OXFORD UNIVERSITY PRESS | JAMIA: Journal of the American Medical Informatics Associati

A Machine Learning Framework to Adjust for Learning Effects in Medical Device Safety Evaluation

Journal:	*Journal of the American Medical Informatics Association*
Manuscript ID	amiajnl-2024-015583
Article Type:	Research and Applications
Keywords:	Post-market safety surveillance, Medical devices, Learning effects, Learning curve, Machine learning

A Machine Learning Framework to Adjust for Learning Effects in Medical Device Safety Evaluation

Jejo D. Koola, MD MS,[1] Karthik Ramesh, BS,[2] Jialin Mao MD, PhD,[3] Minyoung Ahn, BS,[4] Sharon Davis, PhD,[5] Usha Govindarajulu, PhD,[6] Amy M. Perkins, MS,[7,14] Dax Westerman, MS,[5] Henry Ssemaganda, MD MSc,[8] Theodore Speroff, PhD,[7,9] Lucila Ohno-Machado, MD PhD MBA,[10] Craig R. Ramsay, PhD,[11] Art Sedrakyan, MD PhD,[3] Frederic S. Resnic, MD MSc,[8,12,$] Michael E. Matheny, MD MPH MSc[5,7,9,13,$]

[1] Department of Medicine, University of California San Diego, San Diego, CA
[2] School of Medicine, University of California San Diego, San Diego, CA
[3] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY
[4] Jacobs School of Engineering, University of California San Diego, San Diego, CA
[5] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
[6] Center for Biostatistics, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY
[7] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN
[8] Comparative Effectiveness Research Institute, Lahey Hospital and Medical Center, Burlington, MA
[9] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[10] Biomedical Informatics and Data Science, Yale School of Medicine, 100 College Street, New Haven, CT 06510, USA
[11] Health Services Research Unit, University of Aberdeen, Health Sciences Building, Foresterhill, 3rd Floor, Aberdeen AB25 2ZD, UK
[12] Division of Cardiovascular Medicine, Lahey Hospital and Medical Center, Burlington, MA
[13] Geriatric Research Education and Clinical Care Center, Tennessee Valley Healthcare System VA, Nashville, TN
[$] Co-Senior Authors

Correspondence:
Jejo D. Koola
9500 Gilman Dr
MC 0881
La Jolla, CA 92093
Tel: 858-246-2563
Fax: 858-246-2329
jkoola@ucsd.edu

Word Count: 2960

Abbreviations:
ML: Machine Learning
DGP: Data Generating Process
ITE: Individual Treatment Effect
PPV: Positive Predictive Value
NPV: Negative Predictive Value
IPTW: Inverse Probability of Treatment Weighting
CABG: Coronary Artery Bypass Graft
DELTA: Data Extraction and Longitudinal Time Analysis
MSE: Mean Squared Error

Conflict of Interest: We wish to confirm for all authors that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Keywords: Post-market safety surveillance; Medical devices; Learning effects; Learning curve; Machine learning

Financial Support: This work was supported by "Incorporating Learning Effects into Medical Device Active Safety Surveillance Methods" (NHLBI R01 HL149948). Additionally, JM was supported by NHLBI Career Development Award (K01HL159315).

ABSTRACT

Background: Traditional methods for medical device post-market surveillance often fail to accurately account for operator learning effects, leading to biased assessments of device safety. Addressing this critical gap, our study develops a machine learning framework to detect and adjust for operator learning effects, improving the reliability of safety evaluations.

Methods: A gradient-boosted decision tree machine learning method was used to analyze synthetic datasets that replicate the complexity of clinical scenarios involving high-risk medical devices. We designed this process to detect learning effects using a risk-adjusted cumulative sum method, quantify the excess adverse event rate attributable to operator inexperience, and adjust for these alongside patient factors in evaluating device safety signals. To maintain integrity, we employed blinding between data generation and analysis teams.

Results: Analyzing 2,494 synthetic datasets, our framework accurately identified the presence or absence of learning effects in 93.6% of datasets and correctly determined device safety signals in 93.4% of cases. The estimated device odds ratios' 95% confidence intervals encompassed the specified ratios in 94.7% of datasets. In contrast, a comparative model excluding operator learning effects significantly underperformed in detecting device signals and in accuracy.

Conclusions: Demonstrating the capacity of machine learning to overcome complex evaluative challenges, our framework addresses the limitations of traditional statistical methods in current post-market surveillance processes. Future endeavors will extend the framework's applicability to a broader array of device categories and clinical settings using real-world data, with the goal of further enhancing patient safety and healthcare quality.

INTRODUCTION

Medical device failures pose a significant preventable health risk;[1–4] for example, certain faulty cardiovascular devices have cost the U.S. healthcare system ~$1.5 Billion and impacted thousands of lives.[5] The current reliance on spontaneous adverse event reporting, which captures less than 0.5% of device failures,[6,7] underscores the urgent need for more proactive surveillance methodologies. The U.S. Food and Drug Administration's (FDA) interest in active surveillance methodologies offers a promising direction, yet interpreting device safety and effectiveness from real-world data remains fraught with methodological challenges.[8]

Among these, the "learning curve" associated with physician experience significantly influences adverse event rates and may account for up to 30% of adverse events, a factor not adequately addressed by current surveillance strategies.[9–12] Navigating the complex landscape of post-market surveillance, particularly distinguishing between increased adverse events due to learning curves versus those resulting from inherent device faults, is crucial for the healthcare system and regulatory bodies like the FDA. For example, interventions such as mandatory training programs and continuing education may ameliorate learning curve effects, but inherent device faults are more appropriately addressed with regulatory mechanisms such as corrections, recalls, and removals.[13] Pursuing the incorrect remediation pathway may lead to further patient harm.

The statistical methods currently employed to analyze device safety and effectiveness are limited by their reliance on assumptions that often do not hold in the complex landscape of real-world data. In response to these challenges, this work advocates for a methodologic framework utilizing machine learning (ML) techniques that dynamically learn from data and recognize complex patterns, which we evaluated using hierarchical synthetic data.

METHODS

Overview. We developed and validated a machine learning (ML) framework aimed at distinguishing operator learning effects from device-specific safety signals, a critical challenge in post-market device surveillance. Our evaluation utilized synthetic datasets, designed to mirror the complexity of real-world clinical scenarios. These datasets were generated through a process ensuring a clear separation between the data generation team and the data analysis team, to maintain the objectivity and integrity of our findings.

Synthetic Data Generating Process (DGP) Design. We developed a library of 2494 simulated datasets configured with varying pre-specified combinations of device safety signals and provider experiential learning. Using our previously described DGP,[14] these data reflect the complexity of real-world clinical data and inject variations across key features that may impact our ability to detect both learning and device safety signals. Underlying distributions and patient feature correlations were based on clinical data from the Department of Veterans Affairs (VA) between 2005 and 2012. This study was approved as exempt by the Vanderbilt University Medical Center Institutional Review Board (IRB) and the use of VA data was approved by the Tennessee Valley Healthcare System VA Institutional Review Board.

Each dataset contained two devices (the study device, Device B and a reference device, Device A), a varying number of operators, and either 25 or 50 patient features – selected for their established links to adverse clinical outcomes (**Table 1** for an overview). To simulate confounding by indication, each patient received either Device B or A based on their clinical variables. The between device group signal was specified as an odds ratio (OR) of 1.0, for datasets with no safety signal, or weak, moderate, or strong safety signals corresponding to ORs of 1.25, 1.75, or 2.5, respectively.

Table 1: Overview of select patient features in reference population.

Feature	Distribution in reference population
Age in years, median [IQR]	64 [56, 75]
Sex (% male)	96%
Race	
% White	73%
% Black	22%
% Other	2%
% Unkown	3%
Chronic kidney disease	16%
Chronic obstructive pulmonary disease	30%
Hypertension	76%
Myocardial infarction	20%
Insulin use in prior 90 days	19%
Statin use in prior 90 days	49%
Hemoglobin at admission, median [IQR]	12.4 [10.7, 14]
White blood cell count at admission, median [IQR]	7.7 [6, 10]
Encounters in prior year, median [IQR]	34 [18, 61]

Each dataset contained the presence/absence of operator learning. We defined operator experience using the "case order," the number of times the operator had used that specific device prior to each patient. We modeled operator learning curves—using power, exponential, and Weibull distributions—to represent the evolution of provider expertise over time, informed by literature review and previous empirical work (**sTable 1**).[9–12,15] We defined the magnitude of learning effects as either small or large and the learning speed as either slow or fast based on the steady-state adverse event rate. We defined the steady state adverse outcome rate (after all learning is achieved) to be centered at 5% or 20%. **Figure 1** demonstrates the key concepts defined here. An in-depth discussion of the synthetic data generation process is included in **Supplementary Methods A**.

Machine Learning Framework. We constructed a multistep pathway to sequentially assess for learning effects and determine device signal strength after accounting for the learning effects. First, we developed a series of ML models to detect whether operator learning effects were present (**Figure 2A**). If learning was detected, we estimated the excess adverse event rate attributable to inexperience at each case order (**Figure 2B**). We subsequently incorporated these estimates into a new model to estimate the device effect adjusting for learning (**Figure 2C**). We ran the full analytic and evaluation pipeline on 50 bootstraps for each synthetic dataset to obtain confidence intervals. All models used the gradient boosting library XGBoost, version 1.7.5.1,[17] in R software version 4.0.5;[18] we optimized XGBoost hyperparameters on a per-dataset basis using stratified grid search. Comprehensive details are provided in **Supplementary Methods B.**

Learning Identification. We chose to explicitly model two scenarios using two separate XGBoost models fit to each dataset. The first XGBoost model included operator experience as a predictor variable, represented by the case order along with device and patient risk factors ("Learning Included" model). The second XGBoost model included only device and patient risk factors ("Learning Excluded" model). Our intuition was that if experiential learning was not a significant contributor to adverse events for a particular device, then both models should generate roughly equal predictions and exhibit roughly equal performance.

