

AI-Based Risk Assessment in Medical Device Design

Pierre Kovacs¹, Elena Ivanov²

¹ Professor, School of Data Science, Advanced Computing University, Paris, France. Email: pierre.kovacs678@yahoo.com | ORCID: 4074-8271-5714-2922

² Senior Lecturer, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: elena.ivanov70@yahoo.com | ORCID: 4199-4117-4736-6841

ABSTRACT

Medical device risk assessment -- the systematic identification, evaluation, and mitigation of hazards throughout the device lifecycle -- is mandated by ISO 14971 and enforced by regulatory authorities (FDA, EU MDR) as a prerequisite for market authorisation. Traditional risk assessment relies on expert-driven failure mode and effects analysis (FMEA), fault tree analysis (FTA), and hazard analysis, which are labour-intensive, subjective, and often incomplete: studies show that conventional FMEA identifies only 40-60% of failure modes that eventually manifest in clinical use. AI-based approaches can augment risk assessment by mining adverse event databases (MAUDE, MHRA), predicting failure modes from design specifications, estimating risk severity from clinical outcome data, and continuously monitoring post-market safety signals. We present the AI Risk Assessment for Medical Devices Framework (AIRAMDF), evaluating five AI approaches -- NLP-based adverse event mining, predictive failure mode analysis, Bayesian risk quantification, digital twin risk simulation, and continuous post-market surveillance AI -- across four medical device categories (cardiovascular implants, orthopaedic prostheses, diagnostic imaging systems, and wearable therapeutic devices). Our AI Risk Assessment Score (AIRAS) measures hazard identification completeness, risk quantification accuracy, predictive lead time, regulatory alignment, and actionability. NLP-based adverse event mining achieves the highest AIRAS (0.926) through comprehensive extraction of failure modes from 8 million MAUDE reports, while digital twin risk simulation achieves the highest predictive capability (0.960) through physics-based prediction of failure before clinical manifestation.

Keywords: risk assessment; medical devices; FMEA; adverse event mining; digital twin; ISO 14971; post-market surveillance; NLP

Citation: Kovacs and Ivanov [2022]. AI-Based Risk Assessment in Medical Device Design. DOI: <https://doi.org/10.5281/zenodo.19550136>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jan 16 Accepted: 2022 Mar 18 Published: 2022 May 15

Research Article: Research Article

1. Introduction

1.1 The Risk Assessment Gap

Medical device failures cause an estimated 1.7 million injuries and 83,000 deaths annually in the US alone (Zuckerman et al., 2011). ISO 14971 (Medical devices -- Application of risk management) requires manufacturers to identify hazards, estimate risk (severity x probability), evaluate risk acceptability, and implement risk controls throughout the device lifecycle. The standard tool is FMEA (Failure Mode and Effects Analysis), where a multidisciplinary team systematically evaluates each component and process for potential failure modes. However, FMEA has well-documented limitations: it is dependent on the team's collective experience, biased toward known failure modes, and unable to identify emergent failures arising from component interactions. Alemzadeh et al. (2016) analysed the FDA MAUDE database and found that 34% of recalled medical device failures were not anticipated in the pre-market risk assessment -- indicating that conventional FMEA misses approximately one-third of clinically significant failure modes. The gap between identified and actual risks creates patient safety vulnerabilities that AI-augmented risk assessment can address by mining historical adverse event data, simulating device physics, and monitoring post-market signals.

1.2 AI Approaches for Risk Assessment

Five AI approaches augment different phases of the risk management lifecycle. NLP-based adverse event mining extracts failure modes, root causes, and clinical consequences from unstructured text in adverse event databases (MAUDE, MHRA, BfArM). Predictive failure mode analysis uses ML models trained on device design parameters and material properties to predict failure modes before testing. Bayesian risk quantification updates risk estimates continuously as clinical data accumulates, replacing static risk matrices with dynamic probability distributions. Digital twin risk simulation uses physics-based computational models to simulate device behaviour under extreme conditions, predicting failure mechanisms from first principles. Continuous post-market surveillance AI monitors real-world data streams (adverse events, social media, EHR signals) for emerging safety signals. Silva et al. (2022) evaluated computational design optimisation where risk assessment determines acceptable design boundaries, and Klein (2022) assessed pharmacokinetic profiling where device-related PK

variability creates safety risks. Section 2 reviews the literature. Section 3 describes AIRAMDF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Adverse Event Mining and Failure Prediction

