Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations

---

Draft Guidance for Industry and Food and Drug Administration Staff

DRAFT GUIDANCE

This draft guidance document is being distributed for comment purposes only.

Document issued on January 7, 2025.

You should submit comments and suggestions regarding this draft document within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff, Food and Drug Administration, 5630 Fishers Lane, Room 1061, (HFA-305), Rockville, MD 20852-1740. Identify all comments with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions about this document regarding CDRH-regulated devices, contact the Digital Health Center of Excellence at digitalhealth@fda.hhs.gov. For questions about this document regarding CBER-regulated devices, contact the Office of Communication, Outreach, and Development (OCOD) at 1-800-835-4709 or 240-402-8010, or by email at ocod@fda.hhs.gov. For questions about this document regarding CDER-regulated products, contact druginfo@fda.hhs.gov. For questions about this document regarding combination products, contact the Office of Combination Products at combination@fda.gov.

FDA U.S. FOOD & DRUG ADMINISTRATION

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health
Center for Biologics Evaluation and Research
Center for Drug Evaluation and Research
Office of Combination Products in the Office of the Commissioner

Preface

Additional Copies

CDRH

Additional copies are available from the Internet. You may also send an email request to CDRH-Guidance@fda.hhs.gov to receive a copy of the guidance. Please include the document number GUI00007028 and complete title of the guidance in the request.

CBER

Additional copies are available from the Center for Biologics Evaluation and Research (CBER), Office of Communication, Outreach, and Development (OCOD), 10903 New Hampshire Ave., Bldg. 71, Room 3128, Silver Spring, MD 20993-0002, or by calling 1-800-835-4709 or 240-402-8010, by email, ocod@fda.hhs.gov, or from the Internet at https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances.

CDER

Additional copies are available from the Center for Drug Evaluation and Research (CDER), Office of Communication, Division of Drug Information, 10001 New Hampshire Ave., Hillandale Bldg., 4th Floor, Silver Spring, MD 20993-0002, or by calling 855-543-3784 or 301796-3400, or by email, druginfo@fda.hhs.gov, or from the Internet at https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs.

Table of Contents

Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations

---

Draft Guidance for Industry and Food and Drug Administration Staff

This draft guidance, when finalized, will represent the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA staff or Office responsible for this guidance as listed on the title page.

I. Introduction

FDA has long promoted a total product life cycle (TPLC) approach to the oversight of medical devices, including artificial intelligence (AI)-enabled devices, and has committed to developing guidances and resources for such an approach. Some recent efforts include developing guiding principles for good machine learning practice (GMLP)[1] and transparency for machine learning-enabled devices[2] to help promote safe, effective, and high-quality machine learning models; and a public workshop on fostering a patient-centered approach to AI-enabled devices, including discussions of device transparency for users.[3] This guidance intends to continue these efforts, by providing lifecycle management and marketing submission recommendations consistent with a TPLC approach for AI-enabled devices.

This guidance provides recommendations on the contents of marketing submissions for devices that include AI-enabled device software functions including documentation and information that will support FDA's review. To support the development of appropriate documentation for FDA's assessment of devices, this guidance also provides recommendations for the design and

[1] See FDA's website on Good Machine Learning Practice for Medical Device Development: Guiding Principles.
[2] See FDA's website on Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles.
[3] See FDA's website on Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan, the Executive Summary for the "Patient Engagement Advisory Committee Meeting on Artificial Intelligence (AI) and Machine Learning (ML) in Medical Devices," and the website on the Virtual Public Workshop - Transparency of Artificial Intelligence/Machine Learning-enabled Medical Devices.

development of AI-enabled devices that manufacturers may consider using throughout the TPLC. The recommendations reflect a comprehensive approach to lifecycle management of AI-enabled devices throughout the TPLC. Furthermore, the guidance includes FDA's current thinking on strategies to address transparency and bias throughout the TPLC of AI-enabled devices, including by collecting evidence to evaluate whether a device benefits all relevant demographic groups (e.g., race, ethnicity, sex, and age) similarly, to help ensure that these devices remain safe and effective for their intended use.

The emergence of consensus standards related to software has helped to improve the consistency and quality of software development and documentation, particularly with respect to activities such as risk assessment and management. When possible, FDA harmonized the terminology and recommendations in this guidance with software-related consensus standards. The Agency encourages the consideration of such FDA-recognized consensus standards when developing AI-enabled devices and preparing premarket documentation. For the current edition of the FDA-recognized consensus standards referenced in this document, see the FDA Recognized Consensus Standards Database. If submitting a Declaration of Conformity to a recognized standard, we recommend including the appropriate supporting documentation. For more information regarding use of consensus standards in regulatory submissions, refer to the FDA guidance titled "Appropriate Use of Voluntary Consensus Standards in Premarket Submissions for Medical Devices" and "Standards Development and the Use of Standards in Regulatory Submissions Reviewed in the Center for Biologics Evaluation and Research."

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. Scope

For purposes of this guidance, FDA refers to a software function that meets the definition of a device as a "device software function." A "device software function" is a software function that meets the device definition in section 201(h) of the Federal Food, Drug, and Cosmetic Act (FD&C Act).[4] As discussed in other FDA guidance, the term "function" is a distinct purpose of the product, which could be the intended use or a subset of the intended use of the product.[5]

AI-enabled devices are devices that include one or more AI-enabled device software functions (AI-DSFs). An AI-DSF is a device software function that implements one or more "AI models" (referred to as "models" in this guidance) to achieve its intended purpose. A model is a mathematical construct that generates an inference or prediction based on new input data. In this guidance, when "AI-enabled device" is used, it refers to the whole device, whereas when "AI-DSF" is used, it refers only to the function that uses AI. In this guidance, when "model" is used,

[4] Device software functions may include Software as a Medical Device (SaMD) and Software in a Medical Device (SiMD). See FDA's website on Software as a Medical Device (SaMD).
[5] See FDA's guidance titled "Multiple Function Device Products: Policy and Considerations."

it refers only to the mathematical construct.

To continue to support the development of AI enabled devices, this guidance provides recommendations on the documentation and information that should be included in marketing submissions to support FDA's review of devices that include AI-DSFs. For purposes of this guidance, the term "marketing submission" refers to premarket notification (510(k)) submission, De Novo classification request, Premarket Approval (PMA) application, Humanitarian Device Exemption (HDE), or Biologics License Application (BLA).[6] Some of the proposed recommendations in this guidance also may apply to Investigational Device Exemption (IDE) submissions. For AI-enabled devices subject to 510(k) requirements, an AI-enabled device can be found substantially equivalent to a non-AI-enabled device with the same intended use provided, among other things, the AI-enabled device does not introduce different questions of safety and effectiveness compared to the non-AI-enabled device and meets other requirements for a determination of substantial equivalence in accordance with section 513(i) of the FD&C Act.

Generally, the recommendations in this guidance also apply to the device constituent part[7] of a combination product[8] when the device constituent part includes an AI-DSF. In developing an AI-DSF, sponsors should consider the impact of the AI-DSF in the context of the combination product as a whole. For a combination product that includes an AI-DSF, we highly encourage early engagement with the FDA lead review division for the combination product.[9] In accordance with the Inter-Center consult process, the FDA lead review division will consult the appropriate subject matter experts.[10] FDA recommends that sponsors refer to other guidances for recommendations on other aspects of investigational considerations and marketing submissions for combination products.[11]

The recommendations proposed within this guidance are based on FDA's experience with reviewing a variety of AI-enabled devices, as well as current regulatory science research.

While the proposed recommendations are intended to be broadly applicable to AI-enabled devices, many of these recommendations may be specifically relevant to devices that incorporate the subset of AI known as machine learning, particularly deep learning and neural networks. Additional considerations may apply for other forms of AI.

In some cases, this guidance highlights recommendations from other guidances in order to assist manufacturers with applying those recommendations to AI-enabled devices. The inclusion of certain recommendations in this guidance does not negate applicable recommendations in other guidances that may not be included. This guidance should be considered in the context of the

[6] Certain devices are subject to review through a BLA under section 351 of the Public Health Service Act.
[7] See 21 CFR 4.2.
[8] See 21 CFR 3.2(e).
[9] See FDA's guidance titled "Principles of Premarket Pathways for Combination Products."
[10] See FDA's Staff Manual Guide titled "Combination Products: Inter-Center Consult Request Process."
[11] See FDA websites titled "Combination Products Guidance Documents" and "Search for FDA Guidance Documents." See also FDA's website on Combination Products for additional policy information regarding combination products.

FD&C Act, its implementing regulations, and other guidance documents.

This guidance is not intended to provide a complete description of what may be necessary to include in a marketing submission for an AI-enabled device. In particular, this guidance references sections of the FDA guidance titled "Content of Premarket Submissions for Device Software Functions" (hereafter referred to as "Premarket Software Guidance"),which includes significant additional considerations for AI-enabled devices, but does not include references to every section of that guidance. Additionally, this guidance does not address all of the data and information to be submitted in support of a specific indication for an AI-enabled device. FDA recommends that sponsors also refer to other guidances, as applicable to a particular device, for recommendations on other aspects of a marketing submission. [12] Examples of relevant guidances for specific technologies include the FDA guidances titled "Technical Performance Assessment of Quantitative Imaging in Radiological Device Premarket Submissions" and "Technical Considerations for Medical Devices with Physiologic Closed-Loop Control Technology." FDA further encourages sponsors to consider other available resources including consensus standards and publicly available information when preparing their marketing submissions. As with all devices, FDA intends to take a risk-based approach to determining specific testing and applicable recommendations to support marketing submissions for AI-enabled devices.

Early engagement with FDA can help guide product development and submission preparation. In particular, early engagement could be helpful when new and emerging technology is used in the development or design of the device, or when novel methods are used during the validation of the device. FDA encourages sponsors to consider discussing these plans with FDA via the Q-Submission Program.[13]

III. TPLC Approach: General Principles

This guidance acknowledges the importance of a TPLC approach to the management of AI-enabled devices. In addition to recommendations regarding the documentation and information that should be included in marketing submissions, which reflect a comprehensive approach to the management of risk throughout the TPLC, the resources provided in this guidance are also intended to assist with the device development and lifecycle management of AI-enabled devices, which should help support the safety and effectiveness of these devices. This guidance provides both specific recommendations on the information and documentation to support a marketing submission for an AI-enabled device, as well as recommendations for the design, development, deployment, and maintenance of AI-enabled devices, including the performance management.[14]

---

[12] For other guidances with digital health content, please see FDA's website on Guidances with Digital Health Content. For all other guidances, please see FDA's website on Guidance Documents (Medical Devices and Radiation-Emitting Products).

[13] See FDA's guidance titled, "Requests for Feedback and Meetings for Medical Device Submissions: The Q-Submission Program" (hereafter referred to as the "Q-Submission Program").

[14] This guidance is not intended to provide recommendations on reporting to FDA when a device has or may have caused or contributed to a death or serious injury as required by section 519 of the FD&C Act, the Medical Device Reporting (MDR) Regulation in 21 CFR Part 803, or the Medical Device Reports of Corrections and Removals Regulation in 21 CFR Part 806. For an explanation of the current reporting and recordkeeping requirements

This guidance also includes FDA's current thinking on strategies to address transparency and bias throughout the TPLC of AI-enabled devices. These interconnected considerations are important throughout the TPLC and should be incorporated from the earliest stage of device design through decommission to help design transparency and the control of bias into the device and ensure its safety and effectiveness. Transparency involves ensuring that important information is both accessible and functionally comprehensible and is connected both to the sharing of information, and to the usability of a device. AI bias is a potential tendency to produce incorrect results in a systematic, but sometimes unforeseeable way, which can impact safety and effectiveness of the device within all or a subset of the intended use population (e.g., different healthcare settings, different input devices, sex, age, etc.,). A comprehensive approach to transparency and bias is particularly important for AI-enabled devices, which can be hard for users to understand due to the opacity of many models and model reliance on data correlations that may not map directly to biologically plausible mechanisms of action. Recommendations for a design approach to transparency are provided in Appendix B (Transparency Design Considerations). With regards to the control of bias for AI-enabled devices this can include addressing representativeness in data collection for development, testing, and monitoring throughout the product lifecycle, as well as evaluating performance across subgroups of intended use.

Finally, this guidance includes recommendations that address the performance of AI-enabled devices throughout the TPLC, including in the postmarket setting. For example, AI-enabled devices can be sensitive to differences in input data (also referred to as data drift), such as input data used during development as compared to input data in actual deployments. Further, in addition to data drift, which occurs when systems that produce inputs for AI-enabled devices change over times in ways that may impact the performance of the device but may not be evident to users, AI-enabled devices can also be susceptible to changes in performance due to other factors. Sponsors are also encouraged to consider the use of a predetermined change control plan (PCCP), as discussed in FDA guidance titled "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions," which describes an approach for manufacturers to prospectively specify and seek premarket authorization for intended modifications to an AI-DSF (e.g., to improve device performance) without needing to submit additional marketing submissions or obtain further FDA authorization before implementing such modification consistent with the PCCP.

IV. How to Use this Guidance: Overview of AI-Enabled Device Marketing Submission Content Recommendations

This guidance provides recommendations on the documentation and information that should be included in marketing submissions to support FDA's review of devices that include AI-DSFs.

applicable to manufacturers of medical devices, please refer to FDA's guidance titled "Medical Device Reporting for Manufacturers."

There are some differences between the way FDA and the AI community consider the AI-enabled device TPLC and certain terminology. Therefore, this guidance clarifies these differences to facilitate better understanding of the recommendations in this guidance. For example, the AI community often uses the term "validation" to refer to data curation or model tuning that can be combined with the model training phase to optimize the model selection.[15] However, validation is defined in 21 CFR 820.3(z)[16] as "…confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled." This guidance uses the definition in 21 CFR 820.3(z), specifically when addressing the evaluation of performance of the model for its intended use. For clarity, using the term "validation" to refer to the training and tuning process should be avoided in the context of medical device marketing submissions. Also, the term "development" is used throughout this guidance to refer to training, tuning, and tuning evaluation (often referred to as "internal testing" in the AI community). In this guidance, "test data" is used to refer to data that may be used for verification and validation activities, also known as the testing process, and is not used to describe part of the development process. The "FDA Digital Health and Artificial Intelligence Glossary – Educational Resource" provides a compilation of commonly used terms in the artificial intelligence and machine learning space and their definitions.

Sections V through XIII of this guidance describe the marketing submission content recommendations for AI-enabled devices. Specifically, in each section, under, "*Why should it be included in a submission for an AI-enabled device*," an explanation is provided for why certain information should be included in a marketing submission. An explanation of what documentation and information should be included in a marketing submission can be found under "*What sponsors should include in a submission*." Finally, recommendations regarding where sponsors should include the information within each section of a marketing submission can be found under "*Where sponsors should provide it in a submission*." Information regarding recommendations for lifecycle considerations as well as examples of marketing submission materials are provided in the appendices of this guidance.

The recommendations related to marketing submissions are organized according to how they should appear in the submission (See Appendix A (Table of Recommended Documentation)), which does not always align directly with the order of activities in the TPLC. While all referenced submission sections are provided to FDA during premarket review, they include information about what has already been done to develop and validate the device, as well as what a sponsor plans to do in the future to ensure a device's ongoing safety and effectiveness. Some

[15] See International Medical Device Regulators Forum Technical Document N67 titled "Machine Learning-enabled Medical Devices: Key Terms and Definitions."

[16] On February 2, 2024, FDA issued a final rule amending the device Quality System Regulation (QSR), 21 CFR Part 820, to align more closely with international consensus standards for devices (89 FR 7496). This final rule will take effect on February 2, 2026. Once in effect, this rule will withdraw the majority of the current requirements in Part 820 and instead incorporate by reference the 2016 edition of the International Organization for Standardization (ISO) 13485, Medical devices – Quality management systems – Requirements for regulatory purposes, in Part 820. As stated in the final rule, the requirements in ISO 13485 are, when taken in totality, substantially similar to the requirements of the current Part 820, providing a similar level of assurance in a firm's quality management system and ability to consistently manufacture devices that are safe and effective and otherwise in compliance with the FD&C Act. When the final rule takes effect, FDA will also update the references to provisions in 21 CFR Part 820 in this guidance to be consistent with that rule.

sections of the guidance also describe information relevant to multiple steps in the TPLC. One example of how the sections in this guidance may align with the TPLC is included below:

- **Development** – Risk Assessment, Data Management, and Model Description and Development
- **Validation** – Data Management and Validation
- **Description of the Final Device** – Device Description, Model Description and Development, User Interface and Labeling, Public Submission Summary
- **Postmarket Management** – Device Performance Monitoring and Cybersecurity

This guidance generally describes information that would be generated and documented during software development, verification, and validation. However, the information necessary to support market authorization will vary based on the specifics of each AI-enabled device, and during premarket review FDA may request additional information that is needed to evaluate the submission.