By adjusting both models for patient and device factors, we generated risk-adjusted predictions of adverse events. The presence of learning effects was inferred from significant discrepancies in model performances, as evidenced by deviations in predicted probabilities of adverse outcomes. Our evaluation hinged on the construction of risk-adjusted cumulative sum (RA CUSUM) control charts, which utilized the log-likelihood of observed outcomes based on each model's predictions.[19] A cumulative log-likelihood ratio exceeding a pre-established control limit indicated the detection of learning effects. This control limit was set at 1.5 standard

deviations above the baseline log-likelihood values, derived from a training dataset without learning effects.

Learning Effects Estimation. If the control chart identified learning, we estimated the learning curve by first calculating the risk-adjusted predicted probability at each case order, and then subtracting the estimated base rate. We estimated the steady state adverse event rate by identifying a sufficiently flat region of the learning curve. We then calculated the total excess attributable risk at each case order by subtracting the estimated steady state adverse event rate.

Device Effect Estimation. We first constructed a new XGBoost model on the observed data incorporating patient factors and device type, and adjusted for operator experience by integrating the estimated excess adverse event rates linked to operator learning from the Learning Effects Estimation stage. To robustly estimate the device effect, our analysis adopted a counterfactual approach, hypothesizing outcomes under alternate treatment scenarios.[20,21] This method allowed us to generate predicted probabilities of adverse outcomes for each patient as if they had received the opposite device, thereby enabling a direct comparison of the devices' safety. For consistency with existing work on medical device safety, and the U.S. FDA operational standards, we converted the change in adverse event rate to an OR.

Evaluation. We designed our evaluation process to validate the accuracy and robustness of our ML framework in distinguishing between operator learning effects and device-specific safety signals. We conducted a series of evaluations, leveraging synthetic datasets with known parameters to assess our framework's performance across multiple dimensions: identification of learning effects, detection of device safety signals, and accuracy of device effect estimation. We report sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) across the synthetic datasets.

Learning Identification Evaluation: We classified datasets as including learning effects if the risk-adjusted cumulative log-likelihood value exceeded the control limit. We compared this binary evaluation to the ground truth specified in the DGP.

Device Effect Identification Evaluation: We considered a specified device OR > 1.0 from the DGP to be a true signal. After calculating the confidence interval around the estimated OR, we defined signal detection as the lower limit of the confidence interval > 1.0.

Device Effect Accuracy Evaluation: We evaluated accuracy using the coverage statistic, defined as the probability that the estimated device OR 95% confidence interval included the specified device OR.[22] We also measured the mean squared error between the estimated and specified device OR.

Clinically Relevant Threshold Evaluation: Recognizing the FDA's emphasis on balancing device risks against benefits, we conducted pre-specified analyses to evaluate our framework's performance under clinically relevant safety signal thresholds (device signal ORs > 1.5 and > 2.0). These analyses aimed to mirror the decision-making process for regulatory actions, offering a nuanced understanding of the framework's applicability in real-world surveillance.[23]

RESULTS

Our 2,494 synthetic datasets included a median of 1,500 operators and 218,000 observations per dataset. Datasets were evenly distributed across no device safety signal (n=623), weak (n=624), moderate (n=623), or strong (n=624) safety signals corresponding to device ORs of 1.0, 1.25, 1.75, or 2.5, respectively. Of the 2,302 datasets where operator learning was present, the learning curve form was modeled using the Weibull (n=768), exponential (n=767), or power distribution (n=767). Approximately half of the datasets were designed with a large initial magnitude of learning and similarly half of the datasets were

designed with a rapid learning speed. **Table 2** presents a summary of all the relevant dataset characteristics.

Table 2: Characteristics of the 2494 synthetic datasets created by the data generating process.

	All datasets (n=2494)
# of Operators, median (Q1 – Q3)	1500 (792 - 2420)
# of Observations, median (Q1 – Q3)	218k (116k – 353k)
Strength of Device Signal	
Absent	623 (25.0%)
Low (OR ~ 1.25)	624 (25.0%)
Medium (OR ~ 1.75)	623 (25.0%)
High (OR ~ 2.5)	624 (25.0%)
Base Outcome Rate$^{\theta}$	
p ~ 0.05	1246 (50.0%)
p ~ 0.20	1248 (50.0%)
# of Patient Features	
25 (n=1058)	1273 (51.0%)
50 (n=1020)	1221 (49.0%)
Operator Learning	
Absent	192 (7.7%)
Present	2302 (30.8%)
Learning Form	
Exponential form	767 (30.8%)
Power form	767 (30.8%)
Weibull form	768 (30.8%)
Learning Speed$^{\beta}$	
Slow	1152 (46.2%)
Fast	1150 (46.1%)
Learning Magnitude$^{\alpha}$	
Small	1151 (46.2%)
Large	1151 (46.2%)

Note: α: Learning magnitude defined as the absolute increase in adverse event rate at the first case involving B – small and large correspond to 25% and 50% of the mean probability of an adverse outcome due to patient features and device, respectively; β: learning speed defined as the case order by which 95% of learning has occurred – slow and fast correspond to 75 and 25 cases, respectively ; θ: base outcome rate defined as adverse event rate when learning has been saturated

Learning Effect Estimation: The risk-adjusted log-likelihood framework successfully identified the presence of operator learning in 2,196 (sensitivity: 95.4%) of the 2,302 datasets where learning was present (**Table 3**). Of the 192 datasets where operator learning was absent, the algorithm correctly ascertained the absence of learning in 139 (specificity: 72.4%). The sensitivities across all DGP strata of learning form, speed, and magnitude ranged from 91.7% to

Table 3: General performance characteristics of the operator learning identification workflow.

	TP	TN	FP	FN	Sens	Spec	PPV	NPV
All datasets (n=2494)	2196	139	53	106	0.954	0.724	0.976	0.567
Strength of Device Signal								
Absent (n=623)	538	34	14	37	0.936	0.708	0.975	0.479
Low, OR ~ 1.25 (n=624)	550	34	14	26	0.955	0.708	0.975	0.567
Medium, OR ~ 1.75 (n=623)	553	39	9	22	0.962	0.813	0.984	0.639
High, OR ~ 2.5 (n=624)	555	32	16	21	0.964	0.667	0.972	0.604
Base Outcome Rate$^{\theta}$								
p ~ 0.05 (n=1246)	1142	67	29	10	0.991	0.698	0.975	0.870
p ~ 0.20 (n=1248)	1054	72	24	96	0.917	0.750	0.978	0.429
# of Patient Features								
25 (n=1273)	1122	72	27	52	0.956	0.727	0.977	0.581
50 (n=1221)	1074	67	26	54	0.952	0.720	0.976	0.554
Operator Learning								
Absent (n=192)	--	139	53	--	--	0.724	--	0.567
Present (n=2302)	2196	--	--	106	0.954	--	0.976	--
Learning Form								
Exponential (n=767)	754	--	--	13	0.983	--	1.000	--
Power (n=767)	727	--	--	40	0.948	--	1.000	--
Weibull (n=768)	715	--	--	53	0.931	--	1.000	--
Learning Speed$^{\beta}$								
Fast (n=1152)	1094	--	--	58	0.950	--	1.000	--
Slow (n=1150)	1102	--	--	48	0.958	--	1.000	--
Learning Magnitude$^{\alpha}$								
Small (n=1151)	1055	--	--	96	0.917	--	1.000	--
Large (n=1151)	1141	--	--	10	0.991	--	1.000	--

Note: α: Learning magnitude defined as the absolute increase in adverse event rate at the first case involving B – small and large correspond to 25% and 50% of the mean probability of an adverse outcome due to patient features and device, respectively; β: learning speed defined as the case order by which 95% of learning has occurred – slow and fast correspond to 75 and 25 cases, respectively ; θ: base outcome rate defined as adverse event rate when learning has been saturated

99.1%. The sensitivity was lower for Weibull (93.1%) and power (94.8%) learning form specifications compared to the exponential form (98.3%). Learning was also more challenging to detect when the DGP magnitude of learning was small (sensitivity 91.7%) compared to large (99.1%) but nearly the same for DGP fast (95.0%) and slow (95.8%) operator learning speed. Whereas sensitivities did not vary whether the DGP included 25 or 50 patient variables (95.6% and 95.2%, respectively), the sensitivity (99.1%) was greater when the base adverse outcome rate was 5% than at 20% (sensitivity 91.7%). The sensitivity for detecting learning monotonically increased from an absent (OR=1.0) DGP device signal strength (sensitivity 93.6%) to 96.4% with a high (OR=2.5) device effect.

Figure 3 illustrates the outputs from the Learning Effect Estimation and the risk-adjusted cumulative sum procedure. The probability of adverse outcome decreases with increasing case order in the Learning Included Model for Device B; whereas the Learning Excluded Model overestimates the adverse event rate during the early stages of operator experience. These learning curve estimates were included in the modeling of the device effect.

