NISKANEN CENTER

Regulatory Comment

Comments submitted to the Food and Drug Administration in the Matter of:

SOFTWARE PRECERTIFICATION PROGRAM (v0.2)

Dr. Anastasia Greenberg
Technology Policy Fellow
Niskanen Center

Ryan Hagemann
Senior Director for Policy
Niskanen Center

Submitted: July 18, 2018
Docket Number: FDA-2017-N-4301-0001

EXECUTIVE SUMMARY

Ongoing developments in Artificial Intelligence (AI) hold the potential to revolutionize the U.S. health care market by offering more accurate, more personalized, and cheaper diagnostic solutions. The results of these improvements will be enhanced patient care outcomes, alleviation of growing financial pressure on the domestic health care system, and more personalized and efficacious treatment options for patients. Unfortunately, there is currently a gap in the regulatory approval process for AI-powered software-based medical devices. Although Food and Drug Administration (FDA) has taken an important first step in plugging this gap with the release of *Developing Software Precertification Program: A Working Model (v0.2)*, there is much that can be done to improve on the agency's proposal.

To that end, these comments offer feedback and specific recommendations based on the most recent version of the FDA's proposed Software Precertification Pilot Program (Pre-Cert Program). The FDA's current framework for addressing the needs of organization and innovators developing AI-based medical and diagnostic tools is a commendable step in the right direction. In order to fully realize the potential future gains of this technology's application to the medical marketplace, however, there are a number of significant improvements the agency can incorporate into the next version of its Pre-Cert Program. To that end, these comments offer 12 recommendations for the use of AI in medical technologies that can help ensure a strong, flexible, and adaptive regulatory framework that will usher American health care innovation into the 21st century and beyond.

INTRODUCTION

Investment in artificial intelligence (AI) in the health care space has recently exploded, reaching an impressive $794 million in 2016, and is projected to continue growing exponentially.[i] This excitement over the potential of AI in health care is not unfounded: AI can revolutionize the U.S. health care market by offering more accurate, more personalized, and cheaper diagnostic solutions, improving patient care outcomes and alleviating growing financial pressures on the system. These potential benefits warrant increased focus and action from members of Congress and regulators at the Food and Drug Administration to provide clear guidance on the use of AI in medical devices. In order to facilitate ongoing investment and development of these life-enhancing technologies, policymakers need to send clear signals to the private sector. To this effect, the FDA has recently announced its Digital Health Innovation Action Plan, much of which is focused on the new Precertification Program [hereafter "Pre-Cert Program"] for developers working in the Software as a Medical Device (SaMD) arena. As FDA Commissioner Scott Gottlieb has stated: "We envision and seek to develop through the Pre-Cert for Software Pilot a new and pragmatic approach to digital health technology."[ii]

The Pre-Cert Program flips the traditional FDA regulatory pathway on its head by focusing on vetting the software developer, rather than reviewing each SaMD product on its own. Ideally, this would allow for a faster path to market for key AI devices and permit more comprehensive collection of real-world data post-device launch — data that is critical for successful updating and modification of AI-based software. At its root, the goal of the Pre-Cert program is to provide a more flexible and iterative regulatory approach that is much needed for dealing with emerging AI technologies. The program intends to allow precertified organizations to benefit from exemptions to premarket review for certain SaMD products, while providing for streamlined review for other types of SaMDs. The FDA's most recent working model for the program — *Developing Software Precertification Program: A Working Model (v0.2)*[iii] [hereafter *A Working Model v0.2*] — makes explicit reference to SaMDs that "use artificial intelligence and machine learning algorithms." Unfortunately, however, the agency offers no specific recommendations or standards for AI-enabled medical devices, leaving innovators and developers in the dark as to the FDA's thinking on these matters.

In order to address this regulatory gap, the following comments offer feedback and specific recommendations based on *A Working Model v0.2* that provide the FDA with a framework for considering how best to address the needs of organizations and innovators developing AI-based SaMDs. The comments focus on all components of the Pre-Cert Program, organized as follows: eligibility and excellence appraisal, review pathway determination, streamlined premarket review process, and real-world performance analytics.

ELIGIBILITY AND EXCELLENCE APPRAISAL

Recommendation 1: Ensure grant of Pre-Cert status to a diverse portfolio of organizations for the pilot program and beyond.

