Advancements in Clinical Evaluation and Regulatory Frameworks for AI-Driven Software as a Medical Device (SaMD)

Shiau-Ru Yang, Jen-Tzung Chien, *Senior Member, IEEE*, and Chen-Yi Lee

***Abstract*—Owing to the rapid progress in artificial intelligence (AI) and the widespread use of generative learning, the problem of sparse data has been solved effectively in various research fields. The application of AI technologies has resulted in important transformations in healthcare, particularly in radiology. To ensure the high quality, safety, and effectiveness of AI and machine learning (ML) medical devices, the US Food and Drug Administration (FDA) has established regulatory guidelines to support the performance evaluation of medical devices. Furthermore, the FDA has proposed continuous surveillance requirements for AI/ML medical devices. This paper presents a summary of SaMD products that have passed the FDA 510 (k) AI/ML pathway, the challenges associated with the current AI/ML software-as-a-medical-device, and solutions for promoting the development of AI technologies in medicine. We hope to provide valuable information pertaining to medical-device design, development, and monitoring to ultimately achieve safer and more effective personalized medical services.**

***Index Terms*—Software as a medical device (SaMD), AI/ML, computer-aided detection (CADe), computer-aided diagnosis (CADx), computer-aided triage (CADt).**

***Impact Statement*—This paper summarizes SaMD products that have passed the FDA 510 (k) AI/ML pathway, examines the current situation and challenges of AI/ML SaMD, and presents potential solutions.**

I. Introduction

MEDICAL products must be reviewed and approved by national regulatory authorities before they can be marketed. Traditional regulatory frameworks mainly target hardware and pharmaceutical products, whereas software as a medical device (SaMD) faces many regulatory challenges, particularly when featuring self-learning artificial intelligence/machine learning (AI/ML). Any medical device intended for marketing in the USA must undergo review by the US Food and Drug Administration (FDA). Regulatory authorities in different countries often refer FDA's regulatory policies to formulate their own regulations. In April 2019, the FDA published a series of guidelines for AI/ML SaMDs aimed at defining a suitable regulatory framework. In 2013, the International Medical Device Regulators Forum (IMDRF) established a SaMD working group to compile guidelines to assist manufacturers in developing effective and safe SaMDs. Although regulatory frameworks have been gradually proposed in academic circles and by regulatory authorities, many details require further examination and confirmation. Examples include the design and sample size of clinical trials or evaluations, as well as the size and source of training datasets. These details significantly impact the effectiveness and safety of SaMDs in real-world applications.

In 1998, FDA approved the first mammography computer-aided detection (CAD) system. Subsequently, in 2002, the Centers for Medicare & Medicaid Services (CMS) increased payment for CAD, accelerating its development. By 2016, 90% of radiologic diagnosis centers in the US utilized FDA-approved CAD for evaluating mammography images [1]. However, real-world data analysis revealed that CAD did not significantly enhance the precision of radiologist diagnosis [2]. Part of the reason lied in the fact that while CAD systems heightened image sensitivity, they simultaneously reduced specificity, thereby favoring false positive results. This led patients to unnecessary further tests, requiring over US$400 million/year in unnecessary healthcare insurance expenditures. Consequently, the CMS discontinued additional payment for mammography CAD [3]. The continuous advancements in AI and the widespread applications of generative learning are expected to enable personalized precision medicine in disease diagnosis and prediction within the next decade.

This is poised to propel rapid development of the medical industry; however, to avoid replicating experiences like that of CAD breast cancer detection, emphasis should not solely focus on the sensitivity of smart healthcare software.

Recent research has highlighted potential restrictions faced by the FDA during the evaluation of AI/ML for SaMD applications [5], [6], [7], [8], [9], [10]. For instance, most FDA-approved AI medical devices have relied on retrospective studies, with limited details provided in open-access summary reports regarding the sites and sample sizes used to evaluate these devices. It is imperative that AI/ML medical devices undergo evaluation in real-world screening environments with a broad population basis, accompanied by continuous post-marketing surveillance to assess their performance and long-term effects [3], [11].

Received 8 October 2024; revised 20 October 2024; accepted 20 October 2024. Date of publication 23 October 2024; date of current version 22 November 2024. *(Corresponding author: Chen-Yi Lee.)*

The authors are with the Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: cylee@nycu.edu.tw).