Device Effect Estimation: When incorporating operator learning into the model, we correctly ascertained the presence or absence of a device safety signal, with a sensitivity of 96.5% and 99 false positives yielding a specificity of 84.1% (top of **Table 4**). Device effect identification was more challenging for weak device signals (sensitivity of 90.4%) compared to medium (99.2%) and strong signals (100%). Also, sensitivity was greater when the DGP base outcome rate was high (99.3%) than when the outcome rate was low (93.8%). For 623 datasets without a device safety signal, all 99 false positives occurred in the presence of an operator learning effect. False positive rate was correlated with the degree of operator learning present (**Figure 4**). The 95% confidence interval for the estimated device OR included the DGP-specified device effect odds ratio in most datasets (coverage of 94.7%). For datasets without a

device safety signal, coverage was lower (84.1%); otherwise, there were no significant differences in coverage across other DGP subgroups.

Results using a model that did not incorporate operator learning (Learning Excluded Model) are shown in the bottom section of **Table 4**. The overall sensitivity was 99.8% and specificity was 23.3%. Coverage of the DGP-specified device OR was low (26.3%), ranging from 17.5% to 34.1% in the different DGP subgroups.

Clinically Relevant Threshold Evaluation: When we repeated our analysis by setting the threshold for a true device safety signal to an OR of 1.5 or 2.0, we found tradeoffs between sensitivity and specificity (**sTable 2** and **sTable 3**). Our model achieved a specificity of 100% for both clinically relevant thresholds and across all DGP subgroups; however, sensitivity decreased from 96.5% to 69.8% and 55.0% for the 1.5 and 2.0 thresholds, respectively. When using a clinically relevant threshold of OR=1.5, only datasets having a strong device safety signal (specified device effect OR = 2.5) still had a high sensitivity (97.1%).

DISCUSSION

In this study, we proposed and validated a ML framework ensemble to disentangle operator learning effects from device-specific safety signals, to support innovation in the FDA's prioritized active surveillance methods critical needs assessments. Our study showcases the capability of an ML framework, specifically tailored for the complexities inherent in post-market device evaluation, to provide nuanced insights that traditional statistical methods may fail to capture. By leveraging synthetic datasets reflective of real-world scenarios, this research not only underscores the adaptability of ML but also highlights its potential to enhance the precision and reliability of safety signal detection in medical devices.

The high sensitivity and specificity to detect device safety signals achieved by our ML models underline their effectiveness in identifying true safety signals amidst the variability

Table 4: General performance characteristics of the device effect estimation workflow comparing methods that incorporate learning versus methods that do not.

	TP	TN	FP	FN	Sens	Spec	PPV	NPV	Coverage	MSE
Learning Included Model										
All datasets (n=2494)	1806	524	99	65	0.965	0.841	0.948	0.89	0.947	0.114
Strength of Device Signal										
Absent (n=623)	--	524	99	--	--	0.841	--	1.000	0.841	0.035
Low, OR ~ 1.25 (n=624)	564	--	--	60	0.904	--	1.000	--	0.963	0.060
Medium, OR ~ 1.75 (n=623)	618	--	--	5	0.992	--	1.000	--	0.992	0.125
High, OR ~ 2.5 (n=624)	624	--	--	0	1.000	--	1.000	--	0.992	0.237
Base Outcome Rate$^{\theta}$										
p ~ 0.05 (n=1246)	877	281	30	58	0.938	0.904	0.967	0.829	0.959	0.113
p ~ 0.20 (n=1248)	929	243	69	7	0.993	0.779	0.931	0.972	0.935	0.115
# of Patient Features										
25 (n=1273)	916	264	56	37	0.961	0.825	0.942	0.877	0.946	0.138
50 (n=1221)	890	260	43	28	0.969	0.858	0.954	0.903	0.948	0.090
Operator Learning										
Absent (n=192)	132	48	0	12	0.917	1.000	1.000	0.800	0.984	0.012
Present (n=2302)	1674	476	99	53	0.969	0.828	0.944	0.900	0.944	0.123
Learning Form										
Exponential (n=767)	564	154	38	11	0.981	0.802	0.937	0.933	0.934	0.235
Power (n=767)	556	164	27	20	0.965	0.859	0.954	0.891	0.956	0.033
Weibull (n=768)	554	158	34	22	0.962	0.823	0.942	0.878	0.943	0.099
Learning Speed$^{\beta}$										
Fast (n=1152)	824	272	16	40	0.954	0.944	0.981	0.872	0.981	0.020
Slow (n=1150)	850	204	83	13	0.985	0.711	0.911	0.940	0.907	0.226
Learning Magnitude$^{\alpha}$										
Small (n=1151)	834	272	15	30	0.965	0.948	0.982	0.901	0.982	0.026
Large (n=1151)	840	204	84	23	0.973	0.708	0.909	0.899	0.906	0.220
Learning Excluded Model										
	TP	**TN**	**FP**	**FN**	**Sens**	**Spec**	**PPV**	**NPV**	**Coverage**	**MSE**
All datasets (n=2494)	1867	145	478	4	0.998	0.233	0.796	0.973	0.263	0.078
Strength of Device Signal										
Absent (n=623)	--	145	478	--	--	0.233	--	1.000	0.233	0.033
Low, OR ~ 1.25 (n=624)	620	--	--	4	0.994	--	1.000	--	0.296	0.057
Medium, OR ~ 1.75 (n=623)	623	--	--	0	1.000	--	1.000	--	0.287	0.087

High, OR ~ 2.5 (n=624)	624	--	--	0	1.000	--	1.000	--	0.234	0.133
Base Outcome Rate										
p ~ 0.05 (n=1246)	931	102	209	4	0.996	0.328	0.817	0.962	0.348	0.085
p ~ 0.20 (n=1248)	936	43	269	0	1.000	0.138	0.777	1.000	0.177	0.070
# of Patient Features										
25 (n=1273)	952	56	264	1	0.999	0.175	0.783	0.982	0.231	0.101
50 (n=1221)	915	89	214	3	0.997	0.294	0.810	0.967	0.296	0.053
Operator Learning										
Absent (n=192)	144	46	2	0	1.000	0.958	0.986	1.000	0.323	0.037
Present (n=2302)	1723	99	476	4	0.998	0.172	0.784	0.961	0.258	0.081
Learning Form										
Exponential (n=767)	574	18	174	1	0.998	0.094	0.767	0.947	0.198	0.148
Power (n=767)	575	47	144	1	0.998	0.246	0.800	0.979	0.318	0.030
Weibull (n=768)	574	34	158	2	0.997	0.177	0.784	0.944	0.257	0.065
Learning Speed										
Fast (n=1152)	860	74	214	4	0.995	0.257	0.801	0.949	0.320	0.026
Slow (n=1150)	863	25	262	0	1.000	0.087	0.767	1.000	0.195	0.136
Learning Magnitude										
Small (n=1151)	862	84	203	2	0.998	0.293	0.809	0.977	0.341	0.022
Large (n=1151)	861	15	273	2	0.998	0.052	0.759	0.882	0.175	0.140

Notes: α: Learning magnitude defined as the absolute increase in adverse event rate at the first case involving B – small and large correspond to 25% and 50% of the mean probability of an adverse outcome due to patient features and device, respectively; β: learning speed defined as the case order by which 95% of learning has occurred – slow and fast correspond to 75 and 25 cases, respectively ; θ: base outcome rate defined as adverse event rate when learning has been saturated. Coverage refers to the percentage of datasets where the true device odds ratio was contained within the estimated device odds ratio confidence interval. TP: true positive; TN: true negative; FP: false positive; FN: false negative; Sens: sensitivity: Spec: specificity: PPV: positive predictive value; NPV: negative predictive value: MSE: mean-squared error.

introduced by operator learning curves. Prior strategies to account for operator learning when analyzing adverse event rates have included measures such as explicit parametrizations of the learning curve,[24,25] treating operator experience as a categorical variable,[26,27] and splines.[28,29] While the study did not explicitly evaluate learning effects, Ross *et al.* evaluated three different statistical frameworks for detecting implanted medical device adverse outcomes and found their sensitivity for detecting device safety signals to range between 88% - 92%.[30] The precision that our ML model provides is paramount, as it directly influences the ability of

regulatory bodies and healthcare providers to make informed decisions regarding the continued use of medical devices.

Published work suggests that operator experience is a significant contributor to adverse event rates,[11,12,25,31–35] but most research is focused on describing the existence of operator learning without an attempt to better understand model improvement to detect device safety signals when accounting for learning effects. Our study provides a framework that extends these methods to integrate learning identification as part of device safety signal detection. In this study, when not accounting for operator learning, the generated models could only achieve a specificity of 0.232 when attempting to detect a device safety signal. When using more conservative thresholds endorsed by the FDA, our specificity was perfect; however, it increased the chance of false negatives. Our proposed analytic framework would allow regulatory agencies and the medical device industry to self-determine between different cutoffs of sensitivity and specificity. The dangers of false positives in medical device surveillance—ranging from unnecessary alarm to unwarranted regulatory actions—may limit the adoption of any post-market surveillance framework.

A noteworthy aspect of our methodology was the intentional separation between the teams responsible for data generation and data analysis. Concerns have been raised about reproducibility in retrospective analyses of observational cohort data.[36,37] Especially as ML has become prominent in the healthcare data analysis space, the consequences of “p-hacking” may influence published results.[38] Such rigorous standards of data handling and analysis are essential for the credibility of research in the highly scrutinized field of medical device surveillance.

