

Appendix 1

Methods

Ordinal Logistic Regression Sensitivity Analysis

The original survey instrument used a 6-point Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Slightly Disagree, 4 = Slightly Agree, 5 = Agree, 6 = Strongly Agree) for all Q21 items. For the primary analysis, responses were dichotomized at the midpoint (1–3 = “No”/disagree; 4–6 = “Yes”/agree), producing binary targets with the following splits:

- Question 1 (“I struggled with lifestyle changes”): 42.7% disagree vs 57.3% agree (56/75 binary split)
- Question 2 (“I wasn’t sure what actions would help”): 63.4% disagree vs 36.6% agree (83/48 binary split)
- Question 3 (“I felt like a burden to loved ones”): 43.5% disagree vs 56.5% agree (57/74 binary split)

To assess sensitivity to this dichotomization, we fit proportional-odds ordinal logistic regression models (mord Python library, All-Thresholds variant with L2 regularization $\alpha = 1.0$) on the full 6-level ordinal outcome for each question, using the same 31 standardized patient features as the binary models ($N = 131$).

Concordance between ordinal and binary models was evaluated via: (a) Spearman rank correlation of absolute coefficient magnitudes; (b) percentage of features with concordant coefficient sign (direction of effect); and (c) overlap in top 5 ranked features.

Results (see Supplementary [Figure S3](#)):

- Question 1: Spearman $\rho = 0.424$ ($p = 0.017$), 84% direction concordance (26/31 features). Top ordinal predictors: Age, Prior CAD Procedure, Prior PAD Procedure — concordant with the stacked binary model.
- Question 2: Spearman $\rho = 0.604$ ($p < 0.001$), 74% direction concordance (23/31). Top ordinal predictors: Prior CAD related diagnosis, Prior CAD related medical event, Age — concordant with binary model (Age, Annual household income, Recent CAD Procedure).
- Question 3: Spearman $\rho = 0.430$ ($p = 0.016$), 74% direction concordance (23/31). Top ordinal predictors: Prior PAD related diagnosis, Age, Experience with implantable trackers — concordant with binary model findings.

Across all three questions, Age appeared in the top-5 for both ordinal and binary models, and the overall direction of effects was preserved in $\geq 74\%$ of features. These results confirm that the binary dichotomization does not introduce systematic bias in identifying the most impactful patient characteristics; the ordinal model identifies the same core predictors with concordant effect directions.

Likelihood Analysis Formulation

While SHAP values identify which patient characteristics are most predictive of unfavorable responses, they do not directly reveal for whom—that is, which specific subgroups of patients are at elevated risk. The likelihood ratio analysis bridges this gap by quantifying, for each value of each important feature, how much more (or less) likely that value is among patients who endorse unfavorable outcomes versus those who do not.

Formally, for a discrete patient characteristic X taking value x , we define the subgroup-specific likelihood ratio as: $LR(x) = P(X = x \mid Y = 1) / P(X = x \mid Y = 0)$, where $Y = 1$ denotes agreement with unfavorable outcome statements (Likert 4–6) and $Y = 0$ denotes disagreement (Likert 1–3). $LR(x) > 1$ indicates that value x is over-represented among patients reporting unfavorable experiences, while $LR(x) < 1$ indicates under-representation (i.e., a protective factor).

Feature selection for the likelihood analysis was performed using the Kneedle algorithm (17) applied to the mean absolute SHAP importance rankings from the stacked ensemble model. The Kneedle algorithm identifies the point of maximum curvature in the sorted importance curve, providing a data-driven threshold separating high-importance features from diminishing-contribution features.

Conditional probabilities were estimated empirically from the US cohort ($N = 131$). For each selected feature and each of its discrete values x : $P(X = x \mid Y = k) = n(X = x, Y = k) / n(Y = k)$, where n denotes the count of observations satisfying the condition. Continuous features (e.g., age) were pre-categorized into ordinal groups (quintile-based or clinically meaningful cut-points) prior to this analysis, ensuring all features entered as discrete variables with interpretable category labels.