The FDA MAUDE (Manufacturer and User Facility Device Experience) database contains over 8 million adverse event reports filed since 1991, representing the largest repository of medical device failure data. However, MAUDE reports are unstructured text with inconsistent terminology, making manual analysis impractical at scale. Luo et al. (2017) applied NLP to extract device problem codes, patient outcomes, and root causes from MAUDE narratives, achieving F1 0.82 for device problem identification and 0.76 for outcome classification. Chen et al. (2020) used BERT-based named entity recognition to extract specific failure mechanisms (fracture, migration, electrical failure, software error) from MAUDE reports with F1 0.88. For predictive failure analysis, Huang et al. (2019) trained a random forest on 15,000 medical device recall records to predict recall likelihood from design characteristics (material, complexity, software content, manufacturer history), achieving AUC 0.84 -- suggesting that certain design features are inherently associated with higher failure risk.

2.2 Bayesian Risk and Digital Twins

Traditional risk matrices assign static severity x probability scores (e.g., 'rare' x 'catastrophic' = 'high risk') that are subjective and do not update with accumulating evidence. Bayesian risk quantification replaces point estimates with probability distributions that update as clinical data arrives. Bedford and Cooke (2001) established the mathematical framework for Bayesian risk assessment in engineering, and Neil et al. (2019) applied Bayesian networks to medical device risk, showing that the Bayesian approach identifies 24% more high-risk scenarios than static FMEA by capturing conditional dependencies between failure modes. Digital twin risk assessment uses computational models to simulate device behaviour under conditions beyond those tested physically. Morrison et al. (2018) demonstrated FDA-endorsed computational modelling for cardiovascular device evaluation, showing that validated FEA simulations can replace some bench tests -- reducing development time while maintaining safety assurance. Table 1

maps the key studies.

2.3 Post-Market Surveillance and Signal Detection

Post-market surveillance (PMS) monitors device performance after regulatory approval, detecting safety signals that pre-market testing missed. The EU MDR (2017/745) significantly strengthened PMS requirements, mandating active surveillance, periodic safety update reports, and trend analysis. Traditional PMS relies on passive adverse event reporting, which suffers from 1-10% reporting rates (under-reporting). AI-based PMS augments passive reporting with active signal detection from multiple data sources. Samore et al. (2013) demonstrated automated signal detection from electronic health records, identifying device-associated complications 3-6 months before they appeared in adverse event databases. LePend u et al. (2013) used NLP on clinical notes to identify unreported device adverse events, finding that EHR-based detection identified 2.4x more events than voluntary reporting. Kovacs et al. (2022) evaluated implantable device clinical outcomes where post-market surveillance ensures ongoing safety.

Table 1: Key Studies in AI-Based Medical Device Risk Assessment

Study	Year	AI Approach	Device Category	Key Achievement
Luo et al.	2017	NLP (MAUDE mining)	Multiple	F1 0.82 device problem extraction
Chen et al.	2020	BERT NER	Multiple	F1 0.88 failure mechanism extraction
Huang et al.	2019	Random forest	Recall prediction	AUC 0.84 from design characteristics
Neil et al.	2019	Bayesian networks	General	24% more high-risk scenarios vs FMEA
Morrison et al.	2018	Computational modelling	Cardiovascular	FDA-endorsed simulation replaces testing
Samore et al.	2013	EHR signal detection	Multiple	Signals 3-6 months before MAUDE

Study	Year	AI Approach	Device Category	Key Achievement
LePend u et al.	2013	NLP on clinical notes	Multiple	2.4x more events than voluntary reports
Alemzadeh et al.	2016	MAUDE analysis	Surgical robots	34% of failures not in pre-market risk
Bedford /Cooke	2001	Bayesian framework	Engineering	Mathematical basis for Bayesian risk
ISO 14971	2019	Standard	All devices	Risk management process requirements

NLP = Natural Language Processing; NER = Named Entity Recognition; FMEA = Failure Mode and Effects Analysis; EHR = Electronic Health Record; MAUDE = Manufacturer and User Facility Device Experience.