We sought to develop a robust framework with very limited assumptions about the underlying data. A physician’s experience with a particular medical device not only affects the patient’s outcome but also affects the physician’s likelihood of choosing that device for a

procedure. This confounding by indication makes assessing device safety signals, which are often obscured by the interplay of numerous patient and provider variables, challenging. The ML-based counterfactual methodology not only addresses some of these challenges but also offers a more flexible framework that is straightforward to implement regardless of model architecture.

Our study has some other limitations. Synthetic data may not fully represent real-world conditions, although our DGP used real data that preserved complex variable distributions and relationships observed in healthcare data. Real-world data may result in case series without large case volumes and signal detection may be more challenging, particularly with machine learning methods. However, the increasing use of real-world data in cohorts such as All of Us and national active surveillance networks has resulted in datasets with tens or hundreds of thousands of observations.[9,29,39]

There have been prior recalls thought to have resulted from primarily insufficient training related to the medical device.[40] Given this situation, regulatory agencies have highlighted that the ability to detect and adjust learning effects is a critical need for post-market surveillance. The methods presented in this study could be effective in detecting and separating the signals from intrinsic device failures and learning effects. Future work in real-world data sources is needed to integrate these methods into prospective surveillance frameworks, such as the Data Extraction and Longitudinal Time Analysis (DELTA) system and other analytic frameworks, and to assess the potential utility of these methods to support this type of signal detection.[41,42]

In summary, we developed a multi-stage machine learning analytic framework to analyze implantable medical device outcomes that makes few assumptions about the underlying data. Compared to prior parametric methods for post-market surveillance that frequently do not adjust for learning effects, our work shows promise for detecting and adjusting for learning effects in medical device post-market surveillance activities. Future work is needed

in real-world data sources to assess the potential utility of this framework for properly accounting for learning effects in the surveillance of medical devices.

REFERENCES

1 Samore MH, Evans RS, Lassen A, *et al.* Surveillance of Medical Device–Related Hazards and Adverse Events in Hospitalized Patients. *JAMA*. 2004;291:325–34.

2 Garber AM. Modernizing Device Regulation. *N Engl J Med*. 2010;362:1161–3.

3 Maisel WH. Unanswered Questions — Drug-Eluting Stents and the Risk of Late Thrombosis. *New England Journal of Medicine*. 2007;356:981–4.

4 Hauser RG, Kallinen LM, Almquist AK, *et al.* Early failure of a small-diameter high-voltage implantable cardioverter-defibrillator lead. *Heart Rhythm*. 2007;4:892–6.

5 Schulte F, Jewett C. Replacing Faulty Heart Devices Costs Medicare $1.5 Billion in 10 Years. The New York Times. 2017. https://www.nytimes.com/2017/10/02/health/heart-devices-medicare.html (accessed 7 June 2023)

6 O'Shea JC, Kramer JM, Califf RM, *et al.* Part I: Identifying holes in the safety net. *Am Heart J*. 2004;147:977–84.

7 Gross TP, Kessler LG. Medical Device Vigilance at FDA. *Information Exchange for Medical Devices*. IOS Press 1996:17–24. https://doi.org/10.3233/978-1-60750-872-4-17

8 Shuren J, Califf RM. Need for a National Evaluation System for Health Technology. *JAMA*. 2016;316:1153–4.

9 Resnic FS, Wang TY, Arora N, *et al.* Quantifying the Learning Curve in the Use of a Novel Vascular Closure Device: An Analysis of the NCDR (National Cardiovascular Data Registry) CathPCI Registry. *JACC: Cardiovascular Interventions*. 2012;5:82–9.

10 Patrick WL, Iyengar A, Han JJ, *et al.* The learning curve of robotic coronary arterial bypass surgery: A report from the STS database. *Journal of Cardiac Surgery*. 2021;36:4178–86.

11 Kassite I, Bejan-Angoulvant T, Lardy H, *et al.* A systematic review of the learning curve in robotic surgery: range and heterogeneity. *Surg Endosc*. 2019;33:353–65.

12 Arora KS, Khan N, Abboudi H, *et al.* Learning curves for cardiothoracic and vascular surgical procedures – a systematic review. *Postgraduate Medicine*. 2015;127:202–14.

13 Center for Devices and Radiological Health. Recalls, Corrections and Removals (Devices). FDA 2020. https://www.fda.gov/medical-devices/postmarket-requirements-devices/recalls-corrections-and-removals-devices (accessed 21 March 2024)

14 Davis SE, Ssemaganda H, Koola JD, *et al.* Simulating complex patient populations with hierarchical learning effects to support methods development for post-market surveillance. *BMC Medical Research Methodology*. 2023;23:89.

15 Cook JA, Ramsaya CR, Fayers P. Statistical evaluation of learning curve effects in surgical trials. *Clinical Trials*. 2004;1:421–7.

16 Guo Y, Gao L, Zhu Y. ARL Estimation of the Control Chart of Log Likelihood Ratios' Sum for Markov Sequence. *Journal of Mathematics*. 2021;2021:e6649949.

17 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM 2016:785–94. https://doi.org/10.1145/2939672.2939785

18 Team RC. *R: A Language and Environment for Statistical Computing*. Vienna, Austria 2013. http://www.R-project.org

19 Steiner SH, Geyer PL, Wesolowsky GO. Grouped Data Sequential Probability Ratio Tests and Cumulative Sum Control Charts. ;22.

20 Little RJ, Rubin DB. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health*. 2000;21:121–45.

21 Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*. 2011;30:2867–80.

22 Aho KA. *Foundational and Applied Statistics for Biologists Using R*. CRC Press 2013.

23 Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions - Guidance for Industry and Food and Drug Administration Staff. U.S. Department of Health and Human Services, U.S. Food & Drug Administration 2016. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/factors-consider-regarding-benefit-risk-medical-device-product-availability-compliance-and (accessed 7 November 2023)

24 Charland PJ, Robbins T, Rodriguez E, *et al.* Learning curve analysis of mitral valve repair using telemanipulative technology. *The Journal of Thoracic and Cardiovascular Surgery*. 2011;142:404–10.

25 Suri RM, Minha S, Alli O, *et al.* Learning curves for transapical transcatheter aortic valve replacement in the PARTNER-I trial: Technical performance, success, and safety. *The Journal of Thoracic and Cardiovascular Surgery*. 2016;152:773-780.e14.

26 Cai Q, Li Y, Xu G, *et al.* Learning curve for intracranial angioplasty and stenting in single center. *Catheterization and Cardiovascular Interventions*. 2014;83:E94–100.

27 Hemli JM, Henn LW, Panetta CR, *et al.* Defining the Learning Curve for Robotic-Assisted Endoscopic Harvesting of the Left Internal Mammary Artery. *Innovations* *(Phila)*. 2013;8:353–8.

28 Govindarajulu US, Stillo M, Goldfarb D, *et al.* Learning curve estimation in medical devices and procedures: hierarchical modeling. *Statistics in Medicine*. 2017;36:2764–85.

29 Vemulapalli S, Carroll JD, Mack MJ, *et al.* Procedural Volume and Outcomes for Transcatheter Aortic-Valve Replacement. *New England Journal of Medicine*. 2019;380:2541–50.

30 Ross JS, Bates J, Parzynski CS, *et al.* Can machine learning complement traditional medical device surveillance? A case study of dual-chamber implantable cardioverter–defibrillators. *Medical Devices: Evidence and Research*. 2017;10:165–88.

31 Hopkins L, Robinson DBT, Brown C, *et al.* Trauma and Orthopedic Surgery Curriculum Concordance: An Operative Learning Curve Trajectory Perspective. *Journal of Surgical Education*. 2019;76:1569–78.

32 Dai Y, Kusuma S, Greene AT, *et al.* Application-Specific Learning Curve With a Modern Computer-Assisted Orthopedic Surgery System for Joint Arthroplasty. *Journal of Medical Devices*. 2021;15. doi: 10.1115/1.4049545

33 Alli O, Rihal CS, Suri RM, *et al.* Learning curves for transfemoral transcatheter aortic valve replacement in the PARTNER-I trial: Technical performance. *Catheterization and Cardiovascular Interventions*. 2016;87:154–62.

34 Handa N, Kumamaru H, Torikai K, *et al.* Learning Curve for Transcatheter Aortic Valve Implantation Under a Controlled Introduction System — Initial Analysis of a Japanese Nationwide Registry —. *Circulation Journal*. 2018;82:1951–8.

35 Carroll JD, Vemulapalli S, Dai D, *et al.* Procedural Experience for Transcatheter Aortic Valve Replacement and Relation to Outcomes. *Journal of the American College of Cardiology*. 2017;70:29–41.

36 Rotelli MD. Ethical Considerations for Increased Transparency and Reproducibility in the Retrospective Analysis of Health Care Data. *Drug Inf J*. 2015;49:342–7.

37 Shafer SL, Dexter F. Publication Bias, Retrospective Bias, and Reproducibility of Significant Results in Observational Studies. *Anesthesia & Analgesia*. 2012;114:931.

38 Head ML, Holman L, Lanfear R, *et al.* The Extent and Consequences of P-Hacking in Science. *PLOS Biology*. 2015;13:e1002106.

