Harmonizing Safety and Speed: A Human-Algorithm Approach to Enhance the FDA's Medical Device Clearance Policy

Mohammad Zhalechian[a], Soroush Saghafian[b], Omar Robles[c]
[a]Kelley School of Business, Indiana University, Bloomington, IN, mzhale@iu.edu
[b]Harvard Kennedy School, Harvard University, Cambridge, MA, soroush_saghafian@hks.harvard.edu
[c]Emerging Health Consulting, Armonk, NY, orobles@emerginghealthllc.com

The United States Food and Drug Administration's (FDA's) Premarket Notification 510(K) pathway allows manufacturers to gain approval for a medical device by demonstrating its substantial equivalence to another legally marketed device. However, the inherent ambiguity of this regulatory procedure has led to high recall rates for many devices cleared through this pathway. This trend has raised significant concerns regarding the efficacy of the FDA's current approach, prompting a reassessment of the 510(K) regulatory framework. In this paper, we develop a combined human-algorithm approach to assist the FDA in improving its 510(k) medical device clearance process by reducing the risk of potential recalls and the workload imposed on the FDA. We first develop machine learning methods to estimate the risk of recall of 510(k) medical devices based on the information available at the time of submission. We then propose a data-driven clearance policy that recommends acceptance, rejection, or deferral to FDA's committees for in-depth evaluation. We conduct an empirical study using a unique large-scale dataset of over 31,000 medical devices and 12,000 national and international manufacturers from over 65 countries that we assembled based on data sources from the FDA and Centers for Medicare and Medicaid Service (CMS). A conservative evaluation of our proposed policy based on this data shows a 38.9% improvement in the recall rate and a 43.0% reduction in the FDA's workload. Our analyses also indicate that implementing our policy could result in significant annual cost-savings ranging between $2.4 billion and $2.7 billion, which highlights the value of using a holistic and data-driven approach to improve the FDA's current 510(K) medical device evaluation pathway.

Key words: Data-driven policy, machine learning, human-algorithm approach, FDA's 510(k) pathway, medical device evaluation

1. Introduction

The medical device approval pathway is a function of the level of control necessary to provide reasonable assurance of a device's safety and effectiveness (FDA 2018a). In general, devices posing a greater degree of risk are denoted by class and face greater regulatory controls. Although there are four common types of approval pathways (FDA 2022a), the vast majority of medical devices are cleared under the *Premarket Notification 510(k) pathway*. For the year 2022, over three thousand devices were cleared under the 510(k) pathway (FDA 2023a), while less than two hundred

devices were cleared under the Humanitarian Device Exemption, Premarket Approval, and De Novo pathways combined (FDA 2023e, FDA 2023d, and FDA 2023c).

The 510(k) pathway was developed both to reduce the burden for device manufacturers bringing medium-to-low risk (Class II and I) devices to market and to address the limited resources of the United States Food and Drug Administration (FDA) (Kramer and Yeh 2023). Devices cleared under the 510(k) pathway must demonstrate that the device is "substantially equivalent" to a legally marketed device, commonly referred to as a *predicate device* (FDA 2022b). A predicate device can be (a) any device that is legally marketed prior to May 28, 1976 (a preamendments device), for which clinical testing was not required, (b) any device that was reclassified from Class III (high risk) to Class II or Class I (medium-to-low-risk), or (c) any device that was also found to be substantially equivalent under the 510(k) pathway (FDA 2018a). Substantial equivalence, in turn, occurs when the device has the same intended use and technological characteristics, or has different technological characteristics but is as safe and effective as the predicate. The FDA determines whether the device is as safe and effective as the predicate device by reviewing the scientific methods used to evaluate differences in technological characteristics and performance data (FDA 2019).

The approach used by the FDA has been heavily criticized. For example, the National Academy of Medicine stated that "the 510(k) process cannot be transformed into a premarket evaluation of safety and effectiveness as long as the standard for clearance is substantial equivalence to any previously cleared device" (FDA 2022b). In 1996, the U.S. Supreme Court also concluded that "the 510(k) process is focused on equivalence, not safety." (Lohr vs Medtronic 1996). More recently, the FDA has also recognized the shortcomings of its 510(k) pathway indicating the need to add alternative options to demonstrate that a device is as safe and effective as a legally marketed device (Challoner and Senate 2011).

Several studies have also highlighted the potential drawbacks associated with the FDA's 510(k) pathway, indicating that 71% of devices recalled by the FDA for safety concerns between 2005 and 2009 had initially been cleared through the 510(k) pathway (Zuckerman et al. 2011), or that 11% of devices cleared via the 510(k) pathway between 2003 and 2018 had later been subject to Class I or II recalls (Everhart et al. 2023). Despite all these concerns, there is currently a clear lack of evidenced-based understanding on how the FDA can improve its 510(k) pathway. This is evidenced by the FDA's 2023 solicitation for feedback (FDA 2023b and FDA 2023f) and its 2018 Medical Device Safety Action Plan for modernizing the 510(k) pathway (FDA 2018b). Our main goal in this paper is to develop a data-driven approach that can assist FDA in improving its 501(k) pathway. To this end, we note that developing a highly accurate predictive model of recall risk can enable the FDA to pinpoint devices that might require more rigorous examination or necessitate stronger

evidential support before approval, ultimately helping the FDA to mitigate unnecessary harm to patients. Developing such a prediction model requires access to large data containing comprehensive information on both applicant devices and predicates as well as detailed data engineering processes to generate useful insights. This level of detail cannot be directly obtained using publicly available FDA datasets and requires significant additional data collection and pre-processing efforts.

We also note that just developing an accurate prediction model using large-scale data might not be sufficient to develop a policy for improving the FDA's 510(k) pathway. The first challenge is related to the nuanced differences of some 510(k) devices, which may make it hard for a policy fully relying on a machine learning (ML) model to be effective. There is a vast literature in support of the interaction between ML models and humans to achieve better overall performance. Thus, we hypothesize that a combined human-algorithm approach to evaluating medical devices can significantly improve FDA's 510(k) pathway. The second challenge is that any such policy must also consider the limited recourses available to FDA to perform further examination when a device is deemed risky for approval by the model. The FDA receives a high volume of 510(k) submissions each year and has a goal of making a decision within 90 days. The sheer number of submissions can strain the FDA's resources, resulting in a potential backlog and delays in the review process. This backlog may limit the depth and rigor of the review, potentially impacting the thoroughness of the evaluation and the ability to identify potential risks or issues for a variety of devices. Thus, to rigorously improve the 510(k) pathway, one needs to also develop an approval and evaluation policy guideline that not only can benefit from the predicted risk of recall, but can also take into account FDA's limited resources for further evaluation of devices that are not a clear-cut for approval or rejection decisions based on the model's predicted risk.

1.1. Overview and Contributions

We introduce a *data-driven clearance policy* aimed at assisting the FDA in improving its 510(k) pathway. Our proposed policy creates a combined human-algorithm approach by deferring some decisions to human exerts and others to a well-trained ML model. Specifically, our policy generates recommendations for direct approval, rejection, or further evaluation by human experts using a variety of the applicant device's characteristics as well as those of its predicates. Our work is based on close corroboration and extensive discussions with a collaborator (and co-author of this paper) who has been directly involved in numerous related FDA regulatory improvement projects.

Our approach consists of two main steps. In the first step, we collect and assemble FDA data related to 510(k) devices, and develop ML models capable of estimating each device's risk of recall based on the information available at the time of submission. Specifically, employing a text-scraping algorithm, we start by extracting information regarding both applicant devices and their

corresponding predicates, including their characteristics and recall information. Through extensive discussions with our collaborator, we also conduct thorough data engineering to create variables that aid in predicting a recall event. Building on our analysis, we then train and test several ML models and pick the best one, which achieves a cross-validation Area Under the Curve (AUC) score of 0.77. Our findings using this ML model suggest that the number of recalls reported for predicate devices, along with variables relevant to the age of the predicates, hold significant predictive power for a recall event. This validates speculations and initial empirical findings in the recent literature but through an ML lens that is trained on large-scale data involving over 31,000 medical devices produced by 12,000 manufactures from over 65 countries. Interestingly, however, unlike these more known facts, our analysis indicates that the timing of recalls reported for predicates of an applicant device and the age of the latest-approved predicate provide valuable information for predicting a recall event.

In the second step, utilizing the ML model for risk assessment, we proceed to develop a data-driven clearance policy that can assist the FDA in decision-making. To this end, we make use of an optimization approach that sets decision-making thresholds. Our approach takes into account the existence of ranges within the risk spectrum where the predicted risk is not informative enough to make a recommendation, necessitating human expert attention. These thresholds are strategically determined to balance the precision of decision-making with the workload burden imposed on the FDA during the decision-making process. The primary computational challenge lies in enforcing a workload constraint for the FDA, turning the optimization model into a non-convex optimization problem. To address this, we develop an algorithm that provides an approximation by making use of Lagrangian relaxation. This algorithm efficiently identifies effective thresholds on predicted recall risks through an iterative solution procedure. We also emphasize that our policy is designed to be capable of providing additional information (risk labels) for applicant devices that require further evaluation by human experts (e.g., by an FDA committee).

We investigated the potential benefits of our policy compared to the current practice of the FDA using a unique large-scaled dataset that we assembled based on the public FDA datasets (Section 5). Our results indicated a 38.9% improvement in the recall rate and a 43.0% reduction in the FDA's workload. Additionally, we estimated the potential cost-savings from implementing our data-driven policy through estimating medical device replacement costs by medical specialties (Section 5.1). For the very first time, this is done by carefully determining the medical specialty for over 99% of the 1,351 unique Healthcare Common Procedure Coding System (HCPCS) codes/descriptions for the CMS administrative claims data during 2013-2020. Our analysis indicate significant potential annual cost-savings from implementing our policy ranging between $2.4 billion to $2.7 billion, as well as a notable decrease in adverse event outcomes and an enhancement in patient safety.

Our results provide valuable insights for the FDA and address main concerns about the current 510(k) process (Section 6). First, we observe that while there are some easy cases suitable for direct algorithmic decision-making, some other cases are difficult and require more in-depth evaluation and human judgment. This observation confirms our hypothesis that there is a need for a combined human-algorithm approach to integrate the FDA's expertise with quantitative evidence. Next, our results address the FDA's concerns communicated in its recent solicitation announcements as well as its goal to establish best practices to evaluate the safety of medical devices (FDA 2023b). Specifically, we find that best practices should take advantage of the fact that the number of recall events for predicates, particularly recent recalls, is a significant predictor of recall risk for an applicant device. Furthermore, we observe that the age of the latest-approved and earliest-approved predicates, as well as the average age of the predicates are among other important predictors of recall risk. Finally, our results contribute to the recent call for feedback by the FDA on the opportunities and challenges of using artificial intelligence (AI) and ML in the development of drugs and medical devices (FDA 2023f). Our work highlights the importance of crucial considerations in the context of utilizing AI/ML to enhance the evaluation of medical devices, and highlights the importance of employing a combined human-algorithm approach.

1.2. Literature Review

Our paper contributes to four main bodies of research. Below, we concentrate on the most closely related works.

Medical Device Recalls. There is a vast literature focusing on comparing the risk of recall for devices that received approval via the 510(k) pathway and those that received approval via premarket approval (see, e.g., Day et al. 2016, Connor et al. 2017, Janetos et al. 2017, Talati et al. 2018, and Dubin et al. 2021). There are also a few empirical studies (see, e.g., Wowak et al. 2021, Ball et al. 2018, Wowak et al. 2021, and Ball et al. 2018) related to our work that examine factors influencing recall-related events of medical devices. Another related set of studies investigate the relationship between the characteristics of predicate medical devices and the recall events of 510(k) devices (Everhart et al. 2023 and Kadakia et al. 2023). The study of Mukherjee and Sinha 2018, presents a predictive model for risk of recall using the adverse events occurring after the approval time. However, predictive models focusing on estimating the recall risk of an applicant device based on the information available at the time of submission remain scarce. In our study, we address this gap by developing a recall risk prediction model and recommending a clearance policy for the FDA, wherein the policy's recommendations are constrained by the available information at the time of the 510(k) application submission as well as the workload that can be imposed to FDA for further evaluations by human experts prior to decision-making.

Our contributions to this literature are twofold. First, we conduct a thorough examination of factors that prove useful in predicting the risk of recall, and design ML models that can benefit from the available information at the submission time. Second, we propose the first data-driven clearance policy that integrates the predictive power of ML models into the decision-making process in order to improve the FDA's 510(k) regulatory pathway.

Medical Diagnostic Decision. Various studies in the literature focus on medical diagnostic decisions with the goal of optimizing pre-diagnosis or follow-up choices (see, e.g., Ayvaci et al. 2012, Zhang et al. 2012, and Bayati et al. 2018). Our work is closest to the subset of the literature that aims at incorporating a decision support tool in the diagnostic decision and potential biases in the process (see, e.g., Ahsen et al. 2019 and Jussupow et al. 2021). Our work is also closely related to studies that aim at improving quality versus speed trade-offs in diagnostic and triage decisions (see, e.g., Saghafian et al. 2018). Our work differs form this stream of literature as it develops a data-driven policy for improving the FDA's 510(k) pathway in which both ML based predictions and workload considerations are taken into account.

Threshold-Based Decision-Making. There are several existing methods to utilize risk predictions for decision-making by determining a single decision threshold, ranging from utility-based methods (see e.g., Jund et al. 2005, Felder and Mayrhofer 2014, and van Giessen et al. 2018) to receiver operating characteristic (ROC) method (see, e.g., Hajian-Tilaki 2013 and Hong et al. 2021) and Bayesian decision theory (see, e.g., Sheppard and Kaufman 2005 and Weise et al. 2006). Our works is closest to the subset of literature focusing on three-way classifications with three categories of positive, negative, and undecided based on the evidence (see, e.g., Si et al. 2017, Yao 2010, Yao and Zhou 2016, and Garcia et al. 2020). However, our work differs from these studies in two major ways. First, balancing workload plays a significant role in our work. In a three-way classification using two decision thresholds, setting the threshold conservatively often results in achieving a high performance measure in the positive and negative regions. However, it may lead to assigning many other instances to the undecided region which often will require attention from human experts. Consequently, this approach may not significantly reduce the workload in the system, though balancing the workload (speed) and quality of the decisions made is a crucial element in many systems (see, e.g., Saghafian et al. 2018). Second, the existing methods for finding the decision thresholds without workload considerations are less complex and often can be solved using simple heuristics (Si et al. 2017), Bayesian rough sets (Yao 2010, Yao and Zhou 2016), or a linear program (Garcia et al. 2020). In contrast, our methodology requires different techniques to handle the non-linearity in optimizing decision thresholds. Our solution technique, obtained by deriving structural properties of our optimization model and introducing a Lagrangian-based algorithm,

allows us to capture the complexity introduced by workload considerations and identify the decision thresholds efficiently.

Human-in-the-Loop Approaches. An expanding body of research suggests that ML models can outperform humans in making predictions across a wide range of domains (see, e.g., Liu et al. 2018, Shen et al. 2019, Boloori et al. 2022, Ang et al. 2022). While the state-of-the-art ML models exhibit impressive performance, in high-risk domains such as healthcare, there is reluctance to fully embrace automated AI systems and eliminate humans entirely from the loop due to the inherent distrust in AI systems and lack of robustness (Association et al. 2019). Recent literature has highlighted the benefits of incorporating human judgment into the ML model deployment process, leading to the development of human-in-the-loop approaches. The idea of keeping humans in the loop has been implemented in various ways. There have been attempts to design interactive and active ML systems that continuously learn from humans (Wu et al. 2022) or even go beyond human-in-loop mechanisms by systematically incorporating symbiotic learning (Muller 2022, Saghafian 2023). Other attempts include combining separate human and algorithm outputs (e.g., Blattberg and Hoch 1990, Goodwin 2000), introducing systems to elicit human judgment for prediction algorithms (Ibrahim et al. 2021), and learning the human experts' intuition for risk prediction (Orfanoudaki et al. 2022). Another approach to keeping humans in the loop is AI augmentation, where the main idea is to have AI systems work alongside humans and collaborate with them. This idea is different from automation that results in replacing humans with AI (Daugherty and Wilson 2018, Miller 2018). Our work provides further evidence on the latter approach, providing a regulatory policy that allows human experts to concentrate on complex cases with the assistance of AI/ML.

The reminder of the paper is organized as follows. Section 2 provides a summary of our data collection and pre-processing steps. Section 3 describes the development and evaluation of prediction models. Section 4 introduces our data-driven clearance policy, its analytical properties, and our solution methodology. Section 5 provides our empirical results. Finally, we present managerial insights and concluding remarks in Sections 6 and 7, respectively.

2. Setting and Data

In this section, we start by briefly explaining our data collection and pre-processing steps. We then discuss various aspects of our data and provide a data summary.