Digital Object Identifier 10.1109/OJEMB.2024.3485534

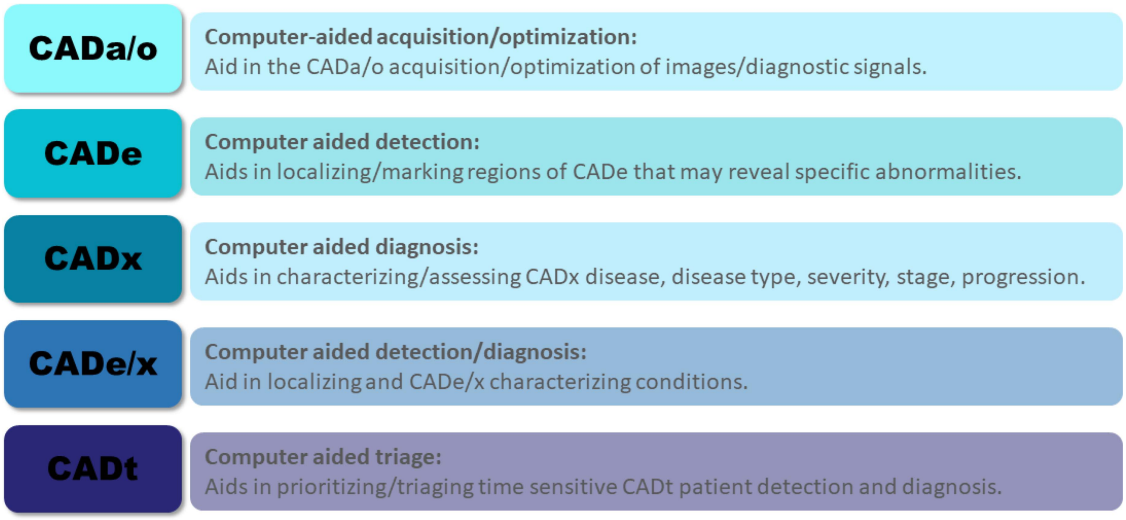

Fig. 1. Based on the intended use of the product, it can be classified into CADa/o, CADe, CADx, CADe/x, and CADt [4].

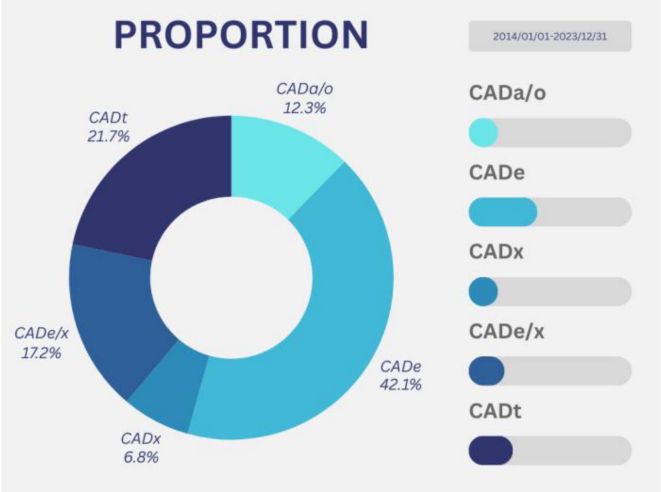

Fig. 2. AI/ML SaMDs approved by the FDA from January 1, 2014, to December 31, 2023. During this period, FDA approved 309 AI/ML SaMDs, comprising 130 CADe (42.1%), 21 CADx (6.8%), 53 CADe/x (17.2%), 67 CADt (21.7%), and 38 CADa/o (12.3%).

To expedite the approval process of AI medical devices by regulatory authorities, while ensuring their effectiveness and safety, in this paper, we offer recommendations for conducting clinical evaluations through small-scale studies before large-scale real-world data assessment. Additionally, we conduct an in-depth analysis of FDA-approved AI medical devices, identifying existing limitations and providing evidence-based support. We hope that our recommendations will significantly aid in the implemntation of AI medical devices.