39 The “All of Us” Research Program. *New England Journal of Medicine*. 2019;381:668–76.

40 Peters W, Pellerin C, Janney C. RESEARCH: Evaluation of Orthopedic Hip Device Recalls by the FDA from 2007 to 2017. *Biomedical Instrumentation & Technology*. 2020;54:418–26.

41 Hickey GL, Bridgewater B, Grant SW, *et al.* National Registry Data and Record Linkage to Inform Postmarket Surveillance of Prosthetic Aortic Valve Models Over 15 Years. *JAMA Internal Medicine*. 2017;177:79–86.

42 Vidi VD, Matheny ME, Donnelly S, *et al.* An evaluation of a distributed medical device safety surveillance system: The DELTA network study. *Contemporary Clinical Trials*. 2011;32:309–17.

Figure 1: Demonstration of the difference between device effects and learning effects. Learning effect: device B demonstrates an improvement in the adverse event rate as operators gain experience. Initial learning effect magnitude: difference between the base outcome rate and the adverse event rate at case order one, after adjusting for patient characteristics. Learning effect speed: number of cases required for the adverse event rate, adjusted for patient characteristics, to decrease by 95% of the initial learning effect magnitude. Device effect: difference in adverse event rate between the two devices, after adjusting for learning effects and patient characteristics.

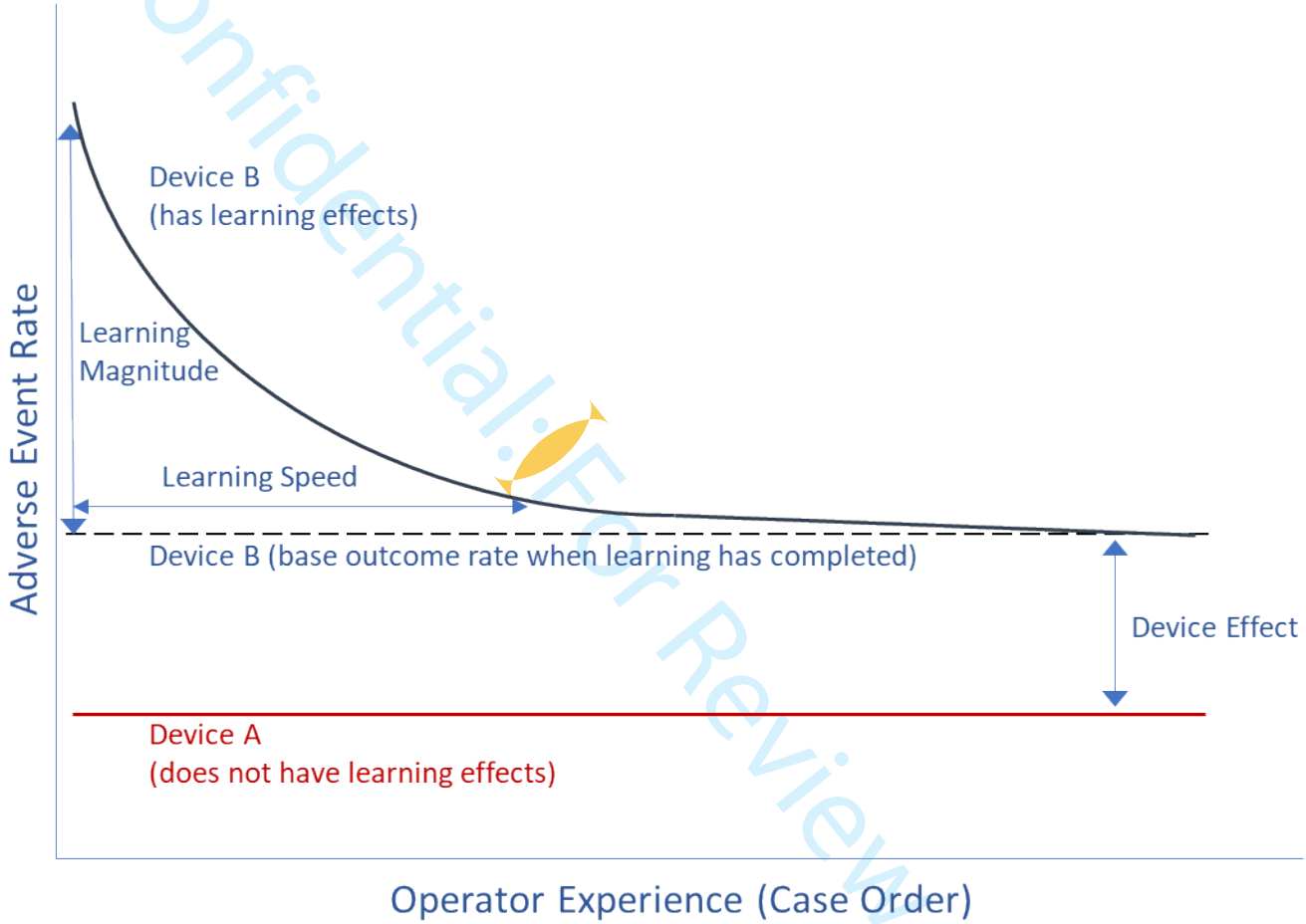


Figure 2: Overall modeling workflow to detect potential of learning effects (**A**), quantify learning effects **(B),** and detect a device safety signal **(C)** that may significantly alter adverse outcome probability.

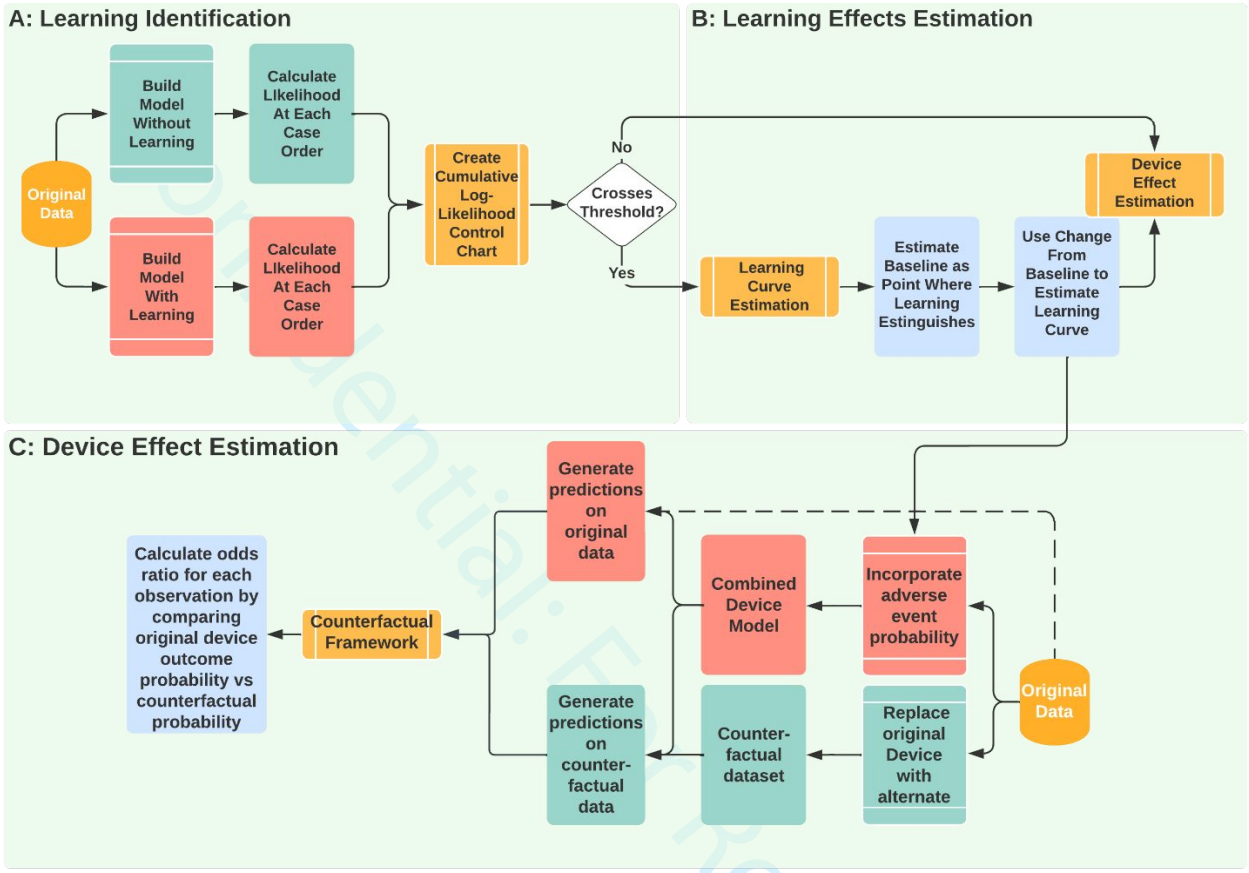


Figure 3: Example dataset demonstrating learning effect identification and estimation workflow. Probabilities from XGBoost models (with and without learning incorporated) are used to estimate the learning curve (**Panel A**). Subsequently, we calculate the likelihood of each XGBoost model at each case order, the likelihood ratio between the learning- and no-learning-XGBoost models, and finally the cumulative sum. When the cumulative sum of log-likelihood ratios crosses the control limit, the system flags potential learning effects (**Panel B**). Confidence bands are calculated using bootstrapping.