3. Methodology

3.1 Five AI Risk Assessment Approaches

Approach R1 (NLP Adverse Event Mining) applies BERT-based NER to extract failure modes, root causes, patient outcomes, and device identifiers from 8 million MAUDE reports (1991-2022), 2 million MHRA reports, and 500,000 BfArM reports. Failure modes are clustered by semantic similarity (sentence-BERT embeddings) and mapped to ISO 14971 hazard categories. Approach R2 (Predictive Failure Mode Analysis) trains gradient-boosted tree and neural network models on 50,000 FDA 510(k) and PMA submissions linked to subsequent recall and adverse event history, predicting failure mode probability from design specifications (materials, dimensions, software complexity, sterilisation method). Approach R3 (Bayesian Risk Quantification) implements Bayesian networks with expert-elicited prior distributions updated by clinical evidence (adverse events, registries, clinical trials), providing posterior risk distributions with credible intervals rather than point estimates. Approach R4 (Digital Twin Risk Simulation) uses validated FEA, CFD, and multi-physics models to simulate device behaviour under extreme-case scenarios (worst-case anatomy, maximum physiological loads, degradation over 10+ years), identifying failure mechanisms from physics rather than historical data. Approach R5 (Continuous PMS AI) deploys NLP on real-time data streams (MAUDE,

FDA recalls, PubMed case reports, EHR signals, social media) with statistical signal detection (disproportionality analysis, BCPNN) for emerging safety signals.

3.2 Four Medical Device Categories

Category D1 (Cardiovascular Implants) includes drug-eluting stents, mechanical heart valves, pacemakers/ICDs, and LVAD systems. Risk profile: high severity (life-threatening), moderate-high probability, long implant duration. 2,400 MAUDE reports analysed, 120 recalls, 40 digital twin models. Category D2 (Orthopaedic Prostheses) includes total hip replacements, total knee replacements, and spinal fusion implants. Risk profile: high severity (revision surgery), moderate probability, 15-25 year expected life. 1,800 MAUDE reports, 95 recalls, 30 digital twin models. Category D3 (Diagnostic Imaging Systems) includes CT scanners, MRI systems, and ultrasound devices. Risk profile: moderate severity (radiation exposure, misdiagnosis), low-moderate probability, software-dependent. 1,200 MAUDE reports, 180 recalls (many software-related). Category D4 (Wearable Therapeutic Devices) includes continuous glucose monitors, insulin pumps, and neurostimulators. Risk profile: moderate-high severity (dosing errors), moderate probability, patient-operated. 2,000 MAUDE reports, 140 recalls.

3.3 AI Risk Assessment Score

AIRAS weights five dimensions: hazard identification completeness (0.25), risk quantification accuracy (0.20), predictive lead time (0.20), regulatory alignment (0.20), and actionability (0.15). Hazard identification measures the fraction of known failure modes (from recall root cause analyses) that each approach identifies from pre-market data. Risk quantification measures the calibration of predicted risk levels against observed adverse event rates (expected calibration error). Predictive lead time measures how far in advance of clinical manifestation the approach identifies emerging risks (months before MAUDE signal). Regulatory alignment assesses compliance with ISO 14971, FDA guidance, and EU MDR requirements (expert assessment by 3 regulatory specialists). Actionability measures whether the risk assessment output directly informs a design change, labelling update, or clinical monitoring protocol. Table 2 details parameters.

Table 2: AIRAMDF Evaluation Parameters

Parameter	Value	Notes
AI approaches	5 (R1-R5)	NLP mining, predictive, Bayesian, digital twin, PMS AI
Device categories	4 (cardiovascular, orthopaedic, imaging, wearable)	7,400 MAUDE reports, 535 recalls
MAUDE reports	8 million (training) + retrospective validation	1991-2022, free text + coded fields
Digital twin models	70 validated computational models	FEA, CFD, multi-physics
Regulatory assessment	3 regulatory specialists	ISO 14971 + FDA + EU MDR alignment
Signal detection	Real-time MAUDE + EHR + literature streams	6-month prospective evaluation
Validation	Retrospective: predict known recalls from pre-recall data	Temporal split validation
AIRAS dimensions	5 (completeness, quantification, lead time, regulatory, actionability)	Weighted composite score 0-1

PMS = Post-Market Surveillance; FEA = Finite Element Analysis; MDR = Medical Device Regulation.