The first step to evaluating an organization's application for precertification is determining eligibility for the program. *A Working Model v0.2* lays out a very broad scope of eligibility: "Any organization that intends to develop or market regulated software in the United States would be considered in-scope for the Software Precertification Program."[iv] We support the FDA's decision to create a broad scope of eligibility, irrespective of an organization's size; both small and large companies are recognized as potentially eligible participants. We encourage the FDA to build on this commitment to equal opportunity for diverse organizations by building a pool of participants from across the spectrum of organizational structure and size as the pilot program grows and into the initial phases of the Pre-Cert Program and beyond.

Recommendation 2: Create a scoring sheet and minimum thresholds for identifying whether each of the excellence principles principles are met, based on demonstrated elements within each organizational domain.

The next step in the precertification process is to evaluate whether a given organization meets the five excellence principles: (1) Product Quality, (2) Patient Safety, (3) Clinical Responsibility, (4) Cybersecurity Responsibility, and (5) Proactive Culture. In Appendix B of *A Working Model v0.2*, the agency lays out a chart specifying proposed organizational domains and elements within each domain that map onto the five excellence principles. The idea is that organizations will discuss in their appraisal application how their organization meets these various elements and which Key Performance Indicators (KPIs) they use to measure each element. The appraisal process is meant to be flexible to allow different types of organizations to demonstrate excellence in their own ways.

While we commend the FDA's commitment to flexibility, we also believe that a minimum level of standardization creates certainty and due process during the appraisal review. The current framework, as laid out in the chart, lacks such certainty. The current mapping of a given element onto a variable number of excellence principles creates confusion for determining whether an organization meets any one, or all, of the excellence principles. For example, the element "developing and maintaining access to highly skilled employees with relevant/applicable clinical knowledge"[v] maps onto all five excellence principles, while the element "buyers and users (patients/physicians) understand expected or minimum support lifetimes and levels"[vi] maps only onto Clinical Responsibility. Furthermore, while many elements in the chart support the Product Quality excellence principle (38 elements), a smaller number of elements support the Cybersecurity Responsibility excellence principle (26 elements). Therefore, it is not clear how many elements, and from which domains, would need to be met in order to satisfy any given excellence principle. It is also not clear whether it is easier to satisfy certain excellence principles by virtue of having a smaller number of mapped elements.

We recommend that the FDA create a tally of the total number of elements within each organizational domain that meet a given excellence principle to facilitate scoring of an organization's excellence across all five principles. For example, under the organizational domain of "Design and Development," there are eight elements. Out of these elements, six meet the Patient Safety principle, Product Quality principle, and the Clinical Responsibility principle, while three meet the Cybersecurity Responsibility principle and five meet the Proactive Culture principle. By scoring an organization on a proportion of principles met within each organizational domain, specific organizational weaknesses can be easily identified, allowing for clear feedback to be provided to the organization during the iterative appraisal process. This scoring method would allow for both standardization in appraisal evaluation, creating more certainty in the process, and a simple method for effectively communicating areas of improvement to the organization. Furthermore, the FDA can identify a minimum threshold to be met for successful appraisal of an organization, taking into account the most important organizational domains and giving an equal weighting to each excellence principle.

Recommendation 3: Include example KPIs for each element used to appraise organizational excellence.

Another issue with the current excellence principles chart is the use of vaguely stated elements, lacking clarity as to how an organization can successfully demonstrate its commitment to each element. We recommend that the FDA include an additional column to the chart to provide sample KPIs that an organization could use to measure each element. For example, the domain "Leadership and Organizational

Support" lists the element "empowering staff to act regarding the decisions or issues impacting users, products, or patient safety." This element is stated in abstract form and it is unclear how to measure the construct "empowering staff." Example KPIs that could be provided for this element might include: (1) ***frequency of weekly meetings between management and staff to discuss and make decisions pertaining to product issues***, and (2) ***level of decision-making ability afforded to lower-ranking staff that does not require management consultation***. These two KPIs would provide proxy measures of how well the organizational leadership "empowers" staff to take actions pertaining to product issues and patient safety. Such KPIs should be provided for each element within each domain in the chart.