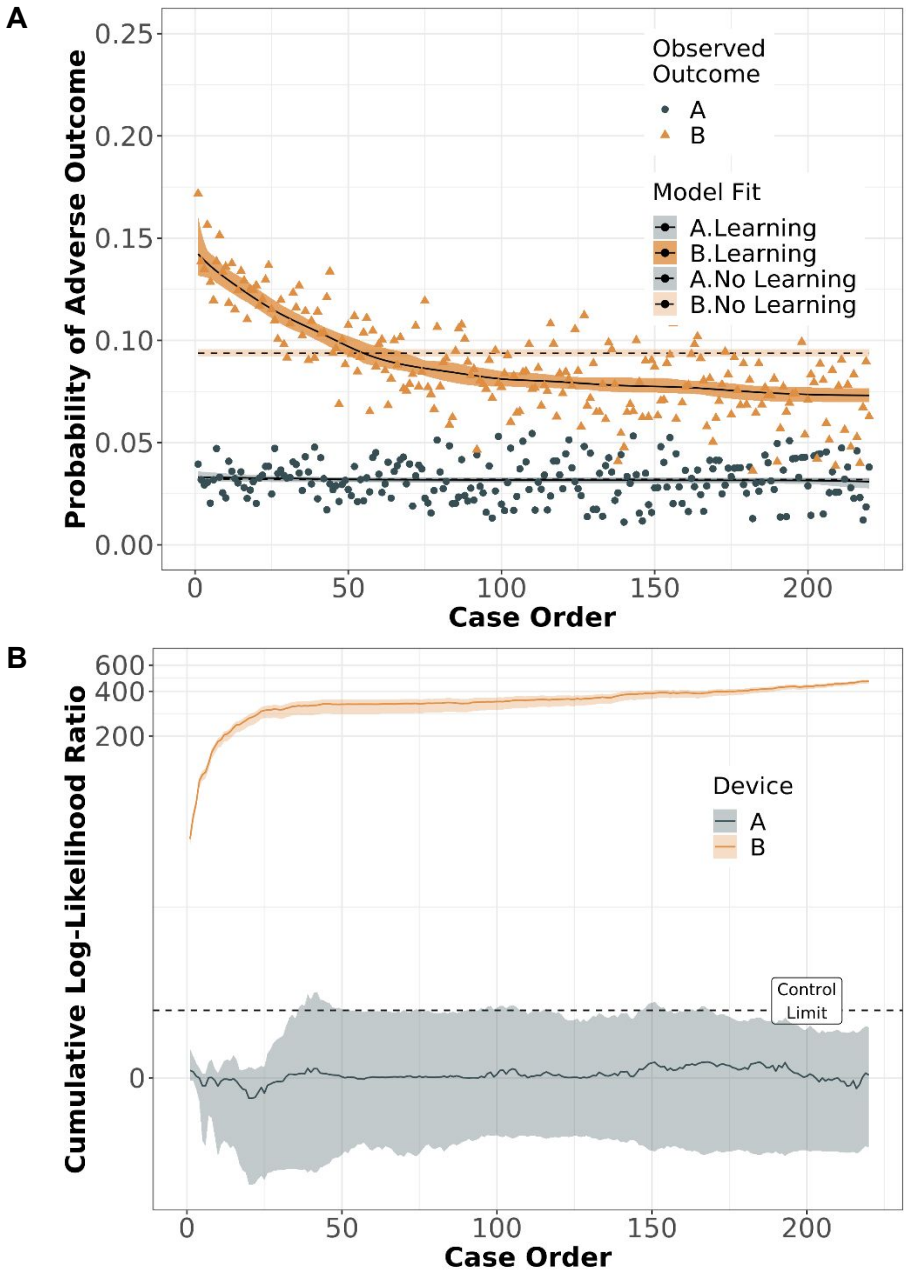
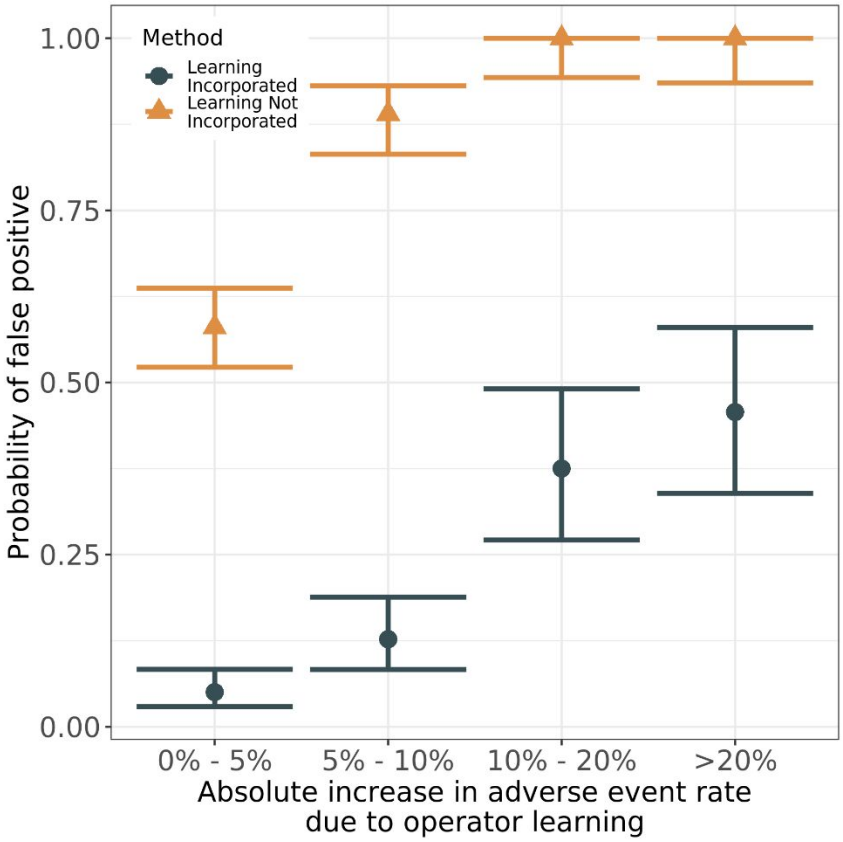


Figure 4: Percentage of false positives for device signal detection within datasets without a ground truth device signal as a function of method and degree of learning present. Orange markers represent a model that does not take learning into account; whereas blue indicates a model that incorporates operator experience. Point estimates represent percentage of total datasets within operator learning adverse event rate group and error bars represent binomial distribution calculated 95% confidence intervals.



A Machine Learning Framework for Detecting and Adjusting for Learning Effects in the Evaluation of Medical Device Safety – Supplementary Materials

CONTENTS

A. DATA GENERATION PROCESS

i. Overview

The data generating process (DGP) was customizable so that each dataset could be tuned to have a varying number of patient features, number of healthcare providers performing the procedures – whom we refer to as operators in this work, number of institutions at which operators perform cases, baseline adverse event rates, device safety signals, presence/absence of operator learning, and the form of operator learning. Learning at the provider level was present in select populations and not present in others. When present, we defined the form, magnitude, and speed of learning. Learning effects were specified as a power, exponential, or Weibull distribution (**Table 1**).

ii. Blinding of the Analysis Team

To minimize experimenter bias, where the ground "truth" is known to the researchers, we established a blinding framework where the data generation team was kept separate from the data analysis team. We developed a process where two separate teams, a data generating team and an analysis team, worked separately. The data generation team created datasets, described below, and handed them to the analysis team. The analysis team subsequently generated reporting metrics, which were handed back to the data generation team, and subsequently information on the accuracy of reporting metrics was the only thing that the analysis team had access to.

We had three distinct phases during this process. During Phase 1, in the early parts of this study, the two teams could communicate openly as we explored the problem space. During Phase 2, the data generation team could provide a limited set of information regarding how the synthetic data was being generated. During Phase 3, which is the phase being reported on in this paper, there was complete blinding of the analysis team so that only the dataset was provided to the analysis team and only evaluation results were returned to the data generation team (Refer to **Figure 1**).

iii. Components of the Synthetic Datasets

Underlying population

Sample of inpatient admissions to U.S. Department of Veterans Affairs facilities in 2005, 2010, 2011, and 2012. Admissions were included if they lasted at least 48 hours and the patient was at least 18 years of age. Admissions were excluded if the patient received hospice care or was admitted to a facility with fewer than 100 admissions per year or did not report key data to the central data warehouse. Features included demographics, vital signs, medications, laboratory values, diagnoses, admission characteristics, and healthcare utilization.

Device options

- Device of interest – Device B in all datasets
- Prevalence – Prevalence of device B set to means of 20% (range 15-25%, sd=5%) or 50% (range 45-55%, sd=5%)
- Feature associations – To simulate confounding by indication, associations between patient features and device assignment randomly selected from a "population" of 5 association set ups.
- Safety signal – Odds ratio for outcome associated with device B set to means 1.0 (no signal, all 1.0), 1.25 (range 1.2-1.3, sd=0.1), 1.75 (range 1.7-1.8, sd=0.1), and 2.5 (range 2.45-2.55, sd=0.1)

Institutions

- Number – Mean 25 (range 20-30, sd=9) and 75 (range 65-85, sd=9)

- Operator distribution – Distribution of the number of operators at each institution was a mixture model with 50% high volume institutions (mean 50 providers, range 40-60, sd=3) and 50% low volume institutions (mean 15 providers, range 10-20, sd=3)
- Learning effects – None at the institutional level

Operators

- Case volumes – Distribution of the annual number of patients treated by an operator will be a mixture model with 50% high volume (mean 100 patients, range 85-115, sd=4) and 50% low volume (mean 30 patients, range 20-40, sd=3).
- Each provider could accumulate experience with both devices, may select patients nonrandomly, and may have a random effect influencing their baseline event rate.
- Entry timing – Annual entry. 50% of each institution's operators begin at the start of the case series, with the remainder divided even to begin their case series at each new year.
- Learning effects – For device B, combinations of each of the following:
    - Presence and absence
    - Form: power, exponential, or Weibull
    - Magnitude: 25% (range 20-30%, sd=5%) or 50% (range 45-55%,sd=5%) of the mean probability of an adverse outcome due to patient features and device assignment among patients assigned to device B
    - Speed: Proficiency achieved over first 25 (range 20-30, sd=8) or 75 (range 70-80, sd=8) cases receiving device B

Outcome

- Feature associations – Associations between patient features and adverse outcomes randomly selected from a "population" of 5 association sets
- Overall event rate – mean of 5% (range 3%-7%, sd=2%), or 20% (range 17%-23%, sd=2%)
- Noise – No extra noise beyond binomial outcome generation from probabilities.

