploits the fact that, with the emergence of AI, some of these chains fork: all devices on the chain remain non-AI devices until one point where a device may be cited by two descendant devices, one incorporating AI features (and becomes an AI device) and the other remaining non-AI. Such forks generate a panel structure, with the forking point dividing the chain into pre-AI and post-AI periods. Our empirical strategy compares changes in device recall probabilities in the pre-AI and post-AI periods on a predicate chain that transitions to AI devices, and contrasts these changes with those observed in a matched chain that remains non-AI throughout. The specifics of this empirical design are discussed below.

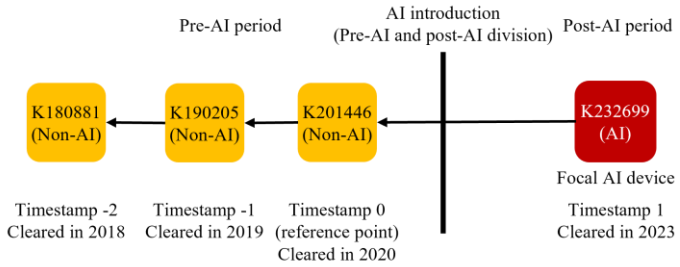
4.1.1. Predicate Chain and Panel Data Construction

The construction of predicate chains serves as the foundation of our DiD setup. Following Zhu et al. (2020) and Everhart et al. (2023), we extract predicate information from each device's 510(k) submission to construct predicate chains. Although applicant devices may cite multiple predicates, since 2011, the FDA has recommended identifying a single "primary predicate," defined as the device whose indications for use and technological characteristics most closely align with those of the applicant device (the primary predicate is the most relevant predicate), to support a timely and well-substantiated clearance decision (FDA 2018). When an applicant device lists multiple predicates, it can note that the "primary predicate is listed first." If no primary predicate is explicitly identified, the first predicate listed is conventionally treated as

the primary predicate. The construction of predicate chains is based on the primary predicate citation relationship between devices, which avoids peripheral or less informative citation branches. When predicate information is absent or the 510(k) document is unavailable[4], the chain terminates at that device. Within each chain, adjacent devices are substantially equivalent, and all devices are temporally ordered by their clearance dates, with each device corresponding to a distinct timestamp on the chain. Thus, the unit of observation is a device in a given chain. If a chain contains any AI devices, the first AI device to appear in temporal order is defined as the *focal AI device*, which marks the introduction of AI on the chain and divides the pre-AI and post-AI periods.

Figure 3 provides an example of a constructed predicate chain, on which the AI device K232699 is the focal AI device of the chain and its immediate predicate is K232699; K232699's immediate predicate is K201466; and so on. Each device is represented by a KNUMBER, a unique device identifier provided by the FDA, starting with the letter "K" followed by six digits. In the example, devices are temporally ordered along the chain by their clearance year, and each device corresponds to a timestamp on the chain (i.e., a relative temporal position. The timestamps on the chain are standardized using the timestamp of the last non-AI device in the pre-AI period (i.e., K201446) as timestamp 0 and incrementally increasing or decreasing by 1 for pre-AI and post-AI timestamps.

Figure 3. Predicate Chain Example



Notes: Each arrow represents a predicate citation relationship, with an applicant device pointing to a predicate device. Detailed information about the devices shown in the figure is provided in Appendix B.

The introduction of AI in specific devices on the predicate chain is arguably not related to pre-existing recall trends for two main reasons. First, the adoption of AI was largely driven by rapid, exogenous advances

[4] Some 510(k) documents are not available either because the documens are missing, or the device is too old to have a document on record, or the device is cleared through a different pathway, such as De Novo (FDA 2025c) or Premarket Approval (FDA 2019).

in general-purpose AI technologies beginning in the early 2010s (Haenlein and Kaplan 2019), coinciding with the subsequent surge in AI devices (Warraich et al. 2024, Windecker et al. 2025). Second, our review of all 510(k) submissions for AI-enabled devices using large language models (OpenAI 2025) reveals no indication that manufacturers introduced AI in response to poor safety performance in earlier generations. Additional details on this analysis are provided in Appendix A.

Across all constructed predicate chains, we identify 53,682 unique devices. These devices comprise the 39,505 devices in our study sample, cleared between 2011 and 2023, as well as 14,177 devices cleared prior to 2011 that do not serve as the focal (or matched) device (i.e., the most recently cleared device) of any constructed predicate chain. Among all 53,682 devices, 23,968 (44.65%) either specify a primary predicate or list only a single predicate; 19,455 (36.24%) list multiple predicates without designating a primary one; and 10,260 (19.11%) lack an accessible 510(k) document.

4.1.2. Treatment Group

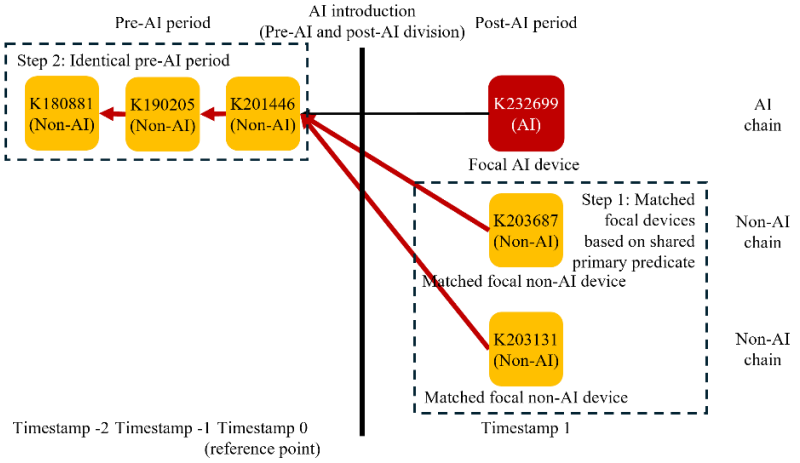
Our treatment group consists of predicate chains with a focal AI device. We focus solely on the chains whose focal AI devices is at the first timestamp in the post-AI period (i.e., a single post-AI timestamp). This empirical design yields a clean DiD comparison by preventing confounding effects of heterogeneous post-AI trajectories within a predicate chain, which can arise when downstream devices introduced after the focal AI device variably adopt or do not adopt AI. Thus, the focal AI device is both the first AI device to appear within the chain and the last device (i.e., newest by clearance date) on that chain included in the DiD estimation. For simplicity, we refer to these predicate chains as "AI chains." Figure 3 in Section 4.1.1 depicts an example of an AI chain in the treatment group: K180881 ← K190205 ← K201446 ← K232699, where the first three devices are non-AI devices, and the last device is the focal AI device.

4.1.3. Control Group

To identify the control group we follow several steps: (1) for the focal AI device on each AI chain, we identify its *matched focal non-AI devices* that cite the same primary predicate as the focal AI device; (2) we extract the full pre-AI portion of the AI chain, ending with the shared primary predicate identified in step (1), and then separately attach each matched focal non-AI device to the end of this portion. These two steps

form a group of chains that constitute the control group, and we refer to them as comparable non-AI chains. Figure 4 provides an example of how these two steps work alongside the DiD setup.

Figure 4. Steps for Identifying the Control Group



Notes: Each arrow represents a predicate citation relationship, with an applicant device pointing to a predicate device. The red-highlighted path represents the matched non-AI chain to the focal AI chain. Detailed information about the devices shown in the figure is provided in Appendix B.

This figure depicts an AI chain: K180881 ← K190205 ← K201446 ← K232699. To identify comparable non-AI chains for this AI chain, we first identify K203687 and K203131, which share the same primary predicate device (K201446) as the focal AI device (K232699). Then we attach both devices, separately, to the pre-AI period of the AI chain to form two comparable non-AI chains: K180881 ← K190205 ← K201446 ← K203687 and K180881 ← K190205 ← K201446 ← K203131. In other words, when we identify multiple comparable non-AI chains, we consider all of them. Similar to the AI chain, we divide the non-AI chains into a pre-AI period (identical to the AI chain's pre-AI period) and a post-AI period (which consists solely of the matched focal non-AI device).

To summarize, this procedure ensures that (1) the focal AI device and its matched focal non-AI device in the post-AI period are substantially similar based on their shared primary predicates[5]; (2) devices on the AI chain and non-AI chains are identical in the pre-AI period; and (3) the resulting matched chains are identical in length. These similarities warrant that AI and non-AI chains are comparable, and concerns that

[5] In Figure 4's example, K203131, K203687, and K232699 are all cardiovascular devices. K203131 and K203687 function as diagnostic/monitoring platforms that provide measured hemodynamic data to clinicians. K232699 similarly supports cardiovascular care but does so by applying AI-based algorithm to ECG data to predict low left ventricular ejection fraction. So the primary difference between these post-AI period devices is whether they use AI. Details of these devices are provided in Appendix B.

manufacturers may strategically choose to use AI in high-technology, low-risk therapeutic categories (He et al. 2019, Benjamens et al. 2020) are addressed within the model framework.

4.1.4. Outcome Variable

The primary outcome variable in our analysis is whether the device at each timestamp of a predicate chain is subject to a serious recall within different year lengths (ranging from one year to five years) following its clearance date. For each predicate chain, we identify the first recorded serious recall – whether ongoing or resolved – associated with the device at each timestamp and code the outcome as 1 if a recall occurs and 0 otherwise. We analyze different recall windows to capture how the effect of AI may evolve with time in the market. In additional analyses, we examine alternative outcome measures, including counts of adverse event reports related to deaths, injuries, and malfunctions, to evaluate the robustness of our findings to different measures of device safety.