Recommendation 4: Include an additional element under the Deployment and Maintenance domain of the excellence appraisal chart to evaluate an organization's ability to collect and analyze post-launch, real-world performance data.

We would suggest additional domains/elements critical to the excellence appraisal be included in the chart. Given that *A Working Model v0.2* emphasizes the critical importance of real-world performance indicators following precertification and SaMD launch, it would be sensible to include a relevant element within the excellence appraisal process itself. As *A Working Model v0.2* states: "In addition to demonstrating excellence, as established through the five excellence principles, precertified organizations would also have a robust mechanism to collect, monitor, and analyze real-world performance of their organization and the products they deliver."[vii] It is unclear how the FDA intends to conduct an excellence appraisal based on Appendix B, as well as how the organization's commitment to real-world performance analysis will be assessed, which will become particularly relevant following precertification. In order to provide clarity to participating organizations, we recommend simplifying the approach by integrating an organization's commitment to real-world performance analysis right into the initial excellence appraisal.

For example, an additional element could be added to the existing organizational domain of "Deployment and Maintenance," which might read: **Ability to collect and analyze real-world performance data following SaMD deployment.** KPIs for this element could include those that measure what mechanisms are integrated into SaMD software to collect and transmit relevant data back to the organization, including adverse-event data; the number of data scientists qualified to evaluate real-world performance data; and so forth.

Recommendation 5: Include a new organizational domain in the excellence appraisal chart called "Artificial Intelligence Development Best Practices" with appropriate elements and KPIs.

Specific to AI-based SaMDs, *A Working Model v0.2* requests input on "elements or domains critical to evaluating the development of software functions using artificial intelligence and machine learning algorithms."[viii]

To that end, we propose the FDA create a new organizational domain — **Artificial Intelligence Development Best Practices** — to be included in the excellence appraisal chart. This domain would focus exclusively on best practices for AI software development. Separation would allow for this entire domain to be crossed off as "Not Applicable" for appraisals of organizations that do not intend to develop AI-based SaMDs. Appendix A provides a recommended list of elements and sample KPIs that could be included under this new domain, based on prevailing best practices for AI software development.[ix] The sample KPIs are given for a single hypothetical SaMD that uses skin biopsy images as the input data for machine learning (ML) algorithm development to detect melanoma.

Recommendation 6: Explicitly allow for outsourcing of core organizational activities in the excellence appraisal process.

An additional issue with the current excellence appraisal under *A Working Model v0.2* is the uncertainty surrounding organizational outsourcing of various activities. Many organizations, especially smaller organizations, do not perform all core organizational functions in-house. Although the FDA has explicitly noted its commitment to equal precertification opportunity for both large and small organizations and allows "companies to identify the boundaries of the organization themselves to determine the business unit…that should be considered for precertification,"[x] it is not currently clear whether smaller organizations/business units that outsource certain activities would face additional barriers in the excellence appraisal process. For example, would an organization that seeks precertification be limited to outsourcing certain core activities only to other entities that are already precertified? If so, we would urge the FDA to reconsider and explicitly permit outsourcing of activities to a broader cohort of actors in order to prevent unnecessary barriers to entry for smaller organizations.

We would further recommend that organizations be permitted to identify outsourced activities within the appraisal application and provide the same KPIs for each outsourced element, just as they would if the activities were performed in-house. Relevant records could be obtained by the organization seeking appraisal from the outsourced entity and provided as evidence to support excellence principles.

Recommendation 7: Develop an excellence appraisal process template and example including estimated timelines for each application stage.

Finally, we strongly encourage the FDA to provide a developed precertification application template with a successful application example. The successful application example could be developed based on actual ongoing precertification pilot processes with all organizational proprietary information removed/altered, or a fictional example. The template should also include timeline estimates for each precertification application stage based on an expected average number of reiterations within the interactive engagement process. These timeline estimates are invaluable for allowing organizations to make strategic business decisions as to which regulatory pathway best meets their needs.

DETERMINING PRECERTIFICATION LEVEL

Recommendation 8: Abolish two levels of precertification and adopt a single precertification standard.