To quantify uncertainty, 95% confidence intervals were computed via the nonparametric bootstrap percentile method ($B = 1,000$ resamples). For each bootstrap iteration b : (1) the full dataset was resampled with replacement, preserving the joint distribution of features and outcomes; (2) the conditional probabilities and resulting likelihood ratio were recomputed for each feature-value pair; and (3) bootstrap replicates with undefined ratios (due to empty conditioning sets) were discarded. The 2.5th and 97.5th percentiles of the valid bootstrap distribution defined the 95% CI bounds.

A subgroup-specific likelihood ratio was considered statistically significant if its 95% bootstrap CI excluded 1.0—that is, if the entire interval lay above 1.0 (indicating significantly elevated risk) or below 1.0 (indicating significant protection). Across all three questions, 29 of 65 unique feature-value pairs (44.6%) achieved significance by this criterion, confirming that the identified subgroup effects are robust to sampling variability.

The resulting likelihood ratios are visualized in [Figure S3](#) with bootstrap confidence intervals shown as error bars. For clinical interpretation: $LR(x) \approx 1$ indicates that the subgroup defined by $X = x$ has the same probability of unfavorable response as the overall population; $LR(x) > 2$ indicates the subgroup is at least twice as likely to report unfavorable outcomes; and $LR(x) < 0.5$ indicates the subgroup is less than half as likely—a strong protective effect. This interpretability is a key advantage over model-internal importance metrics, as it provides actionable, patient-facing information for targeted intervention design.

Appendix 2

Results

Cross National Comparison

The primary analysis was restricted to the U.S. cohort ($N = 131$) due to the HCAHPS-specific framing of the outcome measures. To assess generalizability of the identified predictive structure, we conducted an exploratory cross-national comparison using the full international Beyond Intervention dataset ($N = 1,350$ respondents across 16 markets: USA, UK, Germany, France, Italy, Spain, Japan, China, India, Canada, Australia, South Korea, Mexico, Netherlands, Belgium, and Switzerland).

Because 18 of the 31 U.S. model features are country-specific (e.g., U.S. insurance type, household income brackets) or were derived during U.S.-specific preprocessing, the cross-national comparison was conducted on the 13 universally available clinical and demographic features: Age, Gender, Diabetes, Hypertension, Hypercholesterolemia, Prior CAD Procedure, Recent CAD Procedure, Prior CAD Symptoms, Prior PAD Diagnosis, Prior PAD Procedure, Recent PAD Procedure, Comfort with technology, and Education level. For each market with $N \geq 30$ after listwise deletion (all 16 markets qualified), a Random Forest classifier ($n_estimators = 100$, $ccp_alpha = 0.02$) was trained on the same binary outcome ($Likert \geq 4 = \text{agree}$) with SMOTE oversampling, and feature importance rankings were extracted.

Concordance between each market's feature importance ranking and the U.S. ranking was quantified via Spearman ρ . Across all 15 non-U.S. markets $\times$ 3 questions (42 pairwise comparisons after excluding cases with extreme class imbalance), the mean Spearman ρ was 0.480 ($SD = 0.190$, range: -0.033 to 0.841). Markets with strongest concordance included South Korea (mean $\rho = 0.650$), Japan ($\rho = 0.614$), UK ($\rho = 0.582$), and France ($\rho = 0.533$). Weaker concordance was observed for Canada ($\rho = 0.247$), Switzerland ($\rho = 0.317$), and Australia ($\rho = 0.335$), likely reflecting cultural differences in technology adoption and healthcare access patterns, combined with smaller effective sample sizes ($N = 40\text{--}70$).

Top 5 feature overlap analysis showed that, on average, 2.8/5 features for Question 1, 3.1/5 for Question 2, and 3.5/5 for Question 3 were shared between U.S. and other markets' top 5 rankings. Age and Prior PAD Procedure appeared consistently in the top 5 across the majority of markets, suggesting these are universal predictors of unfavorable patient experiences regardless of healthcare system context.

