Supplementary Materials

Title

Behavioral profiles associated with adherence to adjuvant endocrine therapy in breast cancer: a retrospective population-based cohort study

Journal

Breast Cancer Research and Treatment

Authors

Aaron Dibner-Dunlap, MIA, ORCID 0000-0001-5558-8911[a]; Staci Sutermaster, MPH, ORCID 0000-0002-0100-5722[a], Peter Smittenaar, PhD, ORCID 0000-0002-7017-3439[a], and Sema Sgaier, PhD, ORCID 0000-0002-8311-2686[a,b]

Affiliations

[a] Surgo Health, Washington, DC, USA
[b] Department of Global Health, University of Washington, Seattle, WA, USA

Correspondence to: Sema Sgaier, PhD, 1701 Rhode Island Avenue, NW, Washington, D.C., 20036; semasgaier@surgohealth.com

Appendix: BehavioralPulse Feature Development

BehavioralPulse is a large-scale social and behavioral dataset developed to estimate latent attitudes, beliefs, preferences, and perceived barriers relevant to healthcare behaviors at the individual level across the United States. In this study, BehavioralPulse was used to generate individual-level behavioral propensity scores specifically related to breast cancer worries, mammography beliefs, side-effect concerns, social support, fatalism, and trust in healthcare. These features were then linked to medical claims data using a privacy-preserving record linkage approach.

Primary survey data collection

Breast cancer-specific BehavioralPulse features are derived from a proprietary, nationally representative survey of 6,000 U.S. women of age 40+. The survey was designed using established behavioral science frameworks and oncology expertise to capture constructs that are not observable in administrative or clinical data. These include breast cancer worries, mammography beliefs and barriers, side-effect concerns, social support, fatalism, health literacy, and trust in healthcare.

Surveys were fielded in English and Spanish using high-quality national panels. Survey items were structured to support binary or categorical measurement and downstream predictive modeling. Each item was treated as a distinct behavioral outcome, resulting in 362 candidate breast cancer-relevant variables that served as labeled targets for model training.

Predictive modeling and feature generation

To scale survey responses beyond the sampled survey population, we trained a set of supervised machine learning models, one per behavioral construct, using survey responses as ground truth labels. Predictor variables were drawn from a large, individual-level consumer and demographic dataset covering over 260 million U.S. adults, licensed from a third-party vendor. These predictors include both observed attributes (e.g., age, household composition, residential geography), modeled consumer characteristics, and proprietary neighborhood metrics. Race and ethnicity were not used as predictors in the behavioral models.

For each binarized behavioral outcome, we trained models on a subset of survey respondents (70%) and evaluated on held-out test data (30%). Multiple model classes were evaluated depending on outcome prevalence and complexity, with hyperparameters tuned to optimize out-of-sample discrimination and calibration.

Model performance and quality assurance

All BehavioralPulse models undergo standardized quality assurance prior to inclusion in analytic workflows. Model discrimination was assessed using the area under the receiver operating characteristic curve (ROC-AUC) on held-out test data. For breast cancer-related features used in this study, ROC-AUC values typically ranged from approximately 0.65 to above 0.90, depending on outcome prevalence and construct complexity.

Calibration was assessed using reliability curves to ensure alignment between predicted probabilities and observed response frequencies across the score distribution. Models exhibiting poor calibration or unstable performance were excluded. For selected constructs with available external benchmarks, we compared aggregated predictions against public datasets at appropriate geographic or population levels to assess plausibility.

Population-scale estimation

Validated models were applied to a national consumer dataset to generate behavioral propensity scores for over 45 million U.S. women aged 40+. The final model outputs are individual-level propensity scores ranging from 0 to 1, representing the estimated probability that an individual holds a given belief or concern. This process yielded a dense, individual-level behavioral feature set. These features constitute the BehavioralPulse behavioral layer used in the present analysis. For use in the regression analyses described in the main manuscript, these propensities were binarized at a threshold that results in the known population frequency of the underlying feature to facilitate interpretability of the odds ratios.

BehavioralPulse is designed to reflect population dynamics over time. The underlying consumer dataset is refreshed monthly to capture e.g. residential mobility and household changes, while behavioral models are updated on an as-needed basis.

Data linkage and privacy protection

BehavioralPulse features were linked to medical claims data using a privacy-preserving, hash-based tokenization approach [37, 38]. Individual identifiers were converted into encrypted tokens prior to linkage, ensuring that no personally identifiable information was exchanged or exposed. This linkage enabled the integration of behavioral propensities with observed breast cancer diagnoses while maintaining compliance with applicable data protection and privacy standards.

Interpretation and limitations

BehavioralPulse features are modeled probabilities rather than direct observations of individual beliefs or behaviors. They should therefore be interpreted as estimates of latent propensities that complement, rather than replace, observed clinical and administrative data. Prediction error in these estimates is expected to attenuate associations toward the null, implying that observed effect sizes likely represent conservative estimates of the influence of behavioral factors.