A. Quality System Documentation

When considering the recommendations in Sections V through XIII of this guidance, it may be helpful to consider if the documentation and information that should be included in a marketing submission, under "*What sponsors should include in a submission*," could exist in the Quality System documentation. One source of documentation that may be used as part of demonstrating substantial equivalence or reasonable assurance of safety and effectiveness in the marketing submission for certain AI-enabled devices is documentation related to the ongoing requirements of the Quality System (QS) Regulation.[17] This guidance explains how some documentation that may be relevant for QS regulation compliance for medical devices generally can also be provided premarket to demonstrate how a sponsor or manufacturer is addressing risks associated with AI-enabled devices specifically.

For example, the QS Regulation requires that manufacturers establish design controls for certain finished devices (see 21 CFR 820.30). Specifically, as part of design controls, a manufacturer must "establish and maintain procedures for validating the device design," which "shall ensure that devices conform to defined user needs and intended uses and shall include testing of production units under actual or simulated use conditions" (21 CFR 820.30(g)). In addition, under 21 CFR 820.30(i) a manufacturer must establish and maintain procedures to identify, document, validate or where appropriate verify, review, and approve of design changes before their implementation ("design changes") for all devices, including those automated with

[17] In the postmarket context, design controls may also be important to ensure medical device performance and maintain medical device safety and effectiveness. FDA recommends that device manufacturers implement comprehensive performance risk management programs and documentation consistent with the QS Regulation, including but not limited to management responsibility (21 CFR 820.20), design validation (21 CFR 820.30(g)), design changes (21 CFR 820.30(i)), nonconforming product (21 CFR 820.90), and corrective and preventive action (21 CFR 820.100). While FDA generally does not assess QS Regulation compliance as part of its review of premarket submissions under section 510(k) of the FD&C Act, this guidance is intended to explain how FDA evaluates the performance of the device performance-related outputs of activities that are part and parcel of QS Regulation compliance, and explain how the QS Regulation can be leveraged to demonstrate these performance-related outputs.

software. Similarly, as part of the control of nonconforming product, manufacturers must establish and maintain procedures to "control product that does not conform to specified requirements," including, under some circumstances, user requirements, and to implement corrective and preventative action, including "complaints" and "other sources of quality data" to identify "existing and potential causes of nonconforming product." (21 CFR 820.90(a) and 820.100(a)(1)). Further, manufacturers have ongoing responsibility to manage the quality system and maintain device quality,[18] including by reviewing the "suitability and effectiveness of the quality system at defined intervals and with sufficient frequency according to established procedures" to ensure the quality objectives are being met.[19]

V. Device Description

Why should it be included in a submission for an AI-enabled device: The following section describes information that sponsors should provide in the device description section of their marketing submission to help FDA understand the general characteristics of the AI-enabled device. The following recommendations supplement device-specific recommendations and recommendations provided in the Premarket Software Guidance, where applicable.

The device description supports FDA's understanding of the intended use, expected operational sequence of the device (e.g., clinical workflow of the device), use environment, features of the model, and design of the AI-enabled device. This information is needed for FDA to evaluate the safety and effectiveness of the device. The device description provides important context about what the device does, including how it works, how a user may interact with it, and under what circumstances a device is likely to be used as intended.

For recommendations related to how to include information in the marketing submission about the technical characteristics of the model, and the method by which the model was developed, see Section IX (Model Description and Development) of this guidance.

What sponsors should include in a submission: In general, sponsors should include the following types of information as part of a device description for an AI-enabled device:

- A statement that AI is used in the device.
- A description of the device inputs and device outputs, including whether the inputs are entered manually or automatically, and a list of compatible input devices and acquisition protocols, as applicable.
- An explanation of how AI is used to achieve the device's intended use. For devices with multiple functions, this explanation may include how AI-DSFs interact with each other as well as how they interact with non-AI-DSFs.
- A description of the intended users, their characteristics, and the level and type of training they are expected to have and/or receive. Users include those who will interpret the output. When relevant, list the qualifications or clinical role of the users intended to interpret the output. Users also include all people who interact with the device including

[18] See 21 CFR 820.20.
[19] 21 CFR 820.20(c).

during installation, use, and maintenance. For example, users may include technicians, health care providers, patients, and caregivers, as well as administrators and others involved in decisions about how to deploy medical devices, and how the device fits into clinical care.

- A description of the intended use environment(s) (e.g., clinical setting, home setting).
- A description of the intended workflow for the use of the device (e.g., intended decision-making role), including:
  - A description of the degree of automation that the device provides in comparison to the workflow for the current standard of care;
  - A description of the clinical circumstances that may lead to use; and
  - An explanation of how the outputs will be used in the clinical workflow.
- A description of installation and maintenance procedures.
- A description of any calibration and/or configuration procedures that must be regularly performed by users in order to maintain performance, including when calibration must be performed and how users can identify if calibration is needed again or is incorrect, as applicable.

Additionally, sponsors should include the following types of information as part of a device description for an AI-enabled device that has elements that can be configured by a user:

- A description of all configurable elements of the AI-enabled device, for example:
  - Visualizations that the user can turn on/off (e.g., overlays, quality indicators, or heatmaps);
  - Software inputs;
  - Model parameters when they are configured during use; and/or
  - Alert thresholds.
- A description of how these elements and their settings can be configured, including:
  - A description of the users who make configuration decisions (e.g., clinical user, administrative user, patient) including any necessary qualifications and training needed to make these decisions, as applicable;
  - An explanation of how users know which selections have been made;
  - A description of the level at which the configuration is defined, for example at the patient-, clinical site- or hospital network-level; and
  - A description of customizable pre-defined operating points, their outputs and performance ranges, as applicable. It is also important to specify how the operating points or operating point range(s) were selected based on the indications for use of the device.
- A description of the potential impact of the configurable elements on user decision making.

Finally, if a device contains multiple connected applications with separate interfaces, the device description should address all applications in the device. For example, if there is an application for patients, an application for caregivers, and a data portal for healthcare providers, the device description should include details on all functions across the applications and address how they are connected. Sponsors may also wish to consider enhancing the device description with the use of graphics, diagrams, illustrations, screen captured images, or video demonstrations, including

screen captured video. For more information on how to share elements of the user interface in the marketing submission, see Section VI.A (User Interface).

Where sponsors should provide it in a submission: The AI-enabled device description information should be included in the "**Device Description**" section of the marketing submission.

VI. User Interface and Labeling

The user interface includes all points of interaction between the user and the device, including all elements of the device with which the user interacts (e.g., those parts of the device that users see, hear, touch). It also includes all sources of information transmitted by the device (including packaging and labeling[20]), training, and all physical controls and display elements (including alarms and the logic of operation of each device component and of the user interface system as a whole), as applicable. A user interface might be used throughout many phases of installation and use, such as while the user sets up the device (e.g., unpacking, set up, calibration), uses the device, or performs maintenance on the device (e.g., cleaning, replacing a battery, repairing parts).[21] One way to help support the safety and effectiveness of the device for users is to design the user interface such that important information is provided throughout the course of use, to ensure that the device conforms to defined user needs.[22] An approach that integrates important information throughout the user interface may help ensure that device users have access to information at the right time and in the right location to support safe and effective use, consistent with the intended use of the device. For software or mobile applications, manufacturers may leverage the user interface elements, such as information on the screen or alerts sent to other products, in addition to device labeling, to communicate risks about the device so that the necessary information is provided at the right time.

It is important to provide a holistic understanding of the user interface in a marketing submission to support the agency's understanding of how the device works. If a sponsor references the user interface design in their risk analysis or another section of the submission to control risks, inclusion of the user interface may also support explanations of those risk controls. However, the actual analysis of the efficacy of risk control should be located separately from the description of the user interface. Further information on this topic is described in Section VII (Risk Assessment) and Appendix D (Usability Evaluation Considerations).

With regard to labeling specifically, a device user interface includes, but is not limited to, labeling. Further, within the user interface, labeling is subject to specific regulations. For example, depending on whether the device is for prescription-use or not, manufacturers are required to provide labeling containing adequate directions for use that would ensure that a layman or, for prescription devices, a practitioner licensed by law to administer the device, "can

[20] See section 201(m) of the FD&C Act which defines labeling as "all labels and other written, printed, or graphic matter (1) upon any article or any of its containers or wrappers, or (2) accompanying such article." 21 U.S.C. § 321(m).

[21] See FDA's guidance titled "Applying Human Factors and Usability Engineering to Medical Devices."

[22] See 21 CFR 820.30(g).

use a device safely and for the purposes for which it is intended."[23] One way to satisfy these requirements for AI-enabled devices could be to provide, in the labeling, clear information about the model, its performance characteristics, and how the model is integrated into the device. For example, users may need to know specific information about the model, such as the nature of the data on which the model was trained. These technical characteristics can be critical to the safe and effective use of the device because they can support a user's understanding of how the device should be expected to perform, and what factors may impact performance.

The following sections further detail recommended information on the user interface (Section VI.A), and the labeling (Section VI.B), that should be provided in a marketing submission to support FDA's understanding of what is communicated to users and the elements of the device with which the users interact.

Appendix B (Transparency Design Considerations) of this guidance outlines a recommended approach to transparency, including examples of types of information, modes of communication, and communication styles that may be helpful to consider when designing the user interface (including labeling) of an AI-enabled medical device. It may also be helpful to integrate a model card in the device labeling to clearly communicate information about an AI-enabled device (see Appendix E (Example Model Card)).

Note that inclusion of a unique device identifier (UDI) in the labeling is required for devices, including AI-enabled devices, that are subject to UDI requirements.[24] A new UDI is required when there is a new version and/or model, and for new device packages.[25] See FDA's website on Unique Device Identification System for more information.

A. User Interface

Why should it be included in a submission for an AI-enabled device: It is important for FDA to understand the device's user interface, in order to understand how the device is used. The user interface can convey important information about what the device is intended to do, and how users are intended to interact with it. Seeing the user interface can help FDA understand how the device will be operated and how it will fit into the clinical workflow, which can support the review of a device and help the agency determine whether it is safe and effective.

A representation of the user interface can also serve to support the sponsor's risk assessment and other documentation when the user interface is referenced as an element of those sections. For example, the user interface can communicate important information to users that supports safe and effective use of the device, and the user interface design may play a crucial role in controlling or eliminating risks associated with not knowing or misunderstanding information that is critical to the safe and effective use of the device. While not required, if a sponsor chooses

[23] See 21 CFR 801.5; 21 CFR 801.109(d); FD&C Act section 502(f), 21 U.S.C. § 352(f). Device labeling must comply with the requirements in 21 CFR part 801 and any device specific labeling requirements such as for hearing aids or in special controls.
[24] See 21 CFR 801.20.
[25] See 21 CFR 830.50.

to use elements of the user interface as part of risk control in the risk assessment, the inclusion of the user interface can help further facilitate review. Further information on this topic is described in Section VII. (Risk Assessment) and Appendix D (Usability Evaluation Considerations).

While the user interface does include the printed labeling (e.g., packaging and user manuals) and all elements of the user interface should be designed to collectively support the user's understanding of how to use the device, sponsors should submit labeling separately as described in Section VI.B (Labeling). This section describes how sponsors should provide FDA with an understanding of the remaining elements of the user interface.

What sponsors should include in a submission: Sponsors should provide information about and descriptions of the user interface that makes clear the device workflow, including the information that is provided to users, when the information is provided, and how it is presented. Possible methods to provide this type of information about the user interface include:

- A graphical representation (e.g., photographs, illustrations, wireframes, line drawings) of the device and its user interface. This may include a depiction of the overall device and all components of the user interface with which the user will interact (e.g., display and function screens, alarm speakers, controls).
- A written description of the device user interface.
- An overview of the operational sequence of the device and the user's expected interactions with the user interface. This may include the sequence of user actions performed to use the device and resulting device responses, when appropriate.
- Examples of the output format, including example reports representing a range of expected outcomes.
- A demonstration of the device, for example by providing a recorded video.

Where sponsors can provide it in a submission: The user interface information should be included in the "**Software Description**" in the Software Documentation section of the marketing submission.

B. Labeling

Why should it be included in a submission for an AI-enabled device: A marketing submission must include labeling information in sufficient detail to help FDA determine that the proposed labeling satisfies applicable requirements for the type of marketing submission.[26] Device labeling must satisfy all applicable FDA labeling requirements, including, but not limited to, 21 CFR Part 801, as discussed above.[27] This section of the guidance includes labeling considerations for AI-enabled devices to support compliance with these requirements.

What sponsors should include in a submission: The labeling for an AI-enabled device should address the following types of information in a format and at a reading level that is appropriate for the intended user (e.g., considering characteristics such as age, education or literacy level,

[26] See e.g., 21 CFR 807.87(e) or 21 CFR 814.20(b)(10).

[27] Generally, if the device is an in vitro diagnostic device, the labeling must also satisfy the requirements of 21 CFR 809.10.

sensory or physical impairments, or occupational specialty) to help ensure users can quickly access important information. Tables and graphics may be used to communicate this information.

Inclusion of AI

- Statement that AI is used in the device.
- Explanation of how AI is used to achieve the device's intended use.
    - For devices with multiple functions, this explanation may include how AI-DSFs interact with each other as well as how they interact with non-AI DSFs.

Model Input

- Description of the model inputs (e.g., signals or patterns acquired from other compatible devices, images from an acquisition system (e.g., MRI), or patient-derived samples, which can be input manually or automatically). Related aspects to consider include:
    - For systems incorporating inputs from an electronic interface, information on the necessary system configuration to ensure the inputs are consistent with the design and validation of the AI-enabled device.[28]
    - For systems that require input from other medical devices (e.g., an x-ray device), a list of the specific compatible devices or device specification, along with the acceptable acquisition protocols, as applicable.
    - For systems in which the loss of model inputs may prevent the AI-enabled device from generating an output, an explanation of the potential impact of the lost inputs on the performance of the AI-enabled device.
- Instructions on any steps the user is expected to take to prepare input data for processing by the device, including any expected characteristics (e.g., functional capabilities, experience and knowledge levels, and level of training) of those performing these steps. This information should be consistent with the intended use that was studied in the device validation.

Model Output

- Explanation of what the model output means and how it is intended to be used.

Automation

- Explanation of the intended degree of automation the device exhibits**.**

Model Architecture

- High level description of the methods and architecture used to develop the model(s) implemented in the device.

Model Development Data

- Description of the development data, including:
    - The source(s) of data;
    - Study sites;

[28] For more information, please see FDA guidance titled "Design Considerations and Premarket Submission Recommendations for Interoperable Medical Devices."

591 - Sample size;
592 - Demographic distributions; and
593 - Criteria/expertise used for determining clinical reference standard (ground truth).
594
595 Performance Data
596 - Description of the performance validation data, including:
597 - The source(s) of data;
598 - Study sites;
599 - Sample size;
600 - Other important study design and data structure information (e.g., randomization
601 schemes, repeated measurements, clinical reference standard);
602 - Primary endpoints of the validation study, including pre-specified performance
603 criteria; and
604 - Criteria/expertise used for determining clinical reference standard data.
605
606 Device Performance Metrics
607 - Description of the device performance metrics.
608 - An example of performance metrics may include metrics such as the area under
609 the receiver operating characteristic curve (AUROC), sensitivity and specificity,
610 true/false positive and true/false negative counts (e.g., in a confusion matrix),
611 positive/negative predictive values (PPV/NPV), and positive/negative diagnostic
612 likelihood ratios (PLR/NLR). All performance estimates should be provided with
613 confidence intervals.
614 - Explanation of the device performance across important subgroups. Generally, subgroup
615 analysis by patient characteristics (e.g., sex,[29] age, race, ethnicity,[30] disease severity),
616 geographic sites, and data collection equipment are appropriate.
617 - Description of the corresponding performance for different operating points, including
618 subgroup analysis for each operating point, as applicable.
619
620 Performance Monitoring
621 - Description of any methods or tools to monitor and manage device performance,
622 including instructions for the use of such tools, as applicable when ongoing performance
623 monitoring and management by the user is considered necessary for the safe and effective
624 use of the device.
625
626 Limitations
627 - Description of all known limitations of the AI-enabled device, AI-DSF(s), or model(s).
628 - Some limitations of a model may not reach the degree of severity that would
629 warrant a contraindication, warning, or precaution, but they may still be important
630 to include in labeling. For example, the training dataset may have only included a
631 few patients with a rare presentation of a disease or condition; users may benefit

[29] For more information regarding sex-specific data, please see FDA guidance titled "Evaluation of Sex-Specific Data in Medical Device Clinical Studies."