II. Analysis of FDA 510(k)-Approved AI/ML SaMD

In the US, most SaMDs are currently classified as moderate-risk devices (Class II) and are submitted under FDA 510(k) premarket application. **The applicant must assert that the device has substantial equivalence in terms of safety and effectiveness to existing similar legally approved devices in the market**, thereby demonstrating its suitability for marketing. The 510(k) application is relatively simplified, which saves time and lowers the entry barrier, making it the most common route to enter the US medical device market. We compiled AI/ML SaMDs that were granted 510(k) approval by the FDA and categorized the products based on their intended purpose using FDA classification recommendations. The categories included Computer-Aided Detection (CADe), Computer-Aided Diagnosis (CADx), Computer-Aided Detection/Diagnosis (CADe/x), Computer-Aided Triage (CADt), and Computer-Aided Acquisition/Optimization (CADa/o) (See Fig. 1). This classification aids in subsequent analysis. The IMDRF published a white paper on SaMDs at the end of 2013. Therefore, we compiled AI/ML SaMDs approved by the FDA from January 1, 2014, to December 31, 2023. During this period, FDA approved 309 AI/ML SaMDs, comprising 130 CADe, 21 CADx, 53 CADe/x, 67 CADt, and 38 CADa/o (See Fig. 2).

III. Statistical Design for Clinical Trial or Clinical Evaluation

Clinical evaluation is an indispensable step when submitting a 510(k) application for a SaMD [4], [12]. The objective is to comprehensively evaluate the safety, effectiveness, and performance stability of the device in multiple dimensions so that both users and regulatory authorities can understand the actual performance and potential risks of the product. During the design stage of a clinical trial or evaluation, it is imperative to rigorously define all types of hypotheses that require examination and the statistical methods to be used for analyzing the main results. The objective of the statistical design should be specified, such as comparing superiority, non-inferiority, or equivalence of different treatment methods. These hypotheses will directly influence the sample size, statistical analysis method for the data, and the interpretation and inference of the results. Precise planning of the statistical design can ensure the effectiveness and reliability

TABLE I
CAD CLASSIFICATION STATISTICS

Type	Count	Mentioned in open-access documents													
		Performance		Size of training set		Size of testing set		Sample		Performance & samples present		Multi-center		Observer performance	
CADa/o	38	3	7.89%	0	0.00%	2	5.26%	11	28.95%	3	7.89%	3	7.89%	0	0.00%
CADe	130	58	44.62%	12	9.23%	11	8.46%	62	47.69%	52	40.00%	9	6.92%	10	7.69%
CADx	21	17	80.95%	0	0.00%	0	0.00%	17	80.95%	16	76.19%	4	19.05%	3	14.29%
CADe/x	53	24	45.28%	5	9.43%	4	7.55%	32	60.38%	21	39.62%	8	15.09%	11	20.75%
CADt	67	64	95.52%	3	4.48%	3	4.48%	63	94.03%	62	92.54%	50	74.63%	0	0.00%
Total	**309**	**166**	**53.72%**	**20**	**6.47%**	**20**	**6.47%**	**185**	**59.87%**	**154**	**49.84%**	**74**	**23.95%**	**24**	**7.77%**

TABLE II
COMMON CLINICAL STATISTICAL DESIGN METHODS

Type of comparison	Objective	**Hypothesis A—Head-to-head comparison between the two** Comparison between SaMD and physician	**Hypothesis B—based on observer performance [19]** Comparison between physicians who do and do not use the product
		Primary Endpoints	
Superiority	To prove that the new product is superior to known products in terms of results.	The objective is to prove that the SaMD's results are superior to those of the physician.	The objective is to prove that the performance of physicians who use the product is superior to that of physicians who do not use the product.
Non-inferiority	To prove that the new product is not inferior to known products in terms of results.	The objective is to prove that the SaMD's results are not inferior to those of the physician.	The objective is to prove that the performance of physicians who use the product is not inferior to that of physicians who do not use the product.
Equivalence	To prove that the new product is equivalent to known products in terms of results.	The objective is to prove that the SaMD's results are equivalent to those of the physician.	The objective is to prove that the performance of physicians who use the product is equivalent to that of physicians who do not use the product.

of the device being validated. In the drug or hardware field, there are numerous related professional papers supporting these three clinical trial designs, and guidance manuals have been published by regulatory authorities [13], [14]. However, guidelines for SaMD statistical design are still under development. Given the current situation, we offer relevant recommendations and opinions to enable the industry to provide more substantial scientific evidence when submitting 510(k) applications to the FDA.