2.1. Data Collection

We collected 510(k) applicant device submission data and FDA recall data from public FDA datasets related to years 2008 to 2020. The 510(k) applicant device submission data includes the unique 510(k) number of applicant devices, submission and clearance dates as well as device characteristics, such as medical specialty, product type, and device class. Figure 1 demonstrates

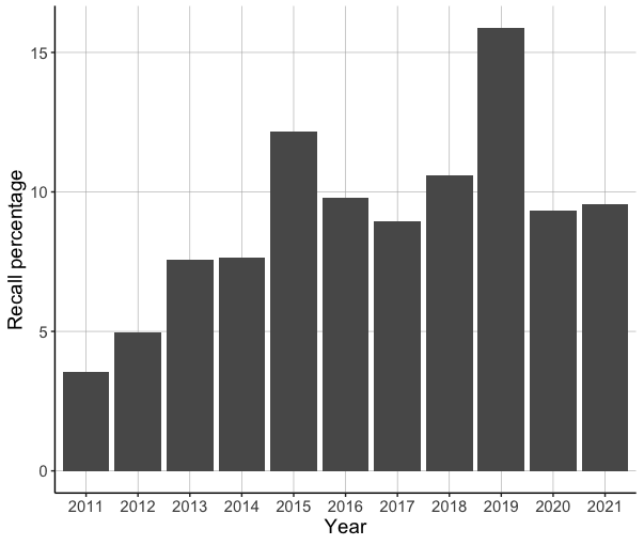


Figure 1 Recall percentage per year

the recall percentage from 2011 to 2021. The FDA recall data contains the unique 510(k) number of recalled devices, events dates, and severity of recalls. The FDA categorizes recall events into three major classes based on the relative degree of health hazard that can be posed to patients. A Class 1 (severe) recall is a situation with a reasonable probability of serious adverse health consequences or death. A Class 2 (moderate) recall is a situation where temporary or medically reversible adverse health consequences are probable. A Class 3 (mild) recall is a situation that is not likely to cause adverse health consequences.

In the publicly available datasets, there is no direct link between an applicant device and its recall information. For each applicant device, we therefore directly investigated whether it had any prior recalls and identified the recall class when a recall event was reported. Using the public evidence summary documents, we developed a text-scraping algorithm to identify the predicate devices for each cleared 510(k) applicant device. Our resulted data included the applicant devices for which the algorithm identified at least one predicate device. We also extracted device characteristics and recall information for all the predicate devices and added them to our dataset.

2.2. Variables and Summary

Our primary outcome is a *binary recall event* indicating whether the applicant device had at least one recall between its FDA clearance date and the end of our study period (2008-2020). A summary of our study sample is provided in Tables 1 and 2.

Applicant Devices' Characteristics. The FDA classifies each device into several medical specialties, such as Orthopedic (OR) or Cardiovascular (CV). *Medical Specialty* refers to the medical specialty of the applicant device identified by the FDA. The FDA classifies medical devices into three classes based on their risks and regulatory controls. Class I devices pose the lowest risk to

Table 1 Summary of the characteristics of the devices in the study sample

	No Recall (N=28,618)	**Recall (N=3,321)**
Medical Specialty (Top and Low Three)		
Radiology (RA)	3,534 (12.3%)	670 (20.2%)
Orthopedic (OR)	5,551 (19.4%)	632 (19.0%)
Cardiovascular (CV)	3,966 (13.9%)	485 (14.6%)
Ear, Nose, Throat (EN)	303 (1.1%)	24 (0.7%)
Clinical Toxicology (TX)	249 (0.9%)	20 (0.6%)
Physical Medicine (PM)	556 (1.9%)	15 (0.5%)
Device Class		
I	934 (3.3%)	26 (0.8%)
II	27,684 (96.7%)	3,295 (99.2%)
Country Code (Top and Low Three)		
United States (US)	19,979 (69.8%)	2,699 (81.3%)
Other	1,887 (6.6%)	179 (5.4%)
Germany (DE)	838 (2.9%)	90 (2.7%)
Switzerland (CH)	392 (1.4%)	31 (0.9%)
Taiwan (TW)	625 (2.2%)	17 (0.5%)
Korea (KR)	839 (2.9%)	15 (0.5%)
Product Code (Top and Low Three)		
Other_OR	3,180 (11.1%)	368 (11.1%)
Other_CV	3,016 (10.5%)	346 (10.4%)
Other_RA	1,495 (5.2%)	238 (7.2%)
IYN	500 (1.7%)	87 (2.6%)
JAK	206 (0.7%)	83 (2.5%)
Other_AN (Anesthesiology)	817 (2.9%)	81 (2.4%)
Implantable		
0	20,666 (72.2%)	2,451 (73.8%)
1	7,952 (27.8%)	870 (26.2%)
Life Sustaining/Supporting		
0	27914 (97.5%)	3,099 (93.3%)
1	704 (2.5%)	222 (6.7%)

Note: Data are presented as number (%)

Other = Countries with a frequency of less than 1%

Other_Y = Product codes within medical specialty Y having a frequency of less than 5%

IYN = Ultrasonic pulsed doppler devices; JAK = Radiology Medical diagnostic X-Ray devices

Table 2 **Summary of the predicates' characteristics and recall information of the devices in the study sample**

	No Recall (N=28,618)	**Recall (N=3,321)**
Predicate Devices' Characteristics		
Num. of Predicate	2.36 (2.29)	2.68 (3.17)
Num. of Unmatched Medical Specialties	0.155 (0.412)	0.153 (0.402)
Num. of Unmatched Product Code	0.507 (0.841)	0.606 (0.987)
Predicate Average Age	4.25 (3.63)	4.04 (3.36)
Predicate Median Age	4.09 (3.62)	3.84 (3.30)
Predicate Newest Age	3.85 (4.72)	3.34 (4.10)
Predicate Oldest Age	7.44 (6.61)	7.39 (6.59)
Recall Information		
Predicate Recalled		
0	23,744 (83.0%)	2,117 (63.7%)
1	4,874 (17.0%)	1,204 (36.3%)
Number of Predicates Recalled	0.228 (0.615)	0.559 (1.11)
Number of Recalls	0.409 (1.88)	1.28 (4.04)
Variance of Recalls	0.040 (0.895)	0.168 (2.00)
Number of Class 1	0.014 (0.202)	0.035 (0.313)
Number of Class 2	0.387 (1.84)	1.23 (3.97)
Number of Class 3	0.009 (0.103)	0.020 (0.175)
Weighted Number of Recalls	0.247 (1.12)	0.867 (2.57)

Note: Data are presented as number (%), or mean (standard deviation)

patients, while Class III devices pose the highest risk. *Device Class* indicates the FDA device class. Most devices in the 510(k) program fall into Class I or Class II. A limited number of applicant devices (preamendments devices) were initially regulated as Class III with the intent that either the FDA would reclassify the device into a lower class or call for the premarket approval application. Due to a limited number of such applicant devices and the lack of a standard protocol, we only included applicant devices of Class I or II in our analyses. *Country Code* indicates the country of origin for a device manufacturer. We reduced the number of country code levels by preserving the most common ones and re-coding the others with a frequency of less than 1% as "Other." The Center for Devices and Radiological Health (CDRH) associates each medical device with a *Product Code* based upon the medical device function. For example, ultrasonic pulsed doppler devices are assigned to "IYN," while Ultrasonic Pulsed Echo imaging devices are assigned to "IYO." We reduced the number of product code levels for each medical specialty by preserving the most common ones and re-coding the others with a frequency of less than 5% as "Other Medical Specialty."

Table 3 Proportion and recall rate of applicant devices with predicate devices having mismatched medical specialty/product code

Num. of Unmatched	Proportion	Recall %
Medical Specialty		
0	63.45	60.43
1	26.53	10.54
2	6.90	12.07
3	2.02	13.95
>4	1.10	18.47
Product Code		
0	86.05	10.42
1	12.69	10.14
>2	1.26	11.72

Implantable is a flag indicating whether a device is designed to be placed into a surgically or naturally formed cavity of the human body. *Life Sustaining/Supporting* is a flag indicating whether a device is essential to restoring or continuing a bodily function.

In the FDA 510(k) program, the *substantial equivalence* is often evaluated based on the similarities between the predicates devices and the applicant device in terms of the composition and design (Zuckerman et al. 2014). In the publicly available datasets, the information of predicate devices is not directly linked to the corresponding applicant devices. Thus, we created several variables corresponding to predicate devices, which can be classified into the following three categories.

The first category accounts for the similarity between an applicant device and predicate devices. *Number of Predicates* in this category is a count of the number of predicate devices identified for an applicant device. *Number of Unmatched Medical Specialties* measures the unique number of medical specialties for predicate devices, corresponding to an applicant device, that do not match the medical specialty of the applicant device. Similarly, *Number of Unmatched Product Codes* measures the unique number of product codes for predicate devices, corresponding to an applicant device, that do not match the product code of the applicant device. Table 3 presents the proportion and recall rate of applicant devices where the medical specialty/product code of predicate devices does not match their medical specialty/product code.

The second category accounts for the age of predicate devices. *Predicate Average Age* is the difference between the average year of approval of predicate devices and the application submission date of the applicant device. *Predicate Median Age* is the difference between the median year of approval of predicate devices and the application submission date of the applicant device. *Predicate Newest Age* indicates the difference between the year of approval of the newest predicate device and the application submission date of the applicant device. Similarly. *Predicate Oldest Age* indicates

the difference between the year of approval of the oldest predicate device and the application submission date of the applicant device.

The third category accounts for the recall status of predicates devices. *Predicate Recalled* is a flag indicating whether at least one of the predicate devices identified for an applicant device is recalled. *Number of Predicates Recalled* is a count of the number of predicate devices with at least one recall. *Number of Recalls* is a count of the total number of recalls for the predicate devices. *Variance of Recalls* is the sample variance of the number of recalls of predicate devices. For example, consider an applicant device with three predicate devices. In the first scenario, the first predicate has six recalls, while the second and third predicates have zero recalls. In the second scenario, each predicate device has two recalls. The sample variance of recalls for the first scenario is 12, while it is zero for the second scenario. We also consider the severity of each recall. *Number of Class 1 Recalls* is a count of the total number of Class 1 recalls among the predicate devices identified for an applicant device. *Number of Class 2 Recalls* and *Number of Class 3 Recalls* can be defined similarly. Another interesting piece of information to consider is the count of recalls with respect to their timing. For example, the importance of a recall event of a predicate device that has occurred many years prior to an applicant device's submission date may differ from a recent recall event. *Weighted Number of Recalls* is the weighted number of recalls for the predicate devices identified for an applicant device. We define a time window of ten years such that a recall event is negligible if it has occurred at or more than ten years prior to the applicant's device submission date. For the other recall events, we assign weights based on the time difference. Thus, a recall event that has occurred at the submission date receives a weight of one, and a recall event that has occurred ten years before the submission date is assigned a weight of zero. For recall events that fall within this ten-year window, weights are calculated proportionally based on their distance from the submission date (e.g., a recall event that happened 4 years ago would be assigned a weight of 0.6).

3. Machine Learning Models for Predicting Recall Events

We now describe the development and evaluation of ML models for predicting recall events. Our ML models are trained on our data described in the previous section which contains information on a variety of medical devices with different characteristics in order to predict recall events of unseen future applicant devices.

3.1. Machine Learning Models

We consider various ML models and compare their performance to pick the best one. The models we consider are regularized logistic regression using both Lasso and Ridge type penalties, decision tree, random forest, and gradient boosting. The logistic regression model is of the form $\log\left(\frac{\mathbb{P}(Y_i=1|X_i)}{1-\mathbb{P}(Y_i=1|X_i)}\right) = \beta_0 + \beta_1^\top X_i$, where X_i is a vector of information (e.g., characteristics of the

applicant device and corresponding predicate devices), β_1 is the unknown coefficient vector, and β_0 is the unknown intercept. Overfitting is one of the biggest causes of the poor performance in working with high-dimensional data such as ours. Regularization is a well-established technique to prevent overfitting by limiting the complexity of the model. The regression coefficients β_0 and β_1 can be estimated by solving the following optimization problem:

$$\max_{\beta_0,\beta_1}\left\{\sum_{i=1}^{N}\left(Y_i\left(\beta_0+\beta_1^\top X_i\right)-\log\left(1+\exp\left(\beta_0+\beta_1^\top X_i\right)\right)\right)-\eta\left\|\beta_1\right\|_p^p\right\},$$

where N is the number of observations in the training set, $\|\cdot\|_p$ denotes p-norm, and η is the regularization tuning parameter. We implement two types of penalties similar to *Lasso* and *Ridge* regression. We do so by setting $p=1$ and $p=2$, respectively. In both models, the value of η is tuned through 10-fold cross validation.

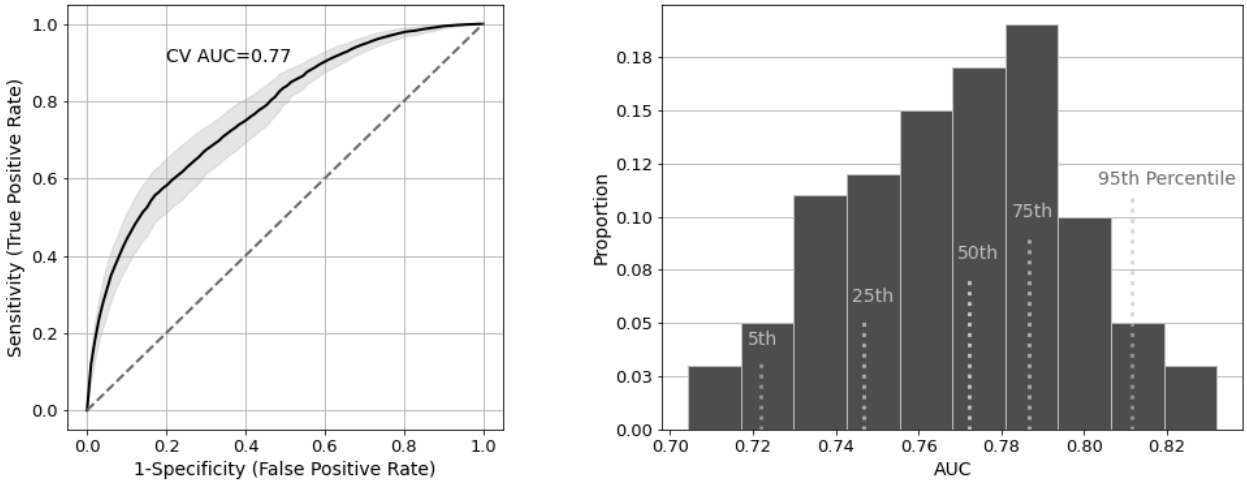
We also implement and train a *decision tree* model, which is better suited to capture nonlinear effects and various interactions in our data. We prune and optimize the depth and the minimum number of samples required to split a node using 10-fold cross validation. In addition, we consider two ensemble learning models built upon *Bagging* and *Boosting* techniques. Specifically, we develop a random forest model, which is a popular bagging method obtained by constructing a multitude of de-correlated decision trees. Boosting involves successively training models such that the next model combined with previous ones minimizes the overall prediction error. We also develop a gradient boosting model, which is a popular boosting method obtained by successively constructing different decision trees. Two parameters for random forest (i.e., the number of trees and the maximum depth of each tree) and gradient boosting (i.e., the number of trees and the learning rate) are optimized through 10-fold cross validation.

3.2. Performance Evaluation of the Machine Learning Models

To compare our ML models, we evaluate their predictive power on unseen 510(k) applicant devices. A 10-fold cross-validation approach is employed to calculate the area under the receiver operating characteristic curve (AUC) for all models. A bootstrapping approach is used to capture statistical fluctuations of AUC scores with respect to random splitting of the data in cross-validation. Table 4 provides a summary of the results, where we report the average AUC for all models, along with the values of standard deviation and quantiles. We find that all models except the decision tree model attain relatively similar performance in terms of the AUC metric, ranging from 0.76 to 0.77. The statistical fluctuations of these four all models are also relatively the same. The lowest standard deviation is 0.027, and the highest 5% and 95% quantiles are 0.72 and 0.81, respectively. The decision tree model is not competitive, which maybe due to overfitting to training sets.

Table 4 Area under the curve for statistical models

Model	Cross-Validation AUC	Standard Deviation	Quantile (5%, 95%)
Log Reg with Ridge penalty	0.76	0.030	(0.71, 0.80)
Log Reg with Lasso penalty	0.77	0.028	(0.72, 0.81)
Decision tree	0.73	0.036	(0.66, 0.78)
Random forest	0.76	0.031	(0.71, 0.81)
Gradient boosting	0.77	0.027	(0.72, 0.81)



(a) Cross-validation AUC (b) Distribution of AUC scores

Figure 2 Out-of-sample performance of the selected model (gradient boosting)

Since gradient boosting has the lowest standard deviation and achieved the highest AUC, we identify it as the best candidate. The receiver operating characteristic (ROC) curve along with the distribution of AUC scores for gradient boosting are depicted in Figure 2. The shaded error bars in Figure 2a correspond to ± 1 standard deviation, and the vertical lines in Figure 2b correspond to different quantiles.