[30] For more information regarding the reporting of age, race, and ethnicity related data, please see FDA guidance titled "Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies."

from knowing the limitations of the data when that rare presentation is suggested by the model as a diagnosis.

Installation and Use

- Information about the installation and implementation instructions, including:
  - Instructions on integrating the AI-enabled device into the site's data systems and clinical workflow; and
  - Instructions for ensuring that any input data are compatible and appropriate for the device.[31]
    - Terms may need to be explicitly defined. For example, a healthcare system and a manufacturer may both have data labeled as "sex," but one may be using sex at birth while the other may be using self-reported sex.

Customization

- Description of and instructions on any customizable features, including:
  - When users or healthcare systems can configure the operating points for the device;
  - When it is appropriate to select different configurations; and
  - When operating points are configurable, how end users can discern the operating point the device is currently operating at.

Metrics and Visualizations

- Explanation of any additional metrics or visualizations used to add context to the model output.

Patient and Caregiver Information

For AI-enabled devices intended for use by patients or caregivers, manufacturers should provide labeling material that is designed for patients and caregivers describing the instructions for use, the device's indication, intended use, risks, and limitations. Patients and caregivers are considered users if they will operate the device, interpret the outcome, or make decisions based on the outcome, even if they are not the only user or the primary operator of the device. This material should be at an appropriate reading level for the intended audience. If patient and caregiver-specific material is not provided, sponsors should provide an explanation of how patients and caregivers will understand how to use the device, including how to make decisions about whether to use the device and how to use the output of the device.

Where sponsors should provide it in a submission: Information regarding the AI-enabled device labeling should be included in the "**Labeling**" section of the marketing submission.

[31] For more information, please see FDA guidance titled, "Design Considerations and Pre-market Submission Recommendations for Interoperable Medical Devices."

ADDITIONAL RESOURCES:

- Appendix B (Transparency Design Considerations) outlines a potential approach to understanding a device's indications for use and a model card, which may aid in the development of the user interface.
- While model cards are not required for presenting information about the labeling or user interface, they may be a helpful tool to organize information. In general, model cards can be adapted to the specific needs and context of each AI-enabled device.
    - Appendix E (Example Model Card) includes an example of a basic model card format intended for users and healthcare providers that conveys information including a summary of the model's intended use and intended users, and evidence supporting safety and effectiveness.
    - Appendix F (Example 510(k) Submission Summary with Model Card) includes an example of a completed basic model card.
- FDA's guidance titled "Device Labeling Guidance #G91-1 (Blue Book Memo)" includes suggestions regarding what information should be included within device labeling.

VII. Risk Assessment

Why should it be included in a submission for an AI-enabled device: A comprehensive risk assessment helps ensure the device is safe and effective. When included in a marketing submission, a comprehensive risk assessment helps FDA understand whether appropriate risks have been identified and how they are controlled. In Section VI.C of the Premarket Software Guidance, FDA recommends that marketing submissions that include device software functions include a risk management file composed of a risk management plan, a risk assessment, and a risk management report. Consistent with this, marketing submissions of AI-enabled devices should include a risk management file that takes into account the recommendations of Premarket Software Guidance and the recommendations of this guidance, in addition to any other applicable guidance.

Sponsors should also refer to the FDA-recognized version of ANSI/AAMI/ISO 14971 *Medical devices - Applications of risk management to medical devices* for additional information on the development and application of a risk management file, which is also applicable to AI-enabled devices. FDA also recognizes that AI-enabled devices can be associated with new or different risks than device software functions generally. Therefore, FDA also recommends that sponsors incorporate the considerations outlined in the FDA-recognized voluntary consensus standard of AAMI CR34971 *Guidance on the Application of ISO 14971 to Artificial Intelligence and Machine Learning*, which is specific to AI-enabled devices.

Risks Across the TPLC

When conducting a risk analysis, the Medical Devices; Current Good Manufacturing Practice (CGMP), final rule (Oct. 7, 1996, 61 FR 52602) states "manufacturers are expected to identify possible hazards associated with the design in both normal and fault conditions. The risks associated with the hazards, including those resulting from user error, should then be calculated in both normal and fault conditions. If any risk is judged unacceptable, it should be reduced to acceptable levels by the appropriate means." This risk assessment should take into account all users, as described in Section VI (User Interface and Labeling) of this guidance, across the

TPLC. FDA recommends that manufacturers follow this approach for AI-enabled devices across their TPLC.

Risks Related to Information in AI-Enabled Devices

One aspect of risk management that can be particularly important for AI-enabled devices is the management of risks that are related to understanding information that is necessary to use or interpret the device, including risks related to lack of information or unclear information. Misunderstood, misused, or unavailable information can impact the safe and effective use of a device. For example, for devices that utilize complex algorithms, including AI-enabled devices, the performance in different disease subtypes may not be apparent to users, or the logic underlying the output information may not be easily understandable, which can negatively affect user understanding and use of the device. Lack of, or unclear information can also make it difficult for different users to understand whether a device is not performing as expected, or how to correctly follow instructions. FDA recommends that consideration of risks related to understanding information should be one part of a comprehensive approach to risk management for an AI-enabled device.

ADDITIONAL RESOURCES:

- ANSI/AAMI HE75 *Human factors engineering - Design of medical devices* includes recommendations on using information in labeling to help control risks.

What sponsors should include in a submission: Sponsors should provide a "Risk Management File" that includes a risk management plan, including a risk assessment. In addition to other considerations, the risk assessment should consider user tasks and knowledge tasks that occur throughout the full continuum of use of the device, including, for example, the process of installing the device, maintaining performance over time, and any risks associated with user interpretation of the results of a device, as appropriate.

In addition to the considerations provided in FDA-recognized voluntary consensus standards[32] and applicable guidances,[33] FDA recommends that sponsors consider the risks related to understanding information during the risk assessment. As with all identified risks, sponsors should provide an explanation of any risk controls, including elements of the user interface, such as labeling, that address the identified risks. Information that may be helpful to discuss such risks and their controls, as applicable, is provided in Appendix D (Usability Evaluation Considerations).

Where sponsors should provide it in a submission: Much of the information on risk assessment for an AI-enabled device should be included in the "**Risk Management File**" in the Software Documentation section of the marketing submission, as recommended by the Premarket Software Guidance.

[32] For more information, see the FDA Recognized Consensus Standards Database.

[33] For more information regarding use of consensus standards in regulatory submissions, refer to the FDA guidances titled "Appropriate Use of Voluntary Consensus Standards in Premarket Submissions for Medical Devices" and "Standards Development and the Use of Standards in Regulatory Submissions Reviewed in the Center for Biologics Evaluation and Research."

ADDITIONAL RESOURCES:

- Appendix B (Transparency Design Considerations) outlines recommendations for a user-centered design approach to developing a device, which may aid in the identification of risks and development of risk controls.
- Appendix D (Usability Evaluation Considerations) provides recommendations on usability testing, which may help sponsors evaluate the efficacy of proposed controls for information related risks.

VIII. Data Management

Why should it be included in a submission for an AI-enabled device: For an AI-enabled device, the model is part of the mechanism of action. Therefore, a clear explanation of the data management, including data management practices (i.e., how data has been or will be collected, processed, annotated, stored, controlled, and used) and characterization of data used in the development and validation of the AI-enabled device is critical for FDA to understand how the device was developed and validated. This understanding helps to enable FDA's evaluation of an AI-enabled device's safety and effectiveness.

The performance and behavior of AI systems rely heavily on the quality, diversity, and quantity of data used to train and tune them. The accuracy and usefulness of a validation of an AI-enabled device also depends on the quality, diversity, and quantity of data used to test it. Thus, FDA reviewers evaluate data management in order to understand whether an AI-enabled device is safe and effective. This includes the alignment of the collection and management of training and test data with the intended use and resulting device requirements.

Data management is also an important means of identifying and mitigating bias. The characterization of sources of bias is necessary to assess the potential for AI bias in the AI-enabled device. AI bias is a potential tendency to produce incorrect results in a systematic, but sometimes unforeseeable, way due to limitations in the training data or erroneous assumptions in the machine learning process. AI bias has been well-documented.[34] For example, during training, models can be over-trained to recognize features of images that are unique to specific scanners, patient subpopulations, or clinical sites but have little to do with generalizable patient anatomy, physiology, or condition, which can lead to AI bias in the resulting model. In another example, underrepresentation of certain populations in datasets could lead to overfitting (i.e., data fitting too closely to the potential biases of the training data) based on demographic characteristics, which can impact the AI-enabled device performance in the underrepresented population.

Using unbiased, representative training data for models promotes generalizability to the intended use population and avoids perpetuating biases or idiosyncrasies from the data itself. For example, in image recognition tasks, confounding may occur when all the diseased cases are imaged with the same instrument, or with a ruler included (e.g., on clinical images of melanoma). Another example of a potential confounding factor is the use of data collected outside the U.S. (OUS) in

[34] See Karen Hao, "This is how AI bias really happens—and why it's so hard to fix," MIT Technology Review 2019.

training, which may bias the model if the OUS population does not reflect the U.S. population due to differences in demographics, practice of medicine, or standard of care. Such confounders in the training data, if not identified and mitigated, can be inadvertently learned by a model, leading to seemingly accurate (but misleading) predictions based on irrelevant characteristics.

The inclusion of representative data in validation datasets may be important, because underrepresentation may impact the ability to identify any performance problems, including understanding performance in underrepresented populations. Although bias may be difficult to eliminate completely, FDA recommends that manufacturers, as a starting point, ensure that the validation data sufficiently represents the intended use (target) population of a medical device. For more information regarding age-, race-, ethnicity-, and sex-specific data please see the FDA guidances titled, "Collection of Race and Ethnicity Data in Clinical Trials and Clinical Studies for FDA-Regulated Medical Products"[35] "Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies," and "Evaluation of Sex-Specific Data in Medical Device Clinical Studies."

If the same confounders are found in the validation data as the development data, it may be particularly difficult to identify the spurious correlations that appear to be leading to correct predictions. Therefore, information about the representativeness of the datasets used in the development and validation of the AI-enabled device is important to help FDA determine substantial equivalence or if there is a reasonable assurance that the device is safe and effective for its intended use.[36] Beyond addressing AI bias, the details of the data management should support the intended use of the device.

To objectively assess the device performance, it is also important for FDA reviewers to understand whether the test data are independent (e.g., sampled from completely different clinical sites) from the training data and are sequestered from the model developers and the model development stage. Appropriate separation of the development and test datasets can help with evaluating the true performance of an AI-enabled device. Data leakage between the validation and development datasets can create uncertainty regarding the true performance of the AI-enabled device.[37]

What sponsors should include in a submission: In a submission, a sponsor should provide the following types of information for both the training and testing data, in the appropriate marketing submission sections. It may be helpful to organize data management information by the sections described below. Generally, information on data collection, development and test data independence, reference standards, and representativeness should be provided. Sponsors should also explain any differences in the data management approach and the characteristics of the data

[35] When final, this guidance will represent FDA's current thinking on this topic.

[36] For more information, see FDA guidance titled "Acceptance of Clinical Data to Support Medical Device Applications and Submissions: Frequently Asked Questions."

[37] Robert F Wolff, Karel GM Moons, Richard D Riley, et al. "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies." Annals of Internal Medicine 170, no. 1 (2019): 51–58 https://doi.org/10.7326/M18-1376; Altman, Douglas G., and Patrick Royston. "What Do We Mean by Validating a Prognostic Model?" Statistics in Medicine 19, no. 4 (2000): 453–73 https://doi.org/10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5.

between the development and validation phases. The submission should include an explanation for the differences and justification for them.

Data Collection

- A description of how data were collected (e.g., clinical study protocols with inclusion/exclusion criteria), including:
  - The names of clinical sites or institutions involved.
    - Sites should be uniquely identified, and they should be referred to consistently throughout the submission.
  - The time period during which the data were acquired.
  - If data were used from a pre-existing database, the appropriateness of the use of this database.
  - If real-world data (RWD) are used, the source and collection of this evidence.
    - If RWD are used, FDA recommends that sponsors provide an assessment of fit-for-purpose data for the selected data source(s) that evaluates both the relevance and reliability of the RWD. FDA encourages sponsors to leverage the Q-Submission Program for obtaining FDA feedback on proposed uses of RWD. For more information regarding RWD, please see the FDA guidance titled "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices."
- A description of the limitations of the dataset.
- A description of the quality assurance processes related to the data, including the controls that were put in place to protect from human error during data acquisition, when applicable.
- A description of the size of each data set.
- A description of the mechanisms used to improve diversity in enrollment within the scope of the study, and how they ensure the generalizability of study results across patient populations and clinical sites.[38] For more information on this topic, please see FDA guidance titled "Collection of Race and Ethnicity Data in Clinical Trials."
- A description of the use of synthetic data. [39]Synthetic data used in support of a regulatory submission should be accompanied by a comprehensive explanation of how the data were generated and why they are fit-for-purpose.

Data Cleaning/Processing

To provide optimum training results, it may be important to clean data used for development, such as by removing incorrect, duplicate, or incomplete data. These processing steps should be

[38] Sponsors may be required to develop or submit information regarding the enrollment of clinical study participants to help improve the strength and generalizability of the study results. For example, the FD&C Act, as amended by section 3601(b) of the Food and Drug Omnibus Reform Act of 2022 (FDORA), enacted as part of the Consolidated Appropriations Act, 2023 (P.L. 117-328)), requires sponsors to submit to FDA diversity action plans for studies of certain devices. See FD&C Act section 520(g)(9), 21 U.S.C. § 360j(g)(9).

[39] For the purposes of this guidance, "synthetic data" is defined as data that have been created artificially (e.g., through statistical modeling, computer simulation) so that new values and/or data elements are generated. Generally, synthetic data are intended to represent the structure, properties and relationships seen in actual patient data, except that they do not contain any real or specific information about individuals. For more information, please see FDA Digital Health and Artificial Intelligence Glossary – Educational Resource | FDA.

described, including data quality factors used, data inclusion/exclusion criteria, treatment of missing data, and whether the steps are internal or external to the AI-DSF.

Testing data, on the other hand, should only be processed in a manner that is representative of the RWD the model will encounter in its intended use. Any such data processing, data quality factors used, data inclusion/exclusion criteria, and treatment of missing data should be justified as aligned with pre-processing implemented in the final AI-DSF.

Reference Standard

For the purposes of this guidance, a reference standard is the best available representative truth that can be used to define the true condition for each patient/case/record.[40] It is possible that a reference standard may be used in device training, device validation, or both. A reference standard is validated by evidence from current practice within the medical and regulatory communities for establishing a patient's true status with respect to a clinical task. The reference standard should reflect the clinical task. Clinical tasks may consist of, for example, classification of a disease or condition, segmentation of contours on medical images, detection by bounding boxes, or localization by markings. The following types of information should be provided regarding the selected reference standard:

- A description of how the reference standard was established.
- A description of the uncertainty inherent in the selected reference standard.
- A description of the strategy for addressing cases where results obtained using a reference standard may be equivocal or missing.
- If the reference standard is based on evaluations from clinicians, provide:
    - The grading protocol used.
    - What data are provided to these clinicians.
    - How the clinicians' evaluations are collected/adjudicated for determining the clinical reference standard, including:
        - blinding protocol; and
        - number of participating clinicians and their qualifications.
    - An assessment of the intra- and/or inter-clinician variability for each task, as applicable, as well as an assessment on whether the observed variability is within commonly accepted standards for a particular measurement task.

Data Annotation

- When data annotation is used, the following types of information should be provided regarding the data annotation approach:
    - A description of the expertise of those performing the data annotation.
    - A description of the specific training, instructions or guidelines provided to data annotators to guide their annotation decisions, including whether annotators are blinded to each other.

[40] For an illustrative example of a reference standard, see FDA guidance titled "Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in Premarket Notification (510(k)) Submissions." This guidance addresses the reference standard for this device. Other device specific guidances and special controls note the appropriate reference standard to be used. For questions about what the appropriate reference standard may be for a device and proposed intended use, consult the appropriate review division via the Q-Submission Program.