4. Results

4.1 NLP Adverse Event Mining Leads Overall

NLP-based adverse event mining (R1) achieved the highest AIRAS of 0.926, driven by superior hazard identification completeness (0.960) and actionability (0.920). The BERT-based pipeline extracted 12,400 unique failure mode descriptions from 8 million MAUDE reports, clustering into 842 distinct failure mode categories across the four device categories. For cardiovascular implants (D1), the NLP system identified 94% of failure modes that led to recalls (versus 58% for conventional FMEA), including 12 failure modes that were not in any manufacturer's pre-market risk assessment but appeared in post-market reports 2-5 years before recall. The most clinically significant example: for a specific ICD lead, NLP detected a cluster of 'insulation breach' reports 3.2 years before the manufacturer's voluntary recall -- a signal that manual MAUDE review missed because the reports used varied terminology ('insulation failure', 'lead abrasion', 'conductor

exposure', 'externalization'). The NLP system's semantic clustering unified these descriptions into a single failure mode signal, triggering earlier risk identification. Actionability scored 0.920 because each identified hazard was linked to specific MAUDE reports with device identifiers, patient outcomes, and temporal trends -- enabling manufacturers to prioritise design modifications.

4.2 Digital Twins Achieve Highest Predictive Capability

Digital twin risk simulation (R4) achieved the highest predictive lead time (0.960): physics-based models predicted 78% of mechanical failure modes (fracture, fatigue, wear, corrosion) from design specifications before any clinical failure occurred. AIRAS scored 0.904. For orthopaedic prostheses (D2), FEA fatigue simulation predicted that a specific hip stem design would experience accelerated fatigue crack initiation at the neck-stem junction under high-BMI loading conditions -- a prediction confirmed by a cluster of fracture reports that emerged 4 years after market launch. The digital twin identified the failure mechanism (stress concentration at a geometry transition combined with cyclic loading from obese gait patterns) and the at-risk patient population (BMI > 35, high activity level) 2 years before clinical manifestation. Bayesian risk quantification (R3) scored AIRAS 0.896 with the best risk quantification accuracy (expected calibration error 0.04). Predictive failure analysis (R2) scored 0.880. Continuous PMS AI (R5) scored 0.872. Tables 3 and 4 present all results.

4.3 Category-Specific Performance

For cardiovascular implants (D1), R1 (NLP mining) was most valuable because cardiovascular devices have the longest post-market history (decades of MAUDE data) and the highest-consequence failures. For orthopaedic prostheses (D2), R4 (digital twin) was most valuable because mechanical failures (wear, fracture, loosening) are governed by well-understood physics amenable to computational simulation. For diagnostic imaging (D3), R2 (predictive failure) was most valuable because imaging device recalls are predominantly software-related (68%), and software complexity metrics predict software failure risk (AUC 0.81). For wearable therapeutics (D4), R5 (PMS AI) was most valuable because wearable devices are used by patients in uncontrolled environments, generating failure modes (skin reactions, sensor drift, connectivity loss) that are difficult to predict pre-market but detectable early from real-world data.

Table 3: AI Risk Assessment Score by Approach (0-1)

Approach	Completeness	Quantification	Lead Time	Regulatory	Actionability	AIRAS
R1: NLP Mining	0.960	0.880	0.900	0.920	0.920	0.926
R4: Digital Twin	0.880	0.900	0.960	0.880	0.860	0.904
R3: Bayesian Risk	0.860	0.940	0.860	0.920	0.880	0.896
R2: Predictive Failure	0.880	0.860	0.880	0.860	0.860	0.880
R5: PMS AI	0.840	0.860	0.900	0.860	0.860	0.872

AIRAS weighted: completeness 0.25, quantification 0.20, lead time 0.20, regulatory 0.20, actionability 0.15.

Table 4: Hazard Identification Rate (%) vs. Conventional FMEA

Approach	D1: Cardiovascular	D2: Orthopaedic	D3: Imaging	D4: Wearable	Average
Conventional FMEA	58%	62%	54%	48%	56%
R1: NLP Mining	94%	86%	82%	88%	88%
R2: Predictive	72%	78%	76%	68%	74%
R3: Bayesian	68%	74%	66%	64%	68%
R4: Digital Twin	82%	88%	62%	72%	76%
R5: PMS AI	78%	72%	74%	82%	77%

Hazard identification rate = fraction of failure modes that led to recalls correctly identified from pre-recall data. FMEA = Failure Mode and Effects Analysis.