The AUC results reported in Table 4 are calculated based on the overall performance of models across all medical specialties. In practice, each medical specialty has a separate committee for evaluating the applicant devices assigned to that medical specialty. The 510(k) process can be viewed as a general rule, but each committee's evaluation criteria may vary. We evaluate the predictive power of gradient boosting separately for each medical specialty. Figure 3 shows the cross-validation AUC per specialty. As can be seen, the model has the best performance in terms of the AUC metric for applicant devices of Clinical Chemistry (CH), Anesthesiology (AN), and

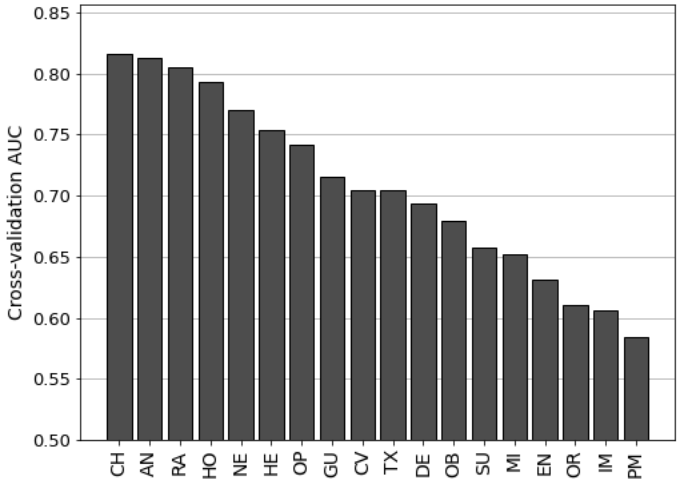


Figure 3 Area under the curve across all specialties

Note: CH = Clinical Chemistry; AN = Anesthesiology; RA = Radiology; HO = General Hospital; NE = Neurology; HE = Hematology; OP = Ophthalmic; GU = Gastroenterology & Urology; CV = Cardiovascular; TX = Clinical Toxicology; DE = Dental; OB = Obstetrics/Gynecology; SU = General & Plastic Surgery; MI = Microbiology; EN = Ear, Nose, & Throat; OR = Orthopedic; IM = Immunology; PM = Physical Medicine (PM)

Radiology (RA). On the other hand, Orthopedic (OR), Immunology (IM), and Physical Medicine (PM) are three medical specialties for which the model has the worst performance.

When it comes to a recall event, the severity of the recall can be relativity identified by the FDA's classification of recalls. Although there is value in detecting each recall class, detecting a high-severity recall class has the highest priority due to its life-threatening nature. We evaluate our gradient boosting model on its ability to detect different classes of recalls. Figure 4 plots the proportion of recalls correctly identified per each recall class versus the overall false recall rate. Overall, the model has better predicting power for a recall Class 1 followed by Class 2 and Class 3. In particular, it achieves the AUC of 0.81, 0.77, and 0.64 in detecting recall Classes 1, 2, and 3, respectively. This is a desired performance, since Class 1 has the highest severity, followed by Class 2 and then Class 3.

Lastly, we investigate the most important variables and their impact when predicating recall risk using our gradient boosting model. We use the shapely additive explanations (SHAP) method (Lundberg and Lee 2017, Lundberg et al. 2020), which leverages a game theory approach to compute the contribution of variables to a predicted value in an additive form. We compute the contribution of each variable to the predicted recall risk in the form of a normalized score ranging between -1 to 1. Figure 5 highlights the 15 most important variables and their impact on the predicted recall

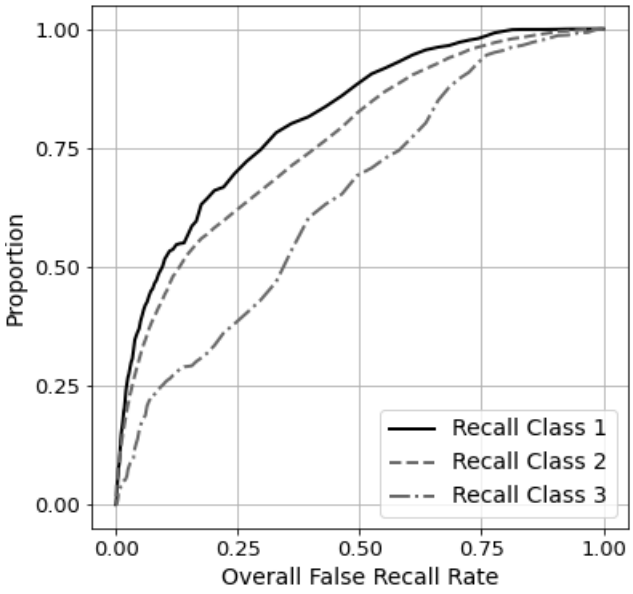


Figure 4 Proportion of recall classes correctly identified by our gradient boosting model

risk. They are ordered by decreasing significance. In Figure 5, each point represents a variable's contribution to the prediction. The value of each contribution is depicted by a color gradient from grey to red, where grey indicates low values and red indicates high values. Our results indicate that *Weighted Num. of Recalls* and the variables relevant to the age of the predicates are among the top five most significant predictors of the recall risk. *Weighted Num. of Recalls* accounts for the number of recall events for predicates where the most recent recalls are prioritized. We observe that a higher value of this variable is associated with a higher recall risk. This observation indicates the importance of the timing of predicate recall events, an aspect that has been overlooked in the literature. It also validates our hypothesis that paying attention to predicate devices with recent recalls can go a long way in raising red flags for an applicant device. In particular, a recall event that has happened long ago may have been fully addressed, but some uncertainties might remain unresolved for more recent recalls. Our results also highlight that *Num. of Recalls* and *Variance of Recalls* are highly predictive of the recall risk.

With regard to the age of predicates, we observe that a higher value of *Predicate Newest Age* is associated with a lower risk of recall. Investigating the distribution of the SHAP values corresponding to this variable, we find that devices for which the *Predicate Newest Age* is less than five years are associated with a significant higher recall risk. This may be because newer predicates might not have undergone extensive usage, which can reveal potential issues over time. We observe that while the *Predicate Oldest Age* is highly predictive of the recall risk, both high and low values of this

variable can lead to a high recall risk. On one hand, older predicates can indicate safety, as they have a history of safe and effective use over time and may be the gold standard for patient care. On the other hand, older predicates may not reflect the advanced technology embedded in new devices, potentially leading to compatibility or performance issues. With regard to the *Predicate Average Age* and *Predicate Median Age*, we observe that, overall, lower values of these predictors are associated with a lower risk of recall.

Among the variables corresponding to the characteristics of applicant devices, medical specialties, country codes, and product codes are highly predictive of the recall risk. The results are consistent with prior literature indicating some heterogeneous effects of these indicators on the risk of recall. Furthermore, our results show that *Life Sustaining/Supporting* is of significant importance. This variable quantifies the risk level associated with an applicant device. Riskier applicant devices are generally under stringent market scrutiny and are more likely to be recalled. Finally, with respect to the similarity between an applicant device and its predicate devices, we observe that neither *Num. of Unmatched Specialties* nor *Num. of Unmatched Product Codes* are among the top 15 most informative variables, suggesting that taking them into account will not be that useful.

4. Design of a Data-Driven Human-Algorithm Decision Support Tool

In this section, we make use of our best predictive model discussed in the previous section, and introduce a data-driven clearance policy that balances increased safety and expeditious evaluation of medical devices. The FDA has designated committees that evaluate devices seeking clearance through the 510(k) pathway. The policy we propose is based on a human-algorithm approach designed to improve and facilitate the 510(k) approval pathway by reducing the risk of potential recalls and the workload of the FDA's designated committees. In Section 5, we perform various empirical investigations to estimate the impact of our proposed policy and further guide the FDA in implementing it.

A typical approach in designing a policy to recommend suitable actions based on an ML model's outputs is to estimate the utility of possible outcomes (recall risk), and impose a fixed threshold on the predicted risk to maximize the utility. Unfortunately, relying solely on a single threshold fails to identify specific regions where risk estimation models perform poorly. In other words, it overlooks the range of risk estimate values where false-positive and false-negative rates are high, which are cases where diagnostic decisions should be delegated to human experts and approached with caution.

We develop an advanced clearance policy by harnessing the power of an ML model. However, in cases where the predicted risk for an applicant device may not be informative enough, our policy has the capability to offer supplementary information for assessment by human experts without

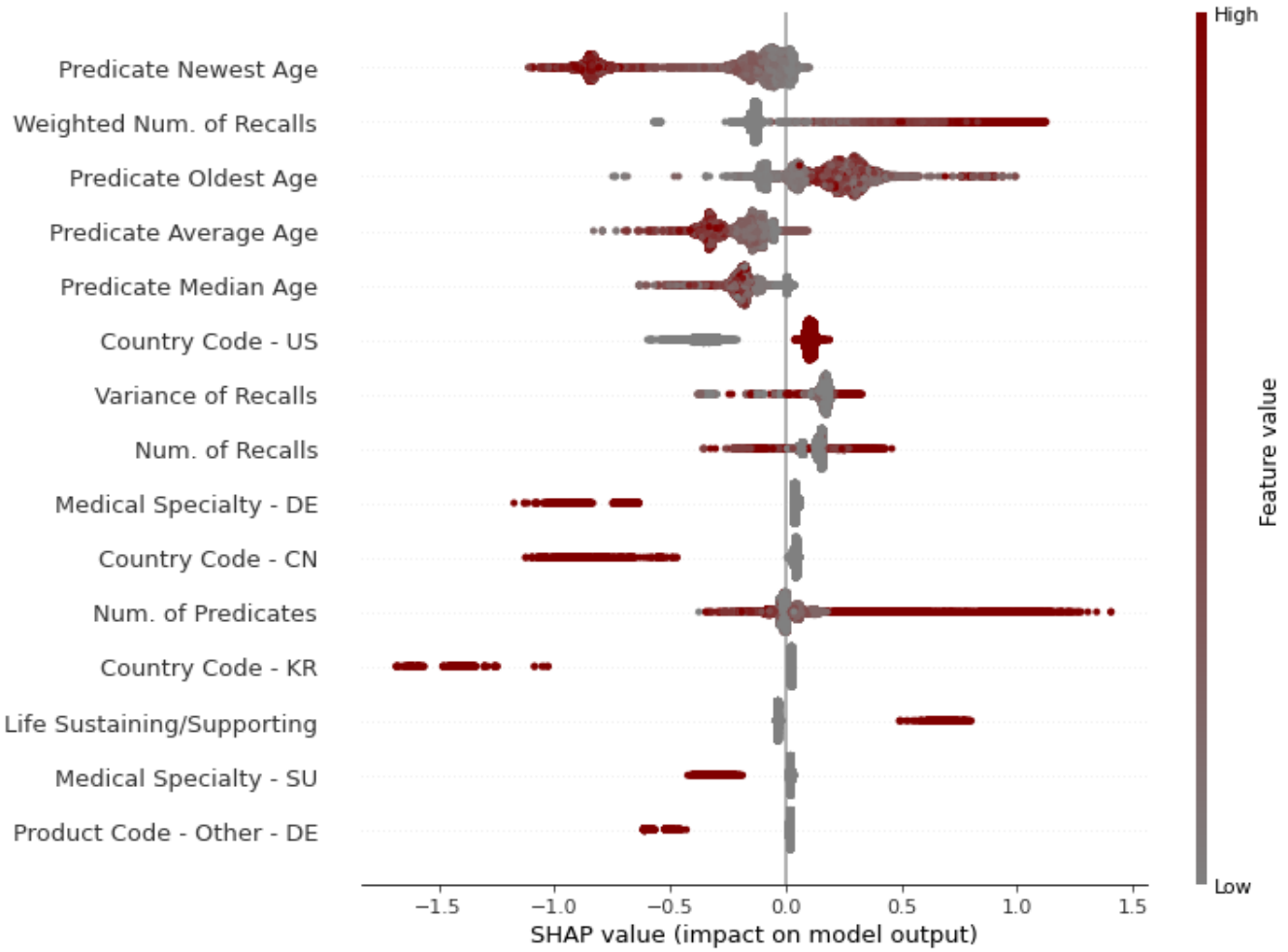


Figure 5 Top 15 predictors of recall risk

Note: US = Unites States; DE = Germany ; CN = China ; KR = Korea; SU = General & Plastic Surgery

presenting a direct recommendation. The evaluation of such intricate cases is delegated to the designated FDA committee, leveraging their expertise to arrive at more informed judgments. The supplementary information are provided as risk labels, which can be generated using the predicted risks by our ML model.

4.1. A Data-Driven Advanced Clearance Policy

Our policy has two main components: (1) an ML predictor to estimate recall risk of a 510(k) device based on the information available to the FDA upon submission by the manufacturer, and (2) an optimization approach that determines whether an applicant device can be quickly accepted/rejected or if it should be deferred and more elaborately evaluated by an FDA assigned committee. Additionally, a classifier can be used to categorize deferred 510(k) devices based on their risk level, thereby providing more supporting information for the assigned committee to guide their decision-making. By combining these main core elements, our policy streamlines the clearance

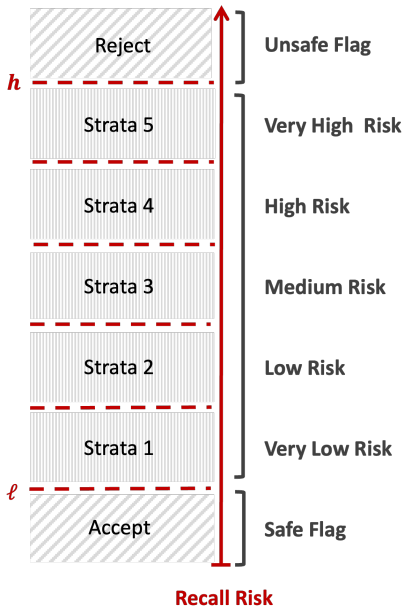


Figure 6 Schematic view of the proposed clearance policy

process, ensuring that devices are assessed with greater efficiency and accuracy while there are enough resources for in-depth evaluation of deferred devices which require more attention.

In our policy, the evaluation occurs through two main phases, as depicted in Figure 6. In the initial phase, the ML model discussed in §3.1 assesses the recall risk of an applicant device based on the available information provided during the submission. If the predicted risk by the ML model is lower than an optimized low threshold (ℓ), the policy recommends accepting the device. If, however, the predicted risk exceeds the high threshold (h), the device is recommended to be rejected. The policy optimizes the specific values of the low and high thresholds to minimize the rates of *acceptance of unsafe devices* and *rejection of safe devices*, while still aligning with the preferences of decision-makers. That is, it ensures that (a) rates of *rejection of unsafe devices* and *acceptance of safe devices* are at least greater than values specified as desired, and (b) the workload imposed to the FDA for more elaborate evaluation of deferred devices does not exceed a preferred level. For devices that are deferred to a FDA committee for more in-depth evaluation, the policy assigns a risk label, ranging from very low risk to very high risk (the number of risk labels illustrated in Figure 6 are for illustrative purposes and can be refined as needed). These risk labels provide additional information to assist the committees in making better judgments.

As mentioned, the first core element of our policy is predicting the recall risk for an applicant device given the information available upon application submission. In Section §3.1, we discussed how we have developed and trained a well-performing ML model using our dataset. Given a vector of inputs containing information on an applicant device and its corresponding predicate devices,

the model generates a probability of recall. Let $f : X \to (0,1)$ be the functional representation of the model, which maps the vector of information on an applicant device and its predicates denoted by X to a probability value in $(0,1)$. The second core element of our policy contains developing a constraint optimization model to identify the low and high thresholds. Let $\delta^+ = f(X|Y=1)$ be the recall risk estimated for an applicant device which has been recalled according to data $(Y=1)$. Similarly, let $\delta^- = f(X|Y=0)$ be the recall risk estimated for an applicant device with no recall in data $(Y=0)$. Using this notation, we next describe the logic behind the optimization model that forms the second core element of our proposed policy.

A well-designed and effective policy is expected to yield low rates of acceptance of unsafe devices $(\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)])$ and rejection of safe devices $(\mathbb{E}[\mathbb{1}(\delta^- \ge h)])$. However, a low value for those measures may yield low rates of rejection of unsafe devices $(\mathbb{E}[\mathbb{1}(\delta^+ \ge h)])$ and acceptance of safe devices $(\mathbb{E}[\mathbb{1}(\delta^- \le \ell)])$. Also, a low workload $(\mathbb{E}[\mathbb{1}(\ell < f(X) < h)])$ for the FDA's committees is desired so only a low ratio of applicant devices should be deferred and judged by them. Due to trade-offs between these measures, identifying the values of low and high risk thresholds (ℓ and h, respectively) is challenging. We develop a non-linear optimization model to solve this challenging problem:

$$\min_{\ell,h} \ \lambda\, \mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] + (1-\lambda)\, \mathbb{E}[\mathbb{1}(\delta^- \ge h)] \tag{1}$$

$$\text{s.t. } \mathbb{E}[\mathbb{1}(\delta^+ \ge h)] \ \ge \ \xi^{ru} \tag{2}$$

$$\mathbb{E}[\mathbb{1}(\delta^- \le \ell)] \ \ge \ \xi^{as} \tag{3}$$

$$\mathbb{E}[\mathbb{1}(\ell < f(X) < h)] \ \le \ \rho \tag{4}$$

$$0 \ \le \ \ell \ \le \ h \ \le \ 1. \tag{5}$$

The objective is to minimize the weighted sum of rates of acceptance of unsafe devices and rejection of safe devices. Although rejecting a safe device may cause a delay in the clearance process for the manufacturer, it is relatively less critical than accepting an unsafe device considering the availability of substantially equivalent devices in the market. However, our framework is general and allows specifying relative importance considerations via the weight $\lambda \in (0,1)$ in (1).