885 - A description of the methods for evaluating quality/consistency of data
886 annotations and adjudicating disagreements (consensus evaluation, sampling).
887 FDA recommends the use of independent assessments by each annotator, without
888 knowledge of the other annotators' decisions, to ensure objective high-quality
889 data annotations; and
890 - A detailed plan for addressing incorrect data annotation.
891
892 Data Storage
893 A description of the data storage of both training and test data. The description should address
894 dataset version control and should ensure the security of the data by addressing the items
895 described in Section XII (Cybersecurity) of this guidance.
896
897 Management and Independence of Data
898 - A description of the development data, including how the development data were split
899 into training, tuning, tuning evaluation, and any additional subsets, and specification of
900 which model development activities were performed using each dataset.
901 - A description of the controls in place to ensure the data used for testing is sequestered
902 from the development process.
903 - A justification of why the data used for validation provides a robust external validation.
904 For example, a description of the sites from which test data originates from, because, in
905 general, test data should come from sites different from those used to develop the AI-
906 DSF.
907
908 Representativeness
909 - An explanation of how the data is representative of the intended use population[41] and
910 indications for use, including:
911 - A description of the relevant population characteristics, when available, including:
912 - Disease conditions (e.g., positive/negative cases, disease severity, disease
913 subtype, comorbidities, distribution of the disease spectrum);
914 - Patient population demographics (e.g., sex,[42] age, race, ethnicity,[43] height,
915 weight);
916 - Data acquisition equipment and conditions (e.g., locations at which data
917 are collected, data acquisition devices/methods, imaging and
918 reconstruction protocols), including any factors that may impact signals
919 analyzed during data acquisition (e.g., patient activities, such as whether a
920

[41] Sponsors may be required to develop or submit information regarding the representativeness of clinical study participants. For example, the FD&C Act, as amended by section 3601(b) of FDORA, enacted as part of the Consolidated Appropriations Act, 2023 (P.L. 117-328)), requires sponsors to submit to FDA diversity action plans for studies of certain devices. See section 520(g)(9) of the FD&C Act, 21 U.S.C. § 360j(g)(9).

[42] For more information regarding sex-specific data, please see FDA guidance titled "Evaluation of Sex-Specific Data in Medical Device Clinical Studies."

[43] For more information regarding age-, race-, and ethnicity-related data, please see FDA guidances titled "Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies," and "Collection of Race and Ethnicity Data in Clinical Trials."

patient is ambulatory, resting, standing; or data acquisition environments, such as intensive care unit, MRI); and
    - Test data collection sites (e.g., clinical sites, institutions). Generally, while a single data collection site may be a useful starting place during initial data assessment phases, reliance on a single site is generally not appropriate for understanding whether the data are representative of the intended use population and indications for use. The use of multiple data collection sites, such as sites in diverse clinical practice settings (e.g., large academic hospital vs. community hospital) may assure a more representative sample of the intended use population. For example, the use of at least three geographically diverse US clinical sites (or health care systems) may be appropriate to clinically validate an AI-enabled device.[44]
  - A characterization of the distribution of data along important covariates, including those corresponding to the population characteristics described above.
  - If any of the relevant population characteristics above were not available for the data, an explanation of why, and a justification of the use of the data without this information. FDA understands that, depending on the source of the patients and/or samples used in the training and test data, some relevant patient characteristic information may not be available.
  - A subgroup analysis or analyses stratified by the identified covariates.
  - If OUS data are used during validation, an explanation regarding how the data compares to the U.S. population and U.S. medical practice in terms of general medical practice, disease presentation, prevalence, and progression as well as the demographic characteristics of patients.[45]
    - Due to the data-driven nature of typical models and the obscurity of their algorithms to end users, their generalized performance on the U.S. target population may not be adequately captured in the clinical study if a significant portion of the validation data are OUS data. AI-enabled devices may also be more sensitive than traditional medical devices to the idiosyncratic patterns in the training or test data. For these reasons, they may require higher proportion of U.S. data in the clinical validation. FDA encourages sponsors to leverage the Q-Submission process for obtaining FDA feedback on proposed uses of OUS data.[46]

Where sponsors should provide it in a submission: The data management information for data used in the development of the model should be included in the "**Software Description**" in the Software Documentation section of the marketing submission, as described in the Premarket Software Guidance.

The data management information for data used in the performance validation (i.e., clinical

[44] For more information regarding site selection, please see FDA guidance titled "Design Considerations for Pivotal Clinical Investigations for Medical Devices."
[45] For more information on the use of OUS data, please see FDA guidance titled "Acceptance of Clinical Data to Support Medical Device Applications and Submissions: Frequently Asked Questions."
[46] For more information on the Q-Submission program, please see FDA guidance titled "Requests for Feedback and Meetings for Medical Device Submissions: The Q-Submission Program."

validation) documentation should be included in the "**Performance Testing**" section of the marketing submission. When the characteristics of data used for model training and validation differ, sponsors should highlight and justify the differences along with the performance validation data management section in the performance testing documentation element.

ADDITIONAL RESOURCES: In addition to the considerations in this guidance, to support the TPLC approach to development, FDA recommends that sponsors and investigators consider the unique characteristics of the AI-enabled device during the study design, conduct, and reporting phases for clinical investigations. Researchers should understand how Investigational Device Exemption (IDE), Protection of Human Subjects and Institutional Review Board regulations,[47] and Good Clinical Practice (GCP) regulations[48] apply to their devices. Resources include consensus guidelines,[49] as well as FDA guidances titled:

- "Significant Risk and Nonsignificant Risk Medical Device Studies"
- "Informed Consent Guidance for IRBs, Clinical Investigators, and Sponsors"
- "Acceptance of Clinical Data to Support Medical Device Applications and Submissions: Frequently Asked Questions"

For more information regarding age-, race-, and ethnicity-specific data, and sex-specific data please see the FDA guidances titled:

- "Collection of Race and Ethnicity Data in Clinical Trials"
- "Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies"
- "Evaluation of Sex-Specific Data in Medical Device Clinical Studies"

IX. Model Description and Development

Why should it be included in a submission for an AI-enabled device: Information about the model (and device) design, including its biases and limitations, supports FDA's ability to assess the safety and effectiveness of an AI-enabled device and determine the device's performance testing specifications.

Section VI.B of the Premarket Software Guidance describes information that should be included as part of a software description in a marketing submission, including the model description. Whereas the device description is broader and provides information about the whole device, how users interact with it, and how it fits into the clinical workflow, the model description, as part of the software description, specifically provides detailed information about the technical characteristics of the model(s) themselves and the algorithms and methods that were used in their development. This information helps FDA understand the basis for the functionality of an AI-enabled device. Understanding the methods used to develop the model also helps FDA identify

[47] See 21 CFR Parts 50 and 56.

[48] See FDA's website on Regulations: Good Clinical Practice and Clinical Trials.

[49] See Liu, Xiaoxuan et al "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Natural Medicine (2020) https://doi.org/10.1038/s41591-020-1034-x; Rivera, Samantha C. et al "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," Lancet (2020) https://doi.org/10.1016/S2589-7500(20)30219-3; Vasey, Baptiste et al Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," BMJ (2022) https://doi.org/10.1136/bmj-2022-070904.

potential limitations, sources of AI bias, and considerations for appropriate device labeling.

What sponsors should include in a submission: In a submission, sponsors should include the information described below for each model in the AI-enabled device.

In situations where multiple models are employed as part of the AI-enabled device, it can be particularly helpful to include a diagram of how model outputs combine to create the device outputs. The description of the algorithms and models should be sufficiently detailed to enable a competent AI practitioner to produce an equivalent model. The use of diagrams in addition to textual descriptions is encouraged to enhance clarity.

Model Description

- An explanation of each model used as part of the AI-enabled device, including but not limited to:
  - Model inputs and outputs;
  - A description of model architecture;
  - A description of features;
  - A description of the feature selection process and any loss function(s) used for model design and optimization, as appropriate; and
  - Model parameters.
- In situations where the AI-enabled device has customizable features involving the model, such as being customizable to operate at multiple pre-defined operating points or with a variable number of inputs, a description of the technical elements of the model that allow for and control customization.
- A description of any quality control criteria or algorithms, including AI-based and third-party ones, for the input data, including how the quality assessment metrics align with the intended use of the device (e.g., intended patient population and use environment).
- A description of any methods applied to the input and/or output data, including:
  - Pre-processing of input data (e.g., normalization);
  - Post-processing of output data; and
  - Data augmentation or synthesis.

Model Development

- A description of how the model was trained, including but not limited to:
  - Optimization methods;
  - Training paradigms (e.g., supervised, unsupervised or semi-supervised learning, federated learning, active learning);
  - Regularization techniques employed;
  - Training hyperparameters (e.g., the loss function learning rate) as applicable; and
  - Summary training performance such as the loss function convergence curves for the different data subsets (such as training, tuning, tuning evaluation).
- If tuning evaluation was conducted, a description of the metrics and results obtained.
- An explanation of any pre-trained models that were used, as applicable.
  - If a pre-trained model was used, specify the dataset that was used for pre-training and how the pre-trained model was obtained.

- A description of the use of ensemble methods (e.g., bagging or boosting), as applicable.
- An explanation of how any thresholds (e.g., operating points) were determined.
- An explanation of any calibration of the model output.

Where sponsors should provide it in a submission: Information on model development, including the model description, and the method for model development, should be included as part of the "**Software Description**" in the Software Documentation section of the marketing submission, as described in the Premarket Software Guidance.

> **ADDITIONAL RESOURCES:** In situations where manufacturers wish to consider development of models that automatically or continuously update, FDA encourages manufacturers to use the Q-Submission Program to discuss considerations related to these AI models early in the development process and review the FDA guidance titled "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions."

X. Validation

For an AI-enabled device, validation includes ensuring that the device, as utilized by users, will perform its intended use safely and effectively, as well as establishing that the relevant performance specifications of the device can be consistently met. For AI-enabled devices, manufacturers should demonstrate users' ability to interact with and understand the device as intended in addition to ensuring the device itself meets relevant performance specifications. To this end, it can be helpful to consider both performance validation (including human factors validation) and an evaluation of usability. Note that, for the purposes of this guidance (in the context of risk controls in the absence of human factors validation), usability describes whether the device can be used safely and effectively by the intended users, including whether users consistently and correctly receive, understand, interpret, and apply information related to the AI-enabled device.

The FDA guidance titled "Applying Human Factors and Usability Engineering to Medical Devices" (hereafter referred to as "Human Factors Guidance"), describes recommendations and requirements for devices and establishes that human factors validation testing encompasses, "all testing conducted at the end of the device development process to assess user interactions with a device user interface to identify use errors that would or could result in serious harm to the patient or user," and is also used "to assess the effectiveness of risk management measures." While the Human Factors Guidance outlines specific recommendations and requirements for human factors validation for devices that have critical tasks, the application of the same or a similar process can also be helpful to demonstrate the appropriate control of other risks. Appendix D (Usability Evaluation Considerations) includes recommendations to help sponsors understand when usability testing may help support the control of risks. The appendix also includes recommendation to help sponsors develop and describe certain types of usability testing in addition to human factors validation, or when human factors validation is not required. The appendix supplements device-specific recommendations and recommendations provided in the Human Factors Guidance where applicable.

Together, performance validation and human factors validation (or an evaluation of usability as appropriate) help provide FDA with information to understand how the device may be used and perform under real world circumstances. Performance validation may employ a variety of testing and monitoring methods to evaluate the statistical performance of the model under testing conditions, and human factors validation testing involves understanding how various users are likely to use a device in context. In other words, performance validation is meant to provide confirmation that device specifications conform to user needs and intended uses, and that performance requirements implemented can be consistently fulfilled, while human factors validation and an evaluation of usability are meant to specifically address whether all intended users can achieve specific goals while using the device and whether users will be able to consistently interact with the device safely and effectively.

Software Version History

Section VI.I of the Premarket Software Guidance describes information that should be included as part of a software description in a marketing submission, including information regarding the software version history. For AI-enabled devices, the software version history includes consideration of the model version and any differences between the tested version of the model and the released version, along with an assessment of the potential effect of the differences on the safety and effectiveness of the device. It is important for FDA to understand what version of the model was tested in order to ensure that all validation activities will be objective, and the model has not been adjusted opportunistically in light of the test data (i.e., post-hoc adjustment) without the Agency's concurrence**.**

New unique device identifiers (UDIs) are required for devices that are required to bear a UDI on its label when there is a new version and/or model, and for new device packages.[50]

A. Performance Validation

Why should it be included in a submission for an AI-enabled device: The performance validation for an AI-enabled device provides objective evidence that the device performs predictably and reliably in the target population according to its intended use. The following recommendations are intended to supplement device-specific recommendations and recommendations provided in other FDA guidances where applicable, including "Design Considerations for Pivotal Clinical Investigations for Medical Devices," "Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests," and "Electronic Source Data in Clinical Investigations."

As part of FDA's evaluation of safety and effectiveness of the device, it is important for FDA to understand how the device performs overall in the intended use population, as well as in subgroups of interest. Acceptable performance in certain subgroups may mask lower performance in other subgroups when the evaluation is performed only for the total population. Poor performance in specific subgroups could make the device unsafe for use in those groups, which may impact the potential scope of the intended use population. Section VIII (Data Management) outlines why stratification and analyses of subgroups of interest is important to FDA's evaluation of safety and effectiveness. An analysis of subgroup performance that supports safe and effective use across the expected intended use population also helps to ensure that

[50] See 21 CFR 830.50.

devices can be used for all intended patients.

While differential performance across subgroups is not unique to AI-enabled devices, the reliance of models on relationships learned from large amounts of data, and the relative opacity of models to users make AI-enabled devices particularly susceptible to unexpected differences in performance. Even when the data used to develop the model is representative during training, models can be over-trained to recognize features of data that are unique to specific characteristics of the study dataset but may be spurious to the identification or treatment of the disease or condition. Spurious learnings could impact performance differentially across characteristics of interest such as disease subtype or patient demographics, especially when data from study participants from different groups tend to be collected at different sites. For example, models may erroneously use demographic information, or another variable corelated with demographic information, as a variable of interest in the model because patients of one demographic tended to be more likely to have a disease in the training data set. This can be particularly difficult to identify with complex models in which the variables of interest may not be understandable to humans. For this reason, the accuracy and usefulness of an evaluation of an AI-enabled device also depends on the quality, diversity, and quantity of data used to test it.

Subgroup analysis provides the tools to evaluate the performance of the device in specific populations and can be helpful in identifying scenarios in which the device performs worse than overall performance. In addition, subgroup analyses are helpful in identifying potential limitations of the device and can contribute to effective labeling by providing end users with additional useful information.

Information on the uncertainty of device outputs is also important because it helps reviewers understand how to interpret device outputs. When not specified for a device type in statute, regulation, or guidance, repeatability and/or reproducibility studies can still help FDA understand and quantify the uncertainty associated with device outputs when provided.

Appendix C (Performance Validation Considerations) of this guidance includes additional recommendations for some common approaches to performance validation. In addition, FDA encourages sponsors to leverage the Q-Submission Program for obtaining FDA feedback on proposed approaches to AI-enabled device development and validation. In particular, early engagement could be helpful to discuss the use of RWD, the use of new and emerging study methods, or the validation of new technologies.

Assessing the Performance of the Human-Device Team

It is important for sponsors to consider the interactions between users and the device when identifying the appropriate methods for performance testing. In the document, "Good Machine Learning Practice: Guiding Principles," Principle Seven discusses placing focus on "the performance of the Human-AI Team." This principle explains that it is important to understand the performance of the "Human-AI team, rather than just the performance of the model in isolation" when a model has a "human in the loop." The intended use and clinical workflow of AI-enabled devices span a continuum of decision-making roles from more autonomous systems to supportive (aid) tools that assist specific users, but rely on the human to interpret the AI outputs and ultimately make clinical decisions.

As the device moves along this spectrum, the nature of the clinical study or other studies (e.g., human factors validation testing) that would be appropriate to support performance evaluation of an AI-based medical device will vary according to the intended use of the model. For some devices, more emphasis may be placed on the model's standalone performance (i.e., Did the actual output match the expected output?). For others, a focus may be assessing the performance of the human-AI team, beyond just the performance of the model in isolation (i.e., Did the intended user working with the new device perform the same or better than the operator alone or with another device?). Sponsors should consider that, in certain scenarios, both standalone and human-device team performance evaluations may support the overall performance evaluation of the AI-enabled device.

Performance evaluation of AI-based medical image analysis systems is an illustrative example of how the clinical study approaches may change as the intended use of the device moves along the spectrum of human-device interactions. Standalone assessments measure the model's performance independently of human interaction, whereas reader studies compare the performance of the intended user both with and without the AI-enabled device (i.e., comparing the human vs. human-device team performance).[51] Reader studies typically serve as the primary performance evaluation for AI-enabled devices that aid in clinical decision-making in medical imaging applications, because they allow sponsors to evaluate the tool's clinical benefit in the hands of the intended user.[52]

What sponsors should include in a submission: The validation testing should provide objective information to characterize the model performance with respect to the intended use. A validation assesses the model's performance on independent datasets. Assessing the robustness of the model to anticipate reasonably foreseeable changes in input data and conditions of use should also be included, as appropriate, based on risk associated with these changes.

