Quantifying Missingness in Real World Data for Enhanced Data Quality Assessment

MSR189

Arielle Marks-Anglin and Yu-Han Kao

Merck & Co., Inc., Rahway, NJ, USA

Background and objective

- Missingness in real-world data (RWD) is a prevalent issue that can lead to biased results and reduced precision in estimates.
- Researchers often address the impact of missing data during the analytic phase, after the sample has already formed. However, missing data can also impact sample selection, especially when considering inclusion/exclusion criteria. Many studies adopt a complete case analysis approach during the cohort attrition stage, often without evaluating the potential effect of missing data on the inferences drawn.
- This study aims to quantify the impact of missing data at the sample formation stage by comparing patient characteristics and overall survival using the original sample and an imputed sample. The assessment is conducted within the context of overall survival for a breast cancer cohort, focusing on the availability of BRCA1 and BRCA2 gene testing data.

Methods

- Patients diagnosed with early-stage breast cancer were selected from Syapse's enriched breast cancer cohort. Patients were required to have known HR status and HER2 status.
- BRCA testing status and results were extracted from the patients' EHR record. Tested and untested patients were compared using Wilcoxon rank-sum (for age) and Chi-square tests.
- A random forest model was trained on the subset of patients that received testing for the BRCA1 and BRCA2 genes. The model predicted BRCA positive vs. negative status. Cross-validation was performed for model tuning.
- Predictors include demographic and tumor characteristics (staging, size, histology, grade, ER/PR & HER2 status), comorbidities, family history and time to recurrence or metastasis.
- Overall survival (OS) was compared between BRCA-positive and BRCA-negative patients, using (1) the original tested results and (2) the original + imputed test results, with BRCA status imputed for patients who did not receive testing

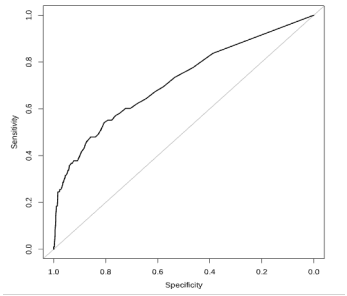
Results

- 23,071 adult patients diagnosed with stage I, II, or III breast cancer between January 1, 2016, and April 11, 2024 were identified in the enriched breast cancer cohort.
- 9,475 (41%) of the cohort received BRCA testing, and 322 (3.%) of those tested were BRCA positive. Among 3,721 patients indicated for BRCA testing under current guidelines, 2,724 (73%) were tested for BRCA
- Patients tested were on average younger, more likely to be married, have private insurance, higher household income, be ER/PR negative and have a family history of breast and ovarian cancer.

Table 1: Comparing patients tested and not tested for BRCA1/BRCA2 genetic variant

Characteristic, n (%)	Not tested, N=13,596	Tested, N=9,475	p-value[a]
Mean age at diagnosis (yrs) (SD)	66 (12)	57 (13)	<0.001
Female	13,533 (100%)	9,350 (99%)	<0.001
Race			0.012
White	10,847 (80%)	7,666 (81%)	
Black/African American	2,035 (15%)	1,293 (14%)	
American Indian/Alaska Native	40 (0.3%)	30 (0.3%)	
Asian	486 (3.6%)	378 (4.0%)	
Other/Not Provided	188 (1.4%)	108 (1.1%)	
Ethnicity			<0.001
Hispanic/Latino	711 (5.2%)	590 (6.2%)	
non-Hispanic/non-Latino	12,117 (89%)	8,550 (90%)	
Unknown	768 (5.6%)	335 (3.5%)	
Marital status			<0.001
Married	6,786 (50%)	5,569 (59%)	
Single	2,086 (15%)	1,598 (17%)	
Other or Unknown	4,724 (35%)	2,308 (24%)	
Median household income			<0.001
Q1 (<$46,458)	1,458 (11%)	798 (8.5%)	
Q2 (>=$46,458 and <$57,955)	1,703 (13%)	1,091 (12%)	
Q3 (>=$57,955 and <$74,086)	3,419 (25%)	2,198 (23%)	
Q4 (>=$74,086)	6,887 (51%)	5,295 (56%)	
Unknown	129	93	
Insurance			<0.001
Medicare/Medicaid	7,615 (56%)	3,422 (36%)	
Private	5,010 (37%)	5,328 (56%)	
Other	586 (4.3%)	541 (5.7%)	
Not insured/Unknown	385 (2.8%)	184 (1.9%)	
Clinical stage group at diagnosis			<0.001
I	10,159 (75%)	6,580 (69%)	
II	2,565 (19%)	2,099 (22%)	
III	872 (6.4%)	796 (8.4%)	
ER negative at diagnosis	1,694 (12%)	1,985 (21%)	<0.001
PR negative at diagnosis	3,325 (24%)	2,868 (30%)	<0.001
HER2 negative at diagnosis	11,664 (86%)	8,055 (85%)	0.003
Family history of breast cancer	5,459 (61%)	6,377 (78%)	<0.001
Unknown	4,612	1,278	
Family history of ovarian cancer	717 (8.4%)	1,307 (17%)	<0.001
Unknown	5,017	1,980	
Patient has other cancer	2,392 (18%)	1,838 (19%)	<0.001
Unknown	32	12	
Metastatic	425 (3.1%)	469 (4.9%)	<0.001
Unknown	4	0	