Constraint (2) ensures that the rate of rejection of unsafe devices is greater than a threshold (ξ^{ru}). Constraint (3) requires that the rate of acceptance of safe devices is greater than a threshold (ξ^{as}). Finally, constraint (4) mandates that the workload (measured via the rate of deferred devices) is less than a desired threshold (ρ). The modeling parameters λ, ξ^{ru}, ξ^{as}, and ρ are specified based on the preference of the decision-maker. They may vary from year to year, depending on the FDA's resources and other factors.

4.2. Structural Properties

The optimization problem (1)-(5) is challenging to solve due to its non-convexity. In this section, we derive important structural properties of the optimization problem, allowing us to find a closed-form solution in many instances and develop an efficient heuristic to find near-optimal results in other instances.

In the optimization problem, computation of low and high thresholds are tangled due to the presence of the FDA's workload constraint, along with the additional requirement that the low threshold must be less than the high threshold. To gain a deeper understanding of the underlying structural properties, we now focus on a relaxed version of the optimization problem. The relaxed problem is defined by removing the constraint regarding the FDA's workload (constraint (4)) from the optimization problem:

Relaxed Problem

$$\min_{\ell,h} \ \lambda\,\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] + (1-\lambda)\,\mathbb{E}[\mathbb{1}(\delta^- \ge h)] \tag{6}$$

$$\text{s.t. } \mathbb{E}[\mathbb{1}(\delta^+ \ge h)] \ \ge \ \xi^{ru} \tag{7}$$

$$\mathbb{E}[\mathbb{1}(\delta^- \le \ell)] \ \ge \ \xi^{as} \tag{8}$$

$$0 \ \le \ \ell \ \le \ h \ \le \ 1. \tag{9}$$

In the Relaxed Problem, there is an interplay between different components. Specifically, if the rate of rejection for unsafe devices is desired to be high, there would be a corresponding escalation in the rate of rejection for safe devices, resulting in an increase in the second term of the objective function. This interplay stems from the fact that a more stringent rejection criterion for unsafe devices inherently involves an elevated level of algorithmic scrutiny, leading to more rejection of safe devices as well.

Conversely, when the rate of acceptance of safe devices is desired to be high, it results in an increase in the rate of acceptance of unsafe devices, yielding an increase in the first term of the objective function. This connection stems from the fact that a higher acceptance rate for safe devices implies a looser algorithmic screening process, inadvertently leading to a higher acceptance rate for potentially unsafe devices.

To solve the Relaxed Problem, we introduce an auxiliary problem for which we can derive a closed-form solution. We then demonstrate how this auxiliary problem is in essence equivalent to the relaxed problem. The auxiliary problem is:

Auxiliary Problem

$$\max_{\ell,h} \ -\theta\,\ell + (1-\theta)\,h \tag{10}$$

$$\text{s.t. } (7)-(9), \tag{11}$$

where $\theta \in (0,1)$.

The following lemma demonstrates the equivalence between a linear program and the Auxiliary Problem, which is non-linear in general. This equivalence, in turn, allows deriving a closed-form solution for the Auxiliary Problem.

LEMMA 1 **(LP Equivalence)**. *The Auxiliary Problem is equivalent to the following linear program:*

$$\max_{\ell,h} \; -\theta\,\ell + (1-\theta)\,h \tag{12}$$

$$s.t.\; h \le h(\xi^{ru}) \tag{13}$$

$$\ell \ge \ell(\xi^{as}) \tag{14}$$

$$0 \le \ell \le h \le 1, \tag{15}$$

where $h(\xi^{ru}) = \sup\{h \in [0,1] : \mathbb{P}(\delta^+ \ge h) \ge \xi^{ru}\}$ *denotes the highest value of* h *for which the true negative rate is greater than or equal to* ξ^{ru}*, and* $\ell(\xi^{as}) = \inf\{\ell \in [0,1] : \mathbb{P}(\delta^- \le \ell) \ge \xi^{as}\}$ *denotes the smallest value of* ℓ *for which the true positive rate is greater than or equal to* ξ^{as}*.*

Using Lemma 1, we next derive a closed-form solution for the Auxiliary Problem.

PROPOSITION 1 **(Closed-Form Solution)**. *For any* $\theta \in (0,1)$ *and a pair of* $h(\xi^{ru})$ *and* $\ell(\xi^{as})$ *in the Auxiliary Problem, we have:*

(a) *if* $h(\xi^{ru}) > \ell(\xi^{as})$*, then* $(\ell(\xi^{as}), h(\xi^{ru}))$ *is the unique optimal solution,*

(b) *if* $h(\xi^{ru}) < \ell(\xi^{as})$*, then the problem is infeasible,*

(c) *if* $h(\xi^{ru}) = \ell(\xi^{as})$*, then* $\ell = h = h(\xi^{ru}) = \ell(\xi^{as})$ *is the single threshold optimal solution.*

This proposition establishes that when the Auxiliary Problem is feasible, there exists a closed-form optimal solution. Additionally, it indicates that the Auxiliary Problem is decomposable and its optimal solution is independent of the value of the parameter θ, which is not generally true for the optimization problem (1)-(5).

Lastly, we establish a connection between the Relaxed Problem and the Auxiliary Problem by demonstrating that the solution of the Auxiliary Problem can be used to solve the Relaxed Problem.

THEOREM 1 **(Connecting the Relaxed and Auxiliary Problems)**. *For any* $\theta \in (0,1)$ *and* $\lambda \in (0,1)$*, and a pair of* $h(\xi^{ru})$ *and* $\ell(\xi^{as})$*, we have:*

(a) *if* $h(\xi^{ru}) > \ell(\xi^{as})$*, then the two-thresholds optimal solution of the Auxiliary Problem is optimal in the Relaxed Problem,*

(b) *if* $h(\xi^{ru}) < \ell(\xi^{as})$*, then both problems are infeasible,*

(c) *if* $h(\xi^{ru}) = \ell(\xi^{as})$*, then the single threshold optimal solution of the Auxiliary Problem is optimal in the Relaxed Problem.*

4.3. Lagrangian Relaxation

Given the structural properties derived in the prior section, the optimization problem proposed in §4.1 is still challenging to solve due to the workload constraint, which couples the computations of the low and high thresholds. In this section, we present a Lagrangian relaxation method that effectively relaxes the constraints related to workload and thresholds feasibility. To handle violations of the workload constraints and infeasible thresholds, we incorporate Lagrangian penalty terms into the formulation. Lagrangian relaxations have been extensively explored in the literature (see e.g., Brown and Smith 2020, Brown and Zhang 2022, and Liu et al. 2022). A key innovation in the Lagrangian relaxation we consider is the capability to simplify the optimization problem significantly by decomposing it into two straightforward sub-problems.

We start off by defining the primal problem which is equivalent to the main optimization proposed in §4.1:

Primal Problem

$$\begin{aligned}
\min_{\ell,h}\ & \lambda\,\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] + (1-\lambda)\,\mathbb{E}[\mathbb{1}(\delta^- \ge h)] \\
\text{s.t. }\ & h \le h(\xi^{ru}) \\
& \ell \ge \ell(\xi^{as}) \\
& \phi(h) - \phi(\ell) \le p \\
& 0 \le \ell \le h \le 1,
\end{aligned}$$

where $\phi(y) = \mathbb{P}(f(X) \le y)$ is the empirical cumulative distribution function (CDF) of $f(X)$. We note that the workload constraint can also be computed as a function of the four indicators introduced in §4.1. Nevertheless, we continue using the empirical CDF for notation simplicity.

Next, we define the Lagrangian problem by dropping the FDA's workload constraint and the feasibility requirement on the low and high thresholds, and by penalizing their violation in the objective function with Lagrangian multipliers γ_1 and γ_2, respectively.

Lagrangian Primal Problem

For a fixed $\gamma_1 \ge 0$ and $\gamma_2 \ge 0$:

$$\begin{aligned}
\min_{\ell,h}\ & \mathcal{L}(\ell,h,\gamma_1,\gamma_2) = \lambda\,\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] + (1-\lambda)\,\mathbb{E}[\mathbb{1}(\delta^- \ge h)] + \gamma_1\,(\phi(h) - \phi(\ell) - p) + \gamma_2\,(\ell - h) \\
\text{s.t. }\ & h \le h(\xi^{ru}) \\
& \ell \ge \ell(\xi^{as}) \\
& 0 \le \ell, h \le 1.
\end{aligned}$$

The following result provides some useful properties of the Lagrangian Primal Problem.

PROPOSITION 2. *For any non-negative Lagrange multipliers γ_1 and γ_2:*

(a) *The optimal solution of the Lagrangian Primal Problem $(\ell^*_{L-P}, h^*_{L-P})$ can be computed as:*

$$\ell^*_{L-P} = \inf_{\ell \in [\ell(\xi^{as}),\, h(\xi^{ru})]} \left\{\lambda\, \mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] - \gamma_1\, \phi(\ell) + \gamma_2\, \ell\right\},$$
$$h^*_{L-P} = \inf_{h \in [\ell(\xi^{as}),\, h(\xi^{ru})]} \left\{(1-\lambda)\, \mathbb{E}[\mathbb{1}(\delta^- \ge h)] + \gamma_1\, \phi(h) - \gamma_2\, h\right\}.$$

(b) *Let the optimal value of the objective function of the Primal Problem and the Lagrangian Primal Problem be $\mathcal{Z}^*_P$ and $\mathcal{Z}^*_{L-P}(\gamma_1, \gamma_2)$, respectively. Then, we have:*

$$\mathcal{Z}^*_{L-P}(\gamma_1, \gamma_2) \le \mathcal{Z}^*_P.$$

Proposition 2 states that the Lagrangian Primal Problem provides a lower bound for the optimal value function for given Lagrange multipliers. Now it remains to identify Lagrange multipliers that provide the tightest possible lower bound. This can be done by solving the following dual problem:

Lagrangian Dual Problem

$$\begin{aligned} \max_{\gamma_1,\gamma_2} \min_{\ell,h}\ & \mathcal{L}(\ell, h, \gamma_1, \gamma_2) \\ \text{s.t. } & h \le h(\xi^{ru}) \\ & \ell \ge \ell(\xi^{as}) \\ & 0 \le \ell, h \le 1 \\ & \gamma_1, \gamma_2 \ge 0. \end{aligned}$$

Let $\mathcal{Z}^*_{L-D}$ be the optimal objective value function of the Lagrangian Dual Problem. By weak duality, we have, $\mathcal{Z}^*_{L-D} \le \mathcal{Z}^*_P$. Using this fact, we propose a simple and efficient algorithm (Algorithm 1) to iteratively solve the dual problem and find a tight lower bound. We next describe this algorithm.

Description of the Algorithm. The algorithm starts with an initialization phase (steps 1-3) and has three main parts. In the first part (steps 5-7), it solves the Lagrangian Primal Problem via the two sub-problems introduced in Proposition 2. Each sub-problem can be solved efficiently using a grid search. The algorithm then updates the lower bound ($\mathcal{Z}^{(n+1)}_{LB}$) regardless of the feasibility of the optimal solution (ℓ^*_{L-P}, h^*_{L-D}) obtained for the problem. In the second part (steps 8-15), the algorithm first updates the upper bound ($\mathcal{Z}^{(n+1)}_{UB}$) and the current best solutions ($\ell^{(best)}$, $h^{(best)}$) if the optimal solution of the relaxed problem is feasible. Next, it checks the termination condition, where it stops and returns the current best solution if the condition is met. In the last part (step 16), a step size is computed, and the Lagrange multipliers are updated accordingly. The chosen step size is widely employed in practice and has consistently shown empirical success (Fisher 1981,

Algorithm 1 Lagrangian Relaxation Algorithm

Input: a feasible solution $(\ell^{(1)}, h^{(1)})$, tolerance level $\epsilon > 0$, maximum number of iterations N.
Initialize lower bound $\mathcal{Z}_{LB}^{(1)} = 10^{-3}$ and upper bound $\mathcal{Z}_{UB}^{(1)} = \mathcal{Z}_P(\ell^{(1)}, h^{(1)})$.
Initialize Lagrange multipliers $\gamma_1^{(1)} = 0$ and $\gamma_2^{(1)} = 0$.
for $n = 1, 2, \ldots, N$ **do**
    Compute $(\ell^*_{L-P}, h^*_{L-P})$ by solving the following two sub-problems:

$$\ell^*_{L-P} = \inf_{\ell \in [\ell(\xi^{as}),\, h(\xi^{ru})]} \left\{ \lambda\, \mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] - \gamma_1^{(n)}\, \phi(\ell) + \gamma_2^{(n)}\, \ell \right\},$$
$$h^*_{L-P} = \inf_{h \in [\ell(\xi^{as}),\, h(\xi^{ru})]} \left\{ (1-\lambda)\, \mathbb{E}[\mathbb{1}(\delta^- \ge h)] + \gamma_1^{(n)}\, \phi(h) - \gamma_2^{(n)}\, h \right\}.$$

    Set $\mathcal{Z}_{L-P}(\gamma_1^{(n)}, \gamma_2^{(n)}) = \lambda\, \mathbb{E}[\mathbb{1}(\delta^+ \le \ell^*_{L-P})] + (1-\lambda)\, \mathbb{E}[\mathbb{1}(\delta^- \ge h^*_{L-P})] + \gamma_1^{(n)}\, g_1^{(n)} + \gamma_2^{(n)}\, g_2^{(n)}$, where $g_1^{(n)} = \phi(h^*_{L-P}) - \phi(\ell^*_{L-P}) - p$ and $g_2^{(n)} = (\ell^*_{L-P} - h^*_{L-P})$.
    Set $\mathcal{Z}_{LB}^{(n+1)} = \max\left\{ \mathcal{Z}_{LB}^{(n)},\, \mathcal{Z}_{L-P}(\gamma_1^{(n)}, \gamma_2^{(n)}) \right\}$.
    if $g_1^{(n)} \le 0$ and $g_2^{(n)} \le 0$ **then**
        Set $\mathcal{Z}_{UB}^{(n+1)} = \min\left\{ \mathcal{Z}_{UB}^{(n)},\, \lambda\, \mathbb{E}[\mathbb{1}(\delta^+ \le \ell^*_{L-P})] + (1-\lambda)\, \mathbb{E}[\mathbb{1}(\delta^- \ge h^*_{L-P})] \right\}$.
        if $\lambda\, \mathbb{E}[\mathbb{1}(\delta^+ \le \ell^*_{L-P})] + (1-\lambda)\, \mathbb{E}[\mathbb{1}(\delta^- \ge h^*_{L-P})] \le \mathcal{Z}_{UB}^{(n)}$ **then**
            Set $\ell^{(\text{best})} = \ell^*_{L-P}$ and $h^{(\text{best})} = h^*_{L-P}$.
    else
        Set $\mathcal{Z}_{UB}^{(n+1)} = \mathcal{Z}_{UB}^{(n)}$.
    if $(\mathcal{Z}_{UB}^{(n+1)} - \mathcal{Z}_{LB}^{(n+1)}) / \mathcal{Z}_{LB}^{(n+1)} \le \epsilon$ **then**
        Stop and return the best solution $(\ell^{(\text{best})}, h^{(\text{best})})$.
    Update Lagrange multipliers:

$$\gamma_1^{(n+1)} = \max\left\{0, \gamma_1^{(n)} + \alpha^{(n)}\, g_1^{(n)}\right\}, \quad \gamma_2^{(n+1)} = \max\left\{0, \gamma_2^{(n)} + \alpha^{(n)}\, g_2^{(n)}\right\},$$

    where $\alpha^{(n)} = c\, (\mathcal{Z}_{UB}^{(1)} - \mathcal{Z}_{L-P}(\gamma_1^{(n)}, \gamma_2^{(n)})) / \left\| (g_1^{(n)}, g_2^{(n)}) \right\|^2$ and $c \in (0, 2]$.

Held et al. 1974). The updating mechanism is designed intuitively so that if the FDA's workload constraint is violated ($g_1^{(n)} > 0$), then $\gamma_1^{(n+1)}$ increases to impose a higher penalty for violating this constraint. Similarly, if the FDA's workload constraint is not violated ($g_1^{(n)} \le 0$), then $\gamma_1^{(n+1)}$ decreases so as to obtain a better solution that reduces the objective function. The same logic applies when updating $\gamma_2^{(n+1)}$.