Validation methods differ depending on the intended use of a device. For example:

- Devices estimating defined measurements otherwise performed by accepted reference methods may need a precision study to adequately assess their repeatability and reproducibility.
- Devices monitoring time-series patient data and needing periodic re-calibrations may need a stability study and a change tracking study to assess their dynamic responses.
- Devices similar to survey instruments measuring less well-defined patient parameters may need additional evidence of construct validity (i.e., the extent to which a test measures what it is proposed to measure).
- Prognostic clinical decision support devices may need longitudinal data with survival analysis, calibration analysis, and/or discrimination analysis (e.g., risk stratification analysis), among other methods.

[51] For more information, see FDA's guidance titled, "Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data Premarket Notification [510(k)] Submissions."

[52] For more information on computer-assisted detection devices, please see FDA guidance titled, "Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in Premarket Notification (510(k)) Submissions."

Depending on the specific AI-enabled device, this evidence could come from non-clinical bench or analytical studies, pre-clinical animal studies, clinical performance studies, clinical outcome studies, or some combination thereof.

Study Protocols

To support performance validation, sponsors should include information regarding all study protocols including statistical analysis plans. The statistical analysis plans should include study design and analysis details. Important aspects for these documents to cover include:

- Study design details, including:
  - A study design description (e.g., prospective, comparative study design with a sufficient statistical power to demonstrate the key clinical performance metric).
    - For a prospective study, procedures and methods that will be followed, a description of the operators involved in these procedures and methods, and any tools or equipment to be used.
    - For a retrospective study, plans on how to handle, prepare, process, and select archived data or material.
  - A description of the data recording mechanisms that will be used to record the version or state of the AI-enabled device used during the study for a given patient.
    - To ensure accuracy, automated collection of these data implemented in an electronic case report form (eCRF)[53] or electronic data capture (EDC) system may be appropriate.
  - A description of the procedures and methods for blinding of the device outputs from the clinical reference standard determination process, masking of the clinical reference standard from the users/interpreters of the device outputs, and masking of the test data from the model developers and clinical team (to avoid opportunistic tweaking or bias in the study design), as applicable.
  - A description of the controls in place to address any risks posed to the patient or user by the AI-enabled device during the study.
  - If the protocol is altered during the execution of the clinical study, the applicant should explain the changes, and identify which changes are deemed minor and major, providing adequate justification for any repeated tests or tests with deviations from the pre-specified plans. The study protocol should be followed and all types of protocol deviations, including those deemed minor, should be minimized.
  - A full accounting of all enrolled subjects (with an accountability table).
  - A description of baseline distributions of the study population and other important factors in the dataset such as data acquisition equipment, device configurations, and disease status or conditions, and a justification of their representativeness. For more information on representativeness in AI-enabled medical devices, refer to Section VIII (Data Management) of this guidance.
- Statistical analysis plans, including:

[53] For more information, see FDA guidance titled "Electronic Source Data in Clinical Investigations."

- A description of the primary endpoint(s) or outcome(s), which should be reflective of the primary objective of the study.
- Pre-specified study success/failure criteria with respect to each of multiple primary endpoints (e.g., performance goals) that are clinically justified (e.g., supported by literature or prior investigations).
- An explanation of the statistical hypotheses, such as null hypothesis, and the alternative (working) hypothesis.
- A sample size justification that ensures adequate study power.
- An explanation of the statistical analysis of the primary endpoint(s), including information to justify the sample size calculation.
- An explanation of the pre-specified, appropriate statistical approaches for handling multiplicity issues and controlling for overall Type I error rates;
- A description of the appropriate statistical methodology.
- A subgroup analysis plan.
  - The appropriate subgroups are informed by the intended use of the device, but should generally include patient sex,[54] age, race, ethnicity,[55] disease variables, clinical data site, data acquisition equipment (e.g., camera brand), and, if applicable, conditions for use (including skill level of the user when relevant), device configurations, and other relevant confounding factors that may impact the device performance.
  - When a specific performance claim is made with respect to a subgroup, the subgroup analysis should be statistically significant, including the inclusion of appropriately powered subgroups. However, when specific subgroup performance claims are not made, subgroup performance does not need to be statistically powered for each subgroup, but effort should be made to include reasonable numbers of patients for each subgroup so that any reported results have meaning and context.

Study Results

To support performance validation, sponsors should include information regarding the study results. Important aspects for these documents to cover include:

- An explanation of the pre-specified results for each test, including subgroup analyses.
- An explanation of the results with adequate subgroup analyses for relevant subgroups as described above.
  - If demographic information is not available for the study data, an explanation of the reasons it is not available, why performance evaluation can be supported without demographic subgroup analysis, and how risks associated with the lack of demographic subgroup analyses have been controlled.
- When feasible, and appropriate, an evaluation of the device repeatability and reproducibility. The specifics of how these studies are conducted will depend on the

[54] For more information on sex-specific data, please see FDA guidance titled "Evaluation of Sex-Specific Data in Medical Device Clinical Studies."

[55] For more information on age-, race-, and ethnicity-specific data, please see FDA guidance titled "Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies."

specific device being evaluated, and may include phantom, simulated, contrived. or clinical data.

Where sponsors should provide it in a submission: Information on the non-clinical or clinical testing of the device should be included in the appropriate sections of the marketing submission. For example, clinical study findings should go in the clinical section of the marketing submission. Information on the software verification and software validation of the model should be included in the "**Software testing as part of Verification and Validation**" in the Software Documentation section of the marketing submission, as described in the Premarket Software Guidance.

ADDITIONAL RESOURCES: Appendix C (Performance Validation Considerations) includes recommendations to help develop and analyze a performance validation study and its data. Appendix D (Usability Evaluation Considerations) includes information to help sponsors evaluate usability risk controls for AI-enabled device submissions.

FDA encourages sponsors to use the Q-Submission Program for obtaining FDA feedback on proposed approaches for AI-enabled device development and validation. If real world evidence is used, sponsors may also wish to refer to FDA guidance titled "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices."

XI. Device Performance Monitoring

Why should it be included in a submission for an AI-enabled device: The performance of AI-enabled devices deployed in a real-world environment (i.e., marketed AI-enabled devices following approval or clearance) may change or degrade over time, presenting a risk to patients. In general, as part of the quality system for a medical device, including an AI-enabled device, manufacturers should have a postmarket performance monitoring plan to help identify and respond to changes in performance in a postmarket setting. The inclusion of a performance monitoring plan in the marketing submission may help to reduce uncertainty and support FDA's evaluation of risk controls.

As part of their ongoing management of AI-enabled devices manufacturers should proactively monitor, identify, and address device performance changes, as well as changes to device inputs and the context in which the device is used that could lead to changes in device performance. In addition, sponsors must develop and implement plans for comprehensive risk analysis programs and documentation consistent with the Quality System Regulation (21 CFR Part 820) to manage risks related to undesirable changes in device performance for AI-enabled devices.[56] These regulations include, but are not limited to, management responsibility (21 CFR 820.20), design validation (21 CFR 820.30(g)), design changes (21 CFR 820.30(i)), nonconforming product (21 CFR 820.90), and corrective and preventive action (21 CFR 820.100). Further, manufacturers must monitor device performance and report to FDA information about deaths, serious injuries, and malfunctions in accordance with 21 CFR Parts 803 and 806.

[56] When the final rule amending the device QSR, 21 CFR Part 820, takes effect on February 2, 2026, the term "risk analysis" will be replaced with "risk management."

FDA generally does not assess quality system regulation compliance as part of its review of marketing submissions under section 510(k) of the FD&C Act. However, in some cases, it may be appropriate for FDA to review details from the sponsor's quality system in the marketing submission to ensure adequate ongoing performance. Such a review may help support a determination of substantial equivalence.

Ongoing performance monitoring is important for AI-enabled devices because, as described above, models are highly dependent on the characteristics of data used to train them, and as such, their performance can be particularly sensitive to changes in data inputs. Changes in device performance may originate from many factors, such as changes in patient populations over time, disease patterns, or data drift from other changes. When performance changes do occur, users may be less likely to identify them in AI-enabled devices if, for example, the devices are part of a highly automated process with limited on-going human interaction, or if the output is prognostic such that different healthcare professionals may be involved in the use of the device and in confirmatory follow-up interactions with the patient. Because the performance of AI-enabled devices can change as aspects of the environments in which they are approved or cleared for use in may change over time, it may not be possible to completely control risks with development and testing activities performed premarket (prior to device authorization and deployment).

FDA recognizes that the environments where medical devices are deployed cannot be completely controlled by the device manufacturer. Further, the presence of factors that may lead to changes in device performance may not always raise concerns about patient harm. Rather, as part of ongoing risk management, it is important for device manufacturers to consider the impact of these factors (e.g., data drift) on the safety and effectiveness of the device. Additional information about performance management processes may be helpful for FDA to determine whether risks have been adequately identified, addressed and controlled.

What sponsors should include in a submission: Sponsors of AI-enabled devices that elect to employ proactive performance monitoring as a means of risk control and to provide reasonable assurance of the device's safety and effectiveness, should include information regarding their performance monitoring plans as part of the premarket submission. Sponsors are encouraged to obtain FDA feedback on the plan through the Q-Submission Program. For a 510(k) submission, FDA generally does not require such plans for devices, absent certain circumstances, for which a performance monitoring plan is not a special control for the particular device type (i.e., the applicable classification regulation). For a De Novo classification request, such a plan may be necessary to control risks posed by the particular device type and so FDA may establish a special control for the device type going forward. For a PMA, a performance monitoring plan may be a condition of approval.[57] However, sponsors may opt to include information regarding the performance monitoring plan in any submission for an AI-enabled device.

Performance monitoring plans should identify and respond to, in a timely fashion, performance changes or conditions that may lead to performance change or degradation. A robust

[57] See 21 CFR 814.44 and 21 CFR 814.82.

performance monitoring plan includes proactive efforts to capture device performance after deployment. Components of such a plan may include:

- A description of the data collection and analysis methods for:
  - Identifying, characterizing, and assessing changes in model performance, including assessing the results from performance monitoring on safety and effectiveness.
  - Monitoring potential causes of undesirable changes in performance, such as:
    - Changes in patient demographics or disease prevalence;
    - Shifts in input data;
    - Changes to input data due to corruption in the data pipeline (input data integrity), such as missing values, duplicate records, data type mismatches; and
    - Changes in users' behavior or in user demographics.
- A description of robust software lifecycle processes that include mechanisms for monitoring in the deployment environment.
- A plan for deploying updates, mitigations, and corrective actions that address changing performance in a timely manner.
- FDA notes that some actions taken to address performance changes may not require a marketing submission or authorization (21 CFR 807.81(a)(3) and 21 CFR 814.39(a)) prior to being taken. Please refer to FDA guidances titled, "Deciding When to Submit a 510(k) for a Change to an Existing Device," and "Deciding When to Submit a 510(k) for a Software Change to an Existing Device" to help assess whether a particular change may require a premarket submission to FDA. Sponsors may also wish to consider the use of a PCCP, as appropriate. [58]
  - This plan does not replace applicable statutory or regulatory requirements, including the requirements to report to FDA information about certain adverse events, and corrections and removals, under 21 CFR Parts 803 and 806.
- A description of the procedures for communicating the results of performance monitoring and any mitigations to device users.

Where sponsors should provide it in a submission: When appropriate, a device performance monitoring plan should be included in the "**Risk Management File**" in the Software Documentation section of the marketing submission.

XII. Cybersecurity

Why should it be included in a submission for an AI-enabled device: As with any digital or software component integrated into a medical device, AI can present cybersecurity risks. FDA's general recommendations for designing and maintaining cybersecurity as well as relevant marketing submission documentation are provided in the guidance document titled "Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions" (hereafter referred to as the "2023 Premarket Cybersecurity Guidance"). The 2023 Premarket Cybersecurity Guidance identifies security objectives that may be relevant for medical

[58] See FDA's guidance titled "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions."

devices, including AI-enabled devices: authenticity, which includes integrity; authorization; availability; confidentiality; and secure and timely updatability and patchability.

For AI-enabled devices that meet the definition of a "cyber device" under section 524B(c) of the FD&C Act, the recommendations in this section of the guidance are intended to help manufacturers meet their obligations under section 524B of the FD&C Act. Examples of AI risks which can be impacted by cybersecurity threats include, but are not limited to:

- *Data Poisoning:* Cyber threats could lead to data poisoning by deliberately injecting inauthentic or maliciously modified data, risking outcomes in areas like medical diagnosis.
- *Model inversion/stealing:* Cyber threats could intentionally use forged or altered data to infer details from or replicate models. These pose risks to continued model performance as well as intellectual property and privacy breaches.
- *Model Evasion:* Cyber attackers could intentionally craft or modify input samples to deceive models, leading them to incorrect classifications. These pose risks to the reliability and integrity of model predictions, potentially undermining trust in AI-enabled devices and exposing them to malicious exploitation.
- *Data leakage:* Cyber threats could exploit vulnerabilities to access sensitive training or inference data in models.
- *Overfitting:* Cyber threats could deliberately "overfit" a model, exposing the AI components to adversarial attacks as these components struggle to adapt effectively to modified patient data.
- *Model Bias:* Cyber threats could lead to manipulation of training data to introduce or accentuate biases. They could exploit known biases using adversarial examples, embed backdoors during training to later trigger biased behaviors, or leverage pre-trained models with inherent biases, amplifying them with skewed fine-tuning data.
- *Performance Drift:* Cyber threats could lead to model performance drift by changing the underlying data distribution, which degrades model performance. Cyber threats could slightly shift the input data over time or exploit vulnerabilities in dynamic environments, causing the model to make inaccurate predictions or become more susceptible to adversarial attacks.

What sponsors should include in a submission: Consistent with the submission documentation recommended in the 2023 Premarket Cybersecurity Guidance regarding the cybersecurity controls and security risk management relevant to the AI components or features, sponsors should include the following types of information:

- Any additional elements in the cybersecurity risk management report, threat modeling, cybersecurity risk assessment, labeling, and other deliverables, where there are unique considerations related to AI cybersecurity.
- An explanation regarding how the cybersecurity testing is appropriate to address the risks associated with the model, including, at minimum the following tests:
  - Malformed input (fuzz) testing; and
  - Penetration testing.
- A Security Use Case View(s) that covers the AI-enabled considerations for the device.
- A description of controls implemented to address data vulnerability and preventing data

leakage, including:
- Access controls;
- Any data encryption; and
- Anonymization or de-identification of sensitive data.

Sponsors should refer to the control recommendations in Appendix 1 of the 2023 Premarket Cybersecurity Guidance for how they may wish to address the specific risks above. Example approaches to controlling cybersecurity risks related to AI-enabled devices include:

- For data poisoning attacks, consider:
  - Validating, authenticating, and cleansing data.
  - Employing anomaly detection and data integrity checks (e.g., cryptographic hashes).
  - Applying adversarial training, which is a method used to improve the robustness and security of models.
- For cyber threats using forged data to introduce overfitting, model bias, etc., consider:
  - Adopting differential privacy, which is a technique to protect the privacy of individual data points in a dataset. When utilizing differential privacy, sponsors should be cognizant of potential trade-offs between privacy and factors such as model accuracy, utility, and efficiency, and provide information on how the trade-offs are addressed.
  - Engaging in secure multi-party computation (MPC), which is a technique that can allow multiple parties to collaboratively train a model without revealing their local datasets to each other.
  - Employing data authentication and integrity protections.
  - Introducing watermarking, which involves embedding hidden watermarks into AI models to prove ownership.
  - Applying continuous model performance monitoring.
- For model evasion, consider adversarial training to enhance model robustness and implement strict input verification checks to ensure data conforms to expected patterns. When deploying adversarial training techniques, sponsors should be cognizant of the trade-offs that may arise between enhanced robustness to attacks and the potential negative impact on model performance (e.g., accuracy), and provide information on how the trade-offs are managed.

Where sponsors should provide it in a submission: The cybersecurity information should be included in the "**Cybersecurity/Interoperability**" section of the marketing submission, as described in the 2023 Premarket Cybersecurity Guidance.

ADDITIONAL RESOURCES: Sponsors may also refer to other FDA guidance documents for additional recommendations relevant to cybersecurity:

- Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions
- Postmarket Management of Cybersecurity in Medical Devices
- Cybersecurity for Networked Medical Devices Containing Off-the-Shelf (OTS) Software
- Off-The-Shelf Software Use in Medical Devices

XIII. Public Submission Summary

Why should it be included in a submission for an AI-enabled device: Transparency is a key component of premarket authorization and is important to patient care. This is especially important for AI-enabled devices, which are heavily data driven and incorporate algorithms exhibiting a degree of opacity. In public workshops and comments, including the October 14, 2021 virtual public workshop on the transparency of AI-enabled devices titled "Transparency of Artificial Intelligence/Machine Learning-enabled Medical Devices," patients noted concerns with the use of AI in their care. The public has consistently called for additional information about how FDA makes authorization decisions about AI-enabled devices, as well as more information about the design and validation of these devices. The public submission summary should include specific information describing the characteristics of these devices to support transparency, which can contribute to public health by increasing understanding of AI-enabled devices and developing public trust.