In our analysis of 309 devices, only 166 (accounting for 53.72%) explicitly recorded performance data in the public summary (see Table I). Among these, CADa/o had 3 (7.89%), CADe had 58 (44.62%), CADx had 17 (80.95%), CADe/x had 24 (45.28%), and CADt had 64 (95.52%). Of these 309 devices, only 24 (7.77%) mentioned a comparative analysis of performance by healthcare professionals before and after using smart medical devices, with CADa/o having 0, CADe having 10 (7.69%), CADx having 3 (14.29%), CADe/x having 11 (20.75%), and CADt having 0. From this, we infer that 282 devices likely only underwent a unilateral performance comparison, as assumed under Hypothesis A in Table II.

Usually, in academic studies, the performance of AI algorithms is compared with the diagnosis results of physicians to evaluate the reliability of the algorithms. After compiling literature on existing smart healthcare software, Eric J. Topol [15] pointed out that the accuracy of the algorithm is not synonymous with its clinical efficacy. If a comparison (see Hypothesis B in Table I) can be made between physicians who use and do not use the product during clinical trials or evaluation, it will aid in comprehensively confirming the effectiveness and safety of the medical device and ensuring its safe promotion in the real world. In the following sections, we will explain why observer performance studies are necessary based on two dimensions.

A. The Current Primary Objective of AI Medical Devices is Assistance

In the drug and medical device field, the development of new products typically hinges on the concept of replacing old products. Therefore, Hypothesis A—head-to-head comparison between the two—is employed. For instance, the aim of newly developed drugs is to supplant existing drugs on the market, and that of new AI smart healthcare devices is to replace outdated ones. The crux of product comparison lies in evaluating the conformity between two different medical devices [16], [17]. However, considering legal, ethical, and responsibility attribution factors, the objective of medical devices is not to replace physicians for independent diagnosis but rather to serve as aids in improving the efficacy of physicians or medical professionals during diagnosis and treatment. Examples include assisting in disease diagnosis, disease prediction, or triaging for clinical

procedures, as clearly stated in the intended use for most smart healthcare devices, which are articulated as "product X is not intended to provide clinical decisions, medical advice, or evaluations of radiation plans or treatment procedures."

Therefore, if only Hypothesis A—head-to-head comparison between the two devices—is conducted, it would be akin to completing only half the work in clinical evaluation, as only product accuracy is confirmed, while leaving unattended the assessment of whether the device can significantly enhance diagnostic precision when actually employed by physicians or professional users.

B. Future Trends of AI Medical Devices: Adaptive Learning and Man-Machine Collaboration

One of the primary strengths of AI medical devices is their self-learning capacity, which enables them to adapt and enhance their performance. It is anticipated that there will be an increase in 510(k) accreditation applications for such devices, and high-risk smart medical devices will become more prevalent. However, the complexity of these systems, continuous advancements in learning capacity, and user-machine interactions are all factors that must be considered in the clinical evaluation process. Regulatory authorities should broaden the scope of evaluation from a single product to the entire system to ensure the safety and effectiveness of these devices in clinical practice [18]. Physicians with varying years of work experience and specialties may exhibit differences in performance. When actually employed, these devices are often operated by professionally trained medical staff. If the smart healthcare device cannot ensure 100% accuracy, clinical evaluation should comprehensively consider consistency across all users when operating the device to mitigate the risk of medical errors stemming from the variable levels of user experience. The evaluation process should ensure that device design and usage can accommodate operators with varying experience levels while maintaining high performance and safety.

Diagnostic errors are a common cause of medical claims in the USA [20], [21], [22]. We believe that manufacturers should not only focus on the technical precision of the product during the development of AI medical devices but also prioritize collaboration with healthcare professionals for comprehensive clinical evaluation. This collaborative approach will ensure that users with varying experience levels can receive consistent assistance from these devices, thereby mitigating the risk of diagnostic errors in real-world settings. Among the 309 devices analyzed, 271 were non-CADa/o (The risk is relatively low.) that may pose a diagnostic risk. However, only 24 stated that performance analysis was conducted on the product by users

IV. Data Size and Sources for Training and Testing Should Be Statedusing Office Word

Out of the 309 devices considered, 20 provided information on the training set size and source, 185 disclosed the clinical evaluation size and source, and only 19 released both the training and clinical evaluation data.