Utilizing the structural properties outlined in §4.2 and exploited in Algorithm 1, we present our main theoretical result, which introduces a systematic approach to solving the optimization problem (1)-(5).

THEOREM 2. *Let $\phi(y) = \mathbb{P}(f(X) \le y)$ be the empirical CDF of $f(X)$. Then, for the optimization problem (1)-(5), we have:*

(a) *If $h(\xi^{ru}) > \ell(\xi^{as})$ and $\phi(h(\xi^{ru})) - \phi(\ell(\xi^{as})) \le p$, then $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution.*

(b) *If $h(\xi^{ru}) > \ell(\xi^{as})$ and $\phi(h(\xi^{ru})) - \phi(\ell(\xi^{as})) > p$, then an approximate solution can be obtained by Algorithm 1.*

(c) *If $h(\xi^{ru}) < \ell(\xi^{as})$, then the problem is infeasible.*

(d) *If $h(\xi^{ru}) = \ell(\xi^{as})$, then the problem has a single threshold solution $\ell^* = h^* = h(\xi^{ru}) = \ell(\xi^{as})$.*

5. Policy Recommendation and Impact Evaluation

Using our results from the previous section, we now propose a data-driven clearance policy that leverages a human-algorithm approach to assist the FDA in its decision-making. We also leverage the dataset we have assembled (see Section 2) and investigate the effectiveness of our policy vis-a-vis the FDA's current practice. To this end, we randomly split the data into training data (70%) and testing data (30%), and we use the ML model presented in 3.2 to provide the inputs required for the optimization procedures discussed in the previous section (see Algorithm 1).

When making use of our policy, three parameters should be set by the decision-maker: ξ^{ru}, ξ^{as}, and p, which correspond to the rates of rejection of unsafe devices, the acceptance of safe devices, and the upper bound on the workload, respectively. We employ a cross-validation approach to investigate the impact of these input parameters on various metrics. This can assist decision-makers in selecting the right input parameters that align with their criteria. Table 5 summarizes our results by showing the impact of these parameters on the acceptance and rejection rates of both safe and unsafe devices as well as the FDA's workload which currently stands at 100% as all devices are evaluated by the FDA's committees. In this table, each input parameter is considered at three levels: low (L), medium (M), and high (H). For the parameters ξ^{ru} and ξ^{as}, we have $L = 0.3$, $M = 0.5$, and $H = 0.7$. For the parameter p, the values are $L = 0.4$, $M = 0.6$, and $H = 0.8$.

To provide a clear performance evaluation, we consider a conservative scenario where we assume that the risk labels assigned to deferred devices do not contribute to enhancing the evaluation process conducted by FDA committees. Specifically, we assume that our policy achieves the same recall rate as the current practice for all deferred devices. Thus, our reported estimates of potential impact are conservative since it is likely that the risk labels assigned can themselves enable the FDA's human experts to improve their decisions.

As an example, consider the case where $\xi^{ru} = M$, $\xi^{as} = L$, and $p = M$ (i.e., M-L-M combination). We observe that using our policy results in rejecting 50.1% of unsafe devices and accepting 30.7% of safe devices. Among the accepted devices, 8.3% would experience a future recall, while 19.9%

of the rejected devices would face no future recall. It is worth noting that the rejection rate of unsafe devices, reported in Table 5, can also be interpreted as the percentage improvement in the recall rate achieved by the proposed policy in comparison to the recall rate under the FDA's current practice. This interpretation is valid because the number of unsafe devices rejected under our policy is equivalent to the difference between the number of unsafe devices accepted under the FDA's current practice and the number of unsafe devices accepted under our policy. Accordingly, this policy leads to 50.1% improvement in the recall rate compared to the FDA's current practice. Additionally, it results in a 51.5% reduction in the workload of the FDA's committees, because only 48.5% of the devices will be forwarded to the FDA's committees for in-depth evaluation and the remaining 51.5% will be automatically accepted or rejected.

From Table 5, we observe that there is an interplay between the threshold parameters and their influence on the acceptance and rejection rates for safe and unsafe devices. A higher value of ξ^{ru} leads to a more stringent policy, resulting in an increased rate of rejection for safe devices. Similarly, a higher value of ξ^{as} is associated with a corresponding increase in the acceptance rate of unsafe devices. This highlights the intricate balance required between safety and efficiency in the policy. Additionally, the optimality gap shown in the last column of Table 5 is very low, showcasing our Lagrangian-based algorithm's ability in discovering near-optimal solutions. Finally, we note that the "N/A" values in the last three rows indicate that the corresponding input parameters render the problem infeasible. The infeasibility of these cases stems from the fulfillment of condition (c) in Theorem 1.

We observe substantial gains in certain scenarios, such as the L-L-L combination, which results in a remarkable 74.1% workload reduction and a significant 72.2% recall rate improvement (rejection of unsafe devices). However, it is essential to note a vital caveat. Some combinations that offer significant improvements in workload reduction and recall rate also lead to an unreasonably high rate of rejection of safe devices. This phenomenon arises in the specific setting of low and high thresholds that, in certain cases, reject a substantial proportion of safe devices. Therefore, the selection of input parameters must strike a delicate balance between reducing workload and maintaining an acceptable rate of rejection for safe devices.

A Representative Policy. To gain deeper insights, we next focus on the L-L-M combination as a representative setting, and investigate various aspects of the resulted policy. Figure 7 illustrates the optimized low and high thresholds, as well as the distribution of predicated risk values for unrecalled and recalled devices. The right-hatched region to the left of the low threshold represents the proportion of safe devices accepted by the policy. Similarly, the left-hatched region to the right side of the high threshold represents the proportion of unsafe devices rejected by the policy. Additionally, the overlapping hatched region to the left of the low threshold depicts the proportion

Table 5 Comparative analysis of key metrics across combinations of input parameters

ξ^{ru}	ξ^{as}	p	Reject Unsafe %	Accept Safe %	Accept Unsafe %	Reject Safe %	Workload %	Gap %
L	L	L	72.2	30.7	8.3	42.6	25.9	23.7
L	L	M	34.8	30.7	8.3	9.9	59.1	2.0
L	L	H	29.9	30.7	8.3	7.6	61.7	0.0
L	M	L	35.7	50.4	22.5	10.2	39.6	0.4
L	M	M	29.9	50.4	22.5	7.6	42.6	0.0
L	M	H	29.9	50.4	22.5	7.6	42.6	0.0
L	H	L	29.9	69.3	38.0	7.6	24.1	0.0
L	H	M	29.9	69.3	38.0	7.6	24.1	0.0
L	H	H	29.9	69.3	38.0	7.6	24.1	0.0
M	L	L	72.6	30.7	8.3	44.4	24.3	26.2
M	L	M	50.1	30.7	8.3	19.9	48.5	0.0
M	L	H	50.1	30.7	8.3	19.9	48.5	0.0
M	M	L	50.1	50.4	22.5	19.9	29.4	0.0
M	M	M	50.1	50.4	22.5	19.9	29.4	0.0
M	M	H	50.1	50.4	22.5	19.9	29.4	0.0
M	H	L	50.1	69.3	38.0	19.9	10.9	0.0
M	H	M	50.1	69.3	38.0	19.9	10.9	0.0
M	H	H	50.1	69.3	38.0	19.9	10.9	0.0
H	L	L	69.5	30.7	8.3	38.9	29.5	0.0
H	L	M	69.5	30.7	8.3	38.9	29.5	0.0
H	L	H	69.5	30.7	8.3	38.9	29.5	0.0
H	M	L	69.5	50.4	22.5	38.9	10.3	0.0
H	M	M	69.5	50.4	22.5	38.9	10.3	0.0
H	M	H	69.5	50.4	22.5	38.9	10.3	0.0
H	H	L	“N/A”	“N/A”	“N/A”	“N/A”	“N/A”	“N/A”
H	H	M	“N/A”	“N/A”	“N/A”	“N/A”	“N/A”	“N/A”
H	H	H	“N/A”	“N/A”	“N/A”	“N/A”	“N/A”	“N/A”

of devices falsely accepted, while the overlapping hatched region to the right of the high threshold depicts the proportion of devices falsely rejected.

Table 6 shows the impact of the representative policy. As it can be seen, using this policy leads to a 43.0% reduction in the workload of the FDA's committees and a 38.9% improvement in the recall rate percentage. Investigating the policy, we observe that applicant devices with a recall risk estimated to be below 0.16 are accepted, while applicant devices with a recall risk estimated to be higher than 0.46 are rejected. The other applicant devices are categorized by a risk label and

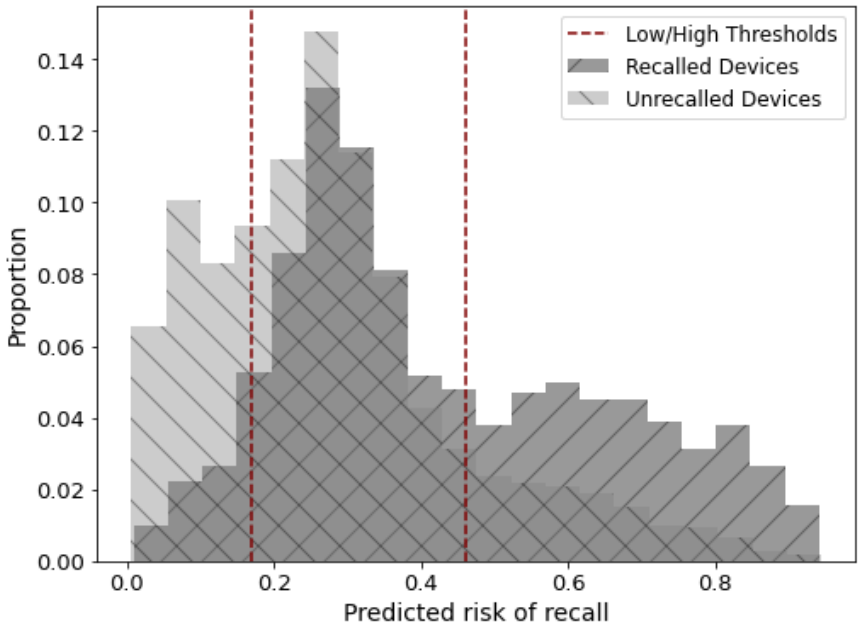


Figure 7 **Optimized low and high thresholds corresponding to $\xi^{ru} = L$, $\xi^{as} = L$, and $p = M$. Distributions of predicted risk values for the unrecalled devices and recalled devices are shown with right- and left-hatched patterns, respectively.**

Table 6 **Impact of the representative policy on workload reduction and improvement in recall rate percentage compared to the FDA's current practice**

ℓ	h	Workload Reduction (%)	Recall Rate Pct. Improvement (%)
0.16	0.46	43.0	38.9

deferred for judgment by an FDA's committee of human experts. Under our conservative evaluation, following this policy results in the rejection of 38.9% of unsafe devices and the acceptance of 28.8% of safe devices. Among the accepted devices, 7.9% will experience a recall in the future, while 13.7% of the rejected devices will have no recalls in the future.

A sensitivity analysis is also conducted to assess the impact of the assumption that risk labels do not enhance the evaluation process conducted by FDA committees. The results highlight the further improvements that risk labels can offer (see Section EC.2 for details).

5.1. Policy Impact: Costs

In the previous sections, we evaluated the impact of our proposed policy in terms of measures such as correct acceptance and rejection rates as well as the resulted workload. In this section, we examine its impact on costs. To this end, we note that a number of independent studies have shown that recalled or prematurely failed devices have likely cost Medicare billions of dollars in recall-related health care expenditures (HHS 2017). In one year alone, the FDA received reports of

nearly three-thousand potential device-related deaths, over one-hundred thousand potential device-related injuries, and over two-hundred thousand adverse event reports concerning medical devices (Zuckerman et al. 2011). Nonetheless, the full extent of injury to patients that is attributable to recalled devices is unknown. Thus, we conservatively assess the cost-savings conforming to the replacement costs of a recalled device. However, it is worth noting that the true cost of a recall may far exceed our calculations (i.e., the replacement costs incurred by the manufacturers) because it will also include potential cost of related injuries. In 2016, for example, a settlement of various state and federal personal injury litigations from recalled pelvic mesh products totaled $121 million (Medtronic 2023). In a separate case in 2016, personal injury claims concerning a bone graft product were settled for $26 million (Medtronic 2016). The fact that these amounts reflect settlements, not affirmative decisions by courts, suggests that full compensation for personal injury could have been even higher.

To our knowledge, there are no previous scholarly publications that have systematically assessed the costs of recalling the spectrum of medical devices. Our review of the literature uncovered two notable publications on this topic. The first publication was a 2017 report by the Office of Inspector General (OIG). Consistent with our own literature review, this report noted that there is no reliable up-to-date estimate of Medicare costs associated with recalled medical devices (HHS 2017). The second publication was a white paper published by the McKinsey Center for Government and cited by the FDA in a presentation discussing the benefit of reducing medical device failure cost (Tack 2021). McKinsey estimated that non-routine events "such as major observations, recalls, warning letters, and consent decrees, along with associated warranties and lawsuits" cost the industry between $2 billion and $5 billion per year on average. The total cost includes $1.5 billion to $3 billion per year on non-routine costs, plus $1 billion to $2 billion in lost sales of new and existing products. This suggests that annual non-routine costs of recalls can range between $0.5 billion to $3 billion per year (Fuhr et al. 2013).

Our assessment of the replacement costs for recalled devices is based on administrative claims data titled "Medicare Durable Medical Equipment, Devices & Supplies" (MDMEDS) for the years 2013-2020. This data is published annually by the Centers for Medicare and Medicaid Services (CMS) and contains information on usage, payments, and submitted charges organized by National Provider Identifier (NPI) and Healthcare Common Procedure Coding System (HCPCS) code (CMS 2021b). The dataset is based on information gathered from CMS administrative claims data for Original Medicare Part B beneficiaries available from the CMS Chronic Conditions Data Warehouse. The data are summarized from 100% final-action Durable Medical Equipment, Prosthetic, Orthotics and Supplies (DMEPOS) non-institutional claim line items (CMS 2021b).

We calculate the replacement costs by medical specialty using the HCPCS codes and descriptions reported in the MDMEDS data. As noted by the OIG report, the replacement cost of recalled medical devices cannot be tracked to individual devices solely with claims data as Medicare claim forms do not contain a field for reporting medical device-specific information (HHS 2017). Consequently, our estimate of replacement costs is specific to the medical specialty, not individual devices. We determined the medical specialty for over 99% of the 1,351 unique HCPCS codes/descriptions in the MDMEDS 2013-2020 data by first identifying keywords in the device classification names available in the 510(k) submission data for each medical specialty and then matching those keywords to the HCPCS descriptions in the MDMEDS data. For example, an HCPCS description containing the word "ostomy," which is a procedure used to treat various diseases of the urinary or digestive systems, was classified as having the medical specialty Gastroenterology/Urology. In a slightly more intricate case, HCPCS descriptions containing the words "glucose monitor" or "glucose" and "monitor" were classified as having the medical specialty Clinical Chemistry. A detailed table showing the crosswalk between keywords and medical specialties is available in Table EC.1.

Once the medical specialties for the HCPCS codes were established, we calculated the average Medicare allowed amount per medical specialty. This calculation was based on the total supplier claims, which reflects the number of products ordered by the referring provider. The Medicare allowed amount includes the amount Medicare paid, the deductible and coinsurance amounts owed by the beneficiary, as well as any amount owed by a third-party payer (CMS 2021a).

We were able to calculate the average Medicare allowed amount for over 75% of the devices in our test dataset based on their respective medical specialties. However, for the devices for which we were unable to compute the average Medicare allowed amount, we faced a challenge in assigning a specific medical specialty to them. Among them, Radiology devices account for 13.7% of all devices in the test dataset and the remaining specialties cumulatively account for less than 8.5% of devices. We believe that the chief reason we could not calculate Medicare costs for these specialties is that the underlying 510(k) devices are not single-use devices, expended on a single patient. In the case of radiology, we examined the majority of 510(k) cleared products between 2013 and 2020 and determined that these devices were either imaging software or multi-use equipment used in providing radiology services such as radiosurgery or radiotherapy. For the purposes of our impact assessment, we created a low and high estimate of the average Medicare allowed amount when we could not directly calculate the average Medicare allowed amount. The low estimate is the lowest average Medicare allowed amount across all of the specialties. The high estimate is the weighted average Medicare allowed amount across all specialties, with weights derived from the frequency of each medical specialty in our testing dataset.

Finally, we calculated the potential cost-savings from implementing our data-driven policy derived from the L-L-M combination (see Figure 7) on the testing set with 9,582 medical devices. In Step 1, we identified 317 recalled devices that were cleared by the FDA but our policy would have rejected if implemented. In Step 2, we determined the number of units being recalled for each of the aforementioned devices using the FDA recall data, which amounted to approximately 64.8 million units. In Step 3, we used the low and high average Medicare allowed amounts to determine the low and high estimated cost-savings that would have occurred had these devices been rejected pursuant to our proposed policy.