Public submission summaries are required and available on the FDA website for most marketing authorization decisions.[59] These summaries describe the device and the information supporting regulatory decision-making. Where a public summary is required, details about the AI-enabled device must be included in sufficient detail in the public-facing documents to support transparency to users of FDA's determination of substantial equivalence or reasonable assurance of safety and effectiveness for the device.[60,61,62] To ensure public access to important information on authorized AI-enabled devices, this section describes the types of information sponsors should include in the public submission summary as well as a possible format for such information.

For AI-enabled devices submitted through the PMA, HDE, De Novo, BLA, or 510(k) pathways, FDA recommends that the information discussed in this section be included in the relevant public submission summary, or the 510(k) Summary (in the section prepared in compliance with 21 CFR 807.92(a)(4)), as applicable. Sponsors should provide the recommended information excluding any patient identifiers, trade secrets, and confidential commercial information. For sponsors submitting a 510(k) Statement (21 CFR 807.93), FDA recommends providing the same information in the submission excluding any patient identifiers, trade secrets, and confidential

[59] See FDA's website titled "CDRH Transparency: Premarket Submissions." See 21 CFR 807.92 for requirements on the form and content of a 510(k) Summary. See 21 CFR 807.93 for requirements on the content and format of a 510(k) Statement. See 21 CFR 814.9(e) for requirements on a PMA decision summary.

[60] In accordance with 21 CFR 807.92, "a 510(k) summary shall be in sufficient detail to provide an understanding of the basis for a determination of substantial equivalence." See 21 CFR 807.92 for requirements on the content and format of a 510(k) Summary. If a sponsor chooses to submit a 510(k) Statement rather than 510(k) Summary, the sponsor should provide information that supports FDA's determination of substantial equivalence. See 21 CFR 807.93 for requirements on the content and format of a 510(k) Statement.

[61] In accordance with 21 CFR 814.9(e), "FDA will make available to the public … a detailed summary of information submitted to FDA respecting the safety and effectiveness of the device that is the subject of the PMA and that is the basis for the order." See 21 CFR 814.9(e) for requirements on a PMA decision summary.

[62] The De Novo decision summary is intended to present an objective and balanced summary of the scientific evidence that served as the basis for the FDA's decision to grant a De Novo request. For more information on De Novo decision summary documents, please see FDA's website on De Novo Classification Request.

commercial information.[63]

While not required, the use of a model card may be one way to communicate information about AI-enabled devices because they are a means to consistently summarize the key aspects of AI-enabled devices and can be used to concisely describe their characteristics, performance, and limitations. Appendix E (Example Model Card) provides recommendations for the contents and formatting of a model card. Research has demonstrated that the use of a model card can increase user trust and understanding. The use of a model card as part of a public submission summary specifically is one way to support clear and consistent communication about an AI-enabled device to the interested parties in the public as well as to users, such as patients, clinicians, regulators, and researchers. The use of the model card can address the challenges associated with determining the best approach to communicate important information about the AI-enabled device.

What sponsors should include in a submission: Sponsors must comply with the submission regulations for their particular submission.[64] In addition, sponsors should consider FDA recommendations for the relevant marketing submission type. Sponsors should also provide the following types of information excluding any patient identifiers, trade secrets, and confidential commercial information:

- A statement that AI is used in the device;
- An explanation of how AI is used as part of the device's intended use. For devices with multiple functions, this explanation may include how AI-DSFs interact with each other as well as how they interact with non-AI DSFs;
- A description of the class of model (e.g., convolutional neural network, recurrent neural network, support vector machine, transformers) and limitations of the model within the device description;
- A description of the development and validation datasets (size, source of data), including information about the demographic characteristics in the training and validation data, along with information about the demographic characteristics in the population(s) of intended use. The description should also compare the training dataset to the validation dataset and model data inputs expected in the intended use. The comparison should describe how independence of test data from training data was ensured;
- A description of the statistical confidence level of predictions, including any other descriptions or metrics that describe statistical confidence and uncertainty, as applicable; and
- A description of how the model will be updated and maintained over time, if applicable.

Sponsors should consider using a model card to organize information. Appendix E (Example Model Card) includes recommendations on the elements that may be included within a model card. While the example model card includes recommended elements and format for a model

[63] For more information, see FDA guidance, "The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)]."

[64] For more information regarding the requirements for PMA, see 21 CFR Part 814. For more information regarding the requirements for 510(k), see 21 CFR 807.81 – 807.100. For more information regarding the requirements for De Novo, see 21 CFR 860.200 – 860.260. For more information regarding the requirements for HDE, see 21 CFR 814.100 – 814.126. For more information regarding the requirements for BLA, see 21 CFR Part 600 – 680.

card, sponsors may include additional information and/or follow a different format. In the absence of the model card structure, sponsors should still consider including the information a model card contains.

Where sponsors should provide it in a submission: The public submission summary should be included in the "**Administrative Documentation**" section of the marketing submission.

> **ADDITIONAL RESOURCES:** Appendix E (Example Model Card) of this guidance provides one example of the format of a model card. Appendix F (Example 510(k) Summary with Model Card) of this guidance provides an example of a public submission summary for a product, including a completed model card.

Appendix A: Table of Recommended Documentation

Sections V-XIII of this guidance provide recommendations regarding the documentation that may be included within a marketing submission for AI-enabled devices. The table below summarizes recommended locations within the marketing submission to provide discussed documentation. One way this documentation may be submitted is through the eSTAR Program. Specifically, eSTAR is an interactive PDF form that guides applicants through the process of preparing a comprehensive medical device submission.[65] eSTAR is free and is required for all 510(k) submissions, unless exempted.

Guidance Section and Recommended Information	**Recommended Section in Sponsor's Marketing Submission**
Section V Device Description	Device Description
Section VI.A User Interface	Software Description
Section VI.B Labeling	Labeling
Section VII Risk Assessment	Risk Management File of Software Documentation
Section VIII Data Management	Data for development: Software Description of Software Documentation Data for testing: Performance Testing
Section IX Model Description and Development	Software Description
Section X.A Performance Validation	Clinical and non-clinical testing: Performance Testing Software verification and software validation: Software testing as part of verification and validation of Software Documentation
Section XI Device Performance Monitoring	Risk Management File of Software Documentation
Section XII Cybersecurity	Cybersecurity
Section XIII Public Submission Summary	Administration Information

[65] For more information on eSTAR, please see FDA's website on eSTAR Program.

Appendix B: Transparency Design Considerations

This appendix contains recommendations for developing a transparent device centered around users. These recommendations are intended to help sponsors develop safe and effective medical devices and high-quality marketing submissions. While sponsors may identify alternate approaches that support FDA's evaluation of safety and effectiveness, they should integrate transparency considerations starting at the design phase of the TPLC to ensure the availability of information to support the marketing submission. It can be difficult to integrate transparency into a device in later stages of the TPLC when changes to the device might require additional testing. In this guidance, transparency refers to clearly communicating the contextually relevant performance and design information of a device to the appropriate stakeholders in a manner that they can understand and act on. Transparency involves ensuring that important information is both accessible and functionally comprehensible and is connected both to the sharing of information, and to the usability of a device. As such, a user-centered approach to transparency design helps support the safe and effective use of AI-enabled devices. Including appropriate transparency information has also been shown to more than double willingness to use a device.

Transparency by Design Across the TPLC

Sponsors should take a holistic approach to identifying relevant contextual factors for device use and how those factors impact device performance when determining what information should be communicated. Sponsors should consider transparency throughout the full continuum of implementation through use, maintenance, and decommission of the AI-enabled device, and should design the device with transparency in mind from the beginning.

The user interface is another area where transparency principles should be used, when appropriate. The information in other elements of the user interface can complement the printed labeling (e.g., packaging and user manuals) to support the user's understanding of how to use the device by providing timely and contextually relevant information throughout the use process, as described in Section VI (User Interface and Labeling). Examples of points of interaction include alerts generated by a device and displayed on the device or pushed to another product, components of associated hardware, and display screens. Effective transparency planning identifies the necessary information for the intended user(s) and context of use, as well as the optimal mediums, timing, and strategies for successful communication of the necessary information.

Generally, the transparency design process should begin with a holistic approach to obtain an understanding of the context in which a product is used, followed by identifying user tasks, and possible risks associated with communication of information during those tasks. This can be accomplished by determining how and when information is needed, integrating contextually appropriate risk controls into the design of the product, and finally validating that the intended users receive and can functionally understand the key information in relevant use contexts. This process may be iterative and may not flow linearly.

Transparency is contextually dependent, so appropriate information will vary across the range of AI-enabled devices and depend upon their benefit/risk profiles and the needs of intended users. The considerations in this appendix are not exhaustive and are intended to help sponsors identify

information about the context in which the device will be used and the needs of the users for the purpose of developing a consistent approach to understanding the transparency needs for their AI-enabled device. It is also important to note that while transparency can help to address certain device risks, particularly those related to misunderstanding or misusing information output by a device, providing transparency about the existence of a significant clinical risk, including a significant risk related to performance in subpopulations of intended users, alone may not be an adequate risk control.

The Right Information at the Right Time

Consider what information the users might need, when they might need it to facilitate decision-making, and the potential risks if the users do not have the appropriate information at the right time, at all, or if it is misunderstood. It is important to focus on the tasks that each user has to perform, and what the user needs to know to perform them in concurrence with the intended use. To identify what information needs to be gained and is critical for users, consider the intended use comprehensively with questions, such as:

- Who needs the information and what is the most effective method of communication?
- When does the user need to understand information to support safe and effective use?
- What is the context of use? Examples of questions about the context of use include:
    - Where will the device be used and what are the conditions in that space?
    - What else might users be doing at the same time?
    - How timely is the application of the information?
    - In what settings will the device output be viewed?
    - Will users who interpret and apply the output be the same as those who operate the device?

Information should be communicated at the time that it is needed. Some examples of elements of the user interface that could be used to communicate transparency information include:

- Packaging,
- Labeling,
- User Training,
- Controls,
- Display elements
- Outputs/ reports,
- Alarms/ warnings, and
- Logic of operation of each device component and of the user interface system as a whole.

Understanding User Characteristics and Needs

The ability of a user to operate an AI-enabled device depends on their personal characteristics and the device use environment. The environments in which AI-enabled devices are used may also influence a user interface design. As part of design inputs, consider the needs of users in the context of use. Understanding users and their needs and limitations should occur early in the development process for the AI-enabled device and may be repeated as the design process continues. Users may include, for example:

- Patients,
- Purchasers,

- Administrators,
- Healthcare Professionals,
- Caregivers, and
- Maintenance Technicians.

It is important to consider the characteristics of each user that may impact the user needs, including appropriate content and format for communication. Considerations may include:

- The user's functional capabilities, including cognitive, physical and sensory capabilities;
- The experience and knowledge levels of the users, including their educational backgrounds;
- The frequency at which the user will interact with the device;
- The level of training users are expected to receive; and
- The similarities and differences of the new information as compared to information the users have utilized in the past.

Communication Style and Format

It is also important to consider the format used for communication. The format should be clear and appropriate for each user at each user task. Factors may include:

- The reading level of the user.
- The location of information.
- Design elements such as:
    - Hierarchy,
    - Visualizations, and
    - Dynamic labeling.

The selection of the timing, mode, and format of communication should be incorporated early to allow for iterative design.

Explainability Information and Visualizations

It is also important to consider when additional information may detract from understanding, rather than add to it. For example, explainability tools or visualizations can be valuable in increasing model transparency and a user's confidence in a model's output and could be developed as part of the user interface. However, if not well designed and validated for the target user group, explainability tools or visualizations could also significantly mislead users. Therefore, sponsors should develop and validate explainability metrics and visualizations through appropriate testing.

Appendix C: Performance Validation Considerations

This appendix contains recommendations for some aspects of clinical performance validation in AI-enabled devices, which are intended to help sponsors develop safe and effective medical devices. While sponsors may identify alternate approaches that support FDA's evaluation of safety and effectiveness, they should rigorously test the device to establish the device's performance, and integrate that planning early in the design and development process to ensure the collection of appropriate data to support the device's intended use. It can be difficult, for example, to gather additional supportive data after the completion of the pivotal clinical study. Sponsors should also follow the recommendations found in other FDA guidances regarding specific clinical study considerations. For example, additional information on evaluating and reporting results for AI-enabled devices can be found in the FDA guidances "Design Considerations for Pivotal Clinical Investigations for Medical Devices," "Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests," and "Electronic Source Data in Clinical Investigations." These recommendations may not apply to all device types.

Pre-specification of Study Protocols and Statistical Analysis Plan

Post-hoc analysis may bias the performance assessment. Therefore, to accurately evaluate the performance of the device, study protocols and statistical analysis plans should be pre-specified. Regardless of whether data are collected prospectively or retrospectively, study design elements (such as sample size justification, and plans on how to handle, prepare, process, and select archived data or material) should be specified prior to beginning the validation study.

Study Reports

All performance and usability assessments should be objective, and the model should not be tweaked opportunistically in light of the test data results (i.e., no post-hoc adjustment). In general, proceeding to execute the study protocol only after a sound validation plan (study protocol and statistical analysis plan) is documented and finalized helps avoid these post-hoc adjustments. Execution of the plan includes collecting the required data, conducting the pre-specified analysis, and reporting the study results. Validation study reports should specify the associated protocol version and adequate justifications should be provided for any repeated tests or tests with deviations from the pre-specified plans.

Masking Protocol

For diagnostic devices, a masking protocol in the clinical study ensures that the user of the test is "blinded/masked" to the clinical reference standard result while the provider of the clinical reference standard result is "blinded/masked" to the test result. The masking protocol also ensures that model developers and the clinical team are completely masked from the test data during the model development process.

For therapeutic devices, masking is sometimes implemented through a randomized-controlled study with two arms (e.g., placebo/sham device arm and subject device arm), when ethically appropriate such as with non-invasive diagnostic devices. This ensures patients and care providers are blinded to the actual treatment assignment. The placebo arm may not have any measurement but only serve as a blinding tool (e.g., so that caretakers will not provide differentiated care in different arms). When such a two-arms study design is not feasible, there

may be potential bias in the performance assessment due to placebo effects.

Model Precision: Repeatability and Reproducibility

An AI-enabled device may often be intended to measure physiological signals when the device is placed on a particular anatomical location. It is important to know how robust the device output is due to potential variations in the measurement system (e.g., whether repeated tests by users will generate significantly different device output due to operator difference and signal variation). A precision study gauges the variability of a device output when making repeated measurements on the same patient, either with the same operator and device (repeatability), or with different operators and devices (reproducibility). More generally, repeatability is the closeness of agreement of repeated measurements taken under the same conditions; and reproducibility is the closeness of agreement of repeated measurements taken under different, pre-specified conditions.

It is important to note that not every diagnostic device needs a precision study, due to clinical and feasibility considerations. For example, there is a feasibility concern when a device may be too harmful on the patients with repeated use (e.g., for radiation or invasive devices). Another example is a monitoring device that tracks a patient's changing physiological status (e.g., hemodynamic parameters) in real-time, where repeated observations of the same truth are not possible.

Key statistics to summarize the repeatability and reproducibility, based on a variance component analysis using a model's continuous metric (e.g., a probability score), are the subject-level standard deviation (SD) and the percent coefficient of variation (%CV). Improving the model to reduce SD or %CV may provide a low-cost way to improve product quality and the success likelihood of a future pivotal clinical study. This is, in part, because the clinical reference standard (i.e., the best available method for establishing the presence or absence of the target condition) is not measured in a precision study. Depending on the product, additional factors may be considered in the precision study. In image classification tasks, a model may be sensitive to data perturbation (e.g., image translation/rotation, light intensity change, random noise). This phenomenon could be abundant for an AI-enabled medical device software running on a generic smartphone using its camera to capture measurement data (e.g., skin lesion analyzers).

Study Endpoints and Acceptance Criteria

Primary endpoints are usually assessed using pre-specified acceptance criteria within a statistical hypothesis testing framework. This approach necessitates an adequate sample size to ensure sufficient study power (i.e., acceptable type II error rate). Secondary and exploratory endpoints may also be used to inform the effectiveness of the device and are part of the totality of evidence that inform regulatory decisions. The evaluations of primary endpoints are typically based on their 95% two-sided confidence intervals (so that type I error can be protected at 5% for two-sided testing; and at 2.5% for one-sided testing). The validation of all outputs should be addressed, appropriately by type (e.g., continuous, categorical, risk scores).