A Working Model v.02 attempts to make a clarification from the previous version (v0.1) regarding precertification level of organizations with or without track records in SaMD development. Unfortunately, the updated language actually creates more confusion as to the eligibility for the two levels of precertification. Previously, only organizations that had demonstrated an excellence track record with respect to SaMD products were considered eligible for Level 2 precertification, while organizations without a track record in SaMD development would only be eligible for Level 1 precertification. In the updated version, the FDA states:

> *Our current thinking reflects the belief that an organization of any **size without a medical device or SaMD currently on the market** should have the opportunity to deliver products for medical purposes as a precertified organization. We believe organizations that have objectively demonstrated excellence*

> *in product-development elements in all five excellence principles and have successfully marketed and maintained products can achieve Level 2 Pre-Cert*. (emphasis added)[xi]

This would suggest that a track record specifically in SaMD products is no longer a differentiating factor for the determination of precertification level. However, *A Working Model v0.2* indicates that Level 1 certification "would be awarded to an organization that has objectively demonstrated excellence in product development in all five excellence principles, with a limited track record in developing, delivering and maintaining *products in the health care space*" (emphasis added).[xii] This seems to imply that a track record in SaMD still remains the essential distinguishing factor for a Level 1 precertification determination. Level 2 precertification, the document declares, would be awarded to "an organization that has objectively demonstrated excellence in product development in all five excellence principles, with a track record in successfully marketing and maintaining products to suggest a level of assurance in the development of safe and effective software."[xiii]

There is no mention of health care experience in the description of Level 2 certification. While Level 2 certification presumes greater trust in an organization and permits more regulatory freedom than Level 1 certification, there is no indication as to whether SaMD track record development is, or is not, necessary for differential precertification levels. As the document currently stands, counterintuitively, it appears that while a limited track record of SaMD excellence is required for Level 1 precertification, no such health care specific track record is now required for Level 2 precertification.

Unclear eligibility standards aside, the agency offers no explanation as to why two distinct precertification levels would be necessary or beneficial. As currently constructed, the bifurcation of precertification levels simply serves as a source of further confusion and tension between certification processes for organizations of different sizes and prominence in the SaMD space. In order to commit to Least-Burdensome regulatory principles[xiv] and to facilitate a path to market for new innovative players in the digital health care space, we recommend that the FDA abolish the two precertification levels and focus on implementation of the program with a single, clear, and efficient path to successful precertification.

REVIEW PATHWAY DETERMINATION

Recommendation 9: Simplify the SaMD risk categorization framework by adopting a single formula yielding a single number that is categorized into three risk levels: low, moderate, and high.

A Working Model v0.2 envisions a framework where precertified organizations would be permitted to market low-risk SaMDs without further regulatory approval from the FDA, while requiring high-risk SaMDs to undergo a Streamlined Premarket Review (SR). To determine which SaMDs require a SR, the FDA relies on the risk categorization for SaMDs developed by the International Medical Device Regulators Forum (IMDRF).[xv] The IMDRF establishes SaMD risk level based on the state of a health care condition that the SaMD will be intended for and the significance of information provided by the SaMD. This framework consists of three levels of "significance of information:" (1) treat or diagnose, (2) drive clinical management, and (3) inform clinical management, and three levels of "state of health care condition": (1) critical, (2) serious, and (3) nonserious. A matrix of these three-by-three possibilities creates nine subtypes of risk, which are further grouped by the IMDRF into four risk types. *A Working Model v0.2* includes a chart to show which of the nine SaMD subtypes of risk would require a SR, based on product development stage and precertification level. This creates a very complex and burdensome process for an organization to determine whether any given SaMD product at any given development stage requires a SR.

The current risk categorization framework is unnecessarily complex and based on arbitrary and theoretical conjectures of how risky an SaMD product is expected to be. From a practical standpoint, many AI-based SaMDs would not easily fit into such a framework. For example, if an AI SaMD is developed to process and analyze skin biopsy images, would the device be considered to diagnose a skin condition such as melanoma or to drive clinical management? Such an SaMD would most certainly be trained on data that was labeled as either positive or negative for melanoma. When providing an output as to whether melanoma is likely present or not, it could be seen as *driving* clinical management by allowing a physician to examine the biopsy further and to send the patient for extra testing, or it could be seen as *diagnosing* the condition outright. In either case, such an approach is highly subjective, and fails to provide clarity for applicants seeking an understanding of what distinguishes *diagnosing* from *driving* in the context of AI-based SaMDs.