These results demonstrate moderate-to-strong preservation of the core predictive structure across diverse international healthcare systems when evaluated on shared clinical features. However, this comparison is limited by: (a) the restriction to 13 of 31 features (the digital health engagement features that drove much of the U.S. model's discrimination could not be evaluated cross-nationally); (b) per-market sample sizes of 30–157, which limit statistical power; and (c) the absence of country-

level random effects to formally model healthcare system heterogeneity. A full multi-level cross-national analysis incorporating country-specific digital health features remains an important direction for future work.

Stratified Digital Health × Comorbidity Analysis

A potential concern is that digital health engagement features (experience with telemedicine, wearable trackers, implantable trackers, mobile apps, remote device monitoring, and digital health coaching) may serve as proxies for disease severity—patients with more complex conditions may have more exposure to health technologies through their care. To assess whether the identified digital health associations reflect genuine technology-related predictors or are confounded by comorbidity burden, we conducted a stratified analysis with explicit correlation testing.

Comorbidity burden was operationalized as the sum of 15 binary cardiovascular indicators: Diabetes, Hypertension, Hypercholesterolemia, Prior CAD Diagnosis, Prior CAD Procedure, Recent CAD Procedure, Prior PAD Diagnosis, Prior PAD Procedure, Recent PAD Procedure, and six additional CAD/PAD-related diagnosis, event, and symptom indicators (mean burden = 9.5, median = 10, range 3–14). Patients were split at the median into high-burden (≥ 10 , $N = 71$) and low-burden (< 10 , $N = 60$) subgroups.

Outcome rates differed substantially between strata, confirming that comorbidity burden is associated with more unfavorable experiences: Question 1 (lifestyle changes): 69.0% vs 43.3% agreement (high vs low burden); Question 2 (recovery uncertainty): 45.1% vs 26.7%; Question 3 (caregiver burden): 63.4% vs 48.3%.

Pearson correlations between each digital health feature and total comorbidity burden were uniformly weak: Experience with digital health coaching ($r = 0.369$, $p < 0.001$), Experience with remote device monitoring ($r = 0.337$, $p < 0.001$), Experience with mobile apps ($r = 0.291$, $p < 0.001$), Experience with wearable trackers ($r = 0.281$, $p = 0.001$), Level of comfort with technology ($r = 0.214$, $p = 0.014$), Experience with implantable trackers ($r = 0.189$, $p = 0.031$), and Experience with telemedicine ($r = 0.181$, $p = 0.039$). Willingness to use technology showed no correlation ($r = 0.008$, $p = 0.924$). The mean $|r|$ across all 8 digital health features was 0.234, with maximum $|r| = 0.369$ —corresponding to only 5–14% shared variance (R^2).

These weak correlations confirm that while sicker patients have marginally more digital health exposure (expected given greater healthcare system contact), the relationship is far too weak to constitute confounding. Moreover, the SHAP framework used in our primary analysis inherently conditions on all features simultaneously when computing each feature's marginal contribution. When the model attributes importance to a digital health feature, this importance is computed after accounting for the patient's full comorbidity profile. The co-selection of both comorbidity features (e.g., Prior PAD related diagnosis) and digital health features (e.g., Experience with implantable trackers) in the same top 5 SHAP rankings—with distinct contribution patterns—provides direct empirical evidence that they capture non-redundant predictive information, supporting the interpretation that digital health engagement represents a clinically meaningful predictor dimension independent of disease severity.

SHAP Feature Ranking Comparison Across Base Models

A key motivation for the stacked ensemble architecture is that individual base classifiers (Random Forest, SVM) can produce substantially different feature importance rankings due to their distinct inductive biases—RF captures non-linear interactions via recursive partitioning while SVM operates in a transformed kernel space. If these disagreements are large, relying on any single model's SHAP values would yield model-dependent, potentially misleading conclusions about which patient characteristics matter most.

To quantify this disagreement, we computed Spearman rank correlations of mean $|SHAP|$ importance rankings between base classifiers for Question 2: RF $\leftrightarrow$ SVM $\rho = 0.538$, RF $\leftrightarrow$ GBM $\rho = 0.736$, SVM $\leftrightarrow$ GBM $\rho = 0.506$. These moderate correlations confirm substantial disagreement—particularly between RF and SVM ($\rho = 0.538$), where nearly half of the ranking variance is unshared.