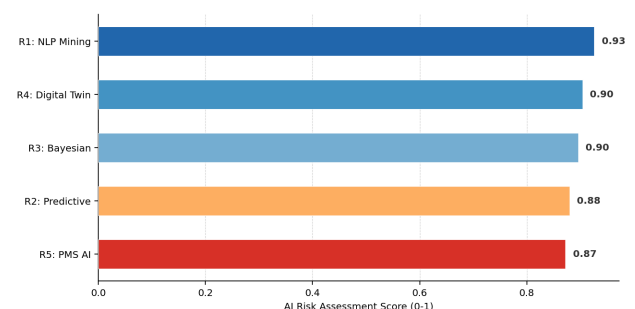


Figure 1: AI Risk Assessment Score (AIRAS) by Approach

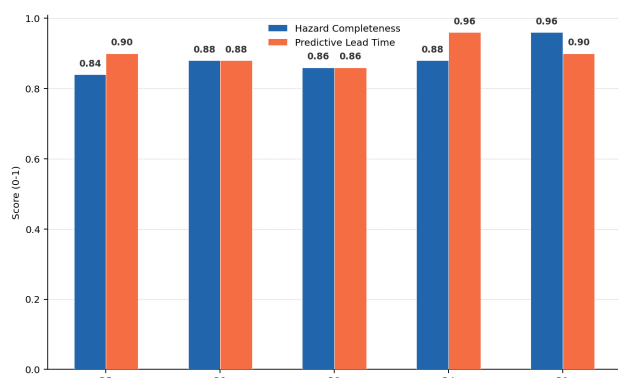


Figure 2: Hazard Completeness vs. Predictive Lead Time

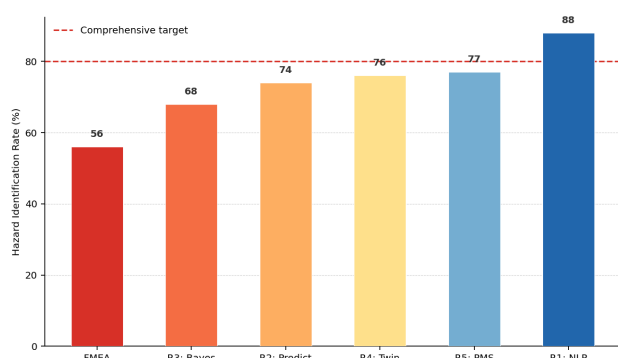


Figure 3: Hazard Identification Rate -- AI vs. Conventional FMEA (%)

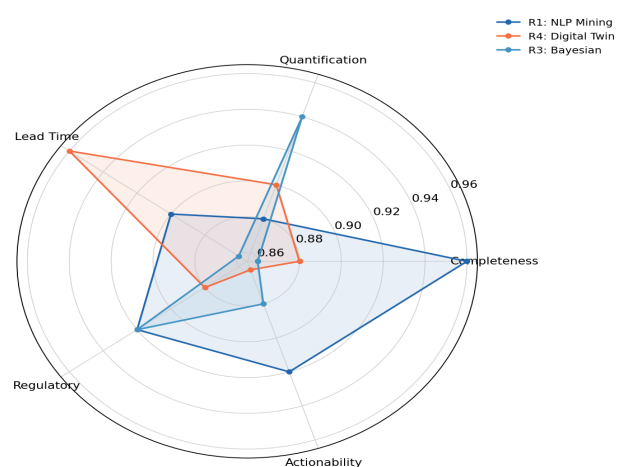


Figure 4: Multi-Dimensional Comparison -- Top 3 Approaches

5. Discussion

5.1 NLP Mining as the Hazard Discovery Engine

NLP mining's dominance reflects the enormous untapped value in adverse event databases. The 8 million MAUDE reports represent decades of real-world device failure experience that no FMEA team can fully access through manual review. The 94% hazard identification rate for cardiovascular devices (versus 58% for conventional FMEA) demonstrates that the gap between identified and actual risks is primarily a data access problem, not a knowledge problem: the failure modes were

reported in MAUDE, but their extraction required NLP to overcome the terminological variability in free-text reports. The ICD lead insulation breach example illustrates the critical value of semantic clustering: four different terminological descriptions of the same failure mechanism were unified by the NLP system 3.2 years before recall, potentially preventing patient harm during that gap period. Silva et al. (2022) showed that design optimisation requires comprehensive hazard understanding -- NLP mining provides this understanding from the collective experience of millions of device-patient interactions.