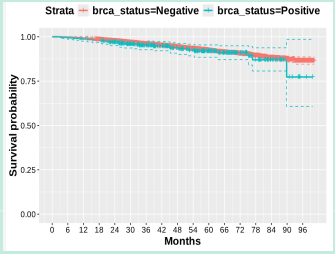
Figure 1: ROC curve for random forest model predicting BRCA status (positive vs. negative)



- The final random forest model had an AUC-ROC of 0.69.
- Youden's method was used to select an optimal probability threshold, resulting in 54% sensitivity and 81% specificity in the model test-set.

Tested patients only

Figure 2a: Overall survival from diagnosis date, by BRCA status (Tested patients only)

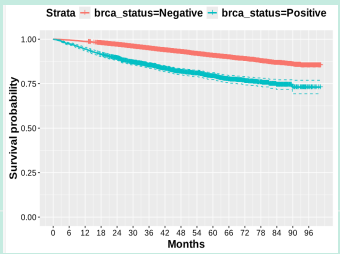


Hazard ratio[b] using tested patients

0.83 (95% CI: 0.55, 1.26)

All patients (with imputed BRCA status for non-tested patients)

Figure 2b: Overall survival from diagnosis date, by BRCA status (All patients)



Hazard ratio[b] using all patients

1.21 (95% CI: 1.07, 1.37)

- 627 (6.7%) of patients in the tested sample died during follow-up, compared to 2,180 (9.5%) in the full sample
- Difference in overall survival is substantially wider when imputing BRCA status for non-tested patients, compared to when using BRCA tested patients only.

[a] Wilcoxon rank-sum test for continuous variables; Chi-square test for categorical variables
[b] Cox PH model adjusted for age at diagnosis, race and ethnicity, Charleson comorbidity score at diagnosis, and clinical stage group at diagnosis

Discussion and next steps

- Differential missingness at the sample selection stage can have meaningful impacts on study results. Attention should be given to the missingness mechanism when interpreting estimates.
- Imputation of BRCA variant status resulted in increased power and a significant change in estimated effect of BRCA status on overall survival. This sensitivity analysis and comparison helps to contextualize the results from the tested sample.
- While the imputation model identified 54% of 'missing' cases in the validation sample, positive predictive value was fairly low (though higher than random chance). Further development and exploration of alternative models (e.g. deep learning) may lead to improved discrimination of positive and negative BRCA cases. The known misclassification rates (Sensitivity, specificity, PPV and NPV) rates can futher be used in a two-stage process involving probabilistic bias analysis[1].

Acknowledgments
Yezhou Sun for his prior work on the enriched breast cancer cohort and study cohort identification

Contact information
Arielle Marks-Anglin, arielle.marks-anglin@merck.com; Yu-Han Kao, yu-han.kao@merck.com

References

1. Hunnicutt, J. N., Ulbricht, C. M., Chrysanthopoulou, S. A., & Lapane, K. L. (2016). Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review. *Pharmacoepidemiology and drug safety*, 25(12), 1343-1353.

Presented at ISPOR EU, Barcelona, November 20, 2024