4.2. Main Model and Results

For each recall window, we run separate linear probability models using the ordinary least squares (OLS) using the following DiD specification (we use logistic regression as a robustness check in Section 4.5.2):

$$Recall_{it} = \beta \cdot AIChain_i \cdot PostAI_{it} + \alpha \cdot AIChain_i + \gamma \cdot PostAI_{it} + \theta_{it} + \eta_i + \tau_t + \varepsilon_{it} \quad (1)$$

where the recall indicator ($Recall_{it}$, coded 1 if a device on chain i at timestamp t is recalled, otherwise coded 0) is estimated as a function of the interaction term between i's AI indicator ($AIChain_i$, coded 1 if i is an AI chain, otherwise coded 0) and the post-period indicator at t ($PostAI_t$, coded 1 if t is in the post-AI period, otherwise coded 0). β is the estimated DiD coefficient that measures the effect of AI integration on medical device recall probability. The model also includes binary indicators for $AIChain_i$ and $PostAI_{it}$, and controls for a set of covariates θ_{it} (including information on device review time length, clearance year, manufacturer, product code, and medical specialty), chain fixed effects η_i (controlling for any time-invariant variables at the chain level), and timestamp fixed effects τ_t (controlling for any unobserved systematic trend across timestamps), with $\varepsilon_{it.}$ representing the error term.

Our constructed chains contain 796 AI chains and 38,709 non-AI chains. Based on these chains, we

implement several data cleaning steps to ensure the validity of the treatment and control groups, including removing (1) 117 AI chains and 5,463 non-AI chains with only one device (i.e., the chain only contains an applicant device and we do not have its predicate information); (2) 166 AI chains without a pre-AI period (i.e., all devices on the chain are AI devices). After removal of these chains, we have 513 AI chains and 33,246 non-AI chains. Our identification procedure introduced in Section 4.1.3 is able to find 258 AI chains (50.29% of all 513 AI chains in the study sample) with at least one comparable non-AI chain.

We report the parameter estimates in Table 3 and find that across one-year to three-year recall windows (i.e., from Model 1 to Model 3), the interaction terms between the AI chain and the post-AI indicators are positive and significant, indicating a greater increase in recall probability for focal AI devices than for focal non-AI deivces during the post-AI period. Specifically, in the post-AI period, the recall probabilities of focal AI devices on AI chains increase by 13.5 to 18.2 percentage points more than those of the matched focal non-AI devices on non-AI chains.

The effect magnitude is the largest when considering a two-year recall window, indicating the greatest effect of AI on device safety. This effect becomes insignificant when the recall window is longer than three years. Such a trend aligns with the fact that device recall risk decreases as devices remain longer on the market. In our study sample, the percentages of all device recalls against recall windows follow a convex curve (Figure 5), indicating a riskier period in years closer to device clearance (especially the first two years, when the number of recalls rapidly increases) than in years farther away (when the increase in the number of recalls slows down). These results suggest that the use of AI in medical devices significantly increases the devices' recall probability over non-AI devices, particularly in the first three years after clearance.

Further analyses on the predicted outcomes using linear probability model show that compared to pre-AI period, the recall probability of focal non-AI devices in the post-AI period substantially reduced (aligning with the general recall trend after 2016); however, the recall probability of focal AI devices in the post-AI period substantially increased (analysis details are provide in Appendix C). The result suggests that although medical devices have become safer over the years, the introduction of new technologies, such as AI, can introduce new risks, unintended consequences, and unforeseen complexities in their use (Lyytinen

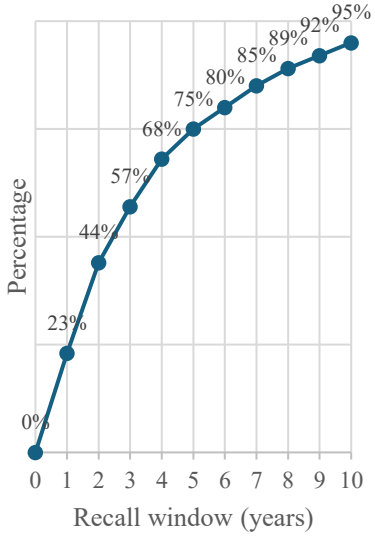
and Rose 2003, Mayer et al. 2020), increasing the probability of recalls.

Table 3. Difference-in-Differences Estimates: Effect of AI Integration on Medical Device Recall Probability

	Linear probability model using OLS				
	Recall (binary indicator) within *n* years after device clearance				
Variable	Model 1 (*n* = 1)	Model 2 (*n* = 2)	Model 3 (*n* = 3)	Model 4 (*n* = 4)	Model 5 (*n* = 5)
AI chain × Post AI (DiD estimator)	0.156***	0.182***	0.135**	0.049	0.062
	(0.039)	(0.049)	(0.053)	(0.050)	(0.050)
AI chain	0.003	0.003	0.006	-0.000	-0.000
	(0.006)	(0.009)	(0.010)	(0.011)	(0.011)
Post AI	-0.067**	-0.059**	-0.026	-0.051	-0.110**
	(0.018)	(0.028)	(0.031)	(0.031)	(0.035)
Covariates	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	4,048	3,965	3,880	3,764	3,568
Number of chains	746	746	744	744	729
R-squared	0.352	0.414	0.433	0.460	0.482

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

Figure 5. Cumulative Percentage of Recalls Across Recall Windows



4.3. Parallel Trend Test

The DiD framework relies on the parallel trends assumption, which requires that no treatment effect be present during the pre-treatment period. Satisfying this condition validates the dividing point between the pre- and post-AI periods and supports the interpretation that any observed differences in outcomes arise only after AI introduction (Angrist and Pischke 2009). To test this assumption and ensure there is no significant difference between the recall probability of devices on AI chains and non-AI chains in the pre-AI period, we follow the standard relative time model in the form:

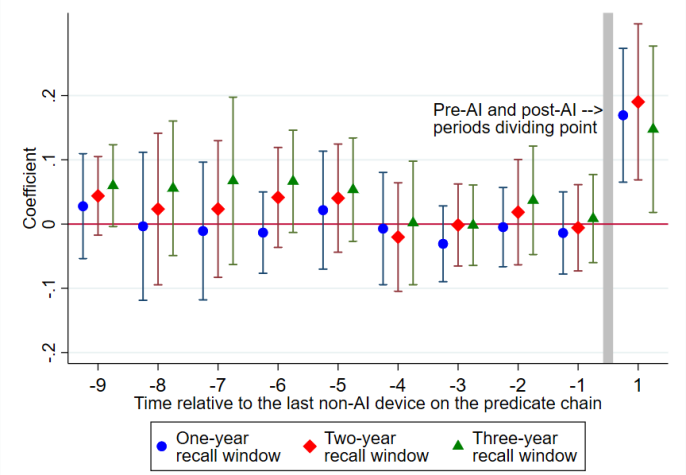
$$Recall_{it} = \sum_{t=1}^{T}(\beta_t \cdot AIChain_i \cdot Timestamp_t) + \alpha \cdot AIChain_i + \sum_{t=1}^{T}(\gamma_t \cdot Timestamp_t) + \theta_{it} + \eta_i + \tau_t + \varepsilon_{it} \quad (2)$$

where all variables are the same as in Equation (1), except that instead of interacting $AIChain_{it}$ with the post-AI indicator, we interact the AI chain indicator with each chain timestamp dummy ($Timestamp_t$) to capture AI's effect over all chain timestamps (T). The changes in coefficient β_t at each timestamp provides insights into how the recall probability changes over time.

We present the results for one-year, two-year, and three-year recall windows in a single Figure 6. For all recall windows, the coefficients of all interaction terms between the AI chain indicator and chain

timestamp dummies are insignificant in the pre-AI period, indicating a flat trend and suggesting no impact on recall probability during this period. In contrast, the coefficients increase in magnitude and become positively significant in the post-AI period. The positive coefficient observed only in the post-AI period confirms that the introduction of AI increases the probability of device recall. Intuitively, since the pre-AI parts of the predicate chain for the treatment and control groups are the same (based on our matching procedure described in Section 4.1.3), no difference in recall probability is present during the pre-AI period, and all differences arise in the post-AI period.

Figure 6. Relative Time Model Results



Notes: Each point represents the coefficient from the OLS regression that models the effect of AI use on the device recall probability over time. The regressors are the interaction terms between the AI chain indicator and chain timestamp dummies in Equation (2). Error bars show 95% confidence intervals relative to the mean values.

4.4. Placebo Test

We also conduct a placebo test to ensure that the estimated effect in the main model is not spurious and that the identified pre- and post-AI period dividing point is valid. For this purpose, we shift the dividing point for all chains to the 25%, 50%, and 75% quantiles of the pre-AI period timestamps separately, and rerun the DiD regression using only pre-period data. Similar to the relative time model, the purpose is to rule out the possibility that the estimated effect was present before the dividing point. If the main results are reliable, no significant effect should be observed in this test. As shown in Table 4, we do not find a significant effect using this test with pre-AI period data, which confirms the reliability of the main results.