The stacked ensemble resolves this by combining base model predictions through a logistic regression meta-learner, effectively averaging over the disagreements. The resulting SHAP values from the stacked model reflect a consensus importance that is less sensitive to any single classifier's biases. Top-ranked features in the stacked model's SHAP analysis were: Age (consistently #1–2 across all questions), Prior PAD Procedure, Prior CAD-related symptoms, and Experience with implantable trackers—features that emerged as important regardless of the base classifier, confirming their robustness.

Supplementary Figure S2 visualizes this comparison, plotting the SHAP importance ranks from RF against SVM for each feature. The scatter pattern confirms that while the top-2–3 features are generally concordant across models, mid-ranked features exhibit considerable rank shuffling—precisely the regime where ensemble averaging provides the most value for reliable clinical interpretation.

Logistic Regression Benchmark

To validate that the identified feature associations are not artifacts of the stacked ensemble architecture, we benchmarked a standard logistic regression (LR) model with L2 regularization ($C=1.0$) and pairwise interaction terms among the top-3 features per question. The same 5-fold stratified cross-validation protocol with within-fold SMOTE was applied.

Results: Question 1 (8 Kneedles features + 3 interactions): $AUC = 0.778 \pm 0.094$; Question 2 (31 features + 3 interactions): $AUC = 0.626 \pm 0.148$; Question 3 (5 Kneedles features + 3 interactions): $AUC = 0.710 \pm 0.121$. By comparison, the stacked ensemble achieved $AUC = 0.774, 0.761, \text{ and } 0.717$, respectively.

For Questions 1 and 3, the LR model achieves comparable AUC (within 0.01), confirming that the identified associations (age, prior PAD procedure, digital health experience) are robust to model choice and do not depend on the non-linear ensemble architecture. For Question 2, the stacked ensemble provides a substantial advantage ($\Delta AUC = +0.135$), attributable to the lower base rate (36.6%) and more complex feature interactions requiring the non-linear capacity of RF and SVM base learners.

Critically, the LR model identified the same dominant predictors across all three questions (Age, Prior PAD Procedure, digital health features), with concordant coefficient signs. This convergent validity—across LR, ordinal regression, and the stacked ensemble—demonstrates that the reported associations are genuine data patterns rather than model-specific artifacts.

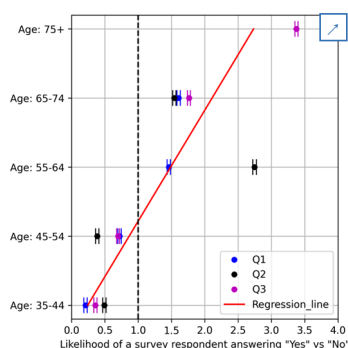


Figure S1 Age vs. Likelihood Univariate Trend. Age group vs. likelihood of unfavorable response across all three questions. A linear regression line ($R^2 = 0.724$) is overlaid showing the monotonic decrease in likelihood ratio with increasing age. Younger patients (35–44) show substantially elevated LR; older cohorts (65–74, 75+) consistently fall below a likelihood ratio of 1.0, indicating protective associations with age.

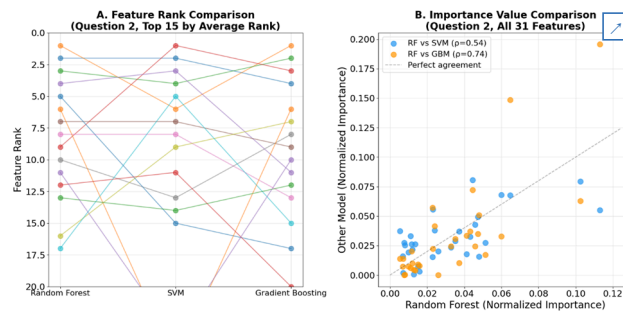


Figure S2 SHAP Feature Ranking Comparison. Feature importance ranking comparison across base classifiers (RF, SVM, GBM) for Question 2 (all 31 features retained). Panel A: bump chart showing rank trajectories of the top 15 features across models, illustrating rank instability in mid-tier features. Panel B: scatter plot of normalized importance values between model pairs, with Spearman correlations (RF↔SVM $\rho = 0.54$, RF↔GBM $\rho = 0.74$, SVM↔GBM $\rho = 0.51$). Moderate inter-model agreement confirms that no single classifier produces universally stable rankings, motivating the stacked ensemble approach to reduce model-dependent bias in feature attribution.