The size of the training dataset profoundly impacts system performance during the development of deep learning systems. However, in addition to data quantity and quality, the inclusion or exclusion criteria during the selection process also significantly affect model generalization. These factors collectively determine whether the training set can adequately reflect the diversity of the intended use scope of the product [23]. During the submission of the 510(k) applications, the applicant must demonstrate that the safety and effectiveness of the device have substantial equivalence to similar approved products on the market. If complete information, such as training and clinical evaluation data size and sources or product performance metrics (sensitivity and specificity), are unavailable for similar products, the manufacturer may lack sufficient data for evaluation.

For these 309 products, the average time to obtain 510(k) clearance was 140 days. Among them, 120 products were upgraded by the same manufacturer, and both the previous and the current generations were AI/ML products (we compared generations where both were AI/ML products). Out of these 120 products, 70 had a shorter 510(K) clearance time than their predecessors, averaging 83 days, with an average reduction of 63 days (the difference between the previous generation's 510(K) clearance time and the new generation's 510(K) clearance time). Having comprehensive product information available for comparison within the same manufacturer facilitated an accelerated acquisition of 510(K) clearance.

V. Conclusion

Comprehensive clinical evaluation is required when developing AI and ML-driven SaMD in addition to emphasis on its sensitivity. This process should include hypothesis evaluation in observer performance studies to ensure the effectiveness and safety of the product. With regard to AI/ML SaMD, public information should include complete product information, including the data scale and source used for training, as well as clinical evaluation and product performance indicators (such as sensitivity and specificity). After obtaining 510(k) accreditation, large-scale real-world data (RWD) surveillance and optimization are still required for the device as they will not only improve the practicality of the medical device but also improve the overall medical-service quality.

Owing to the widespread application of RWD, real-world evidence (RWE) has become an important basis for medical device performance or safety data. This method is recognized by medical-device regulatory authorities in different countries [24], [25], [26]. To ensure that RWE can effectively support the performance or safety of medical devices, one must formulate hypotheses and objectives during the construction process and design suitable study protocols and statistical-analysis methods. Existing RWE or RWE to be generated in the future should be obtained based on the formulated study methods. Additionally, the correlation between RWE and expected evidence objectives, data reliability, and the appropriateness of study designs for converting RWD to RWE as well as the implementation rigor must be considered when the data are used. The quality of RWE is vital for determining if it can be an important basis for medical-device review [27].

After appropriate RWE is generated, the data provide substantial benefits to the entire life cycle of the medical device. This life cycle includes pre-marketing research and development, changes in intended use, establishment of historical control group, supplementation of clinical data, and post-marketing safety surveillance or study. The quality of RWD is the foundation for RWE generation. Additionally, the reliability of study results and their generalizability to the target population depend on the data-acquisition method and statistical analysis.

The US FDA emphasizes the importance of post-marketing data acquisition in promoting the comprehensive evaluation of medical-device safety and performance. This process not only ensures the effectiveness of the product in actual usage but can also identify potential safety issues in a prompt manner, thereby providing greater assurance to patients. Therefore, continuous data monitoring and analysis is essential to AI/ML SaMD developers as it not only increases the market competitiveness of the product but also promotes improvement to the medical industry.

In summary, AI/ML SaMD development requires a systematic process from product design to clinical evaluation and then post-marketing monitoring, and each step should be performed in a stringent manner. The safety and performance of medical devices will be ensured more comprehensively through the effective use of RWD, which will ultimately improve patient health and medical-service quality.

REFERENCES

[1] J. D. Keen, J. M. Keen, and J. E. Keen, "Utilization of computer-aided detection for digital screening mammography in the United States, 2008 to 2016," *J. Amer. College Radiol.*, vol. 15, no. 1, pp. 44–48, 2018.

[2] C. D. Lehman et al., "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA Intern. Med.*, vol. 175, no. 11, pp. 1828–1837, 2015.

[3] J. G. Elmore and C. I. Lee, "Artificial intelligence in medical imaging—Learning from past mistakes in mammography," *JAMA Health Forum*, vol. 3, no. 2, 2022, Art. no. e215207.

[4] U.S. Food and Drug Administration, "Software as a medical device (SAMD): Clinical evaluation," US Food & Drug Administration, Richmond, VA, USA, 2017. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation

[5] E. Wu, K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou, "How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals," *Nat. Med.*, vol. 27, no. 4, pp. 582–584, 2021.

[6] E. Petersen et al., "Responsible and regulatory conform machine learning for medicine: A survey of challenges and solutions," *IEEE Access*, vol. 10, pp. 58375–58418, 2022.