Table 4. Placebo Tests: Falsification Using Artificial Treatment Timing in Pre-AI Period

	OLS model								
	Recall (binary indicator) within *n* years after device clearance								
	Dividing point at 25% quantile			Dividing point at 50% quantile			Dividing point at 75% quantile		
Variable	Model 1 (*n* = 1)	Model 2 (*n* = 2)	Model 3 (*n* = 3)	Model 4 (*n* = 1)	Model 5 (*n* = 2)	Model 6 (*n* = 3)	Model 7 (*n* = 1)	Model 8 (*n* = 2)	Model 9 (*n* = 3)
AI chain × Post AI	-0.001	0.007	-0.001	-0.009	-0.019	-0.022	-0.006	-0.009	-0.013
(Placebo DiD estimator)	(0.033)	(0.034)	(0.026)	(0.019)	(0.021)	(0.026)	(0.021)	(0.024)	(0.025)
AI chain	0.001	-0.007	0.001	0.009	0.019	0.022	0.003	0.004	0.007
	(0.033)	(0.034)	(0.026)	(0.019)	(0.021)	(0.026)	(0.011)	(0.012)	(0.012)
Post AI	0.191	0.334**	0.369*	0.194	0.344**	0.376*	0.192	0.340**	0.372*
	(0.154)	(0.167)	(0.196)	(0.154)	(0.166)	(0.197)	(0.153)	(0.166)	(0.196)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,302	3,302	3,298	3,302	3,302	3,298	3,302	3,302	3,298
Number of chains	746	746	744	746	746	744	746	746	744
R-squared	0.448	0.526	0.555	0.448	0.526	0.555	0.448	0.526	0.555

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

4.5. Robustness Checks

4.5.1. Complexity of Device Relationships on Predicate Chains

We conduct several analyses to check the robustness of the main results. First, as discussed in Section 4.1.1 and Figure 3, the predicate chains we created are based on the primary predicate device citation relationship between devices. While this criterion ensures substantial equivalence between devices within each chain, they may still differ in some characteristics. As such, devices on a predicate chain reflect not the evolution of one device, but a series of similar device models. To ensure that our main results are robust to the subtle variations in the device relationships within each predicate chain, we categorize the predicate chains into three distict types: In Type 1 chains, all devices have the same device name and manufacturer name (10,068 chains or 25.49% of all chains); In Type 2 chains, at least one pair of adjacent devices differ in device name *or* manufacturer name (5,544 chains or 14.03% of all chains); In Type 3 chains, at least one pair of adjacent devices differ in device name *and* manufacturer name (23,893 chains or 60.48% of all chains). Among our 513 AI chains, 17 are Type 1, 64 are Type 2, and 432 are Type 3; among our 33,246 non-AI chains, 4,432 are Type 1, 5,438 are Type 2, and 23,376 are Type 3.

From Type 1 to Type 3, the device names and manufacturer names of two adjacent devices become more distinct, increasing the intricacy of the relationship between devices on the predicate chain. In the raw

data, many names representing the same device or manufacturer are recorded with a slight difference (i.e., not standardized). We use an algorithm based on fuzzy matching (Fellegi and Sunter 1969, Bailey et al. 2018) to standardize similar names and provide details in Appendix C. Table 5 provides an example for each type of predicate chains.

Table 5. Examples of Predicate Chain Types by Device Name and Manufacturer Consistency

Chain type	Generation 3 primary predicate	Generation 2 primary predicate	Generation 1 primary predicate	Focal AI or matched non-AI device of the chain
Type 1 chain	-	**KNUMBER**: K011342 **Device name:** CELLULAR ANALYSIS SYSTEM **Manufacturer name:** BECKMAN COULTER	**KNUMBER**: K081930 **Device name:** CELLULAR ANALYSIS SYSTEM **Manufacturer name:** BECKMAN COULTER	**KNUMBER**: K120771 (AI) **Device name:** CELLULAR ANALYSIS SYSTEM **Manufacturer name:** BECKMAN COULTER
Type 2 chain	**KNUMBER**: K180238 **Device name:** CARTO 3 EP NAVIGATION SYSTEM, VERSION 6.0 AND ACCESSORIES **Manufacturer name:** BIOSENSE WEBSTER	**KNUMBER**: K191660 **Device name:** CARTO 3 EP NAVIGATION SYSTEM, VERSION 6.0 AND ACCESSORIES **Manufacturer name:** BIOSENSE WEBSTER	**KNUMBER**: K213264 **Device name:** CARTO RMT EP NAVIGATION SYSTEM V8 **Manufacturer name:** BIOSENSE WEBSTER	**KNUMBER**:K223733(non-AI) **Device name:** CARTO 3 EP NAVIGATION SYSTEM WITH ADVANCED FOCUS MAPPING (AFM) **Manufacturer name:** BIOSENSE WEBSTER
Type 3 chain	**KNUMBER**: K180881 **Device name:** HEMOSPHERE ADVANCED MONITOR **Manufacturer name:** EDWARDS LIFESCIENCESS LLC	**KNUMBER**: K190205 **Device name:** HEMOSPHERE ADVANCED MONITOR **Manufacturer name:** EDWARDS LIFESCIENCESS LLC	**KNUMBER**: K201446 **Device name:** HEMOSPHERE ADVANCED MONITOR **Manufacturer name:** EDWARDS LIFESCIENCESS LLC	**KNUMBER**: K232699 (AI) **Device name:** LOW EJECTION FRACTION AI-ECG ALGORITHM **Manufacturer name:** ANUMANA, INC

Table 6. Robustness of AI Effect Across Predicate Chain Types - Stratified Analysis

	Linear probability model using OLS									
	Recall (binary indicator) within *n* years after device clearance									
	Type 1 + Type 2 chains					**Type 3 chains**				
Variable	Model 1 (*n* = 1)	Model 2 (*n* = 2)	Model 3 (*n* = 3)	Model 4 (*n* = 4)	Model 5 (*n* = 5)	Model 6 (*n* = 6)	Model 7 (*n* = 7)	Model 8 (*n* = 8)	Model 9 (*n* = 9)	Model 10 (*n* = 10)
AI chain × Post AI (DiD estimator)	0.214**	0.262**	0.169*	-0.097	-0.088	0.147***	0.171***	0.096*	0.013	0.053
	(0.088)	(0.103)	(0.107)	(0.117)	(0.158)	(0.053)	(0.063)	(0.073)	(0.067)	(0.074)
AI chain	-0.045	0.008	0.007	-0.001	-0.001	0.019	-0.146*	-0.230**	-0.220	-0.311
	(0.109)	(0.013)	(0.014)	(0.001)	(0.001)	(0.085)	(0.083)	(0.091)	(0.169)	(0.214)
Post AI	-0.043	-0.196	-0.227	-0.004	-0.055	-0.031	-0.005	0.018	-0.079*	-0.166***
	(0.108)	(0.197)	(0.189)	(0.092)	(0.102)	(0.029)	(0.042)	(0.047)	(0.048)	(0.060)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,082	1,071	1,053	1,047	1,028	2,966	2,894	2,827	2,717	2,540
Number of chains	280	280	279	279	276	636	625	616	608	577
R-squared	0.594	0.620	0.705	0.682	0.675	0.426	0.487	0.507	0.536	0.550

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

We stratify our DiD model estimate by the types of predicate chains. Because the number of Type 1 chains is too small to yield meaningful results, we combine Type 1 and Type 2 chains in one set of models

and use Type 3 chains alone in another set. As shown in Table 6, the coefficients for the interaction term between the AI chain and the post-AI indicators are positive and significant, similar to the main results (except for Models 3 and 8 that consider a three-year recall window; these models' results are significant at the 90% confidence interval). The effect magnitudes for Type 1 and Type 2 chains are larger than those for Type 3 chains, indicating that the effect of AI use on increasing device recall probability can be further amplified when the relationship between devices in the considered predicate chains is more concordant in terms of their device names and manufacturer names. These results indicate that the type of constructed predicate chain does not influence the estimated effect.

4.5.2. Logistic Regression Model

Second, to test the sensitivity of our findings to empirical specification, we estimate logistic regression instead of OLS. Unlike OLS, logistic regression estimates the difference in the odds of recall between devices on AI chains and devices on non-AI chains. We report the coefficient estimates (odds are obtained by converting these estimates) in Table 7, where we find that in the post-AI period, the odds of recall for devices on the AI chains are about 3.38 (i.e., $e^{1.479}$-1; Model 3) to 6.34 (i.e., $e^{1.994}$-1; Model 2) times higher than those on the non-AI chains. Although these results are not comparable to those obtained using the OLS model in terms of absolute magnitude, we find that the effect's direction, significance, and the trend in magnitude are similar to those in the main analysis. As such, we confirm the robustness of our main results to the choice of estimation method.

Table 7. Robustness Check: Logistic Regression

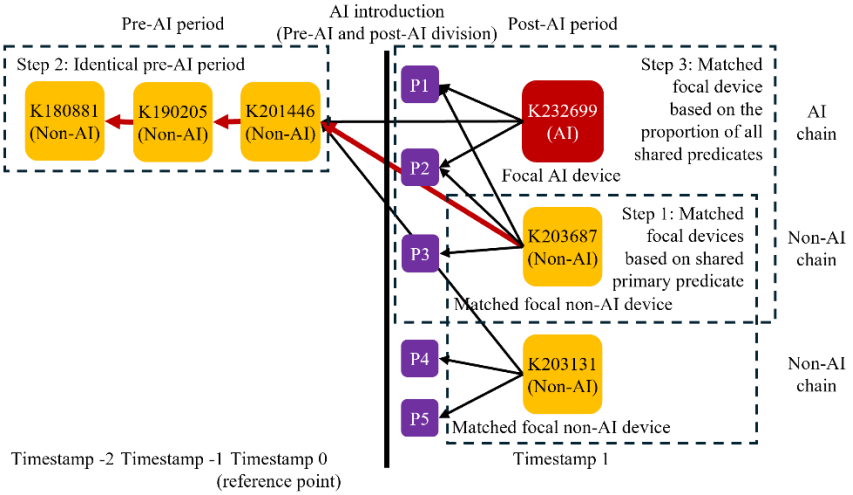
	Logistic regression model				
	Recall (binary indicator) within *n* years after device clearance				
Variable	Model 1 *(n* = 1)	Model 2 *(n* = 2)	Model 3 *(n* = 3)	Model 4 *(n* = 4)	Model 5 *(n* = 5)
AI chain × Post AI	1.586***	1.994***	1.479**	0.521	1.450
(DiD estimator)	(0.517)	(0.514)	(0.567)	(0.726)	(1.106)
AI chain	0.045	-0.298	0.134	0.046	0.038
	(0.141)	(1.929)	(0.146)	(0.140)	(0.139)
Post AI	-0.897***	-2.230***	-0.734*	-1.110***	-2.224***
	(0.321)	(0.426)	(0.411)	(0.426)	(0.476)
Covariates	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	4,048	3,965	3,880	3,764	3,568
Number of chains	746	746	744	744	729

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

4.5.3. Stricter Procedure for Identifying the Treatment Group

Third, our main model identifies comparable non-AI predicate chains solely by requiring that the timestamp-1 devices in the post-AI period share the same single primary predicate. Because the FDA recommends, but does not mandate, that applicants list only one predicate in their 510(k) filing, we conduct further robustness analysis. The idea we incorporate is that in matching devices with multiple predicates listed, only a high percentage of shared predicate devices guarantees the similarity of the two devices. To enfore this stricter matching criterion, we refine the procedure to account for all shared predicates – not just the primary one – and retain only those matched non-AI devices that have a high fraction of predicates in common with the focal AI device. This refinement corresponds to the additional step (3) in Figure 7.