Other

- Missingness – None
- Omitted variables – None
- Timeframe – 3 years of cases

B. DATA ANALYSIS PROCESS

i. Learning Identification

We fit two separate XGBoost models to each dataset; the first model (the "Learning Model") a priori included learning as a predictor variable, represented by the case order, and the second model excluded this predictor (the "No Learning Model"). Our intuition was that if experiential learning was not a significant contributor to adverse events for a particular device, then both models should generate roughly equal predictions and exhibit roughly equal performance (accounting for some noise). We calculated the log-likelihood of the adverse event given the patient risk factors, device assignment, and case order based on the learning and no-learning models. After fitting the data, for each observation, we calculated the predicted probability of an adverse event using the two separate $XGBoost^{learning}$ and $XGBoost^{no\ learning}$ models, with and without learning incorporated.

$$p_i^{learning} = XGBoost_i^{learning}\left(X_{i_{device}}, X_{i_{patient\ factors}}, X_{i_{case\ order}}\right) \#(1)\#$$

$$p_i^{no\ learning} = XGBoost_i^{no\ learning}\left(X_{i_{device}}, X_{i_{patient\ factors}}\right) \#(2)\#$$

Risk Adjustment for Patient Factors and Device. We risk-adjusted for patient factors and device by calculating a partial dependence plot (PDP) for the case order and device variables. Partial dependence plots are a frequently used mechanism to assess the dependence of a machine-learning model's output on a subset of the predictor variables.[1,2] Given the output *f(x)* of a machine learning algorithm – in our case a predicted probability – the partial dependence of *f* on a subset of predictor variables of interest, X_S, is the expectation of *f* over the marginal distribution of all variables other than the predictors of interest. In practice we estimate the PDP by fixing the predictor variables of interest and averaging over all the observed values of the other variables in the dataset.

We took the original dataset and created *n* copies where in each copy of the dataset we replaced the case order variable with *1..n*. Similar to partial dependence plots, we averaged the predicted probability across all observations at each case order. We performed the same procedure for the "no learning incorporated" model.

We subsequently calculated the likelihood of the adverse event given the estimated probability of adverse outcome using the learning and no-learning models, separately, and calculated the log-likelihood ratio. Assuming *j=1..m* where *m* is the maximum case order and Y_i are the outcomes for observations *i=1..n* at that specific case order.

$$CUMSUM = \sum_{j=1}^{\max case\ order} \log\left(LR_j\right) \#(3)$$

$$LR_j = \frac{\mathcal{L}_i^{learning}}{\mathcal{L}_i^{no\ learning}} = \frac{Pr\left(\boldsymbol{Y_i} \middle| p_i^{learning}\right)}{Pr\left(\boldsymbol{Y_i} \middle| p_i^{no\ learning}\right)}$$

$$= \frac{\prod_i^n p_i^{learning}\left(1 - p_i^{learning}\right)^{1 - Y_i}}{\prod_i^n p_i^{no\ learning}\left(1 - p_i^{no\ learning}\right)^{1 - Y_i}} \#(4)$$

We subsequently calculated the risk-adjusted cumulative sum (CUSUM) of log-likelihood ratios as the charting statistic and signaled learning identified when the cumulative sum crossed the control limit. We defined the control limit, the threshold above which the system alerts to potential performance deviation, as 1.5σ of the cumulative log-likelihood values from a training dataset of device-observations generated without learning.

ii. Learning Curve Estimation

If learning was identified by the control chart, we estimated the learning curve form by averaging the predicted probability for observations at each case order from the XGBoost Learning Model ($p_{learning}$, *Eq. 1)*. We then estimated the steady state adverse event rate by identifying the region of the learning curve that was sufficiently flat, which we defined using the numerically approximated first derivative. More specifically we looked for the region of the curve where the absolute value of the first derivative was < 0.0001. We calculated the total excess attributable risk at each case order by subtracting the estimated steady state adverse event rate.

iii. Device Effect Estimation

Suppose there are N subjects. For each individual subject i with baseline x_i composed of some covariates (e.g., basic information, laboratory tests, and image tests, etc.), let $Y_{i,D=B}(x_i)$ and $Y_{i,D=A}(x_i)$ represent the potential outcomes for the two devices $D = \{A,B\}$. Given $X_i = x$, the ITE, $\tau(x)$, for subject i is defined as the conditional mean difference in potential outcomes

$$\tau_i(x) = \mathbb{E}\left[Y_i \middle| D = B, X_i = x\right] - \mathbb{E}\left[Y_i \middle| D = A, X_i = x\right] \#(5)$$

However, Eq. 5 is not estimable because in observational studies (and for most randomized trials) the potential outcomes $\{Y_{i,D=B}, Y_{i,D=A}\}$ are hypothetical and only the actual outcome Y_i from the actual treatment assignment is observed. A widely used assumption is the assumption of strongly ignorable treatment assignment (SITA), which assumes that treatment assignment is conditionally independent of the potential outcomes given the variables. We can, however, use the counterfactual, or potential outcomes, framework[3] to hypothesize what would have happened if an individual i had received both treatments.

To generate the two components in Eq. 5, $\mathbb{E}[Y_i | D = B, X_i = x]$ and $\mathbb{E}[Y_i | D = A, X_i = x]$, we used a process similar to Foster (2011)[4] to build an XGBoost model on the observed data to predict outcomes Y_i based on X_i, D_i, L_i, where *X, D,* and $L|X,D$ represent the patient factors, device, and operator, respectively. Unlike our Learning Identification model, where operator experience was represented by the numeric case order, we used the estimated additional adverse event rate due to learning, or lack thereof, as obtained from our Learning Identification model. Therefore L_i is conditioned on the device and patient factors.

To obtain the counterfactual estimate for subject i, we took the original dataset and replaced the Device variable with the opposite device and reran the observation through the XGBoost model for the counterfactual predicted probability. Therefore, for each subject i, we would obtain two probabilities $p_i|D = A$, $p_i|D = B$, from which we could estimate the change in adverse event rate for Device B compared to Device A, while controlling for patient factors and operator experience. We converted this to an odds ratio for each patient/observation in the standard way:

$$OR_i = \frac{\frac{p_i^{Device\ B}}{1 - p_i^{Device\ B}}}{\frac{p_i^{Device\ A}}{1 - p_i^{Device\ A}}}$$

To obtain the device odds ratio for the entire dataset, we took the median across all patients/observations. We obtained confidence intervals for the dataset device odds ratio using bootstrapping as described in the main methods.

iv. Model Fit

To evaluate the XGBoost model framework's ability to capture the underlying risk associations within the synthetic datasets, we also report on discrimination and calibration performance (Refer to the Appendix for procedure details and results). discrimination and calibration model performance for each dataset using the area under the receiver operator characteristic curve (AUC) and the Estimated Calibration Index (ECI). The ECI looks at the squared difference between the predicted probability from the model and an estimated observed probability of adverse outcome, ranging between 0 and 1, with 0 meaning perfect calibration.[5] We report the

median values and 25th and 75th percentiles for the AUC and ECI across all 2494 synthetic datasets. For each dataset we have a pair of performance measures, AUC from the Learning Included model and an AUC from the Learning Excluded model (and similarly from the ECI). To test whether incorporating learning as a predictor into the model produces a model with better discrimination and calibration performance, we used the Wilcoxon rank sum test to assess for statistically significant differences between the Learning Included and Learning Excluded XGBoost models.

v. Clinically Relevant Threshold Evaluation

In post-market surveillance activities, the FDA and other regulatory agencies have stressed the importance of weighing a non-conforming device's potential risks against benefits and commonly require a clinically relevant safety signal to take strong regulatory action or enforce a recall.[6–8] The FDA has stressed the importance of weighing a non-conforming device's potential risks compared to the device's benefits provided to the patient.[9] While regulatory procedures do not specify a blanket threshold for initiating a recall, in general a decision to initiate regulatory action needs to weigh multiple factors, including the magnitude and type of patient harm, the likelihood of harm, the inherent benefit of the device, and uncertainty of evidence. In practice, commonly used thresholds are odds ratios of 1.5 or 2.0. We conducted two additional pre-specified analyses where we assessed device effect estimation performance when a true signal was defined as a specified device odds ratio > 1.5 or > 2.0. For each of these analyses we defined signal detection as the lower limit of the confidence interval not crossing 1.5 and 2.0, respectively.