The problems with this risk categorization framework are further exacerbated when considering this same example in light of the "state of health care condition" variable. Is an AI-based SaMD that was trained on skin biopsy images considered critical, serious, or nonserious? If an AI-based SaMD were developed to detect early signs of problematic skin coloration, with a high rate of successful and risk-free treatment options (e.g., skin removal under local anesthetic), this could easily be considered nonserious under the condition category. On the other hand, if such an AI-based SaMD were trained on a large dataset of skin biopsy images and was sensitive enough to detect nonserious skin coloration issues, it would most certainly also detect much rarer serious or even critical conditions such as advanced stage melanoma. Therefore, given the power of ML models to find nuanced and complex patterns in data without explicit rule-based programming, it would be very difficult to easily fit such an SaMD into the imagined nine SaMD subtypes. It also appears that the current risk categorization framework would provide an unjustified tendency to consistently categorize AI-based SaMDs into high risk categories because of the power of AI to provide health analytics for serious and critical health conditions, without taking into account that a high level of accuracy would mitigate the overall device risk and actually reduce severe patient outcomes.

In an attempt to resolve these issues, we propose a simplified framework for SaMD risk categorization that would be generalizable to all SaMD product types. We propose a simple formula for calculating an SaMD risk which would yield a single risk number. This risk number could then be categorized into three levels: **low**-, **moderate**-, and **high-risk**. For a given SaMD, the formula takes into account the severity of the most serious health condition that could be detected by the device, the proportion of that health condition in the representative population of device users, and the error rate of the SaMD. This formula would yield a number between 0 and 1: 0 representing the lowest level of SaMD risk and 1 indicating a maximal level of SaMD risk. Therefore, three risk levels could be obtained by splitting scores into three bins; for example: 0 to 0.3 indicating low risk, 0.3 to 0.5 indicating moderate risk, and 0.5 to 1 indicating high risk. The proposed formula is as follows:

> **SaMD risk** = (Severity of most serious health condition + Proportion in user population + Error rate) / 3

Severity of most serious health condition: Value between 0 and 1: 0 represents no health consequence and 1 represents potentially lethal condition.

- *Example*: An AI-based SaMD was trained on skin biopsy images to detect skin abnormalities. It is able to detect anything from harmless skin discoloration to melanoma (the latter being the most serious condition that the SaMD is able to detect). The melanoma survival rate in the U.S. is very high for early stage detection (99 percent) and moderate for the latest stages of detection (40 percent).[xvi] Therefore, the severity will be approximated as 0.7 out of 1.

Proportion in user population: Value between 0 and 1: 0 represents inexistence of the most severe health condition in the user population, and 1 represents all users having the given condition.

- *Example*: The most serious health condition that an SaMD could detect is melanoma. The device will be used by physicians on the general U.S. population. The rate of melanoma is the general U.S. population is 2.3 percent. Therefore, as a proportion this equals 0.023.

Error rate: Value between 0 and 1: 0 represents perfect accuracy of the SaMD in detecting/classifying the health condition and 1 represents complete mismatch in classification/detection.

- *Example*: The SaMD was trained on skin biopsy images and the final classifier has an F1 (accuracy) score of 0.95. Therefore, we convert the accuracy score into an error score (1-0.95) to get an error score of 0.05.

When plugging in these sample numbers into the formula we get:

SaMD risk = (**0.7** + **0.023** + **0.05**) / 3

= 0.26 → low risk SaMD

Based on our proposed risk categorization (0 to 0.3 → low-risk, 0.3 to 0.5 → moderate-risk, 0.5 to 1 → high-risk), this example would fall into the low-risk category. Even though the hypothetical SaMD can detect melanoma, a potentially lethal condition, the high accuracy of the AI-based SaMD helps to counterbalance the overall device risk. Note that in our formula, all three variables are given an equal weighting: severity, proportion in user population, and error rate. This means that even if an SaMD is designed only to detect a low-risk health condition, if the error rate of the SaMD is high, the final risk categorization would also be higher than otherwise expected. Furthermore, all things being equal, the greater the total number of people who could be diagnosed with a severe condition, the higher the total SaMD risk. The formula also accounts for the vulnerability of the user population. If an SaMD is intended for use on patients who are already predisposed to the health condition in question, this is reflected in a higher relative value for "proportion of user population with the condition," thereby increasing the overall SaMD risk.