Figure 7. Additional Step for Identifying the Control Group



Notes: Each arrow represents a predicate citation relationship, with an applicant device pointing to a predicate device. The red-highlighted path represents the matched non-AI chain to the focal AI chain. Appendix B provides information about the devices shown. Predicates P1-P5 do not reflect real devices and are imagined for illustrative purposes only.

To determine the percentage of shared predicates, we compute the Jaccard similarity scores between the two devices' sets of predicates. This score computes the percentage of common elements in two sets relative to the total number of unique elements in those sets (Jaccard 1912, Arts et al. 2018). Specifically, the score is computed between the device at timestamp 1 on the AI chain and every potential device at timestamp 1 on the non-AI chain. If the score exceeds a threshold θ (we set $\theta = 0.5$ to indicate majority), we select the device at timestamp 1 on the respective non-AI chain; otherwise, the device is not selected. In Figure 7, K232699's predicates are {K201446, P1, P2}, K203687's predicates are {K201446, P1, P2,

P3}, K203131's predicates are {K201446, P4, P5}. K232699's and K203687's predicate-based Jaccard similarity score is 0.75, and K232699's and K203131's predicate-based Jaccard similarity score is 0.20. As such, K203687 is chosen and the comparable non-AI chain for the focal AI chain is K180881 ← K190205 ← K201446 ← K203687. An AI-chain can have multiple comparable non-AI chains, as long as the predicat-based Jaccard similarity score is above θ. If at timestamp 1 in the post-AI period, there are no non-AI devices citing the last non-AI device in the pre-period, or the predicate-based Jaccard similarity score between the device at timestamp 1 on the AI chain and all potential devices at timestamp 1 on the non-AI chain is below θ, we remove the focal AI chain in our study sample.

Because this criterion is stricter, only 129 of the 513 AI chains in the study sample are identified as having at least one comparable non-AI chain (25.15%, down from 50.29% when only the previous, more relaxed matching criterion was used). We rerun the DiD model using panel data based on the new comparable chains and present the results in Table 8. Although the magnitudes of the effect are smaller than those of the main results, the direction and trend are similar. The smaller magnitudes likely reflect the reduced sample size, which omits many AI chains that can find comparable non-AI chains under a relaxer criterion and would otherwise contribute to identifying the effect. Another explanation is that the stricter criterion yields chains that are so similar that they remove much of the meaningful variation needed for causal identification, making any incremental effect of AI inherently more difficult to detect.

Table 8. Robustness Check: Stricter Procedure for Identifying Comparable Non-AI Chains

	Linear probability model using OLS				
	Recall (binary indicator) within *n* years after device clearance				
Variable	Model 1 *(n* = 1)	Model 2 *(n* = 2)	Model 3 *(n* = 3)	Model 4 *(n* = 4)	Model 5 *(n* = 5)
AI chain × Post AI	0.092**	0.121***	-0.064*	0.177*	0.098
(DiD estimator)	(0.036)	(0.045)	(0.097)	(0.077)	(0.094)
AI chain	-0.081	0.377*	-0.157	-0.796***	-0.856***
	(0.164)	(0.194)	(0.139)	(0.207)	(0.234)
Post AI	0.405***	-0.164	-0.017	-0.167	0.114
	(0.102)	(0.123)	(0.212)	(0.201)	(0.222)
Covariates	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	1,448	1,405	1,369	1,325	1,263
Number of chains	310	310	308	308	294
R-squared	0.410	0.607	0.687	0.791	0.761

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

4.5.4. Alternative Outcome Variables

Fourth, while the main results rely on probabilities of serious recalls, examining additional safety outcomes can provide complementary evidence and help demonstrate robustness. To assess this, we use FDA adverse-event reports (deaths, injuries, and malfunctions) as alternative outcomes, applying the same post-clearance time windows as in the main analysis. As shown in Table 9 the estimates closely align with the main results when considering the number of malfunction reports and all types of adverse event reports. However, a significant effect is not observed when considering only the number of deaths or serious injury reports (reported in Appendix E). One explanation for this difference by adverse event type is that the number of deaths reported due to AI devices to date is too small to yield an effect (only 0.2% of AI devices and 1.0% of non-AI devices in our study sample are associated with a death outcome, too small to yield a statistically significant effect). Another potential reason is that AI devices incur more reportable malfunctions (as they are software-driven and issues are more visible to caregivers) without inflicting severe patient harm, such as death and injury (Muralidharan et al. 2024, Lin et al. 2025).

Table 9. Robustness Check: Number of Adverse Event Reports

	Linear probability model using OLS									
	Recall (binary indicator) within *n* years after device clearance									
	Number of malfunction reports					**Number of all adverse event reports**				
Variable	Model 1 ($n = 1$)	Model 2 ($n = 2$)	Model 3 ($n = 3$)	Model 4 ($n = 4$)	Model 5 ($n = 5$)	Model 6 ($n = 6$)	Model 7 ($n = 7$)	Model 8 ($n = 8$)	Model 9 ($n = 9$)	Model 10 ($n = 10$)
AI chain × Post AI (DiD estimator)	1.899***	2.344***	2.512	1.752	1.007	1.822**	2.196*	2.158	0.405	-2.158
	(0.674)	(0.885)	(2.274)	(5.154)	(10.340)	(0.716)	(1.222)	(4.026)	(9.097)	(17.109)
AI chain	-2.630**	-2.307	0.075	4.263	7.760	-1.202	-0.137	4.692	11.566	17.751
	(1.238)	(1.501)	(2.854)	(5.202)	(8.815)	(1.312)	(1.874)	(4.884)	(9.370)	(14.832)
Post AI	-0.602	-1.870**	-5.291*	-9.704*	-17.214**	-1.794**	-4.076**	-9.986*	-17.248*	-27.135*
	(0.605)	(0.939)	(2.839)	(5.434)	(8.770)	(0.684)	(1.505)	(5.136)	(9.661)	(14.545)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,048	3,965	3,880	3,764	3,568	4,048	3,965	3,880	3,764	3,568
Number of chains	746	746	744	744	729	746	746	744	744	729
R-squared	0.585	0.457	0.236	0.213	0.215	0.566	0.334	0.206	0.199	0.201

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

5. Mechanisms Linking AI Functions to Device Recall

The preceding analyses reveal that integrating AI into medical devices can increase the likelihood of device recalls, highlighting the need to examine the mechanisms driving this effect. Because AI applications vary

in their clinical and technical roles, different AI functions may influence device safety in distinct ways. For example, AI systems used for administrative or workflow support (e.g., image triage, scheduling, or documentation) typically operate upstream from clinical decision-making, so errors are more likely to cause delays or inefficiencies rather than serious issues (Rajkomar et al. 2019). In contrast, AI functions involved in diagnosis or treatment recommendation interact more directly with clinical judgments, increasing the likelihood that erroneous outputs may shape clinical actions and, consequently, patient outcomes (Kelly et al. 2019). Thus, safety outcomes depend not only on whether a device uses AI, but also on the specific role the AI plays within the clinical workflow. To uncover the mechanisms through which AI contributes to recall risk – and to move beyond treating AI as a uniform technological feature – we distinguish between two pathways by which AI may elevate recall probability: the AI clinical function mechanism and the AI technical function mechanism.

5.1. Impact of AI Clinical Functions on Recalls

As shown in Figure 1, AI clinical functions in medical devices can be broadly classified into two categories: those supporting patient assessment (e.g., diagnosis or monitoring) and those supporting patient intervention (e.g., surgical assistance or treatment guidance) (Singh et al. 2025). Assessment-oriented AI operates early in the clinical workflow to influence subsequent testing, triage, and treatment choices, amplifying the consequences of incorrect outputs (Rajkomar et al. 2019). These devices must interpret heterogeneous, high-dimensional patient data, making its failure modes more complex and less predictable (Topol 2019a; Kelly et al. 2019). Errors at this stage often manifest as clinically plausible classifications, which can be difficult for manufacturers to detect during pre-market testing because they may only emerge under real-world variations that are hard to fully simulate. By contrast, AI used for patient intervention typically functions within tightly constrained, highly supervised workflows, such as surgical assistance or treatment planning, where system behavior is more observable and deviations from expected performance are easier to identify (Sendak et al. 2020). This difference suggests a potential mechanism behind the higher recall probability for AI devices: devices employing assessment-oriented AI may be more prone to undetected issues at the pre-market stage because their outputs are harder to comprehensively validate,

increasing the likelihood that latent defects surface only after deployment.