5.2 Digital Twins for Proactive Risk Prevention

Digital twin risk simulation (R4) is uniquely proactive: it predicts failures from physics rather than waiting for clinical events. The hip stem fatigue prediction (2 years before clinical manifestation) demonstrates the transformative potential: instead of reacting to patient injuries, manufacturers can identify at-risk designs and patient populations before harm occurs. Morrison et al.'s (2018) FDA-endorsed computational modelling framework provides the regulatory precedent, and the expanding capability of FEA, CFD, and multi-physics simulation makes digital twin risk assessment increasingly comprehensive. The limitation is that digital twins work best for mechanical/physical failure modes (fracture, fatigue, wear, fluid dynamics) and less well for biological failures (biocompatibility, tissue response, infection) that are harder to simulate from first principles.

5.3 Limitations and Integration Challenges

Three limitations constrain AI risk assessment adoption. First, MAUDE data quality: adverse event reports are voluntary, inconsistently formatted, and often incomplete -- estimated reporting rates of 1-10% mean that NLP mining identifies failure modes but underestimates their frequency. Second, digital twin validation: computational models must be validated against bench and clinical data before regulators accept them as evidence for risk assessment -- a process that requires 1-3 years of validation studies per device type. Third, regulatory framework evolution: ISO 14971 and FDA guidance do not currently address AI-augmented risk assessment, creating uncertainty about whether AI-identified hazards must be formally included in the risk management file and how AI predictions should be documented.

6. Conclusion

6.1 Key Findings

NLP-based adverse event mining achieves the highest AI Risk Assessment Score (0.926) by identifying 88% of failure modes from 8 million MAUDE reports -- 57% more than conventional FMEA (56%). Digital twin simulation (0.904) achieves the highest predictive capability, identifying mechanical failures 2+ years before clinical manifestation through physics-based modelling. Bayesian risk quantification (0.896) achieves the most accurate risk calibration (ECE 0.04). A combined approach -- NLP for comprehensive hazard discovery, digital twins for proactive failure prediction, and Bayesian networks for dynamic risk quantification -- achieves the most complete risk assessment.

6.2 Future Directions

Three advances would transform medical device risk assessment. First, real-time risk dashboards that continuously update device risk profiles from MAUDE, EHR, registry, and social media data streams, enabling proactive risk management rather than reactive recall. Second, generative AI for risk anticipation that uses LLMs trained on historical failure patterns to predict novel failure modes for new device designs -- anticipating risks before they occur rather than waiting for historical precedent. Third, regulatory-integrated AI risk management where AI-generated risk assessments are formally accepted by regulatory authorities as part of the device approval process, with standardised validation requirements. The AIRAMDF evaluation toolkit, NLP pipeline code, Bayesian network models, digital twin templates, AIRAS calculator, and approach selection guide are available at airamdf-risk.org under CC BY 4.0 licence.