This simple formula allows for simplification, systemization, and standardization of risk categorization that provides a fair risk assessment for all types of SaMDs without an unfair disadvantage for AI-based diagnostic devices.

STREAMLINE PREMARKET REVIEW PROCESS

Recommendation 10: Require precertified organizations to proceed to Streamlined Review only for high-risk SaMDs.

Currently, *A Working Model v0.2* includes a chart describing which SaMDs would require a SR depending on the SaMD subtype (based on the IMDRF risk categorization) and on the organization's precertification level. Given our recommendations for simplifying the FDA's current risk categorization and precertification approach, this chart would no longer be required. Instead, we recommend that based on our three risk categories, only high-risk SaMDs should require SR. This would create a much simpler process for organizations to determine when a SR is necessary. For each new SaMD, the organization would simply apply the risk formula above and those devices that fall within the high-risk classification would proceed to SR.

Recommendation 11: Clarify the Streamlined Premarket Review process requirements.

We also recommend that the FDA clarify the SR process itself. For example, under the traditional regulatory pathway, a request can be submitted for Breakthrough Device Designation which allows for "priority review, flexible clinical design, and an iterative review process …"[xvii] This description is reminiscent of the SR process described in *A Working Model v0.2*. The FDA should clarify whether the SR process is identical to the existing Breakthrough Device Designation. If the two are identical, the agency should note as much; if they are intended to be distinct, the agency should explain what differentiates the two processes.

Additionally, *A Working Model v0.2* specifies that for those SaMDs that require SR, "the current review standard requires that a new device be substantially equivalent to a legally marketed predicate device,"[xviii] in reference to the 510(k) process under the Federal Food, Drug, and Cosmetic (FD&C) Act.[xix] The FDA intends to streamline the 510(k) review under the SR process for precertified organizations by reducing the number of required 510(k) documentation and focusing instead on "essential" documentation. We would request the FDA clarify which specific 510(k) documentation would be considered "essential" under the SR process and which documentation would no longer be required. In a similar vein, according to the FD&C Act, devices that do not fall under the 510(k) process could be evaluated under the De Novo process.[xx] Although this possibility is not mentioned in *A Working Model v0.2*, we recommend the FDA allow SR to include both a simplified 510(k) and De Novo option for precertified organizations.

REAL-WORLD PERFORMANCE ANALYTICS

Recommendation 12: Include a domain that measures dynamic physician reliance on SaMD tools under User Experience Analytics for Real-World Performance Analytics assessments.

A major goal of the Pre-Cert Program is to shift some of the safety and effectiveness verifications away from the premarketing stages of new SaMD products and onto postmarketing Real-World Performance Analytics (RWPA). But the current RWPA framework neglects an important aspect of RWPA specific to AI-based diagnostic devices.

A Working Model v0.2 lays out three analytic types for assessing RWPA: Real World Health Analytics, User Experience Analytics, and Product Performance Analytics, along with various domains under each type. Under the User Experience Analytics type, the currently proposed domains are: user satisfaction, issue resolution, user feedback channels, and user engagement. However, the User Experience Analytics domain is missing a measure of the dynamic means by which a physician relies on AI-based tools. For example, when a physician uses an AI diagnostic device that is intended to detect signs of melanoma in skin biopsy images, the physician has the option to review the skin biopsy image and either accept or reject the suggestion provided by the SaMD. A domain should be included to measure the rate over time at which physicians using the SaMD accept or reject the diagnosis/recommendation and this information should be mapped onto long-term health outcomes for the patient. There is likely to be an interaction between SaMD accuracy and physician trust of the SaMD. For example, if the SaMD provides highly accurate diagnostic outcomes, but the physician tends not to rely on its output, long-term patient outcomes would suffer. On the other hand, if a physician generally relies on the SaMD but steers away from accepting its diagnosis/recommendation for specific difficult cases, the physician's judgment may be contributing to more positive long-term patient outcomes by correctly rejecting rare SaMD misdiagnoses. This RWPA domain is key for tracking physician trust and reliance on SaMD tools over time and would provide invaluable insights into whether physicians are becoming too reliant on an AI-based SaMD, not reliant enough on an SaMD, or reliant to an optimal extent.