Figure 8 shows the percentage of recalls avoided by our proposed policy per medical specialty in the testing set (Step 1). For example, 55.1% for the GU medical specialty indicates that 109 recalls out of 198 recalls corresponding to the GU medical specialty in the testing set were avoided by our proposed policy. As can be seen, Radiology (RA) devices have the highest frequency, followed by Anesthesiology (AN) and Hematology (HE). On the other spectrum, Obstetrics/Gynecology (OB), Clinical Toxicology (TX), and Physical Medicine (PM) have the lowest frequency (zero). Comparing with the recall rate per medical specialty in our dataset (2008-2020), we observe that OB, PM, and TX are among the top four medical specialties in terms of having low recall rates. The alignment between low avoided recalls and low overall recall rates suggests that our policy is functioning as intended, particularly in areas where it is most needed. Table 7 reports the total average cost-savings per medical specialty (Step 3). Among them, Orthopedic (OR), Cardiovascular (CV), and Gastroenterology & Urology (GU) constitute the top three medical specialties with the highest average cost savings, respectively. The results based on our testing set show that the overall cost-savings from implementing our policy range between $7.7 billion and $8.7 billion during the several years in our study period. Given that there are approximately 3,000 510(k) submissions annually (Dubin et al. 2021 and Kadakia et al. 2023), by extrapolating from the total cost-savings in our testing set, a rough estimation of the resulted annual cost-savings is between $2.4 billion and $2.7 billion.

5.2. Post-Hoc Analysis

In this section, we delve deeper into assessing the performance of our proposed policy via post-hoc analyses. We start by examining the characteristics of the deferred medical devices based on some of the top predictors of recall risk. The results are summarized in Table 8. In this table, we also compare devices that are deferred by our proposed policy for more in-depth evaluation to those that have been correctly accepted or rejected. It is evident that the risk estimates are higher for unrecalled devices deferred by our policy compared to those that have been correctly accepted. Similarly, the risk estimates are lower for recalled devices deferred by the policy compared to those

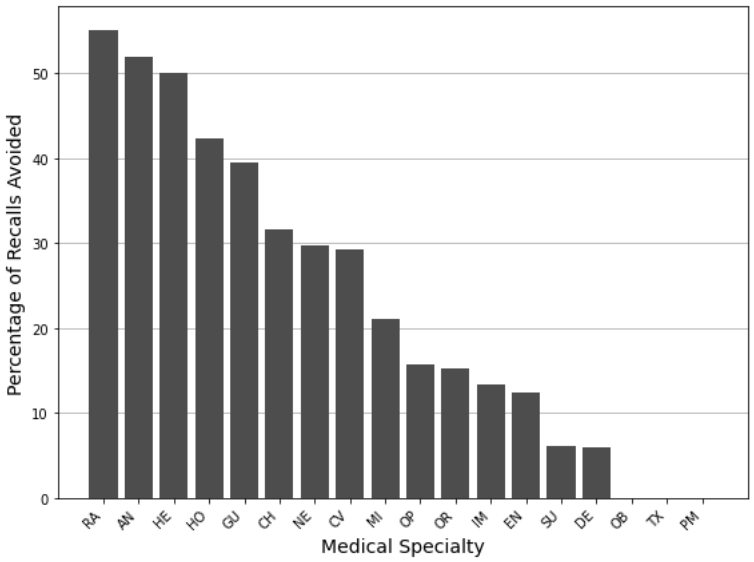


Figure 8 **Impact of our proposed policy in terms of avoiding recalls in the testing set**

Table 7 **Total cost-savings associated with six most cost-saving medical specialties in the testing set**

Medical specialty	Recalls Avoided %	Average Product Quantity	Average Cost-Savings (in 1,000 U.S dollars)
Orthopedic (OR)	15.3	349,325.3	$ 4,723,951
Cardiovascular (CV)	29.3	17,865.0	$ 1,810,240
Gastroenterology Urology (GU)	39.5	1,432,739.3	$ 837,178
Hematology (HE)	50.0	147,978.9	$ 462,295
General Hospital (HO)	42.4	141,564.6	$ 222,750
Ophthalmic (OP)	15.8	39,984.7	$ 46,843.2

that have been correctly rejected. Additionally, significant differences can be observed in variables such as Predicate Average Age, Predicate Newest Age, Number of Recalls, Weighted Number of Recalls, and Variance of Recalls between the deferred devices and the correctly diagnosed ones. These results confirm our hypothesis that while there are some easy cases suitable for direct algorithmic decision-making, some other cases are difficult and require more in-depth evaluation and human expert judgment.

Our post-hoc analysis also provide various other practice-relevant insights. For example, they highlight that while the predicted risk is highly effective in distinguishing between easy and hard cases, the risk estimate alone may not be sufficient for diagnosing the safety of deferred devices. Specifically, our results indicate that deferred devices with higher values for Num. of Recalls and Weighted Num. of Recalls are more likely to be recalled. This aligns with the recent protocols suggested by the FDA (FDA 2023b) in selecting predicates that continue to perform safely, and

Table 8 Comparison of devices correctly evaluated and deferred by our policy in the testing set

	Unrecalled Devices		Recalled Devices	
	Accepted	Deferred	Rejected	Deferred
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Risk Estimate	0.09 (0.04)	0.29 (0.07)	0.67 (0.13)	0.30 (0.07)
Num. of Unmatched Specialties	0.18 (0.44)	0.14 (0.39)	0.11 (0.37)	0.16 (0.39)
Predicate Average Age	5.93 (5.28)	5.36 (4.84)	4.95 (4.07)	5.06 (4.56)
Predicate Newest Age	4.34 (5.21)	3.72 (4.55)	3.05 (3.31)	3.24 (4.01)
Predicate Oldest Age	7.74 (6.62)	7.21 (6.48)	7.29 (6.27)	7.32 (6.80)
Num. of Recalls	0.15 (1.29)	0.18 (0.68)	2.30 (3.10)	0.48 (3.33)
Weighted Num. of Recalls	0.08 (0.79)	1.30 (3.18)	1.66 (2.23)	0.28 (1.83)
Variance of Recalls	0.01 (0.15)	0.01 (0.04)	0.18 (0.58)	0.02 (0.19)

highlights that not only the number of recalls matters for selected predicates but also the timing of their recalls.

Table 9 presents a comparison of devices accepted by our proposed policy and those accepted based on the FDA's current practice in the testing set. As our policy does not directly diagnose the deferred devices, this analysis is conducted under the assumption that the deferred devices are evaluated following the FDA's current practice. We observe from our results that the risk estimate for devices accepted by our proposed policy is lower compared to devices accepted by the FDA's current practice. The lower risk estimate indicates that our policy is cautious about accepting devices with a potentially higher risk. In addition, the age of the latest-approved predicate for devices accepted under our policy is slightly higher compared to devices accepted under the FDA's current practice, while the age of the earliest-approved predicate is slightly lower. This indicates a careful balance between leveraging proven safety records and embracing new technology. Furthermore, the number of recalls, weighted number of recalls, and variance of predicates' recalls for devices accepted by our policy is significantly lower than those accepted by FDA's current practice. This is a substantial difference and indicates that our policy has the tendency to accept devices with predicates with fewer historical recalls, which aligns with the existing literature suggesting the correlation between the chance of recall and the number of recalls for predicates. Finally, we observe that the number of unmatched specialties and the oldest age of predicates do not exhibit a significant difference between devices accepted by our policy and the current practice.

6. Policy and Managerial Implications

Our work investigates the degree to which a data-driven policy can assist the FDA in improving its 510(k) medical device clearance process. We hypothesized that our combined human-algorithm approach to evaluating medical devices will result in a reduction of recall rates and the workload

Table 9 Comparison of accepted devices by our policy and current practice in the testing set

	Our Policy	Current Practice
	Mean (SD)	Mean (SD)
Risk Estimate	0.23 (0.11)	0.29 (0.19)
Num. of Unmatched Specialties	0.15 (0.40)	0.15 (0.40)
Predicate Average Age	5.53 (4.98)	5.52 (4.91)
Predicate Newest Age	3.89 (4.76)	3.79 (4.62)
Predicate Oldest Age	7.40 (6.57)	7.50 (6.62)
Num. of Recalls	0.19 (1.29)	0.51 (2.40)
Weighted Num. of Recalls	0.10 (0.72)	0.32 (1.45)
Variance of Recalls	0.01 (0.14)	0.05 (1.11)

burden imposed on the FDA. Conducting an in-depth evaluation of the performance of our policy as well as the FDA's current practice, we found that while the predicted risk is highly effective in distinguishing between easy and hard cases, utilizing it alone may not be sufficient. This confirms our hypothesis that there is a need for a combined human-algorithm approach, where devices with a mid-range predicted risk of recall (non-easy cases) are deferred to human experts for further evaluation. That is, integrating the FDA's expertise with quantitative evidence is required to improve the 510(k) medical device clearance process. A conservative evaluation of our proposed policy based on our data showed a 38.9% improvement in the recall rate and a 43.0% reduction in the FDA's workload. Our cost analyses projected that implementing our policy could result in significant annual cost-savings ranging between $2.4 billion and $2.7 billion. Overall, these findings imply that our proposed policy is effective in significantly reducing recall rates, the FDA's workload, and costs incurred due to recalls.

We believe that our work addresses some of the main concerns about the current 510(k) process. Most recently, the FDA has responded to these concerns with modernization efforts "to improve the safety of medical devices while continuing to create more efficient pathways to bring critical devices to patients" (FDA 2018b). In 2018, the FDA proposed "promoting innovation and improving safety by driving innovators toward reliance on more modern predicate devices." That is, the newer devices should be compared to the benefits and risks of more modern technology as the as older predicates might not reflect the advanced technology embedded in new devices or align with the FDA's current understanding of device benefits and risks (FDA 2018c). However, the FDA ultimately acknowledged that its initial proposal "may not optimally promote safer and more effective devices" in all instances since devices that use modern rapidly-evolving technology could benefit from comparison to more recent predicates whereas older devices may establish a history of safety and effective use in other cases. We found that a higher value of the newest predicate age is

linked to a reduced recall risk. Specifically, devices with a predicate age of less than five years are associated with a significantly higher risk of recall, possibly due to their limited real-world usage that may reveal potential issues over time. Regarding the oldest predicate age, our results indicated that both high and low values can predict a high recall risk. This aligns with discussions in the literature regarding the benefits and risks of using older predicates. Older predicates may signify safety and be a standard for patient care, but they might not reflect the latest technology, leading to compatibility issues (FDA 2018b, FDA 2018c). We also found that lower values of average and median predicate age generally correlate with a lower recall risk.

Recently, the FDA developed a new draft guidance in 2023 on best practices for selecting predicate devices based on device characteristics rather than just age (FDA 2023b). Three of the FDA's proposed best practices include (1) selecting predicates that continue to perform safely and as intended, (2) selecting predicates that do not have unmitigated use-related or design-related safety issues, and (3) selecting predicates that have not been subject to a design-related recall. We found that the number of recall events for predicates is a significant predictor of recall risk for an applicant device. This finding specifically aligns with the FDA's proposed practice regarding the selection of predicates that continue to perform safely. Additionally, our results emphasize the importance of the timing of predicates' recall events. These insights call for a more targeted approach towards predicates with recent recalls. Our hypothesis is that manufacturers might not have had sufficient time to address potential issues with these predicates. Consequently, applicant devices resembling such predicates may face recall due to similar issues. Shifting focus to the characteristics of the applicant devices, we observed that some product codes, country codes, and medical specialties are crucial predictors of recall risk. The results are consistent with prior literature suggesting the heterogeneous effects of these indicators on the risk of recall. Furthermore, our study confirms the importance of variables that quantify risk, such as life-sustaining. Our conjecture is that riskier applicant devices are generally under stringent market scrutiny and are more likely to be recalled.

Our research also addresses another of the FDA's recent inquiries on the use of AI/ML. In 2023, the FDA published a discussion paper seeking feedback on the opportunities and challenges of using of AI/ML in the development of drugs and medical devices intended to be used with drugs (FDA 2023f). Our research addresses the crucial considerations relevant to model development, performance, monitoring, and validation within the context of utilizing AI/ML to enhance the evaluation of medical devices.

One of the specific benefits of our proposed policy is that it has a transparent structure (two phases based on clear rules), which allows for continued transparency of the 510(k) review process. Second, it benefits from explainability, which is an important element in allowing stakeholders, such as regulators and health providers, to sanity check models beyond mere performance (Amann

et al. 2020, Zech et al. 2018, Saghafian and Hopp 2019). Third, in developing and evaluating our proposed policy, we paid specific attention to selecting appropriate metrics, such as false positive and false native rates and the FDA's workload. We highlight that relying solely on one of these metrics may not provide a complete picture of the reality. For example, false positives and false negatives can have significantly different consequences. False positives may lead to unnecessary rejection of safe devices, while false negatives could result in missed diagnoses of unsafe devices.

Finally, our proposed policy follows findings from the recent literature that suggest combining human intuition and judgment with the power of AI/ML algorithms by creating human-algorithm "centaurs" can go a long way (Orfanoudaki et al. 2022, Saghafian 2023). Our proposed policy creates a combined human-algorithm approach by deferring some decisions to human exerts and others to a well-trained algorithm. Our results show that this approach can be very effective, mainly because it makes use of both the expertise of FDA committees when needed and the power of an algorithm that can accurately predict future recall risks for a variety (but not all) of devices.

It is worth discussing possible concerns regarding the practicality of implementing our proposed policy, including (i) FDA's ability to leverage our data-driven clearance policy, and (ii) reaction of applicants to the 510(k) pathway. Regarding point (i), we note that the FDA's recent draft guidance on best practices for selecting a predicate device indicates that our policy would align with the FDA's avowed commitment "to improve the predictability, consistency and transparency" of the 510(k) pathway while not proposing changes to applicable statutory and regulatory standards, such as how the FDA evaluates substantial equivalence, or the applicable requirements, including the requirement for valid scientific evidence" (FDA 2018b). As discussed earlier, three of the FDA's proposed best practices are an attempt to improve the safety of medical devices by mitigating the use of predicates with safety issues. Our data-driven policy uses established AI/ML methods to provide additional scientific evidence, namely an estimated recall risk of a 510(k) applicant device based, in part, on safety issues present in predicate devices. Moreover, our clearance policy is not intended to be a substitute for the FDA's expertise in assessing applicant devices. It is a tool that the FDA can leverage, as it sees fit, to increase device safety while reducing workload. The FDA, among other approaches, can choose to use our tool as a guide for identifying devices that warrant higher priority concerns rather than outright rejection.

Regarding point (ii), we note that similar to the FDA, applicants will have the capability to replicate our model. This proactive response can, in turn, decrease the workload on the FDA during the review process and reduce the risk of recalls after a device is cleared. If, on the other hand, applicants begin to only mask factors that increase the risk of rejection (e.g., changing the country of manufacture with no other substantive change), our ML predictor will adapt accordingly as the masking becomes more prominent in the training data over time. Furthermore, as discussed earlier,

our clearance policy is neither intended to be a substitute for the FDA's expertise in assessing applicant devices, nor is it intended to be a blackbox. We expect that the FDA will continue its commitment to transparency when communicating the reasons for a rejection, including the prediction and variable importance in our model. Moreover, FDA rejections of 510(k) applicants are specific to the application under review. Applicants are further able to seek clearance by submitting a new application that addresses the FDA's concerns.

7. Conclusion

This research aims to enhance safety and expedite clearance procedures within the 510(k) pathway. Our primary focus is on introducing a data-driven clearance policy intended to assist the FDA in refining the 510(k) medical device clearance process. Our methodology and approach are informed by comprehensive discussions with our collaborator, who possesses substantial experience in various FDA regulatory consulting projects. Our modeling framework enables the FDA to integrate its expertise with quantitative evidence. While it does not prescribe a specific course of action for devices that warrant further evaluation, it allows for the incorporation of the FDA experts' judgment when necessary. Focusing the FDA's expertise on devices requiring the most attention can significantly enhance the evaluation process, improving patient safety and reducing unnecessary workload for the FDA.

Our results suggest that that our methodology can lead to significant potential improvements in the recall rates and the FDA's workload. Despite this, our projections indicate a relatively low percentage of acceptance of unsafe devices. Our analysis also suggests that using our policy can lead to a substantial reduction in costs, alongside a decrease in adverse event outcomes and an enhancement in patient safety.

We believe future studies are required to further investigate the impacts of implementing our proposed policy. For example, as discussed earlier, we only provide conservative estimates on the cost-savings due to our proposed policy. Specifically, our assessment of the replacement costs for recalled devices likely underestimates the true cost of a recall by excluding the potential cost of adverse events. An expansion on our current work can assess the amount of injury and death that could have been averted under our proposed policy. Finally, we believe a future study could run an experiment, where 510(k) applicant devices are randomly assigned to either our proposed policy or the current FDA procedure. Data collected on such a randomized experiment can further inform the FDA about the advantages and disadvantages of our policy, and shed light on needed modifications prior to full implementation.