To probe whether this differential functional use serves as a plausible mechanism behind the higher recall rates observed for AI devices, we draw on the medical device AI use taxonomy[6] (which categorizes AI devices by their clinical and technical function) provided in Singh et al. (2025) to identify the AI devices whose AI use is for clinical assessment. Then, we give a predicate chain an indicator of AI function in clinical assessment if the focal AI device of this chain is one of the identified AI devices for clinical assessment and replace the DiD term in the main model with a difference-in-differences-in-differences (DDD) term that captures the interaction between the AI chain, the post-AI period, and the AI function indicators. As before, the unit of analysis is a device within a chain, with all devices on the chain being substantially equivalent. Specifically, we leverage the following estimation model:

$$\begin{aligned} Recall_{it} = {} & \beta \cdot AIChain_i \cdot PostAI_{it} \cdot AIforPatientAssessment_i \\ & + \alpha \cdot AIChain_i \cdot PostAI_{it} \\ & + \gamma \cdot AIChain_i \cdot AIforPatientAssessment_i + \delta \cdot PostAI_{it} \cdot AIforPatientAssessment_i \\ & + \phi \cdot AIChain_i + \varphi \cdot PostAI_{it} + \omega \cdot AIforPatientAssessment_i \\ & + \theta_{it} + \eta_i + \tau_t + \varepsilon_{it} \end{aligned} \quad (3)$$

where all shared variables represent the same as in Equation (1). The additional subscript v indexes the key factor we study to understand the AI function moderating effects; $AIforPatientAssessment_i$ a binary variable that is coded 1 if the focal AI device on the AI chain *i* uses AI for clinical assessment, otherwise coded 0; $AIChain_i \cdot PostAI_{it} \cdot AIFunction_{vi}$ is the DDD term for chan *i* at timestamp *t* with respect to the AI function v. The coefficient β is the estimate of the moderating effect of AI's patient assessment function on device recall probability. The model also includes two-way interaction terms between each pair of the three variables and the independent terms of each of the three variables.

As shown in Table 10, the DDD coefficients are positively significant in both the one-year and two-year recall window models, suggesting that AI used for clinical assessment drives the elevated recall probability observed among AI devices. This finding indicates that assessment-oriented AI functions in medical devices used for diagnosis or monitoring merit heightened safety scrutiny.

[6] The creation of this taxonomy is based on an rigorous review process (involving at least two human experts) of the decription of the device product code 510(k) submission documents (Singh et al. 2025).

Table 10. Mechanism Analysis: Moderating Effect of AI Clinical Function Type on Device Recall

	Linear probability model using OLS		
	Recall (binary indicator) within *n* years after device clearance		
Variable	Model 1 (*n* = 1)	Model 2 (*n* = 2)	Model 3 (*n* = 3)
AI chain × Post AI × AI for patient assessment (DDD estimator)	0.224**	0.209**	0.168*
	(0.100)	(0.115)	(0.117)
AI chain × Post AI	-0.024	-0.003	-0.015
	(0.084)	(0.099)	(0.098)
AI chain × AI for patient assessment	0.072***	0.043**	0.065***
	(0.018)	(0.018)	(0.024)
Post × AI for patient assessment	0.001	-0.012	-0.027
	(0.042)	(0.059)	(0.070)
AI chain	-0.011**	0.002	0.005
	(0.005)	(0.007)	(0.009)
Post AI	0.212	0.504***	0.619***
	(0.135)	(0.155)	(0.174)
AI for patient assessment	-0.094	0.103	-0.247**
	(0.087)	(0.108)	(0.106)
Covariates	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes
Observations	4,048	3,965	3,880
Number of chains	746	746	744
R-squared	0.403	0.475	0.512

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

5.2. Impact of AI Technical Functions on Recalls

AI functions in medical devices can also be classified based on AI's technical characteristics (Figure 1). Broadly, the technical functions of AI in medical devices can be categorized into those used for data generation (e.g., image enhancement, synthetic data simulation) and those used for data analysis (e.g., disease exploration or prediction) (Singh et al. 2025). Data-generation functions may pose greater safety and reliability challenges because they directly modify or create clinical inputs, making it more difficult for manufacturers to anticipate all potential failure modes during pre-market design and validation. For example, prior work in medical imaging cautions that AI-driven image modification (e.g., distortion removal or contrast enhancement) can introduce subtle artifacts or suppress clinically relevant signals (Bluemke et al. 2020, Kulkarni et al. 2021). These issues are especially difficult to detect in advance because generated or altered data often retain realistic visual characteristics, making deviations from ground truth hard to identify and increasing the risk of silent failure and automation bias (Roberts et al. 2021). As a result, manufacturers may not be able to exhaustively validate the behavior of generative models across the full range of real-world variation.

By contrast, AI used for data analysis typically operates on unaltered clinical inputs and produces outputs that can be benchmarked against known reference standards, facilitating more predictable performance evaluation (Sendak et al. 2020). This difference suggests a potential mechanism behind the higher recall probability observed among AI devices: devices employing data-generation functions may be more prone to undetected issues at the pre-market stage because their outputs are harder to verify comprehensively, increasing the likelihood that latent defects surface only after deployment.

To examine this explanation, we rely on the same taxonomy of medical device AI usage in Section 5.1 to identify the AI devices whose AI is used for data generation. Then, we give a predicate chain an indicator of AI function for data generation if the focal AI device of this chain is one of the identified AI devices for data generation and re-estimate Equation (3), changing the moderating variable of interest to $AIforDataGeneration_i$. As reported in Table 11, we find that the DDD coefficients are positively significant in models considering one-year to three-year recall windows, indicating that AI in medical devices used for data generation is the driving factor for higher recall probability of AI devices. This finding suggests that redcuing patient harms requires improvements in the use of AI in data generation functions .

Table 11. Mechanism Analysis: Moderating Effect of AI Technical Function Type on Device Recall

	Linear probability model using OLS		
	Recall (binary indicator) within *n* years after device clearance		
Variable	Model 1 (n = 1)	Model 2 (n = 2)	Model 3 (n = 3)
AI chain × Post AI × AI for data generation (DDD estimator)	0.311***	0.329***	0.283**
	(0.099)	(0.124)	(0.141)
AI chain × Post AI	0.039	0.043	0.024
	(0.030)	(0.045)	(0.052)
AI chain × AI for data generation	0.103***	0.091***	-0.062*
	(0.027)	(0.028)	(0.032)
Post × AI for data generation	0.081	-0.009	0.017
	(0.050)	(0.075)	(0.081)
AI chain	0.053***	0.044***	0.065***
	(0.019)	(0.016)	(0.022)
Post AI	0.119	0.435***	0.536***
	(0.123)	(0.160)	(0.182)
AI for data generation	-0.016	-0.052	0.214**
	(0.087)	(0.095)	(0.106)
Covariates	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes
Observations	4,048	3,965	3,880
Number of chains	746	746	744
R-squared	0.416	0.479	0.514

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

6. Heterogeneity Analysis

To obtain a comprehensive understanding of variations in the identified effect and to provide more practical guidelines for improving medical device safety, we investigate heterogeneity in the causal effects of AI use across clinical AI functions, technical AI characteristics, device specialties, and recall classifications.

6.1. Subsample Analysis by AI Function

As discussed in Section 5, AI in medical devices can be categorized by its clinical functions (patient assessment vs. patient intervention) and its technical functions (data generation vs. data analysis). Examining how the main effects differ across these categories helps identify which specific AI functions – or combinations of functions – drive the observed increase in recall probability, thereby enabling stakeholders to target oversight and regulatory efforts more precisely. To do so, we first partition the AI-device sample by clinical function and estimate separate DiD models using each subsample of focal AI devices and their matched focal non-AI devices, following the main model specification. We then repeat this approach by splitting the sample according to AI technical function. As shown in Table 12, only the subsamples corresponding to patient assessment (Models 1 and 2) and data generation (Models 7 and 8) exhibit a positive and significant effect, whereas the other functional subcategories do not.

Table 12. Effect Heterogeneity Analysis: Variation by AI Function Type

	Linear probability model using OLS											
	Recall (binary indicator) within *n* years after device clearance											
	AI clinical function						**AI technical function**					
	Patient assessment Num. AI chains: 206			Patient intervention Num. AI chains: 52			Data Generation Num. AI chains: 65			Data Analysis Num. AI chains: 203		
Variable	Model 1 (*n*=1)	Model 2 (*n*=2)	Model 3 (*n*=3)	Model 4 (*n*=1)	Model 5 (*n*=2)	Model 6 (*n*=3)	Model 7 (*n*=1)	Model 8 (*n*=2)	Model 9 (*n*=3)	Model 10 (*n*=1)	Model 11 (*n*=2)	Model 12 (*n*=3)
AI chain × Post AI	0.197***	0.169***	0.120*	0.004	-0.091	-0.030	0.337***	0.236**	0.047	0.042	0.053	0.042
	(0.054)	(0.066)	(0.071)	(0.037)	(0.082)	(0.082)	(0.101)	(0.099)	(0.117)	(0.036)	(0.045)	(0.049)
AI chain	0.063***	0.044***	0.068***	0.011	0.013*	0.001	0.114***	0.140***	0.093***	0.037***	0.016	0.035**
	(0.015)	(0.016)	(0.021)	(0.007)	(0.007)	(0.002)	(0.031)	(0.043)	(0.030)	(0.014)	(0.012)	(0.015)
Post AI	0.036	0.330*	0.381*	-0.031	-0.031	0.040	-0.013	-0.017	0.058	0.133	0.527***	0.669***
	(0.142)	(0.180)	(0.199)	(0.027)	(0.039)	(0.066)	(0.061)	(0.093)	(0.084)	(0.110)	(0.171)	(0.196)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,107	3,041	2,984	941	924	896	906	896	879	3,256	3,179	3,107
Num. chains	566	566	566	180	180	178	170	170	170	605	605	603
R-squared	0.431	0.481	0.525	0.937	0.920	0.893	0.546	0.622	0.672	0.487	0.528	0.564

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp. The samples for patient assessment and patient intervention are mutually exclusive, but the samples for data generation and data analysis overlap across ten devices.