CONCLUSION

The FDA's Digital Health Innovation Action Plan and Pre-Cert Program hold immense promise for laying the appropriate regulatory groundwork to support burgeoning innovation in the U.S. digital health care space. However, the current iteration of the Pre-Cert Program requires significant work in order to actualize the benefits of AI-based medical device innovation. The FDA should consider these 12 recommendations in order to help build a strong and flexible regulatory scaffold that will usher American health care innovation into the 21st century and beyond.

We would like to thank the FDA for the opportunity to comment on this issue and look forward to continued engagement on this and other topics.

[i] Peter Stone et. al., *Artificial Intelligence and Life in 2030*, One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, (Stanford University, Sept. 2016), http://ai100.stanford.edu/2016-report.

[ii] "FDA Announces New Steps to Empower Consumers and Advance Digital Health Care," FDA Voice, 27 July 2017, https://blogs.fda.gov/fdavoice/?p=6161.

[iii] *Developing Software Precertification Program: A Working Model v0.2* (U.S. Food and Drug Administration, June 2018) [hereafter *A Working model (v0.2)*].

[iv] *Ibid.*, p. 13.

[v] *Ibid.*, p. 37.

[vi] *Ibid.*

[vii] *A Working model (v0.2)*, p. 10.

[viii] *Ibid.*, p. 15.

[ix] Andrew Ng, “Advice for Applying Machine Learning”, Stanford University. Retrieved Online: http://cs229.stanford.edu/materials/ML-advice.pdf.

[x] *Ibid.*, p. 13.

[xi] *Ibid.*, p. 19.

[xii] *Ibid.*

[xiii] *Ibid.*

[xiv] “Advice for Applying Machine Learning,” p. 6.

[xv] Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. *IMDRF Software as a Medical Device (SaMD) Working Group* (U.S. FDA, 18 Sept. 2014).

[xvi] “Survival Rates for Melanoma Skin Cancer, by Stage,” (American Cancer Society, 20 May 2016), http://bit.ly/2NYyEIS.

[xvii] *Breakthrough Devices Program*, Draft Guidance for Industry and Food and Drug Administration Staff, Docket No. FDA-2017-D-5966-0001 (U.S. FDA, 25 Oct. 2017), http://bit.ly/2LcDTpV.

[xviii] *A Working Model v0.2*, p. 26.

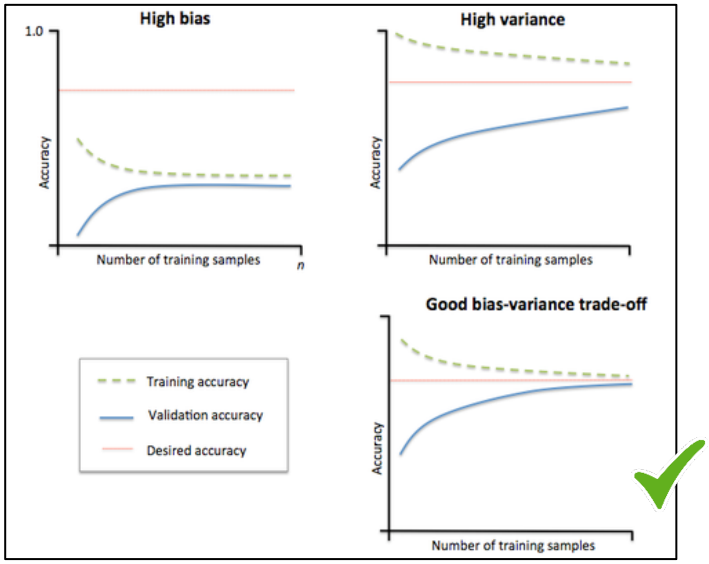
[xix] 21 C.F.R. § 807.81(a)(3).

[xx] 21 U.S.C. § 360c(a)(1) (2018), https://www.law.cornell.edu/uscode/text/21/360c; FD&C Act § 513(f)(2).