Acknowledgments

The authors would like to thank Anders Olsen for his assistance with data curation. They also thank George Ball for his thoughtful feedback on the manuscript. Additionally, the authors are grateful for the constructive comments and helpful suggestions received during the presentation of this work at several conferences.

References

Ahsen ME, Ayvaci MUS, Raghunathan S (2019) When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Information Systems Research* 30(1):97–116.

Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Consortium P (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making* 20:1–9.

Ang YQ, Chia A, Saghafian S (2022) *Using machine learning to demystify startups' funding, post-money valuation, and success* (Springer).

Association AM, et al. (2019) Augmented intelligence in health care. *AMA https://www. ama-assn. org/system/files/2019-08/ai-2018-board-policy-summary. pdf* .

Ayvaci MU, Alagoz O, Burnside ES (2012) The effect of budgetary restrictions on breast cancer diagnostic decisions. *Manufacturing & Service Operations Management* 14(4):600–617.

Ball G, Macher JT, Stern AD (2018) Responding strategically to competitors' failures: Evidence from medical device recalls & new product submissions .

Bayati M, Bhaskar S, Montanari A (2018) Statistical analysis of a low cost method for multiple disease prediction. *Statistical methods in medical research* 27(8):2312–2328.

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model+ 50% manager. *Management science* 36(8):887–899.

Boloori A, Saghafian S, Traub S (2022) Understanding the opioid epidemic: Human-based versus algorithmic-based perceptions, treatments, and guidelines. *Treatments, and Guidelines (December 9, 2022)* .

Brown DB, Smith JE (2020) Index policies and performance bounds for dynamic selection problems. *Management Science* 66(7):3029–3050.

Brown DB, Zhang J (2022) Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Operations Research* 70(5):3015–3033.

Challoner DR, Senate U (2011) Medical devices and the public's health: the fda 510 (k) clearance process at 35 years. *Written Statement before the Committee on Health, Education, Labor, and Pensions US Senate, Institute of Medicine of the National Academics, Washington* .

CMS (2021a) Medicare durable medical equipment, devices & supplies - by referring provider and service. URL https://data.cms.gov/resources/medicare-durable-medical-equipment-devices-supplies-by-referring-provider-and-service-data-dictionary, accessed on June 12, 2024.

CMS (2021b) Medicare durable medical equipment, devices & supplies - by referring provider and service - centers for medicare & medicaid services data. URL https://data.cms.gov/provider-summary-by-type-of-service/medicare-durable-medical-equipment-devices-supplies/medicare-durable-medical-equipment-devices-supplies-by-referring-provider-and-service, accessed on June 12, 2024.

Connor MJ, Tringale K, Moiseenko V, Marshall DC, Moore K, Cervino L, Atwood T, Brown D, Mundt AJ, Pawlicki T, et al. (2017) Medical device recalls in radiation oncology: analysis of us food and drug administration data, 2002-2015. *International Journal of Radiation Oncology* Biology* Physics* 98(2):438–446.

Daugherty PR, Wilson HJ (2018) *Human+ machine: Reimagining work in the age of AI* (Harvard Business Press).

Day CS, Park DJ, Rozenshteyn FS, Owusu-Sarpong N, Gonzalez A (2016) Analysis of fda-approved orthopaedic devices and their recalls. *JBJS* 98(6):517–524.

Dubin JR, Simon SD, Norrell K, Perera J, Gowen J, Cil A (2021) Risk of recall among medical devices undergoing us food and drug administration 510 (k) clearance and premarket approval, 2008-2017. *JAMA Network Open* 4(5):e217274–e217274.

Everhart AO, Sen S, Stern AD, Zhu Y, Karaca-Mandic P (2023) Association between regulatory submission characteristics and recalls of medical devices receiving 510 (k) clearance. *JAMA* 329(2):144–156.

FDA (2018a) The device development process - pathway to approval. URL https://www.fda.gov/patients/device-development-process/step-3-pathway-approval, accessed on June 12, 2024.

FDA (2018b) Medical device safety action plan: Protecting patients, promoting public health. URL https://www.fda.gov/files/about%20fda/published/Medical-Device-Safety-Action-Plan--Protecting-Patients--Promoting-Public-Health-%28PDF%29.pdf, accessed on June 12, 2024.

FDA (2018c) Statement from an fda commissioner on transformative new steps to modernize fda's 510(k) program to advance the review of the safety and effectiveness of medical devices. URL https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-and-jeff-shuren-md-director-center-devices-and, accessed on June 12, 2024.

FDA (2019) Safety and performance-based pathway - guidance for industry and food and drug administration. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/safety-and-performance-based-pathway, accessed on June 12, 2024.

FDA (2022a) How to study and market your device. URL https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/how-study-and-market-your-device, accessed on June 12, 2024.

FDA (2022b) Premarket notification 510(k). URL https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k, accessed on June 12, 2024.

FDA (2023a) 510(k) devices cleared in 2022. URL https://www.fda.gov/medical-devices/510k-clearances/510k-devices-cleared-2022, accessed on June 12, 2024.

FDA (2023b) Best practices for selecting a predicate device to support premarket notification 510(k) submission. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/best-practices-selecting-predicate-device-support-premarket-notification-510k-submission, accessed on June 12, 2024.

FDA (2023c) Device classification under section 513(f)(2)(de novo). URL https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/denovo.cfm, accessed on June 12, 2024.

FDA (2023d) Devices approved in 2022. URL https://www.fda.gov/medical-devices/pma-approvals/devices-approved-2022, accessed on June 12, 2024.

FDA (2023e) Humanitarian device exemption (hde). URL https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfHDE/hde.cfm, accessed on June 12, 2024.

FDA (2023f) Using artificial intelligence and machine learning in the development of drug and biological products. URL https://www.federalregister.gov/documents/2023/05/11/2023-09985/using-artificial-intelligence-and-machine-learning-in-the-development-of-drug-and-biological, accessed on June 12, 2024.

Felder S, Mayrhofer T (2014) Risk preferences: consequences for test and treatment thresholds and optimal cutoffs. *Medical Decision Making* 34(1):33–41.

Fisher ML (1981) The lagrangian relaxation method for solving integer programming problems. *Management science* 27(1):1–18.

Fuhr T, George K, Pai J (2013) The business case for medical device quality. *McKinsey Center for Government* .

Garcia GGP, Lavieri MS, Jiang R, McCrea MA, McAllister TW, Broglio SP, Investigators CC, et al. (2020) Data-driven stochastic optimization approaches to determine decision thresholds for risk estimation models. *IISE transactions* 52(10):1098–1121.

Goodwin P (2000) Correct or combine? mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting* 16(2):261–275.

Hajian-Tilaki K (2013) Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine* 4(2):627.

Held M, Wolfe P, Crowder HP (1974) Validation of subgradient optimization. *Mathematical programming* 6:62–88.

HHS (2017) Shortcomings of device claims data complicate and potentially increase medicare costs for recalled and prematurely-failed devices. URL https://oig.hhs.gov/oas/reports/region1/11500504.asp, accessed on June 12, 2024.

Hong H, Guo C, Liu ZH, Wang BJ, Zhou SZ, Mu DL, Wang DX (2021) The diagnostic threshold of cornell assessment of pediatric delirium in detection of postoperative delirium in pediatric surgical patients. *BMC pediatrics* 21:1–8.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Janetos TM, Ghobadi CW, Xu S, Walter JR (2017) Overview of high-risk medical device recalls in obstetrics and gynecology from 2002 through 2016: implications for device safety. *American journal of obstetrics and gynecology* 217(1):42–46.

Jund J, Rabilloud M, Wallon M, Ecochard R (2005) Methods to estimate the optimal threshold for normally or log-normally distributed biological tests. *Medical decision making* 25(4):406–415.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* 32(3):713–735.

Kadakia KT, Dhruva SS, Caraballo C, Ross JS, Krumholz HM (2023) Use of recalled devices in new device authorizations under the us food and drug administration's 510 (k) pathway and risk of subsequent recalls. *JAMA* 329(2):136–143.

Kramer DB, Yeh RW (2023) Quantitative analyses of regulatory policies for medical devices: Matching the methods to the moment. *JAMA* 329(6):467–469.

Liu S, Shen ZJM, Ji X (2022) Urban bike lane planning with bike trajectories: Models, algorithms, and a real-world case study. *Manufacturing & Service Operations Management* 24(5):2500–2515.

Liu Y, Zhang H, Zeng L, Wu W, Zhang C (2018) Mlbench: benchmarking machine learning services against human experts. *Proceedings of the VLDB Endowment* 11(10):1220–1232.

Lohr vs Medtronic (1996) 56 f3d 1335 (11th cir 1995). URL https://supreme.justia.com/cases/federal/us/518/470/case.pdf, accessed on June 12, 2024.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* 2(1):56–67.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Medtronic (2016) Form 10-k for the fiscal year ended april 29.

Medtronic (2023) Form 10-k for the fiscal year ended april 28.

Miller SM (2018) Ai: Augmentation, more so than automation .

Mukherjee UK, Sinha KK (2018) Product recall decisions in medical device supply chains: a big data analytic approach to evaluating judgment bias. *Production and Operations Management* 27(10):1816–1833.

Muller E (2022) How ai-human symbiotes may reinvent innovation and what the new centaurs will mean for cities. *Technology and Investment* 13(1):1–19.

Orfanoudaki A, Saghafian S, Song K, Chakkera HA, Cook C (2022) Algorithm, human, or the centaur: How to enhance clinical care? *Available at SSRN 4302002* .

Saghafian S (2023) Effective generative ai: The human-algorithm centaur. *Available at SSRN 4587250* .

Saghafian S, Hopp WJ (2019) The role of quality transparency in health care: Challenges and potential solutions. *NAM perspectives* 2019.

Saghafian S, Hopp WJ, Iravani SM, Cheng Y, Diermeier D (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang PH, Ming WK, et al. (2019) Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR medical informatics* 7(3):e10010.

Sheppard JW, Kaufman MA (2005) A bayesian approach to diagnosis and prognosis using built-in test. *IEEE Transactions on Instrumentation and Measurement* 54(3):1003–1018.

Si B, Yakushev I, Li J (2017) A sequential tree-based classifier for personalized biomarker testing of alzheimer's disease risk. *IISE Transactions on Healthcare Systems Engineering* 7(4):248–260.

Tack LC (2021) Uncovering and maximizing the value of fda inspections. URL https://fda.report/media/150417/8.+Uncovering+and+Maximizing+the+Value+of+FDA+Inspections_+-+LT+Colin+Tack.pdf, accessed on June 12, 2024.

Talati RK, Gupta AS, Xu S, Ghobadi CW (2018) Major fda medical device recalls in ophthalmology from 2003 to 2015. *Canadian Journal of Ophthalmology* 53(2):98–103.

van Giessen A, de Wit GA, Moons KG, Dorresteijn JA, Koffijberg H (2018) An alternative approach identified optimal risk thresholds for treatment indication: an illustration in coronary heart disease. *Journal of clinical epidemiology* 94:122–131.

Weise K, Hübel K, Rose E, Schläger M, Schrammel D, Täschner M, Michel R (2006) Bayesian decision threshold, detection limit and confidence limits in ionising-radiation measurement. *Radiation protection dosimetry* 121(1):52–63.

Wowak KD, Ball GP, Post C, Ketchen Jr DJ (2021) The influence of female directors on product recall decisions. *Manufacturing & Service Operations Management* 23(4):895–913.

Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L (2022) A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135:364–381.

Yao Y (2010) Three-way decisions with probabilistic rough sets. *Information sciences* 180(3):341–353.

Yao Y, Zhou B (2016) Two bayesian approaches to rough sets. *European Journal of Operational Research* 251(3):904–917.

Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* 15(11):e1002683.

Zhang J, Denton BT, Balasubramanian H, Shah ND, Inman BA (2012) Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management* 14(4):529–547.

Zuckerman D, Brown P, Das A (2014) Lack of publicly available scientific evidence on the safety and effectiveness of implanted medical devices. *JAMA Internal Medicine* 174(11):1781–1787.

Zuckerman DM, Brown P, Nissen SE (2011) Medical device recalls and the fda approval process. *Archives of internal medicine* 171(11):1006–1011.

Electronic Companion

EC.1. Proofs

All proofs for lemmas, propositions, and theorems are given below.

LEMMA 1. *The Auxiliary Problem is equivalent to the following linear program:*

$$\begin{aligned} \max_{\ell,h} \; & -\theta\,\ell + (1-\theta)\,h \\ s.t. \; & h \le h(\xi^{ru}) \\ & \ell \ge \ell(\xi^{as}) \\ & 0 \le \ell \le h \le 1, \end{aligned}$$

where $h(\xi^{ru}) = \sup\{h \in [0,1] : \mathbb{P}(\delta^+ \ge h) \ge \xi^{ru}\}$ *denotes the highest value of* h *for which the true negative rate is greater than or equal to* ξ^{ru}*, and* $\ell(\xi^{as}) = \inf\{\ell \in [0,1] : \mathbb{P}(\delta^- \le \ell) \ge \xi^{as}\}$ *denotes the smallest value of* ℓ *for which the true positive rate is greater than or equal to* ξ^{as}*.*

Proof of Lemma 1: We need to show that constraints (7)-(9) can be written as linear functions. First, we rewrite the constraints (7)-(9) as chance constraints:

$$\begin{aligned} \mathbb{E}[\mathbb{1}(\delta^+ \ge h)] \ge \xi^{ru} \iff \mathbb{P}(\delta^+ \ge h) \ge \xi^{ru}, \\ \mathbb{E}[\mathbb{1}(\delta^- \le \ell)] \ge \xi^{as} \iff \mathbb{P}(\delta^- \le \ell) \ge \xi^{as}. \end{aligned}$$

Next, we define the following notation:

$$\begin{aligned} h(\xi^{ru}) &= \sup\left\{h \in [0,1] : \mathbb{P}(\delta^+ \ge h) \ge \xi^{ru}\right\}, \\ \ell(\xi^{as}) &= \inf\left\{\ell \in [0,1] : \mathbb{P}(\delta^- \le \ell) \ge \xi^{as}\right\}. \end{aligned}$$

Note that $h, \ell \in [0,1]$ and interval $[0,1]$ is convex and compact. Supremum is attained since $\mathbb{P}(\delta^+ \ge h)$ is continuous and weakly decreasing in h. Similarly, infimum is attained since $\mathbb{P}(\delta^- \le \ell)$ is continuous and weakly increasing in ℓ.

Accordingly, the proof is completed by the following results:

$$\mathbb{P}(\delta^+ \ge h) \ge \xi^{ru} \iff h \le h(\xi^{ru}), \text{ and } \mathbb{P}(\delta^- \le \ell) \ge \xi^{as} \iff \ell \ge \ell(\xi^{as}).$$

Q.E.D.

PROPOSITION 1. *For any $\theta \in (0,1)$ and a pair of $h(\xi^{ru})$ and $\ell(\xi^{as})$ in the Auxiliary Problem, we have:*

(a) *if $h(\xi^{ru}) > \ell(\xi^{as})$, then $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution,*

(b) *if $h(\xi^{ru}) < \ell(\xi^{as})$, then the problem is infeasible,*

(c) *if $h(\xi^{ru}) = \ell(\xi^{as})$, then $\ell = h = h(\xi^{ru}) = \ell(\xi^{as})$ is the single threshold optimal solution.*

Proof of Proposition 1: We prove each case separately.

Case (a). We first find an optimal solution, and then we prove its uniqueness. When $h(\xi^{ru}) > \ell(\xi^{as})$, the polyhedral feasible region of the problem has three extreme points: $(\ell(\xi^{as}), \ell(\xi^{as}))$, $(\ell(\xi^{as}), h(\xi^{ru}))$, and $(h(\xi^{ru}), h(\xi^{ru}))$. The objective values corresponding to these extreme points are $\ell(\xi^{as})(1-2\theta)$, $-\theta\,\ell(\xi^{as}) + (1-\theta)h(\xi^{ru})$, and $h(\xi^{ru})(1-2\theta)$, respectively. For any $\theta \in (0,1)$, we have $h(\xi^{as})(1-2\theta) > \ell(\xi^{as})(1-2\theta)$ because $h(\xi^{ru}) > \ell(\xi^{as})$. Also, for any $\theta \in (0,1)$, we have $-\theta\,\ell(\xi^{as}) + (1-\theta)h(\xi^{ru}) > h(\xi^{ru})(1-2\theta)$ when $h(\xi^{ru}) > \ell(\xi^{as})$. Accordingly, $(\ell(\xi^{as}), h(\xi^{ru}))$ is an optimal solution for the problem.