An AI-enabled medical device can produce a variety of outputs, such as diagnostic/prognostic predictions, or treatment triaging/priority ranking/selection/planning. The validation of these outputs may involve an analytic study (e.g., precision, bench, simulation study), literature

review, a diagnostic performance study, a reader study (e.g., multi-reader multi-case imaging study), or a clinical outcome study (e.g., based on a study or randomized-controlled trial design).

When specifically considering an AI-enabled diagnostic device, the key performance assessment is its diagnostic accuracy, which is evaluated in a pivotal diagnostic performance study. Due to sampling variation, the uncertainties of the accuracy estimates are typically quantified, usually in the form of 95% two-sided confidence intervals. The study acceptance criteria can be based on statistical inferences using hypothesis testing methods (e.g., comparing a lower/upper confidence limit to a pre-specified performance goal). Note that inferences based on point estimates ignores the statistical uncertainty of the estimates and is not generally acceptable in the primary analysis. It is always compared to a comparator that can be tested and evaluated on the same patient/data as the device. This comparator can be the clinician, another device that is adequately validated for the same intended use, or standard of care. The evaluation on the same patient/data is key to mitigate differences in the task difficulty levels and disease spectrum due to sampling variation.

Depending on the nature of the diagnostic output (i.e., binary, polychotomous, or continuous), different evaluation metrics are possible.

- For binary diagnostic output, evaluation may be based on, sensitivity, specificity, positive/negative predictive values (PPV/NPV), and positive/negative diagnostic likelihood ratios (LR+/LR-).
- For risk stratification output that classifies a patient into one of multiple risk groups and that may often be found in prognostic models, some evaluation metrics are pre/post-test risks and likelihood ratios.
- For an output that evaluates a patient's disease risk with a continuous score, some risk evaluation methods are calibration plot, receiver operating characteristic (ROC) curve, and decision curve analysis. In the context of biomarker evaluation, the predictiveness curve analysis may be used.
- For continuous score, agreement study methods using MAE (mean absolute error), RMSE (root mean squared error), scatter plot, Deming regression, and Bland-Altman analysis are often used.

When the test data consists of multiple observations per patient, the within-patient correlations should be accounted for in the calculation of confidence intervals. Failure to account for the repeated measurements appropriately in the statistical analysis may lead to biased estimates and incorrect narrow confidence intervals, which may hinder objective evaluation of the device performance. Statistical techniques that account for patient-level repeated measurements include the bootstrap resampling method and analytic methods for clustered data.

Validation of AI-based Pre-processing Steps

Some models may include a quality control algorithm that discards "low" quality cases from further processing. However, such low-quality cases may actually be truly hard/difficult ones –an example of missing not at random (MNAR), which may lead to skewed diagnostic performance in accuracy metrics (e.g., sensitivity and specificity) but also may be biased (e.g., in the sense that more patients than warranted may not get any results due to declared low-quality events). An

analysis of cases deemed low-quality should be conducted to verify that the quality control algorithm does not discard challenging cases.

For example, compare two hypothetical AI-enabled diagnostic devices (A and B) using cellphone cameras for certain skin disease detection. Assume they use the same diagnostic models, except that A has a more aggressive quality control (QC) algorithm than B in declaring low-quality cases. After excluding those cases that fail the QC algorithm, it may not be surprising to observe that A would have a better diagnostic performance than B, because many low-quality images dropped by A but not by B may in fact be good quality but difficult cases which are not included in the performance assessment for A.

Thus, a good practice is to examine the influence of a QC algorithm by checking the proportion of low-quality dropouts and assessing the results of a sensitivity analysis assuming a worst-case scenario (i.e., assuming the QC failure cases are all difficult ones that the model fails to classify successfully).

Appendix D: Usability Evaluation Considerations

As described in this guidance Section X Validation, sponsors should conduct human factors evaluations as part of design controls (21 CFR 820.30) for every medical device requiring a premarket submission. The Human Factors Guidance outlines analytical approaches to this evaluation as well as specific requirements for human factors validation for devices when one or more critical tasks has been identified. Human factors engineering processes typically begin with preliminary analysis and evaluation of all tasks that identifies critical tasks which, if performed incorrectly or not performed at all, could cause serious harm. [66] Sponsors should perform this analysis to identify whether a device has a critical task. If a critical task is identified, sponsors should refer to the Human Factors Guidance and perform human factors validation. While sponsors of devices that do not have a critical task may not need to submit a human factors validation testing report, they may choose to use the process outlined in the Human Factors Guidance, or another approach of their choosing to evaluate usability,[67] to test their device design, and support the efficacy of risk controls. This appendix is focused on the evaluation of usability to support risk controls when a human factors validation testing report is not required, where usability addresses whether all intended users can achieve specific goals while using the device and whether users will be able to consistently interact with the device safely and effectively. This includes, but is not limited to, whether users can consistently and correctly receive, understand, interpret, and apply information related to the AI-enable device.

While FDA's Human Factors Guidance outlines recommended analytic approaches for evaluating usability, sponsors may choose to utilize alternative approaches for the evaluation of user tasks outside of the scope of that guidance. If this testing is used to support a risk control (as described in Section VII (Risk Assessment)), sponsors should include a description of the pre-specified testing protocols and analysis plans, and a justification for the appropriateness of the assessment method.

For AI-enabled devices, it may be specifically important for sponsors to identify and evaluate risk controls related to user tasks regarding the interpretation and use of information and interactions with novel user interfaces. The application of this information is a particular challenge for users of AI-enabled devices because models developed through AI techniques vary in explainability and interpretability. For example, some models can be explained using a simple decision tree which is generally easy for a user to follow and understand the basis of a model's recommendations. Other models use complex, deep neural networks, where it may not be feasible to explain in a way that allows a user to completely understand the basis of recommendations, even with comprehensive information on its inputs, nodes, and weights. This means that users may not be able to easily and independently verify whether the recommendations and decisions made by an AI-enabled device are appropriate. As such, AI-enabled devices can be prone to errors of device use and information interpretation. The challenges with interpretability and explainability increase when the intended user has limited

[66] For more information regarding critical tasks, please see FDA guidance titled "Applying Human Factors and Usability Engineering to Medical Devices."

[67] See Section X (Validation) for context regarding "usability" for the purpose of this guidance.

training in interpreting the output of models, when the intended use is in situations that require urgent action, when the model has no evident biological mechanism of action, and when the model changes through iterative updates. These errors can cause harm (injury or damage to the health of including the effects of delayed or incorrect clinical intervention, or damage to property or the environment)[68] and impact the safe and effective use of the device.

When sponsors choose to include an evaluation of usability to support the control of risks related to information, as described in Section VII (Risk Assessment), the evaluation should be appropriate to demonstrate that the user can both find and apply the information. In such cases, an impact assessment may be used to determine which user tasks could have an adverse or positive effect on knowing, understanding, and applying information for the device. As appropriate for the AI device, this assessment may include, for example, evaluation of the training program intended for risk control. For more complex AI devices with several sequential risk controls, an example evaluation approach could include use of the device in a clinical feasibility study that includes comprehensive assessment of how the user interpreted the AI outputs and what actions were taken. Ultimately, it is important to evaluate whether the user can operate and interpret the device, including demonstrating that users can understand and apply important information about the use of the device and its output in the actual context of clinical decision-making.

Sponsors may wish to draw on the general structure outlined Section 6.3.1 (Task Analysis) of the Human Factors Guidance, which provides an example of an analysis technique to systemically break down device use into discrete user tasks. However, it is important to understand that while the Human Factors Guidance focuses on "serious harm," sponsors may need to provide documentation evaluating and addressing any potential risk associated with misuse, including misinterpretation, to ensure that the device is safe and effective for its intended use.

Appendix B (Transparency Design Considerations) of this guidance also outlines recommendations to user-centered transparency, which may help with the identification of user tasks and risks related to usability and information interpretation, as well as help sponsors develop design approaches to control these risks.

[68] ANSI/AAMI/ISO 14971 *Medical devices—Application of risk management to medical devices*.

Appendix E: Example Model Card

A model card is a popular format for communicating information about a device that may align with the kind of information that FDA may require, for example in the publicly available 510(k) summaries[69] and labeling.[70] The model card format and content discussed below is intended to serve as an example of possible formatting a sponsor could use to communicate information about the model and the AI-enabled device in the public submission summary and other locations where this information may be shared by the sponsor. It is important to note that FDA does not require the inclusion of a model card or a specific model card format, and this example should not be considered a template. The example model card below has been designed based on user-centered research to present data in an order and format that is useful and easy to understand for non-technical audiences and is provided to sponsors to facilitate the inclusion of a model card.

In general, model cards can be adapted to the specific needs and context of each AI-enabled device. However, for the public summary, we encourage sponsors to follow the general principles for creating model cards outlined in this guidance. Some elements may not be available for some devices.

When model cards are provided in a digital format, research has demonstrated that a dynamic approach to formatting that allows users to expand sections individually as needed makes the information easier to digest. While the public submission summary is provided as a PDF document and the format is static, sponsors should consider the use of dynamic labeling when possible.

DEVICE NAME – Model Card

Device Information:

- Name of the Device
- Version of the Device
- Date when the Device was created (or last updated)
- Model Architecture

Regulatory Status (For model cards used outside of the public submission summary):

- Authorization status
- File number

Description:

- Intended users (e.g., healthcare professionals, caregivers, patients).
- Intended use – The general purpose of the device or its function. This includes the indications for use.
- Indications for use – Describes the disease or condition the device will diagnose, treat,

[69] See 21 CFR 807.92.
[70] See 21 CFR Part 801.

prevent, cure or mitigate, including a description of the target patient population for which the device is intended and the intended use environment (e.g., intensive care unit, step-down unit, home).

- Instructions for Use – Directions and recommendations for optimal use of the model.
- Clinical benefits (e.g., analyze personalized patient information to improve diagnosis, treatment assignment, monitoring, or prevention of a medical condition, risk assessment) and limitations, including whether the device is intended to be used by, or under the supervision of, a healthcare provider.
- Clinical workflow phase (e.g., patient pre-registration, digitization of forms or clinical scales, patients' triage, telehealth & virtual rounds, clinical decision support systems, workflow optimization, evidence-based methods to optimize medical interventions, feedback from users).
- Inputs and outputs of the model and contribution to healthcare decisions or actions.
- Degree of automation compared to the current standard of care, including whether the device supports or automates decision making.

Performance and Limitations:

- Accuracy (e.g., sensitivity, specificity, positive/negative predictive values, and their 95% two-sided confidence intervals).
- Known biases or failure modes.
- Precision (reproducibility) associated with the provided outputs.
- Known gaps in the data characterization, such as patient populations that are not well represented in development (e.g., training) or testing datasets, and therefore, may be at risk of bias.
- Limitations in the model development or performance evaluation.
- Known circumstances where the device input will not align with the data used in development and validation.
- Evidence (e.g., clinical trial number or for published results of a supporting study, the unique reference ID such as Digital Object Identifier, or PubMed Identifier information).
    - Data Characterization for data used to test the device:
        - Data sources (e.g., clinical trials, public or proprietary databases) including details on any devices used to collect data;
        - Data types used (e.g., structured numerical data, structured categorical data, unstructured text, images, time-series data, or a combination); and
        - Relevant details including the sample size, effect size, data quality, reference standard, diversity, and representativeness.
- Methods used to establish and ensure that the model meets the intended use and user requirements (e.g., human factors validation/usability evaluation, user acceptance testing, clinical validation, identification of pre-trained models, other).

Risk Management:

- Potential risks associated with the model, the data, and the outputs (e.g., contraindications, side effects, data privacy risks, cybersecurity risks, bias risks, information gaps).
- Description of information that could impact risks and patient outcomes, across the

product lifecycle.

- Interactions, Deployment, and Updates. When appropriate, provide the:
  - Computational Resources Required.
  - Details regarding how the model is deployed and updated, including:
    - How to conduct local site-specific acceptance testing or validation;
    - Ongoing performance monitoring;
    - Transparent reporting of successes and failures;
    - Change management strategies; and
    - Proactive approaches to address vulnerabilities.
  - Communication to parties of as-needed information.
  - Software quality (specify, standards and regulatory compliance issues, intellectual property issues, risk management and safeguards used, other).

Development:

- Data Characterization of data used to develop the device:
  - Data sources (e.g., clinical trials, public or proprietary databases) including details on any devices used to collect data.
  - Data types used (e.g., structured numerical data, structured categorical data, unstructured text, images, time-series data, or a combination).
  - Relevant details including the sample size, effect size, data quality, reference standard, diversity, and representativeness.

Appendix F: Example 510(k) Summary with Model Card

In general, publicly available summaries must follow the applicable the requirements for the specific marketing submission (e.g., 510(k),[71] De Novo,[72] PMA[73]). The items below are not an exhaustive list of topics that a manufacturer may be expected to cover, and all topics may not apply to all marketing submissions. Likewise, FDA may request additional information to be included in this summary. This appendix serves as only an example of the types of information sponsors should generally provide in a 510(k) summary, including an example of a completed Basic Model Card. Information does not need to be repeated between the model card and other sections of the public summary, but information can be repeated if the sponsor believes that the alternate format provides useful context.

Indications For Use:

The Disease X screening model is software intended to aid in screening for Disease X on patients above the age of 22 by analyzing 12-lead electrocardiogram (ECG) recorded from compatible ECG devices. It is not intended to be a stand-alone diagnostic device for Disease X. However, a positive result may suggest the need for further clinical evaluation in order to establish a diagnosis of Disease X. If the patient is at high risk for Disease X, a negative result should not rule out further non-invasive evaluation. It should not be used to replace the current standard of care methods for diagnosis of Disease X but applied jointly with clinician judgment.

Device Description:

The stand-alone software contains a machine learning model that uses a convolutional neural network to interpret and analyze 10 seconds of a 12-lead resting electrocardiogram acquired from 4 compatible ECG devices (A, B, C, and D) and provide an output on the likelihood of whether a patient has Disease X and further clinical evaluation is required. The software also contains quality checks that will notify the end user on whether the ECG data provided does or does not meet the ECG input requirements to generate a model output. If it does not meet the requirements, an error message will be displayed.

Summary of Technological Characteristics:

	Subject Device	Predicate Device	Comparison
Application Number	KXXXXXX	KXXXXXX	-
Product Codes	XXX	XXX	-

[71] See 21 CFR 807.92. For more information, please see FDA guidance titled "The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)]."
[72] See 21 CFR 860.220. For more information, please see FDA guidance titled "De Novo Classification Process (Evaluation of Automatic Class III Designation)."
[73] See 21 CFR Part 814.9(e).

	Subject Device	Predicate Device	Comparison
Regulation Number	21 CFR XXXX	21 CFR XXXX	-
Rx/OTC	Rx	Rx	Same
Indication for Use	The Disease X screening model is software intended to aid in screening for Disease X on patients above the age of 22 by analyzing 12-lead electrocardiogram recorded from compatible ECG devices. It is not intended to be a stand-alone diagnostic device for Disease X. However, a positive result may suggest the need for further clinical evaluation in order to establish a diagnosis of Disease X. If the patient is at high risk for Disease X, a negative result should not rule out further evaluation. It should not be used to replace the current standard of care methods for diagnosis of Disease X but applied jointly with clinician judgment.	Software intended to be used as an aid in determining if a patient has Disease X in patients 18 years and above. The software analyzes a 12 lead ECG from compatible devices and should not be used as a stand-alone diagnostic device.	Similar. Both devices are used as aids and screening tools for Disease X. The indications for use for the predicate device is for patients 18 years and above whereas the subject device is for patients 22 and above.
Operational Mode	Spot Check/Not to be used as a diagnostic device	Spot Check/Not to be used as a diagnostic device.	Same

	Subject Device	Predicate Device	Comparison
Hardware Inputs	12 Lead ECG from the following compatible devices: A B C D	12 Lead ECG from the following compatible devices: A B	Similar. While both require inputs from a 12 Lead ECG, the subject device allows for more compatible ECG input devices.
Output	The software provides the following outputs: 1. Presence of Disease X. Seek further clinical evaluation to establish a diagnosis of Disease X. 2. Presence of Disease X not likely. However, please use clinical judgment and determine if further evaluation is necessary. 3. Error Message: The 12 lead ECG does not pass the quality checks in place.	Software provides an output on the possibility of Disease X and if further evaluation is needed.	Similar. Both devices identify if there is presence of Disease X and whether further evaluation is needed. Both identify that it should not be used as a stand-alone and clinical judgment should be used if further evaluation is needed for diagnosis of Disease X.
Ground Truth for Model Training	Echocardiogram	Echocardiogram	Same

	Subject Device	Predicate Device	Comparison
Performance	Sensitivity: 87% (83%, 89%) Specificity: 83% (81%, 85%) Positive Predictive Value (PPV): 56%	Sensitivity: 82% (78%, 85%) Specificity: 81% (79%, 84%) Positive Predictive Value (PPV): 53%	Similar. The subject device has better performance than the predicate device in sensitivity, specificity and PPV.