Additionally, we cross-combine the two clinical functions (assessment and intervention) and the two technical functions (data generation and data analysis) to create four subsamples of AI chains whose focal AI device use AI for different combinations of functional purposes. For each subsample, we pair the focal AI devices with their matched focal non-AI devices and estimate separate DiD models following the main specification. Table 12 reports the sample sizes and results. Only the subsample in which devices employ AI for both patient assessment *and* data generation (Models 1 and 2) yields a positive and significant effect; none of the other functional combinations show statistically significant results.

This suggests that devices that use AI for data generation in patient assessment, such as AI-based image processing, exhibit a higher likelihood of recalls and thus warrant greater oversight. These heterogeneous analyses further confirm that patient assessment and data-generation functions are the key drivers of the elevated recall risk associated with AI devices, reinforcing the mechanism-based evidence identified in Section 5.

Table 13. Effect Heterogeneity Analysis: Variation by Cross-combination of AI Functions

	Linear probability model using OLS											
	Recall (binary indicator) within *n* years after device clearance											
	Patient assessment + data generation Num. AI chains: 59			Patient assessment + data analysis Num. AI chains: 154			Patient intervention + data generation Num. AI chains: 6			Patient intervention + data analysis Num. AI chains: 49		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Variable	(*n*=1)	(*n*=2)	(*n*=3)	(*n*=1)	(*n*=2)	(*n*=3)	(*n*=1)	(*n*=2)	(*n*=3)	(*n*=1)	(*n*=2)	(*n*=3)
AI chain × Post AI	0.317***	0.197**	-0.034	0.075*	0.069	0.069	0.000	0.191	0.255	0.051	-0.205	-0.287
	(0.112)	(0.113)	(0.148)	(0.041)	(0.052)	(0.056)	(0.040)	(0.178)	(0.187)	(0.047)	(0.171)	(0.237)
AI chain	0.122***	0.151***	0.105***	0.033***	0.004	0.021	-0.000	-0.032	-0.042	0.011	0.012	0.001
	(0.034)	(0.047)	(0.037)	(0.012)	(0.015)	(0.016)	(0.007)	(0.030)	(0.031)	(0.007)	(0.007)	(0.001)
Post AI	-0.458*	-0.194	0.123	-0.115	0.347*	0.457**	0.111***	0.177**	0.469***	-0.028	-0.050	-0.039
	(0.263)	(0.593)	(0.589)	(0.119)	(0.182)	(0.200)	(0.042)	(0.084)	(0.134)	(0.030)	(0.044)	(0.052)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	818	810	796	2,359	2,299	2,255	88	86	83	897	880	852
Num. chains	154	154	154	431	431	431	16	16	16	174	174	172
R-squared	0.611	0.659	0.710	0.520	0.544	0.583	0.383	0.363	0.483	0.969	0.929	0.941

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

6.2. Subsample Analysis by AI Technical Features

Beyond broad AI functions, it is also important to examine how the identified effects vary across more granular AI technical features used in medical devices. Such insights can guide stakeholders toward more targeted design improvements and regulatory strategies at the feature level to promote safer AI deployment.

Motivated by this, we group the AI chains according to the specific AI technical features employed by their focal AI device (Singh et al. 2025). As presented in Table 14, we observe positive effects only in the subsample of AI chains in which the focal AI device use AI for image enhancement (Models 1 and 2), suggesting that this type of AI features may be the most prone to recalls. Interestingly, the subsample of AI chains whose focal AI device use AI for detection or prediction shows a negative effect (i.e., a decrease in device recall probability), indicating that these AI features actually make the device safer. This insight aligns with our argument in the AI technical function mechanism: medical devices using AI for detection or prediction may face lower recall risk because their outputs map to clear clinical endpoints, making performance issues easier to validate and monitor. These models are typically trained against well-defined ground truth, allowing manufacturers to identify failures early. By contrast, image-enhancement AI works on raw or partially processed data, where correctness is less explicit and subtle distortions may go unnoticed until clinical use, increasing the likelihood of post-market recalls.

Table 14. Effect Heterogeneity Analysis: Variation by AI Technical Features

	Linear probability model using OLS											
	Recall (binary indicator) within *n* years after device clearance											
	Data Generation: 65						Data Analysis: 203					
	Image enhancement Num. AI chains: 58			Acquisition guidance/ synthetic data generation Num. AI chains: 7			Quantification/ feature localization Num. AI chains: 187			Detection/ prediction Num. AI chains: 16		
Variable	Model 1 (*n*=1)	Model 2 (*n*=2)	Model 3 (*n*=3)	Model 4 (*n*=1)	Model 5 (*n*=2)	Model 6 (*n*=3)	Model 7 (*n*=1)	Model 8 (*n*=2)	Model 9 (*n*=3)	Model 10 (*n*=1)	Model 11 (*n*=2)	Model 12 (*n*=3)
AI chain × Post AI	0.352***	0.318***	0.094	-0.021	-0.234	-0.356	0.018	0.021	0.000	-0.095*	-0.177**	-0.184**
	(0.111)	(0.103)	(0.131)	(0.202)	(0.253)	(0.323)	(0.034)	(0.044)	(0.050)	(0.125)	(0.077)	(0.073)
AI chain	0.125***	0.122***	0.083**	-0.311**	-0.261**	-0.042	0.123	0.506***	0.366**	0.052**	0.026**	0.026**
	(0.035)	(0.043)	(0.032)	(0.047)	(0.053)	(0.056)	(0.081)	(0.125)	(0.153)	(0.021)	(0.012)	(0.012)
Post AI	-0.007	-0.102	0.004	0.485***	0.795***	0.741***	0.363***	0.775***	0.913***	-0.120**	-0.242**	-0.264**
	(0.062)	(0.083)	(0.079)	(0.127)	(0.134)	(0.125)	(0.107)	(0.165)	(0.194)	(0.053)	(0.053)	(0.063)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	810	802	789	96	94	90	2,997	2,928	2,862	259	251	245
Num. chains	150	150	150	20	20	20	557	557	555	48	48	48
R-squared	0.608	0.683	0.744	0.270	0.339	0.377	0.495	0.515	0.546	0.879	0.906	0.921

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp. The samples for data generation and data analysis overlap across ten devices.

6.3. Subsample Analysis by Medical Specialties of Devices

Prior research (He et al. 2019; Warraich et al. 2024) and our study sample indicate that AI-enabled medical devices are highly concentrated in two specialties: radiology and cardiovascular medicine. In our data,

88.82% of AI devices cleared between 2011 and 2023 fall into these categories (radiology: 78.77%; cardiovascular: 10.05%). This concentration motivates an examination of whether and how the estimated effects vary by medical specialties. To investigate this, we partition all predicate chains into three subgroups: AI chains whose focal AI devices are radiology devices, chains whose focal AI devices are cardiovascular devices, and all remaining chains. We then estimate the main DiD specification separately for each subgroup.

As presented in Table 15, the positive and significant effect of AI emerges only in the subgroup of radiology devices (Models 1 and 2), with no corresponding effect in the cardiovascular or other-specialty subgroups. This pattern suggests a potential mechanism behind the overall relationship between AI use and recall probability: the elevated recall risk associated with AI devices may be concentrated in radiology, where AI systems interact with complex imaging data and may introduce failure modes that are more difficult for manufacturers to detect and fully validate prior to market entry.

However, we caution that the lack of statistical significance in other categories should not be interpreted as evidence of lower regulatory need, but rather as a reflection of more limited observable adverse outcomes at present. From a policy standpoint, this points to the value of encouraging faster adverse-event reporting in other specialties, should such events arise, while prioritizing more active regulatory oversight of radiology, where AI use is sufficiently widespread to enable effective risk monitoring.

Table 15. Effect Heterogeneity Analysis: Variation by Medical Specialties

	Linear probability model using OLS								
	Recall (binary indicator) within *n* years after device clearance								
	Radiology Number of AI chains: 213			Cardiovascular Number of AI chains: 25			Others Number of AI chains: 20		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Variable	(*n*=1)	(*n*=2)	(*n*=3)	(*n*=1)	(*n*=2)	(*n*=3)	(*n*=1)	(*n*=2)	(*n*=3)
AI chain × Post AI	0.141***	0.123**	0.057	0.319	-0.204	-0.201	-0.028	0.043	0.008
(DiD estimator)	(0.048)	(0.056)	(0.058)	(0.221)	(0.125)	(0.123)	(0.104)	(0.057)	(0.152)
AI chain	0.161**	0.544***	0.557***	-0.007	-0.003	-0.004	-0.059	0.055	-0.493***
	(0.064)	(0.093)	(0.097)	(0.028)	(0.016)	(0.016)	(0.123)	(0.063)	(0.106)
Post AI	0.440***	0.750***	0.770***	-0.369	-0.717***	-0.768***	-0.039	-0.011	-0.003
	(0.139)	(0.163)	(0.180)	(0.446)	(0.239)	(0.250)	(0.179)	(0.042)	(0.183)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,443	3,388	3,318	406	389	384	199	188	178
Number of chains	609	609	607	81	81	81	56	56	56
R-squared	0.415	0.469	0.512	0.828	0.934	0.944	0.590	0.833	0.699

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

6.4. Device Recall Reasons

Although the main analysis shows that AI use increases the likelihood of medical device recalls, it does not distinguish among the underlying causes of these events, as the outcome is modeled simply as a binary indicator. FDA reports indicate that software-related failures are the predominant source of recalls for AI-enabled devices (FDA 2021, 2023b). This pattern reflects the software-intensive nature of AI devices, whose clinical performance depends on complex code, models, and data pipelines rather than physical components. As a result, AI devices are more vulnerable to issues such as coding errors, incorrect parameterization, data-handling failures, or unintended behavior following software updates – problems less common in traditional devices, where recalls often stem from mechanical wear, material defects, or manufacturing faults (FDA 2025b).