[7] P. Clark, J. Kim, and Y. Aphinyanaphongs, "Marketing and US food and drug administration clearance of artificial intelligence and machine learning enabled software in and as medical devices: A systematic review," *JAMA Netw. Open*, vol. 6, no. 7, 2023, Art. no. e2321792.

[8] K. Zhang, B. Khosravi, S. Vahdati, and B. J. Erickson, "FDA review of radiologic AI algorithms: Process and challenges," *Radiology*, vol. 310, no. 1, 2024, Art. no. e230242.

[9] S. L. McNamara, P. H. Yi, and W. Lotter, "The clinician-AI interface: Intended use and explainability in FDA-cleared AI devices for medical image interpretation," *NPJ Digit. Med.*, vol. 7, no. 1, 2024, Art. no. 80.

[10] M. Mashar et al., "Artificial intelligence algorithms in health care: Is the current food and drug administration regulation sufficient?," *JMIR AI*, vol. 2, no. 1, 2023, Art. no. e42940.

[11] K. Cao et al., "Large-scale pancreatic cancer detection via non-contrast CT and deep learning," *Nat. Med.*, vol. 29, no. 12, pp. 3033–3043, 2023.

[12] S. S. Shah and A. Gvozdanovic, "Digital health; what do we mean by clinical validation?," *Expert Rev. Med. Devices*, vol. 18, no. 1, pp. 5–8, 2021.

[13] J. A. Lewis, "Statistical principles for clinical trials (ICH E9): An introductory note on an international guideline," *Statist. Med.*, vol. 18, no. 15, pp. 1903–1942, 1999.

[14] E. Ich, "Statistical principles for clinical trials," in *Proc. Int. Conf. Harmonisation*, 1998.

[15] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019.

[16] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[17] H. X. Barnhart, A. S. Kosinski, and M. J. Haber, "Assessing individual agreement," *J. Biopharmaceut. Statist.*, vol. 17, no. 4, pp. 697–719, 2007.

[18] S. Gerke, B. Babic, T. Evgeniou, and I. G. Cohen, "The need for a system view to regulate artificial intelligence/machine learning-based software as medical device," *NPJ Digit. Med.*, vol. 3, no. 1, 2020, Art. no. 53.

[19] N. A. Obuchowski, M. Meziane, A. H. Dachman, M. L. Lieber, and P. J. Mazzone, "What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance?," *Academic Radiol.*, vol. 17, no. 6, pp. 761–767, 2010.

[20] A. C. Schaffer, A. B. Jena, S. A. Seabury, H. Singh, V. Chalasani, and A. Kachalia, "Rates and characteristics of paid malpractice claims among US physicians by specialty, 1992-2014," *JAMA Intern. Med.*, vol. 177, no. 5, pp. 710–718, 2017.

[21] D. E. Newman-Toker et al., "Burden of serious harms from diagnostic error in the USA," *BMJ Qual. Saf.*, vol. 33, no. 2, pp. 109–120, 2024.

[22] A. S. S. Tehrani et al., "25-Year summary of US malpractice claims for diagnostic errors 1986–2010: An analysis from the National Practitioner Data Bank," *BMJ Qual. Saf.*, vol. 22, no. 8, pp. 672–680, 2013.

[23] A. Kleppe, O.-J. Skrede, S. De Raedt, K. Liestøl, D. J. Kerr, and H. E. Danielsen, "Designing deep learning studies in cancer diagnostics," *Nat. Rev. Cancer*, vol. 21, no. 3, pp. 199–211, 2021.

[24] U.S. Food and Drug Administration, "Examples of real-world evidence (RWE) used in medical device regulatory decisions: Selected examples with file summaries, details on real-world data source, populations, and descriptions of use," *Center Devices Radiological Health. Sel. Examples File Summaries, Details Real-World Data Source, Populations, Descriptions Use*, 2021. [Online]. Available: https://www.fda.gov/media/146258/download

[25] U.S. Food and Drug Administration, "Examples of real-world evidence (RWE) used in medical device regulatory decisions," 2019. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices

[26] U.S. Food and Drug Administration, "Real-world evidence," 2022. [Online]. Available: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

[27] F. Liu and D. Panagiotakos, "Real-world data: A brief review of the methods, applications, challenges and opportunities," *BMC Med. Res. Methodol.*, vol. 22, no. 1, 2022, Art. no. 287.