APPENDIX A:
ARTIFICIAL INTELLIGENCE DEVELOPMENT BEST PRACTICES

Organizational Domain	Elements	Excellence Principles					KPIs
		PS	PQ	ClinR	CybR	PC	
Artificial Intelligence Development Best Practices	Assessment of potential problematic machine learning (ML) features/variables that include biased or non-representative data.	X	X			X	**Descriptive statistics of all features (input variables) and output variables that were used to build the ML classifier/algorithm.** ***Melanoma Example:*** Proportion of male/female in dataset; average and standard deviation of variables such as: age, number of previous health conditions, family history of melanoma; representative example of skin biopsy image and averaged skin biopsy image; number of positive and negative melanoma examples in dataset; etc.
	A strong theoretical justification for how various patient data are related to the diagnosis/treatment of a given disease/condition to help to prevent unintended consequences of complex ML models.	X	X	X		X	**Explanation of reasons for including each feature as a variable for training the algorithm.** ***Melanoma Example:*** Age was included as a variable because likelihood of melanoma is known to increase with age. Family history of melanoma was included as a variable because there is an established heritable component of melanoma; etc.

	Ensuring generalizability of ML model to new data.	X	X	X		X	**Description of how data were split into independent training, validation, and test sets to avoid biasing the final model.** ***Melanoma Example:*** Training examples with skin biopsies are labeled as positive or negative for melanoma. The dataset is obtained across different patients and randomized to split 60 percent into the training set, while saving 20 percent for the validation set and the remaining 20 percent for the test set. During model development, the model/training parameters are only modified based on results from the validation set, and never from the test set.
	Ensuring that the model is measuring what it claims to measure.	X	X	X		X	**Description of how the initial dataset was labeled for training. This includes an explanation of how a "ground truth" was established for labeling the data in supervised models. For example, was the ground truth established based on clinical data? Was multiple observer averaging used to make a final labeling determination?** ***Melanoma Example:*** The skin biopsy images were evaluated by three radiologists for presence or absence of melanoma. Only images with 100 percent agreement across radiologists were included in the model development dataset.
	Assessment for unintended data alterations.		X			X	**Description of any preprocessing and/or data cleaning steps and verifications that these steps did not alter the data in unintended ways.** ***Melanoma Example:*** Principle Component Analysis (PCA) was performed on the skin biopsy images to reduce data dimensionality and a Gaussian filter was applied to smooth the images. Following these processing steps, a random subset of the processed images were re-evaluated by two radiologists

							for melanoma with no disagreement between the unprocessed and processed images, indicating that no unintended alterations have occurred.
	Identification of model building flaws.	X	X	X		X	**An inclusion of the learning curves during model training to allow for identification of major model building issues, such as models that show high bias (i.e., underfitting) or high variance (i.e., overfitting). High bias or high variance is an indication that the model will not generalize well to new data.** ***Melanoma Example:*** Training and validation curves were plotted for melanoma detection model development and the curves are within an acceptable bias/variance trade-off range:  **Figure 1:** Various possible learning curves obtained during machine learning algorithm development. *Source: [https://sebastianraschka.com/faq/docs/ml-solvable.html.](https://sebastianraschka.com/faq/docs/ml-solvable.html)*

	Assessment of model accuracy.	X	X	X			**Reporting of model accuracy scores based on calculations that are appropriate for a given ML situation. For example, F1 scores should be included for models trained on input data with rare events (i.e., low positive-to-negative ratio in the input data -presence of disease is very rare) to avoid misleading accuracy scores.** ***Melanoma Example:*** The dataset of skin biopsy images includes 1 percent positive melanoma cases. After model development was completed, the F1 score was calculated to be 0.95, showing a good level of model prediction accuracy.
	Commitment to model updates based on real-world performance	X	X		X		**Evidence of plans and mechanisms for obtaining new data after SaMD release for use in improvement to model accuracy and generalizability.** ***Melanoma Example:*** Mechanisms are put in place to continuously obtain skin biopsy images from new patients with radiologist establishment of presence/absence of Melanoma. Following substantial amount of new data collection, plans are in place to re-train the model and to retest improvement to model accuracy, as well as update the SaMD when accuracy and generalizability are shown to have improved.