To assess whether the effects vary by recall type, we follow Zhu et al. (2025) and Everhart et al. (2023b) in classifying recall events into software-related versus all other causes. Using recall data (FDA 2023b), we construct two panel datasets following the procedures in Section 4: one capturing only software-related recalls and another capturing all remaining recalls. We then estimate the main DiD specification separately for each dataset. As reported in Table 16, a significant positive effect of AI appears only in the software-related recall panel, with no detectable effect for recalls arising from non-software causes.

This recall-type analysis reinforces that AI itself is the critical component driving the observed recall patterns. Because software failures are precisely where AI functionality is embedded, the emergence of an AI effect exclusively in software-related recalls further indicates that the elevated recall risk is rooted in AI-specific software vulnerabilities rather than broader device characteristics. This finding, while consistent with the software-centric nature of AI devices, underscores the need for strengthened software quality assurance and responsible AI practices to enhance the safety of AI-enabled medical technologies.

7. Discussion

As healthcare providers, manufacturers, and regulators remain uncertain about the real-world safety consequences of AI-enabled technologies and existing regulatory frameworks are still evolving, our results offer timely empirical guidance for governance in this emerging domain.

Table 16. Effect Heterogeneity Analysis: Variation by Recall Reasons

	Linear probability model using OLS					
	Recall (binary indicator) within *n* years after device clearance					
	Panel data is created considering only the *software-related* recall events			Panel data is created considering all recall events *except those related to software*		
Variable	Model 1 (*n*=1)	Model 2 (*n*=2)	Model 3 (*n*=3)	Model 4 (*n*=1)	Model 5 (*n*=2)	Model 6 (*n*=3)
AI chain × Post AI (DiD estimator)	0.144***	0.157***	0.146**	0.033	0.030	-0.012
	(0.042)	(0.054)	(0.061)	(0.030)	(0.035)	(0.038)
AI chain	0.089***	0.086***	0.089***	-0.025***	-0.039***	-0.016
	(0.014)	(0.014)	(0.017)	(0.007)	(0.007)	(0.012)
Post AI	0.228**	0.398***	0.440***	-0.013	0.096	0.156
	(0.103)	(0.135)	(0.150)	(0.090)	(0.096)	(0.114)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,048	3,965	3,880	4,048	3,965	3,880
Number of chains	746	746	744	746	746	744
R-squared	0.444	0.499	0.515	0.341	0.402	0.471

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Covariates include review time length, clearance year, manufacturer, product code, and medical specialty. Fixed effects include chain and timestamp.

Our analysis suggests that problematic uses of AI for patient assessment and data generation are key contributors to elevated recall risk. Such AI devices include, for example, those that assist clinicians in screening high-risk fetal conditions or in generating and enhancing synthetic clinical data for treatment simulation and planning, where errors can have particularly consequential implications. This insight points to the need for targeted regulatory attention to AI systems operating with these functions. Strengthened pre-market evaluation, enhanced post-market surveillance, and clearer standards for validation and monitoring may help improve the safety and reliability of such AI applications. From a compassionate AI perspective, prioritizing governance in these high-stakes contexts is essential for protecting vulnerable patients and maintaining trust in AI-enabled care.

Across medical specialties, radiology emerges as the field with both the highest adoption of AI-enabled devices and the greatest susceptibility to serious recalls. Radiology's rapid embrace of AI reflects an effort to use technology compassionately – to reduce diagnostic burden, accelerate clinical workflows, and improve patient outcomes. At the same time, the elevated recall risk suggests that patients in this specialty may benefit from enhanced regulatory monitoring, faster adverse-event reporting, and more robust post-market surveillance, particularly during the early adoption phase. Our findings thus provide guidance for more tailored, specialty-specific regulatory oversight, allowing governance structures to better align with

heterogeneous patterns of AI use and risk across clinical domains.

Importantly, our results also indicate that AI use appears comparatively safe in many other specialties, offering a promising outlook for continued innovation. Pathology (e.g., risk-prediction tools based on laboratory data), neurology (e.g., seizure-detection systems), and ophthalmology (e.g., diabetic-retinopathy screening) illustrate areas where AI is augmenting clinical practice safely. Nevertheless, the radiology findings underscore the importance of cautious, expertise-informed adoption. Consistent with compassionate AI principles, clinicians, developers, and regulators should ensure that AI tools are designed and deployed to complement – rather than replace – professional judgment, thereby mitigating risk while preserving clinical accountability.

Further, our analysis by recall reason reveals that the increased risk associated with AI devices is concentrated in software-related failures. FDA reports indicate that software failures account for the majority of recalls involving AI-enabled devices (FDA 2018, 2023a). This pattern reflects the software-intensive nature of AI systems, whose clinical performance depends on complex code, model architectures, and data pipelines rather than physical components. Consequently, AI devices are more susceptible to coding errors, incorrect parameterization, data-handling failures, or unintended behavior following software updates – issues that are less prevalent in traditional medical devices, where recalls more often stem from mechanical wear or manufacturing defects (FDA 2025b). Because these failures arise precisely where AI functionality is embedded, the concentration of recall risk in software-related failures indicates that AI-specific software vulnerabilities, rather than broader device characteristics, are driving the observed effects. This finding reinforces the need for strengthened software quality assurance, responsible design, lifecycle monitoring, and regulatory oversight for software as a medical device (SaMD), especially as AI-driven software becomes increasingly central to clinical care.

8. Conclusion and Future Work

Using a causal identification strategy and FDA public data, we provide evidence that integrating AI into medical devices increases the likelihood of serious device recalls, particularly in the early years after market clearance. These findings are robust to the way we contruct predicate chains (central to our empirical

method), estimation models, strictness for identifying the control group, and outcome measures.

Analyzing the safety of AI devices is essential to compassionate AI because it safeguards patient dignity, minimizes preventable harm, and reinforces the trust that compassionate care depends on. Evidence shows that device performance and safety risks are not evenly distributed: pulse oximeters, for example, have much higher error rates in Black patients, leading to delayed detection of hypoxemia (Sjoding et al. 2020). Recent reporting also highlights gender differences in X-ray imaging, where AI systems trained on biased datasets may misinterpret scans for women, and warns that AI-based dermatology tools can underestimate skin cancer in people with darker skin because most training images come from lighter-skinned patients (Mundasad 2024). These disparities illustrate how technical shortcomings can amplify existing health inequities. Evaluating risks across functional areas or therapeutic categories enables regulators to anticipate such failures early, reducing the likelihood that malfunctions or recalls will disproportionately harm marginalized groups. This proactive, equity-centered oversight aligns directly with the ethical commitments of compassionate AI – ensuring that AI-enabled medical technologies support patient well-being, prevent inequitable harm, and benefit those most vulnerable in clinical care.

More broadly, our findings contribute to the ongoing debate over whether the growing AI adoption represents a boon or a bane in healthcare. By empirically demonstrating that AI adoption can introduce safety risks alongside its potential benefits, our study helps reconcile polarized perspectives and provides a more balanced foundation for theory development in medical AI research. Aligned with emerging discussions on compassionate AI (Morrow et al. 2023, Palmieri et al. 2026), our results emphasize that responsible, human-centered design and governance are essential for ensuring that AI enhances clinical decision-making without compromising patient safety. Ultimately, by showing how AI can effectively complement human expertise when thoughtfully managed, this study offers a practical pathway for researchers, practitioners, and regulators seeking to advance medical device safety and foster trustworthy, patient-centered AI in healthcare.

This study focuses on the first instance of AI integration into a generational chain of medical devices, providing a clean identification setting and insight into the early phase of AI adoption. As AI becomes

increasingly embedded across successive generations, future work using richer longitudinal data will be essential for understanding how safety risks evolve over time and whether regulatory interventions effectively mitigate them.

References

Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, Bhavani SV, et al. (2025) AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA* 334(18):1650–1664.

Arts S, Cassiman B, Gomez JC (2018) Text Matching to Measure Patent Similarity. *Strategic Management Journal* 39(1):62–84.

Babic B, Glenn Cohen I, Stern AD, Li Y, Ouellet M (2025) A General Framework for Governing Marketed AI/ML Medical Devices. *npj Digit. Med.* 8(1):328.

Bailey M, Cao R, Kuchler T, Stroebel J (2018) The Economic Effects of Social Networks: Evidence from the Housing Market. *Journal of Political Economy* 126(6):2224–2276.

Basu S, Hassenplug JC (2012) Patient Access to Medical Devices — A Comparison of U.S. and European Review Processes. *New England Journal of Medicine* 367(6):485–488.

Benjamens S, Dhunnoo P, Meskó B (2020) The State of Artificial Intelligence-based FDA-Approved Medical Devices and Algorithms: An Online Database. *npj Digital Medicine* 3(1):118.

Berente N, Gu B, Recker J, Santhanam R (2021) Managing Artificial Intelligence. *MIS Quarterly* 45(3):1433–1450.