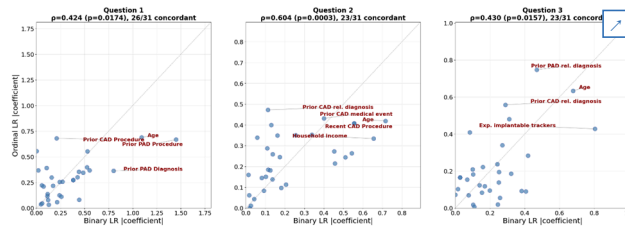


Figure S3 Ordinal vs. Binary Coefficient Comparison. Comparison of feature importance between ordinal (proportional-odds, 6-level outcome) and binary logistic regression models. Each panel plots the absolute standardized coefficient from the ordinal model (y-axis) against the binary model (x-axis) for all 31 patient features. Spearman ρ and direction concordance shown per question. Moderate-to-strong correlations ($\rho = 0.42$ – 0.60 , all $p < 0.02$) confirm that dichotomization does not alter the identified predictive structure.

Table S1 NLP-Based Survey Item Rankings

Rank	Survey Item	Cosine Similarity	Selected
1	I didn't understand what happened to me	0.524	
2	I felt like a burden to loved ones because I had to rely on them	0.523	★
3	I felt alone and overwhelmed in managing my condition	0.509	
4	I felt confident about where to look for information	0.465	
5	I was pleasantly surprised that my symptoms had been reduced	0.446	
6	I felt confused with what to do after my procedure	0.426	
7	I wasn't sure what actions would help most with my recovery	0.404	★

Rank	Survey Item	Cosine Similarity	Selected
8	I was frustrated by my health limitations	0.402	
9	I believed that I was “fixed”	0.386	
10	I was nervous or anxious about the future	0.385	
11	This was an emotional “wake-up call”	0.364	
12	I struggled with the lifestyle changes my physician talked about	0.356	★
13	I was optimistic about my ability to follow through	0.292	
14	I was disappointed that the procedure wasn’t the fix I was hoping for	0.204	

NLP-Based Survey Item Rankings. Beyond Intervention Year 3 post-intervention experience items (N=14) ranked by mean cosine similarity to quality-of-care-relevant HCAHPS domains. Embeddings generated using all-roberta-large-v1 (1024-dimensional) on item text only (no patient response data). Rankings confirmed across five diverse sentence-transformer architectures (pairwise Spearman $\rho = 0.48\text{--}0.74$, mean = 0.57, 8/10 pairs significant at $p < 0.05$). ★ = selected for final analysis via multi-criteria process (NLP similarity + response variation + sample completeness + expert clinical review).

Table S2 Additional patient features

Characteristic	N = 131
Education level	
Some grade school	2 (1.5)
High school diploma	17 (13.0)
Some college (no degree)	5 (3.8)
Associate degree	17 (13.0)
Technical or occupational certificate	18 (13.7)
Bachelor’s degree	38 (29.0)
Master’s degree	31 (23.7)
Doctorate	3 (2.3)
Insurance type	
No insurance	4 (3.1)
Private insurance	59 (45.0)
Government insurance	68 (51.9)

These features were included as model inputs but excluded from the likelihood analysis due to low model impact (below the Kneedle elbow threshold). Together with Tables 1 and 2, they complete the full set of 31 patient characteristics used in modeling. Values are N (%). All data was self-reported by patients. Private insurance includes employer-sponsored and individual plans. Government insurance includes Medicare, Medicaid, and Veterans Affairs coverage.