Model Training Description:

The model was trained from a dataset independent from the test dataset. The model was trained with 30,000 patients that received an ECG and echocardiogram performed within 30 days apart from one another. The echocardiogram was used to establish clinical reference standard (ground truth) in patients. The dataset was collected from clinical databases from 2 diverse hospital networks (Hospital A and Hospital B). Disease X was defined as patients who had a left ventricular wall thickness >= 15 mm based on echocardiographic imaging.
The training dataset contained the following demographic breakdown that was representative of the disease population:

Race	Percentage (%)
White	75.5
Black or African American	13.6
American Indian or Alaska Native	1.3
Asian	6.3
Native Hawaiian or Pacific Islander	0.3
Two or More Races	3.0

The sample consisted of 49.5% male and 50.5% female participants. The average age was 62 years with the following age breakdown below:

Age (years)	Percentage (%)
Under the age of 40	10
40-49	10
50-59	25
60-69	30
70-79	15
Greater than the age of 79	10

Patients with Disease X were 20% of the overall cohort while patients without Disease X (control group) consisted of 80% of the overall cohort. Both groups were split into training (50%), tuning (20%) and tuning evaluation (30%) datasets. The sensitivity and specificity of the model were calculated from the tuning evaluation datasets. The model was able to achieve the following:

- Sensitivity: 87% (83%, 89%)
- Specificity: 83% (81%, 85%)
- Positive Predictive Value (PPV): 56%

Summary of Non-Clinical Performance Data

The model was evaluated taking into account applicable requirements of the FD&C Act and implementing regulations. This included the following testing:

- Human Factors and Usability testing was conducted and documentation was provided as recommended in FDA's guidance document "Applying Human Factors and Usability Engineering to Medical Devices."
- Cybersecurity testing was conducted and documentation was provided as recommended in FDA's guidance document "Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions."
- Software verification and validation testing was conducted and documentation was provided as recommended in the Premarket Software Guidance.

Summary of Clinical Validation:

Study Design

The model was validated in a retrospective study of 25,000 patients and their patient records across 5 different and diverse health systems across the United States. The objective of the study was to establish the performance of the model on screening for the presence of disease X. The inclusion criteria for the model were the following:

- The patients enrolled in the study were greater than the age of 22 with at least one 12-lead resting ECG and an echocardiogram within 30 days following the date of the ECG. The most recent echocardiogram was paired with the most recent ECG for that patient prior to the echocardiogram.
- The following models of ECG devices (A, B, C and D) were used to collect the 12-lead resting ECG data and used as the inputs to the model.
    - The 12-lead ECG duration must be 10 seconds long.

The exclusion criteria for the model were the following:

- The patients enrolled in the study were less than 22 years old.
- Mandatory data were missing (i.e., technical parameters of ECG, age or race demographic, information regarding the conducted ECG and echocardiogram).
- Different device models of 12-lead ECGs were used to collect the ECG data.
- The 12-lead ECG duration is not 10 seconds long.
- The patient has a pacemaker.

Each of the 5 sites contributed around 5,000 patient-ECG pairs to a final pool of 25,000 patient-ECG pairs. The study sample had the following demographic breakdown that was representative of the disease population.

Race	Percentage (%)
White	75.5
Black or African American	13.6
American Indian or Alaska Native	1.3
Asian	6.3
Native Hawaiian or Pacific Islander	0.3
Two or More Races	3.0

The study sample had the following hospital site breakdown:

Hospital Sites	Percentage (%)
A	19.64
B	21.36
C	20.1
D	18.4
E	21.5

The sample consisted of 49.5% male and 50.5% female participants. The average age was 65 years with the following age breakdown below:

Age (years)	Percentage (%)
Under the age of 40	10
40-49	10
50-59	16
60-69	23
70-79	22
Greater than the age of 79	19

The study sample ECG pairs were collected by the following ECG acquisition devices. The breakdown can be found below:

ECG Acquisition Device	Percentage (%)
A	26.6
B	25.1
C	24.9
D	23.3

Primary Endpoints

The co-primary endpoints regarding this study were to have the lower limits of their 95% two-sided confidence intervals be:

- Sensitivity: 75% or higher
- PPV: 50% or higher

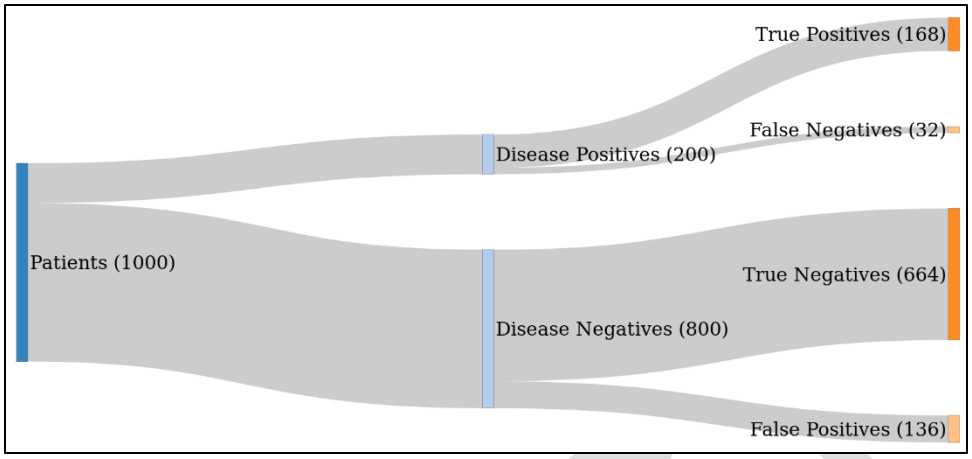
Study Results

The model achieved a sensitivity of 84%, a specificity of 83%, a PPV of 55%, and a negative predictive value (NPV) of 95%. Both the point estimates and their 95% two-sided confidence intervals, along with the confusion matrix, can be reported in a table as shown in the following example.

	Ref. Pos	**Ref. Neg**	**Sum**	**Likelihood Ratio**	**Performance**
Test. Pos	4200	3400	7600	4.9 (4.8, 5.1)	PPV= 55.3% (4.9%,654.1%, 56.4%)
Test. Neg	800	16600	17400	0.2 (0.2, 0.2)	NPV=95.4% (95.1% 95.7%)
Sum	5000	20000	2500	1 (1,1)	Prevalence 20% (19.5%, 20.5%)
Performance	Sensitivity = 84% (82.9%, 85%)	Specificity = 83% (82.5%, 83.5%)			

Plain Language to Interpret the Study Results for Benefit Risk Consideration

Assume the prevalence of Disease X in the intended use population of the device is 20%. Among 1000 patients from the target population, about 168 (1000 × Prevalence × Sensitivity) patients will be correctly classified as having the Disease X (i.e., 168 device true positives out of 200 total reference positive patients), while about 136 (1000 × (1 - Prevalence) × (1 - Specificity)) patients will be wrongly classified as having the Disease X (i.e., 136 device false positives out of 800 total reference negative patients). Thus, each true positive patient comes at the cost of 0.8 (136/168) false positive patients (compares to Y from the standard of care, or 4 (800/200) from a worst-case scenario where every patient is called positive). Furthermore, to identify one extra true positive patient, we need to assess about two patients (considering potential device positive/negative outcomes) since NNP (Number Needed to Predict) = 1 / (PPV + NPV - 1) = 1.97 (compares to the standard of care with NNP of Z, or a perfect device with a NNP of one).



The subgroup analysis for each demographic can be found below.

Please note that while confidence intervals could not be generated for this fictitious example, sponsors should include confidence intervals on all reported results. Placeholders have been included in each cell to represent the confidence interval: (X_{ll}, X_{ul}), where "ll" stands for lower limit and "ul" stands for upper limit.

Race	**Percentage (%)**	**Sensitivity**	**PPV**
White	75.5	85.3 (X_{ll}, X_{ul})	57.3% (X_{ll}, X_{ul})
Black or African American	13.6	82.9 (X_{ll}, X_{ul})	54.4% (X_{ll}, X_{ul})
American Indian and Alaska Native	1.3	81.6 (X_{ll}, X_{ul})	54.8% (X_{ll}, X_{ul})
Asian	6.3	83.9 (X_{ll}, X_{ul})	56.1% (X_{ll}, X_{ul})
Native Hawaiian and Other Pacific Islander alone	.3	83.6 (X_{ll}, X_{ul})	56.5% (X_{ll}, X_{ul})
Two or More Races	3	84.1 (X_{ll}, X_{ul})	55.4% (X_{ll}, X_{ul})
Age	**Percentage (%)**	**Sensitivity**	**PPV**
Under the age of 40	10	84.9 (X_{ll}, X_{ul})	55% (X_{ll}, X_{ul})
40-49	10	85.1 (X_{ll}, X_{ul})	55.4% (X_{ll}, X_{ul})
50-59	16	84.1 (X_{ll}, X_{ul})	55.4% (X_{ll}, X_{ul})
60-69	23	84.5 (X_{ll}, X_{ul})	56% (X_{ll}, X_{ul})
70-79	22	83.6 (X_{ll}, X_{ul})	55.4% (X_{ll}, X_{ul})
Greater than the age of 79	19	82.1 (X_{ll}, X_{ul})	52.7% (X_{ll}, X_{ul})

The subgroup analysis for each ECG acquisition device can be found below:

ECG Acquisition Device	Percentage (%)	Sensitivity	PPV
A	26.6	84.7 *(X_{ll}, X_{ul})*	56.5% *(X_{ll}, X_{ul})*
B	25.1	83.6 *(X_{ll}, X_{ul})*	54.3% *(X_{ll}, X_{ul})*
C	24.9	85.4 *(X_{ll}, X_{ul})*	57.9% *(X_{ll}, X_{ul})*
D	23.3	84.6 *(X_{ll}, X_{ul})*	55.1% *(X_{ll}, X_{ul})*

The subgroup analysis for each hospital site can be found below:

Hospital Sites	Percentage (%)	Sensitivity	PPV
A	19.64	83.6 *(X_{ll}, X_{ul})*	54.3% *(X_{ll}, X_{ul})*
B	21.36	85.1 *(X_{ll}, X_{ul})*	51.4% *(X_{ll}, X_{ul})*
C	20.1	84.1 *(X_{ll}, X_{ul})*	55.4% *(X_{ll}, X_{ul})*
D	18.4	85.4 *(X_{ll}, X_{ul})*	57.9% *(X_{ll}, X_{ul})*
E	21.5	84.7 *(X_{ll}, X_{ul})*	56.5% *(X_{ll}, X_{ul})*

Model Card:

Device Information

- Model Name: Disease X Screening Model
- Model version: version 1.0.1
- Model release date: December 2023
- Model architecture: Convolutional Neural Network

Device Description

- Intended User: Healthcare professionals.
- Indications for Use: The model is software intended to aid in screening for Disease X on patients above the age of 22 by analyzing recordings of 12-lead ECG made on compatible ECG devices. It is not intended to be a stand-alone diagnostic device for Disease X. However, a positive result may suggest the need for further clinical evaluation in order to establish a diagnosis of Disease X. If the patient is at high risk for Disease X, a negative result should not rule out further non-invasive evaluation. It should not be used to replace the current standard of care methods for diagnosis of Disease X but applied jointly with clinician judgment.

- Clinical workflow phases: To be used as an aid and screening tool for further clinical follow-up (e.g., echocardiogram) in order to establish a diagnosis of Disease X.
- Clinical Benefit: To provide point-of-care screening of Disease X where cardiac imaging may not be available.

Performance and Limitations

- Data type: 12-lead electrocardiogram (ECG)
  - Description: 10 second duration of a 12-lead electrocardiogram (ECG) obtained from the following four compatible ECG devices (A, B, C, and D). The compatible ECG devices have a sampling rate of 500 Hz.
- Clinical Reference Standard: An echocardiogram obtained within 30 days of the ECG to establish clinical reference standard.
- Model Validation:
  - Data size and type: A retrospective study of 25,000 patients and their patient records across 5 different and diverse health systems across the United States. Each of the 5 sites contributed 5,000 patient-ECG pairs to a final pool of 25,000 patient-ECG pairs.
  - Exclusion Criteria:
    - The patients enrolled in the study were less than 22 years old.
    - Mandatory data were missing (i.e., technical parameters of ECG, age or race demographic, information regarding the conducted ECG and echocardiogram).
    - ECG data contained either corrupt or missing lead(s).
    - Different models of 12-lead ECGs were used to collect the ECG data.
    - The 12-lead ECG duration is not 10 seconds long.
    - The patient has a pacemaker.
  - Data Results (calculated from test datasets):
    - Sensitivity: 84% (82.9%, 85%)
    - Specificity: 83% (82.5%, 83.5%)
    - PPV: 55.3% (54.1%, 56.4%)
    - NPV: 95.4% (95.1%, 95.7%)
- Non-Clinical Testing:
  - Human Factors and Usability testing was conducted and documentation was provided as recommended in FDA's guidance document "Applying Human Factors and Usability Engineering to Medical Devices."
  - Cybersecurity testing was conducted and documentation was provided as recommended in FDA's guidance document "Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions."
  - Software verification and validation testing was conducted and documentation was provided as recommended in the Premarket Software Guidance.

Risk Management:

Risk management was conducted, and documentation was provided as recommend in the Premarket Software Guidance and in accordance with ANSI/AAMI/ISO 14971 Medical devices - Applications of risk management to medical devices.

- Potential risks associated with the model, the data, and the outputs (e.g., contraindications, side effects, data privacy risks, cybersecurity risks, bias risks, information gaps): The potential risks associated with the model include incorrect follow-up due to a false positive or false negative output, which can occur because of (1) model bias or (2) using the model in an unsupported patient population or with unsupported input/hardware. Furthermore, information gaps may lead to overreliance on the device output for follow-up. Controls for identified risks include clinical validation testing, software verification and validation testing, human factors testing and labeling.
- Description of information that could impact risks and patient outcomes, across the product lifecycle: Model development and clinical validation included only 10% of participants under the age of 40, which may mean that the model's performance on that subgroup is not fully characterized.
- Interactions, Deployment, and Updates: A comprehensive Device Performance Monitoring Plan is in place that is consistent with the Quality System Regulation (21 CFR Part 820) which continuously monitors the deployed model to evaluate site-specific performance, identify vulnerabilities, and ensure transparency of performance and ongoing maintenance to sites and end users.
    - Computational resources required.
    - Details regarding how the model is deployed and updated:
        - How to conduct local site-specific acceptance testing or validation: Prior to use of the model in the site's entire population, the model is deployed, and data is collected for a one-month period in order to understand any issues with integration into the sites' existing systems and measure performance on a subset of the patient population for that site. Through this process, issues with deployment can be addressed prior to exposure to the entire population and can help characterize performance of the model and the need for additional training and development. Alternatively, sites may opt to provide historical data that can be used to assess expected performance at the site.
        - Ongoing performance monitoring: Automated performance calculation is deployed along with the model and calculated every 6 months; if the performance is out of the expected range, an automated e-mail will be sent to the site administrator and sponsor. This will initiate a process for understanding performance issues and a mitigation plan will be put in place to address this.
        - Transparent reporting of successes and failures: All sites will have access to anonymized reports that will include successes and failures of deployed models at various sites, along with site characteristics to contextualize these successes and failures.
        - Change management strategies: Change management will be implemented

consistent with established Quality System procedures if and when issues arise that require a change or if features are requested by sites and users.
        - Proactive approaches to address vulnerabilities: Sites and users are encouraged to report any issues within 48 hours of the issue occurring, which will then follow complaint handling procedures and for which a fix will be issued according to these procedures.
    - Communication to parties of as-needed information: Automated e-mails will be generated by the device when performance is out of the expected range, as described above.
    - Software quality (specify, standards and regulatory compliance issues, intellectual property issues, risk management and safeguards used, other):

Development:

- Model Training:
    - Data size: 30,000 patients that received an ECG and echocardiogram performed within 30 days apart from one another. Dataset collected from clinical databases from 2 diverse hospital networks (Hospital A and Hospital B).
    - Patients with Disease X were 20% of the overall cohort while patients without Disease X (control group) consisted of 80% of the overall cohort. Both groups were split into training (50%), tuning (20%) and tuning evaluation (30%) datasets.
    - Data Results (calculated from tuning evaluation datasets):
        - Sensitivity: 87% (83%, 89%)
        - Specificity: 83% (81%, 85%)

Conclusion:

While there are differences noted in the technological characteristics of the proposed system and the predicate device, the differences do not raise different questions of safety or effectiveness. Based on the information provided in this submission, the subject device demonstrates that it is substantially equivalent to the predicate device through the results of clinical performance and results of non-clinical verification and validation.