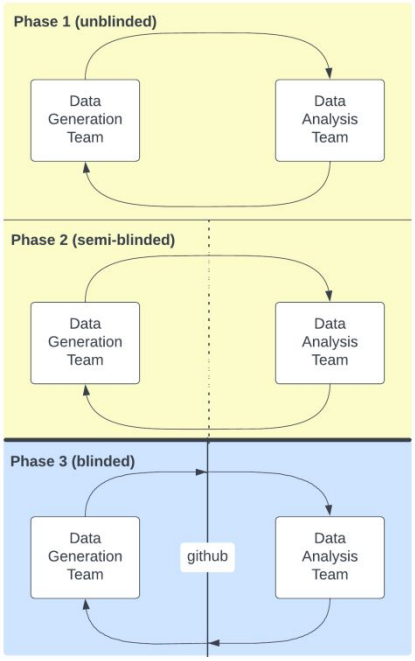
C. EXTENDED RESULTS

i. Model Fit

Across the 2494 datasets, incorporating learning into the XGboost model modestly improved discrimination (AUC = 0.803 [95% CI: 0.755 – 0.892] vs. AUC = 0.755 [0.710 – 0.820], p < 0.001, for learning vs. the no learning model, respectively). Similarly, calibration was modestly improved, ECI of 0.215 [0.223 – 0.355] vs. 0.322 [0.215 – 0.999] for learning vs. no learning, respectively, p=0.045).

D. FIGURE 1: PHASES OF DATA BLINDING TO SEPARATE THE DATA GENERATION AND ANALYSIS TEAMS.

During phase 1, test datasets were generated that were completely unblinded, with the data generation team providing active feedback. During phase 2, the data generation team provided the hyper-parameters (e.g., learning curve form) to the analysis team along with the datasets. During phase 3 (the datasets presented in this paper), only the dataset was provided to the analysis team and only evaluation results were returned to the data generation team for scoring.



E. TABLE 1: FUNCTIONAL FORMS AND STATISTICAL CHARACTERISTICS AMONG POTENTIAL LEARNING RATE EQUATIONS.

	Functional form	Learning rate	Initial performance	Asymptote
Weibull	$a + b\exp\left(-cx^{d}\right)$	$-bcdx^{d-1}\exp\left(-cx^{d}\right)$	$a + b\exp\left(-c\right)$	a
Power law	$a + bx^{-c}$	$-cbx^{-(c+1)}$	$a + b$	a
Exponential	$a + \exp\left(-cx\right)$	$-bc\exp\left(-cx\right)$	$a + b\exp\left(-c\right)$	a

F. TABLE 2: GENERAL PERFORMANCE CHARACTERISTICS OF THE DEVICE EFFECT ESTIMATION WORKFLOW WHEN A CLINICALLY RELEVANT DEVICE SIGNAL ODDS RATIO OF 1.5 IS USED AS THE THRESHOLD.

	TP	TN	FP	FN	Sens	Spec	PPV	NPV	Coverage	MSE
All datasets (n=2494)	871	1247	0	376	0.698	1.000	1.000	0.768	0.947	0.114
Strength of Device Signal										
Absent (n=623)	0	623	0	0	--	1.000	--	1.000	0.841	0.035
Low, OR ~ 1.25 (n=624)	0	624	0	0	--	1.000	--	1.000	0.963	0.060
Medium, OR ~ 1.75 (n=623)	265	0	0	358	0.425	--	1.000	0.000	0.992	0.125
High, OR ~ 2.5 (n=624)	606	0	0	18	0.971	--	1.000	0.000	0.992	0.237
Base Outcome Rate										
p ~ 0.05 (n=1246)	363	623	0	260	0.583	1.000	1.000	0.706	0.959	0.113
p ~ 0.20 (n=1248)	508	624	0	116	0.814	1.000	1.000	0.843	0.935	0.115
# of Patient Features										
25 (n=1273)	434	650	0	189	0.697	1.000	1.000	0.775	0.946	0.138
50 (n=1221)	437	597	0	187	0.700	1.000	1.000	0.761	0.948	0.090
Operator Learning										
Absent (n=192)	54	96	0	42	0.562	1.000	1.000	0.696	0.984	0.012
Present (n=2302)	817	1151	0	334	0.710	1.000	1.000	0.775	0.944	0.123
Learning Form										
Exponential (n=767)	267	384	0	116	0.697	1.000	1.000	0.768	0.934	0.235
Power (n=767)	289	383	0	95	0.753	1.000	1.000	0.801	0.956	0.033
Weibull (n=768)	261	384	0	123	0.680	1.000	1.000	0.757	0.943	0.099
Learning Speed										
Fast (n=1152)	385	576	0	191	0.668	1.000	1.000	0.751	0.981	0.020
Slow (n=1150)	432	575	0	143	0.751	1.000	1.000	0.801	0.907	0.226
Learning Magnitude										
Small (n=1151)	385	575	0	191	0.668	1.000	1.000	0.751	0.982	0.026
Large (n=1151)	432	576	0	143	0.751	1.000	1.000	0.801	0.906	0.220

Notes: Coverage refers to the percentage of datasets where the true device odds ratio was contained within the estimated device odds ratio confidence interval. TP: true positive; TN: true negative; FP: false positive; FN: false negative; Sens: sensitivity: Spec: specificity: PPV: positive predictive value; NPV: negative predictive value: MSE: mean-squared error.

G. TABLE 3: GENERAL PERFORMANCE CHARACTERISTICS OF THE DEVICE EFFECT ESTIMATION WORKFLOW WHEN A CLINICALLY RELEVANT DEVICE SIGNAL ODDS RATIO OF 2.0 IS USED AS THE THRESHOLD.

	TP	TN	FP	FN	Sens	Spec	PPV	NPV	Coverage	MSE
All datasets (n=2494)	343	1870	0	281	0.55	1.000	1.000	0.869	0.947	0.114
Strength of Device Signal										
Absent (n=623)	0	623	0	0	--	1.000	--	1	0.841	0.035
Low, OR ~ 1.25 (n=624)	0	624	0	0	--	1.000	--	1	0.963	0.06
Medium, OR ~ 1.75 (n=623)	0	623	0	0	--	1.000	--	1	0.992	0.125
High, OR ~ 2.5 (n=624)	343	0	0	281	0.55	--	1.000	0	0.992	0.237
Base Outcome Rate										
p ~ 0.05 (n=1246)	109	934	0	203	0.349	1.000	1.000	0.821	0.959	0.113
p ~ 0.20 (n=1248)	234	936	0	78	0.75	1.000	1.000	0.923	0.935	0.115
# of Patient Features										
25 (n=1273)	144	962	0	167	0.463	1.000	1.000	0.852	0.946	0.138
50 (n=1221)	199	908	0	114	0.636	1.000	1.000	0.888	0.948	0.09
Operator Learning										
Absent (n=192)	12	144	0	36	0.25	1.000	1.000	0.8	0.984	0.012
Present (n=2302)	331	1726	0	245	0.575	1.000	1.000	0.876	0.944	0.123
Learning Form										
Exponential (n=767)	113	575	0	79	0.589	1.000	1.000	0.879	0.934	0.235
Power (n=767)	117	575	0	75	0.609	1.000	1.000	0.885	0.956	0.033
Weibull (n=768)	101	576	0	91	0.526	1.000	1.000	0.864	0.943	0.099
Learning Speed										
Fast (n=1152)	145	864	0	143	0.503	1.000	1.000	0.858	0.981	0.02
Slow (n=1150)	186	862	0	102	0.646	1.000	1.000	0.894	0.907	0.226
Learning Magnitude										
Small (n=1151)	151	863	0	137	0.524	1.000	1.000	0.863	0.982	0.026
Large (n=1151)	180	863	0	108	0.625	1.000	1.000	0.889	0.906	0.22

Notes: Coverage refers to the percentage of datasets where the true device odds ratio was contained within the estimated device odds ratio confidence interval. TP: true positive; TN: true negative; FP: false positive; FN: false negative; Sens: sensitivity: Spec: specificity: PPV: positive predictive value; NPV: negative predictive value: MSE: mean-squared error.

H. REFERENCES

1. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).
2. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **24**, 44–65 (2015).
3. Little, R. J. & Rubin, D. B. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Annu. Rev. Public Health* **21**, 121–145 (2000).
4. Foster, J. C., Taylor, J. M. G. & Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Stat. Med.* **30**, 2867–2880 (2011).
5. Van Hoorde, K., Van Huffel, S., Timmerman, D., Bourne, T. & Van Calster, B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J. Biomed. Inform.* **54**, 283–293 (2015).
6. Kadakia, K. T., Beckman, A. L., Ross, J. S. & Krumholz, H. M. Renewing the Call for Reforms to Medical Device Safety—The Case of Penumbra. *JAMA Intern. Med.* **182**, 59–65 (2022).
7. Fargen, K. M. *et al.* The FDA approval process for medical devices: an inherently flawed system or a valuable pathway for innovation? *J. NeuroInterventional Surg.* **5**, 269–275 (2013).
8. Kramer, D. B., Xu, S. & Kesselheim, A. S. How Does Medical Device Regulation Perform in the United States and the European Union? A Systematic Review. *PLOS Med.* **9**, e1001276 (2012).
9. *Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions - Guidance for Industry and Food and Drug Administration Staff*. https://www.fda.gov/regulatory-information/search-fda-guidance-

documents/factors-consider-regarding-benefit-risk-medical-device-product-availability-compliance-and (2016).