Bhargava A, López-Espina C, Schmalz L, Khan S, Watson GL, Urdiales D, Updike L, et al. (2024) FDA-Authorized AI/ML Tool for Sepsis Prediction: Development and Validation. *NEJM AI* 1(12):AIoa2400867.

Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, Halpern EF, Hess CP, Schiebler ML, Weiss CR (2020) Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board. *Radiology* 294(3):487–489.

Bockstedt JC, Buckman JR (2025) Humans' Use of AI Assistance: The Effect of Loss Aversion on Willingness to Delegate Decisions. *Management Science* 72(1):323–342.

Choudhury A, Asan O (2020) Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. *JMIR Medical Informatics* 8(7):e18599.

Christensen J (2024) Philips Reaches $1.1 Billion Settlement for CPAP Machine Lawsuits, Admits No Fault or Liability. *CNN*. Retrieved (January 1, 2026), https://www.cnn.com/2024/04/30/health/philips-1-billion-cpap-settlement-agreement/index.html.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64(3):1155–1170.

Everhart AO, Sen S, Stern AD, Zhu Y, Karaca-Mandic P (2023) Association Between Regulatory Submission Characteristics and Recalls of Medical Devices Receiving 510(k) Clearance. *JAMA* 329(2):144–156.

Everhart AO, Stern AD (2022) Post-market Surveillance of Software Medical Devices: Evidence from Regulatory Data. *Harvard Business School*. Retrieved (January 1, 2026), https://www.hbs.edu/faculty/Pages/item.aspx?num=63330.

FDA (2018) The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications. *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/510k-program-evaluating-substantial-equivalence-premarket-notifications-510k.

FDA (2019) Premarket Approval (PMA). *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-approval-pma.

FDA (2023a) Medical Device Recalls. *FDA*. Retrieved (January 1, 2026), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfres/res.cfm.

FDA (2023b) 510(k) Database. *FDA*. Retrieved (January 1, 2026), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm.

FDA (2023c) 510(k) Clearances. *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/medical-devices/device-approvals-denials-and-clearances/510k-clearances.

FDA (2023d) Manufacturer and User Facility Device Experience (MAUDE). *FDA*. Retrieved (January 1, 2026), https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm.

FDA (2024) Premarket Notification 510(k). *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k.

FDA (2025a) Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.
FDA (2025b) Artificial Intelligence and Machine Learning in Software as a Medical Device. *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
FDA (2025c) De Novo Classification Request. *FDA*. Retrieved (February 1, 2026), https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/de-novo-classification-request.
FDA (2025d) Embolization Device Recall and Correction: Medtronic Removes Unused 027 Compatible Pipeline Vantage Embolization Device with Shield Technology, Updates Use Instructions for 021 Compatible Pipeline Vantage Embolization Devices due to Increased Risk of Incomplete Wall Apposition and Braid Deformation. *FDA*. Retrieved (January 1, 2026), https://www.fda.gov/medical-devices/medical-device-recalls/embolization-device-recall-and-correction-medtronic-removes-unused-027-compatible-pipeline-vantage.
Fellegi IP, Sunter AB (1969) A Theory for Record Linkage. *Journal of the American Statistical Association* 64(328):1183–1210.
Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M (2021) Do as AI Say: Susceptibility in Deployment of Clinical Decision-aids. *npj Digital Medicine* 4(1):31.
Haenlein M, Kaplan A (2019) A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review* 61(4):5–14.
He J, Baxter SL, Xu Jie, Xu Jiming, Zhou X, Zhang K (2019) The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nature medicine* 25(1):30–36.
Hou T, Li M, Tan Y (Ricky), Zhao H (2024) Physician Adoption of AI Assistant. *Manufacturing & Service Operations Management* 26(5):1639–1655.
Institute of Medicine (2011) *Medical Devices and the Public's Health: The FDA 510(k) Clearance Process at 35 Years* (National Academies Press, Washington, D.C.).
Jaccard P (1912) The Distribution of the Flora in the Alpine Zone. *New Phytologist* 11(2):37–50.
Jussupow E, Benbasat I, Heinzl A (2024) An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making1. *MIS Quarterly* 48(4):1575–1590.
Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research* 32(3):713–735.
Kulkarni V, Gawali M, Kharat A (2021) Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice. *JMIR Medical Informatics* 9(9):e28776.
Lee B, Kramer P, Sandri S, Chanda R, Favorito C, Nasef O, Ross JS, Sharfstein J, Dai T (2025) Early Recalls and Clinical Validation Gaps in Artificial Intelligence–Enabled Medical Devices. *JAMA Health Forum* 6(8):e253172.
Lee B, Patel S, Favorito C, Sandri S, Jennings MR, Dai T (2025) Development and Commercialization Pathways of AI Medical Devices in the United States: Implications for Safety and Regulatory Oversight. *NEJM AI* 0(0):AIra2500061.
Lenharo M (2024) The Testing of AI in Medicine Is a Mess. Here's How It Should Be Done. *Nature* 632(8026):722–724.
Lin JC, Jain B, Iyer JM, Rola I, Srinivasan AR, Kang C, Patel H, Parikh RB (2025) Benefit-Risk Reporting for FDA-Cleared Artificial Intelligence−Enabled Medical Devices. *JAMA Health Forum* 6(9):e253351.
Lyytinen K, Rose GM (2003) The Disruptive Nature of Information Technology Innovations: The Case of Internet Computing in Systems Development Organizations. *MIS Quarterly* 27(4):557–596.
Mayer AS, Strich F, University of Passau (Germany), Fiedler M (2020) Unintended Consequences of Introducing AI Systems for Decision Making. *MSQE*:239–257.
Morrow E, Zidaru T, Ross F, Mason C, Patel KD, Ream M, Stockley R (2023) Artificial Intelligence Technologies and Compassion in Healthcare: A Systematic Scoping Review. *Frontiers in Psychology* 13.
Mundasad S (2024) Medical Device Bias Needs Immediate Action - Review. Retrieved (January 1, 2026), https://www.bbc.com/news/health-68533067.
Muralidharan V, Adewale BA, Huang CJ, Nta MT, Ademiju PO, Pathmarajah P, Hang MK, et al. (2024) A Scoping Review of Reporting Gaps in FDA-Approved AI Medical Devices. *npj Digital Medicine* 7(1):273.
Nong P, Adler-Milstein J, Apathy NC, Holmgren AJ, Everson J (2025) Current Use And Evaluation Of Artificial Intelligence And Predictive Models In US Hospitals. *Health Affairs* 44(1):90–98.

Obermeyer Z, Emanuel EJ (2016) Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine* 375(13):1216–1219.
OpenAI (2025) Introducing GPT-5. Retrieved (January 1, 2026), https://openai.com/index/introducing-gpt-5/.
Palmieri S, Robertson CT, Cohen IG (2026) New Guidance on Responsible Use of AI. *JAMA* 335(3):207–208.
Rai A (2020) Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science* 48(1):137–141.
Rajkomar A, Dean J, Kohane I (2019) Machine Learning in Medicine. *New England Journal of Medicine* 380(14):1347–1358.
Rathi VK, Ross JS (2019) Modernizing the FDA's 510(k) Pathway. *New England Journal of Medicine* 381(20):1891–1893.
Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, et al. (2021) Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans. *Nature Machine Intelligence* 3(3):199–217.
Sarker S, Chatterjee S, Xiao X, Elbanna A (2019) The Sociotechnical Axis of Cohesion for the is Discipline: Its Historical Legacy and its Continued Relevance1. *MIS Quarterly* 43(3):695–719.
Sedgwick (2023) 2023 State of the Nation Recall Index: Essential insight and analysis for mitigating recall risk. Retrieved (January 1, 2026), https://marketing.sedgwick.com/acton/media/4952/us-sotn-2023-index-report-download-pr.
Sendak M, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, Ratliff W, Balu S (2020) A Path for Translation of Machine Learning Products into Healthcare Delivery. *EMJ Innovations*.
Singh R, Bapna M, Diab AR, Ruiz ES, Lotter W (2025) How AI is Used in FDA-Authorized Medical Devices: A Taxonomy Across 1,016 Authorizations. *npj Digital Medicine* 8(1):388.
Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS (2020) Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine* 383(25):2477–2478.
Topol E (2019) *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (Basic Books).
Turel O, Kalhan S (2023) Prejudiced Against the Machine? Implicit Associations And the Transience of Algorithm Aversion. *MIS Quarterly* 47(4):1369–1394.
Vo CD, Jiang B, Azad TD, Crawford NR, Bydon A, Theodore N (2020) Robotic Spine Surgery: Current State in Minimally Invasive Surgery. *Global Spine Journal* 10(2):34S-40S.
Warraich HJ, Tazbaz T, Califf RM (2024) FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. *JAMA* 333(3):241–247.
Windecker D, Baj G, Shiri I, Kazaj PM, Kaesmacher J, Gräni C, Siontis GCM (2025) Generalizability of FDA-Approved AI-Enabled Medical Devices for Clinical Use. *JAMA Network Open* 8(4):e258052.
Yin J, Ngiam KY, Tan SSL, Teo HH (2025) Designing AI-Based Work Processes: How the Timing of AI Advice Affects Diagnostic Decision Making. *Management Science* 71(11):9361–9383.
Zhu Y, Everhart A, Mandic PK, Sen S (2020) Using NLP to Extract Predicate History from Medical Device Approvals. *2020 International Conference on Information Systems (ICIS)*
Zhu Y, Sen S, Everhart A, Karaca-Mandic P (2025) A Deep Learning Approach for Predicting FDA's 510(k) Medical Device Recalls Using Device Citation Relationships. *Information Systems Research*:Published Online.