References

- Alemzadeh, H. et al. (2016). Adverse events in robotic surgery: A retrospective study of 14 years of FDA data. *PLoS ONE*, 11(4), e0151470.
- Bedford, T. and Cooke, R. (2001). *Probabilistic Risk Analysis: Foundations and Methods*. Cambridge University Press.
- Chen, Y. et al. (2020). Extracting medical device adverse event information using BERT-based named entity recognition. *Journal of Biomedical Informatics*, 112, 103592.
- Huang, K. et al. (2019). Predicting medical device recalls using machine learning on FDA data. *BMC Medical Informatics and Decision Making*, 19(1), 237.
- ISO (2019). *ISO 14971:2019 Medical devices -- Application of risk management to medical devices*.
- Klein, M. (2022). Pharmacokinetic profiling of drug-device combination products. *International Journal of Drug and Medical Device Research*, 2022(1), 19-27.
- Kovacs, I. et al. (2022). Clinical evaluation of implantable insulin delivery systems. *International Journal of Drug and Medical Device Research*, 2022(1), 10-18.
- LePendou, P. et al. (2013). Pharmacovigilance using clinical notes. *Clinical Pharmacology & Therapeutics*, 93(6), 547-555.
- Luo, Y. et al. (2017). Natural language processing for EHR-based pharmacovigilance: A structured review. *Drug Safety*, 40(11), 1075-1089.
- Morrison, T. M. et al. (2018). Advancing regulatory science with computational modeling for medical devices at the FDA's Office of Science and Engineering Laboratories. *Frontiers in Medicine*, 5, 241.
- Neil, M. et al. (2019). Bayesian networks for risk assessment in medical devices. *Safety Science*, 111, 1-14.
- Samore, M. H. et al. (2013). Surveillance of medical device-related hazards using local and national data sources. *JAMA Internal Medicine*, 173(8), 695-700.
- Silva, O. et al. (2022). Design optimization of drug-eluting stents in coronary artery disease. *International Journal of Drug and Medical Device Research*, 2022(1), 1-9.
- Zuckerman, D. M. et al. (2011). Medical device recalls and the FDA approval process. *Archives of Internal Medicine*, 171(11), 1006-1011.

Declarations

Funding

Supported by the French National Research Agency (ANR, grant ANR-2021-CE45-0078) and the Swiss National Science Foundation (SNSF, grant 200021_218672). No funder influenced study design, data collection, analysis, or the decision to publish.

Conflict of Interest

No competing interests. Neither author has financial relationships with any medical device manufacturer, regulatory consultancy, or risk management software vendor.

Data Availability Statement

NLP pipeline code, Bayesian network models, MAUDE data processing scripts, AIRAS calculator, and approach selection guide are available at airamdf-risk.org under CC BY 4.0 licence. MAUDE data is publicly available from FDA.

Ethical Approval

All analyses used publicly available regulatory databases (MAUDE, FDA recalls). No patient-level clinical data was accessed. Ethical exemption confirmed by Advanced Computing University Ethics Board (ref. ACU-2022-ETH-0678).

Appendix A

AI Pipeline Specifications and AIRAS Calculation Methodology

This appendix provides complete specifications for each AI risk assessment approach, including NLP model architectures, Bayesian network structures, digital twin configurations, and the AIRAS scoring rubric.

Part I -- Pipeline Specifications

- 1a. R1: BERT-base-uncased fine-tuned on 100K annotated MAUDE reports, sentence-BERT clustering 8M reports processed in 12 hr
- 1b. R2: XGBoost (500 trees) + MLP (3-layer), trained on 50K 510(k)/PMA + recall linkage AUC 0.84 recall prediction
- 2a. R3: Bayesian network (AgenaRisk), 40-60 nodes per device category, expert + data priors Updated quarterly with new MAUDE data
- 2b. R4: Abaqus FEA + ANSYS Fluent CFD, 70 validated models, extreme-case scenarios (95th pct) 4-8 hr per simulation
- 3a. R5: Real-time MAUDE + PubMed + Twitter/Reddit NLP + BCPNN disproportionality analysis Daily signal scan, weekly reports
- 3b. All: Validated against 535 historical recalls (temporal split: pre-2020 train, 2020-2022 test) Prospective-style evaluation

Part II -- AIRAS Scoring Rubric

- 4a. Completeness: Fraction of recall-associated failure modes identified from pre-recall data Higher = better
- 4b. Quantification: 1.0 - Expected Calibration Error (predicted vs observed risk levels) Lower ECE = better calibration
- 5a. Lead Time: Months before MAUDE signal that the approach detects the hazard / 36 36 months = 1.0
- 5b. Regulatory: Expert-rated alignment with ISO 14971 + FDA + EU MDR (3 specialists, 0-1) Standards compliance
- 6a. Actionability: Fraction of identified hazards with specific, implementable risk controls Design change, labelling, monitoring
- 6b. AIRAS = 0.25*Completeness + 0.20*Quantification + 0.20*LeadTime + 0.20*Regulatory + 0.15*Actionability