Next, we use a contradiction argument to prove that $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution. Suppose that there is another optimal solution $(\bar{\ell}, \bar{h}) \neq (\ell(\xi^{as}), h(\xi^{ru}))$. According to the polyhedral feasible region, there are two possible scenarios: (1) $\bar{h} < h(\xi^{ru})$ and $\bar{\ell} \geq \ell(\xi^{as})$, or (2) $\bar{h} \leq h(\xi^{ru})$ and $\bar{\ell} > \ell(\xi^{as})$. In both scenarios, we have:

$$-\theta\,\bar{\ell} + (1-\theta)\,\bar{h} < -\theta\,\ell(\xi^{as}) + (1-\theta)\,h(\xi^{ru}).$$

This is a contradiction on the optimality of $(\bar{\ell}, \bar{h})$. Thus, we conclude that $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution.

Case (b). The condition of $h(\xi^{ru}) < \ell(\xi^{as})$, results in $h < \ell$ which contradicts with the requirement of $0 \leq \ell \leq h \leq 1$. Thus, the problem is infeasible.

Case (c). This is a direct result of Case (a).

Q.E.D.

THEOREM 1. *For any $\theta \in (0,1)$ and $\lambda \in (0,1)$, and a pair of $h(\xi^{ru})$ and $\ell(\xi^{as})$, we have:*

(a) *if $h(\xi^{ru}) > \ell(\xi^{as})$, then the two-thresholds optimal solution of the Auxiliary Problem is optimal in the Relaxed Problem,*

(b) *if $h(\xi^{ru}) < \ell(\xi^{as})$, then both problems are infeasible,*

(c) *if $h(\xi^{ru}) = \ell(\xi^{as})$, then the single threshold optimal solution of the Auxiliary Problem is optimal in the Relaxed Problem.*

Proof of Theorem 1: First, we highlight that both problems have the same polyhedral feasible region. Hence, any feasible solution of the Auxiliary Problem is also a feasible solution to the Relaxed Problem.

Next, we show that both problems have the same optimal solutions in each case.

Case (a). By Proposition 1, when $h(\xi^{ru}) > \ell(\xi^{as})$, we have that $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution of the Auxiliary Problem for any $\theta \in (0,1)$. The objective function of the Relaxed Problem is the convex combination of $\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)]$ and $\mathbb{E}[\mathbb{1}(\delta^- \ge h)]$, which are monotone functions. Since $\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)]$ is weakly increasing in ℓ, we have $\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] \ge \mathbb{E}[\mathbb{1}(\delta^+ \le \ell(\xi^{as}))]$ for any $\ell \ge \ell(\xi^{as})$. Similarly, since $\mathbb{E}[\mathbb{1}(\delta^- \ge h)]$ is weakly decreasing in h, we have $\mathbb{E}[\mathbb{1}(\delta^- \ge h)] \ge \mathbb{E}[\mathbb{1}(\delta^- \ge h(\xi^{ru}))]$ for any $h \le h(\xi^{ru})$. Accordingly, for any feasible pair of ℓ and h, we have:

$$\lambda\,\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] + (1-\lambda)\,\mathbb{E}[\mathbb{1}(\delta^- \ge h))] \;\ge\; \lambda\,\mathbb{E}[\mathbb{1}(\delta^+ \le \ell(\xi^{as}))] + (1-\lambda)\,\mathbb{E}[\mathbb{1}(\delta^- \ge h(\xi^{ru})].$$

Therefore, we can conclude that $(\ell(\xi^{as}), h(\xi^{ru}))$ is also the optimal solution of the Relaxed Problem.

Case (b). By Proposition 1, when $h(\xi^{ru}) > \ell(\xi^{as})$, the Auxiliary Problem is infeasible. The Relaxed Problem is also infeasible since both problems are restricted to the same set of constraints.

Case (c). This is a direct result of Case (a).

Q.E.D.

PROPOSITION 2. *For any non-negative Lagrange multipliers γ_1 and γ_2:*

(a) *The optimal solution of the Lagrangian Primal Problem $(\ell^*_{L-P}, h^*_{L-P})$ can be computed as:*

$$\ell^*_{L-P} = \inf_{\ell \in [\ell(\xi^{as}),\, h(\xi^{ru})]} \left\{\lambda\,\mathbb{E}[\mathbb{1}(\delta^+ \le \ell)] - \gamma_1\,\phi(\ell) + \gamma_2\,\ell\right\},$$
$$h^*_{L-P} = \inf_{h \in [\ell(\xi^{as}),\, h(\xi^{ru})]} \left\{(1-\lambda)\,\mathbb{E}[\mathbb{1}(\delta^- \ge h)] + \gamma_1\,\phi(h) - \gamma_2\,h\right\}.$$

(b) *Let the optimal value of the objective function of the Primal Problem and the Lagrangian Primal Problem be $\mathcal{Z}^*_P$ and $\mathcal{Z}^*_{L-P}(\gamma_1,\gamma_2)$, respectively. Then, we have:*

$$\mathcal{Z}^*_{L-P}(\gamma_1,\gamma_2) \le \mathcal{Z}^*_P.$$

Proof of Proposition 2: We prove each case separately.

Claim (a). Easy to see.

Claim (b). Let (ℓ^*, h^*) be the optimal solution of the Primal Problem. First, we observe that $\mathcal{Z}^*_{L-P}(\gamma_1,\gamma_2) \le \mathcal{L}(\ell^*, h^*, \gamma_1, \gamma_2)$ since (ℓ^*, h^*) is the optimal solution of the Primal Problem, not the Lagrangian Primal Problem.

Next, for any feasible solution of the Primal Problem (ℓ, h), we have $\phi(h) - \phi(\ell) \le p$. Therefore, $\mathcal{L}(\ell, h, \gamma_1, \gamma_2) \le \mathcal{Z}_P(\ell, h)$, where $\mathcal{Z}_P(\ell, h)$ is the objective function value of the Primal Problem corresponding to (ℓ, h). Since (ℓ^*, h^*) is the optimal solution and satisfies the feasibility condition, we have $\mathcal{L}(\ell^*, h^*, \gamma_1, \gamma_2) \le \mathcal{Z}^*_P$.

Combining these results, we conclude:

$$\mathcal{Z}^*_{L-P}(\gamma_1,\gamma_2) \le \mathcal{L}(\ell^*, h^*, \gamma_1, \gamma_2) \le \mathcal{Z}^*_P.$$

Q.E.D.

THEOREM 2. *Let $\phi(y) = \mathbb{P}(f(X) \le y)$ be the empirical CDF of $f(X)$. Then, for the optimization problem (1)-(5), we have:*

(a) *If $h(\xi^{ru}) > \ell(\xi^{as})$ and $\phi(h(\xi^{ru})) - \phi(\ell(\xi^{as})) \le p$, then $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution.*

(b) *If $h(\xi^{ru}) > \ell(\xi^{as})$ and $\phi(h(\xi^{ru})) - \phi(\ell(\xi^{as})) > p$, then an approximate solution can be obtained by Algorithm 1.*

(c) *If $h(\xi^{ru}) < \ell(\xi^{as})$, then the problem is infeasible.*

(d) *If $h(\xi^{ru}) = \ell(\xi^{as})$, then the problem has a single threshold solution $\ell^* = h^* = h(\xi^{ru}) = \ell(\xi^{as})$.*

Proof of Theorem 2: First, we rewrite the constraint on workload using the definition of the CDF function:

$$\mathbb{E}[\mathbb{1}(\ell < f(X) < h)] \le p \iff \phi(h) - \phi(\ell) \le p. \tag{EC.1}$$

Next, We prove each case, separately.

Case (a). By (EC.1) and the assumption, $(\ell(\xi^{as}), h(\xi^{ru}))$ meets the FDA's workload constraint. Hence, the optimization problem (1)-(5) becomes equivalent to the Relaxed Problem. By Theorem 1, if $h(\xi^{ru}) > \ell(\xi^{as})$, then the optimal solution of the Auxiliary Problem is optimal in the Relaxed Problem. Also, by Proposition 1, if $h(\xi^{ru}) > \ell(\xi^{as})$, then $(\ell(\xi^{as}), h(\xi^{ru}))$ is the unique optimal solution of the Auxiliary Problem. Consequently, $(\ell(\xi^{as}), h(\xi^{ru}))$ is also the unique optimal solution of the optimization problem (1)-(5).

Case (b). In this case, $(\ell(\xi^{as}), h(\xi^{ru}))$ is not a feasible solution because it violates the FDA's workload constraint by assumption. In particular, the problem has the following feasible region:

$$\Theta = \{(\ell, h) \text{ s.t. } \phi(h) - \phi(\ell) \le p,\ h \le h(\xi^{ru}),\ \ell \ge \ell(\xi^{as})\}.$$

Our Algorithm 1, by design, finds a near-optimal solution within a finite number of steps.

Case (c). The condition of $h(\xi^{ru}) < \ell(\xi^{as})$ results in $h < \ell$, which contradicts with the requirement of $0 \le \ell \le h \le 1$ in optimization problem (1)-(5). Thus, the problem is infeasible.

Case (d). This can be shown by following a closely analogous argument to Case (a).

Q.E.D.

EC.2. Sensitivity Analysis

In all of our analysis, we have followed a conservative evaluation approach by assuming that the risk labels generated for deferred devices do not improve the performance of the FDA's committees in evaluating deferred applicant devices. However, in practice, these risk labels can assist the committees in allocating their limited resources more effectively. The risk labels provide useful supplementary information, ensuring that evaluation efforts are proportional to the risk levels, enabling a more efficient use of resources. For example, they help determining the review depth and the level of scrutiny required for each applicant device. Riskier devices may undergo more rigorous evaluations, involving higher levels of scrutiny and a more thorough analysis of safety and efficacy.

Accordingly, we conduct a sensitivity analysis to compare the performance of our policy with the FDA's current practice. In our analysis, we assume that the risk labels do not hurt the FDA's committees performance in evaluating a deferred device. This assumption implies that all the safe devices deferred will be accepted by the FDA's committees. Furthermore, we assume that the probability that the FDA's committees will fail to detect an unsafe deferred device is as follows:

$$\mathcal{L}(f(X),k) = 2\left(1 - \frac{1}{1+\exp\left(-k(f(X)-\ell^*)\right)}\right),$$

where $f(X)$ is the estimated risk, ℓ^* is the optimized low threshold, and parameter $k \geq 0$ is a scalar where higher values of k correspond to improved performance of the FDA's committees in detecting unsafe devices. When $k=0$, this probability is equal to 1 by our design for any unsafe device that has been deferred. This implies that $k=0$ is the baseline that matches the FDA's current practice in evaluating deferred devices without supplementary information (risk labels).

Figure EC.1 illustrates the predicted risk and the probability of failing to reject an unsafe device for 100 random samples, where samples are ordered based on their predicted risk. As can be seen, the probability of failing to reject an unsafe device decreases as the predicted risk of the deferred device increases. When $k=0$, the FDA's committees will fail to reject an unsafe device with a probability of 1. As k increases, this probability decreases. This observation aligns with our intuition that unsafe devices appearing less risky are more challenging for the FDA's committees to detect. Figure EC.2 illustrates the improvement in the recall rate percentage with respect to different values of k. When $k=0$, the recall rate percentage improvement is 38.9%, which matches the value observed in our conservative evaluation. As k increases, this recall rate percentage improvement increases, potentially exceeding 50% compared to the FDA's current practice for higher values of k. This highlights the significance of the FDA's committees' performance in evaluating deferred devices.

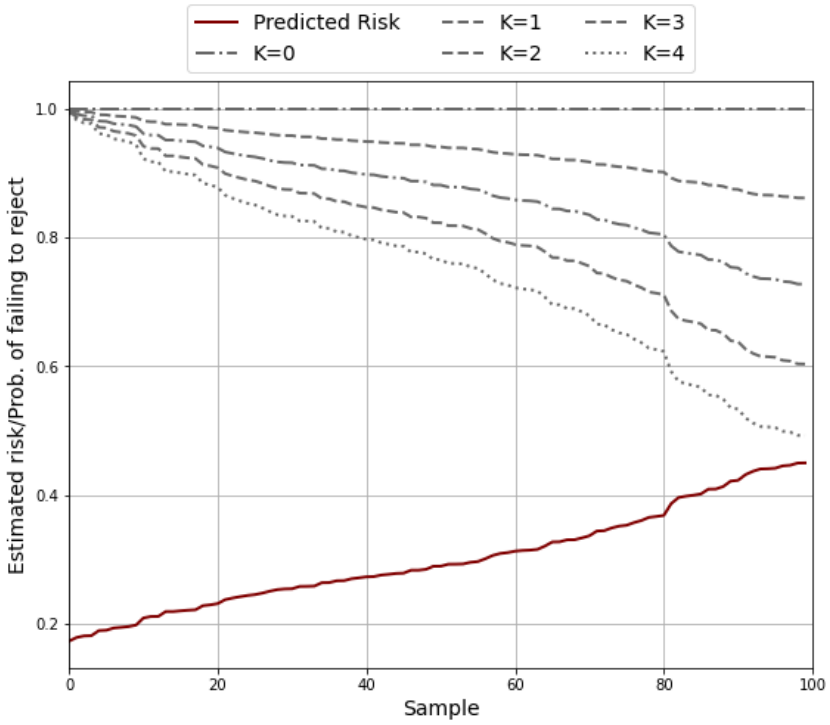


Figure EC.1 Predicted risk and probability of failing to reject unsafe deferred devices

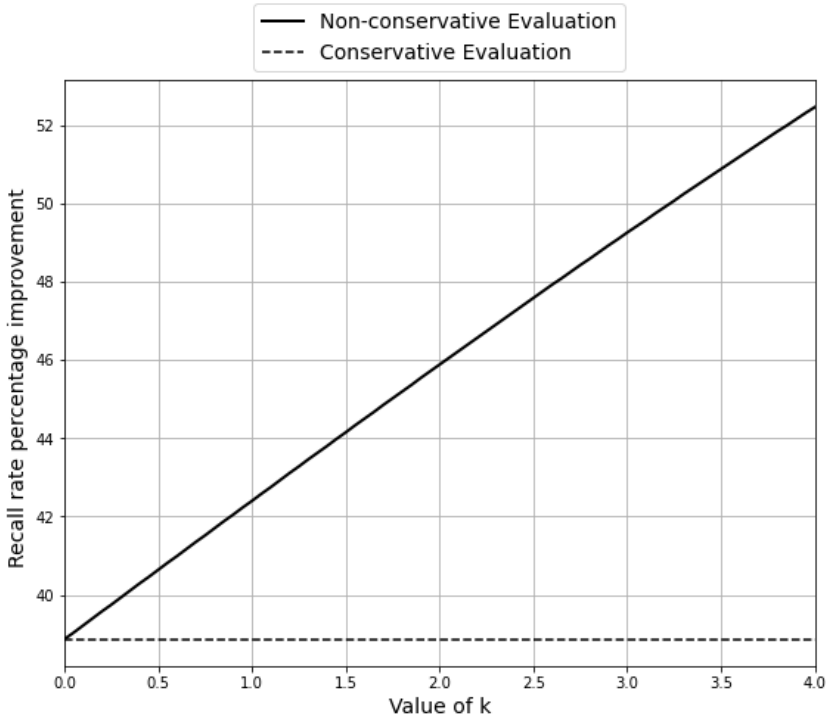


Figure EC.2 Recall rate percentage improvement across different k values

EC.3. Crosswalk between Keywords and Medical Specialties

Table EC.1 Medical Specialties and Keywords

Medical Specialty	Keyword(s)
Anesthesiology	"aerosol" and "compressor", "airway", "breathing circuits", "cough", "face mask", "nasal cannula", "nasal mask", "nebulization", "nebulizer", "oropharyngeal", "oxygen", "positive expiratory pressure", "respiratory", "tracheal suction", "ventilator"
Cardiovascular	"defibrillator", "pneumatic compression device"
Clinical Chemistry	"calibrator solution", "glucose monitor", "glucose" and "monitor"
Dental	"osteogenesis"
Gastroenterology/Urology	"bladder", "cervical", "drainage bag", "indwelling catheter", "insertion tray" and "catheter", "leg strap", "male" and "catheter", "ostomy", "parenteral", "pelvic floor", "stoma cap", "urethral", "urinary"
General & Plastic Surgery	"adhesive", "bandage", "chest wall", "collagen" and "wound", "compression" and "wrap", "dressing", "gauze", "lancet", "skin barrier", "sterile water", "tape", "tubing" and "pump", "ultraviolet" and "therapy", "wound"
General Hospital	"ambulatory infusion pump", "bath", "bed", "canister" and "pump", "chair", "compression stocking", "compression" and "garment", "compressor" and "for equipment", "drug infusion", "footplate", "footrests", "heel loop", "infusion pump", "insulin", "irrigation", "iv pole", "lubricant", "mattress", "transfer device", "urinal" and "jug-type"
Neurology	"conductive garment", "nerve stimulation"
Physical Medicine	"armrest", "cane", "commode chair", "crutches", "electrical" and "stimulator", "flexion", "foot" and "density insert", "foot" and "shoe molded", "heat pad", "inlay" and "shoe", "knee" and "exercise", "leg" and "compressor", "neuromuscular stimulator", "patient lift", "patient support system", "patient transfer", "pneumatic" and "compressor", "rear wheel", "traction" and "cervical", "trapeze", "vehicle", "walker", "wheelchair"