The long-term financial burden of breast cancer: experiences of a diverse cohort of survivors identified through population-based registries

Reshma Jagsi, et al

Supplementary Appendix: Methods of Weighting and Imputation

To allow statistical inferences to be more representative of the original targeted population, we applied complex survey weights to the data collected. First, we created design weights to account for the oversampling of African Americans and Latinas. Second, we created non-response weights to account for the fact that persons with certain characteristics were not as likely to respond to the surveys at each time point.

Design weights compensate for the disproportionate selection across race and SEER sites; survey unit non-response weights compensate for the fact that women with certain characteristics were not as likely to respond to the surveys. Patients who did not respond to both surveys were more likely to be African American—35.2% v. 26.7%, P<.001; to be Latina—17.2% vs. 13.3%, p=0.002; to have stage II-III disease—54.9% v. 37.8%, P<0.001; and to have had mastectomy—37.5% vs. 30.8%, P<0.001.

We utilized multivariable logistic regression analyses to create the non-response weights. We multiplied design weights and non-response weights to create the final overall weights and then normalized those weights so that the sum total of the weights equaled the total number of the respondents. All percentages and p values reported in this manuscript reflect, where appropriate, adjustment using these weights.

In addition, we utilized the multiple imputation technique of fully conditional specification (FCS) to minimize the impact of survey item non-response and incomplete-case-deletion methods in the modeling process. First, we examined whether any variables theoretically selected for inclusion in the two multiple variable models were missing substantial data due to item non-response. We found that item non response was very low for most variables, and uniformly <5% (as depicted in Table 1) for all variables except income. We calculated imputed income, defined as <$50,000 versus $50,000+, using logistic regression models with the following covariates: age at diagnosis, education attainment, race, marital status, and employment status at diagnosis. We imputed the data twenty times, creating 20 distinct samples, on which the same logistic regression models for experiencing worsening financial status at least partly attributed to breast cancer and for experiencing at least one major privation were each estimated with the results combined using Rubin's formula.[1,2] The remaining covariates in these models, with missing information less than 5%, were not imputed, due to the relatively low level of missing data and the increased complexity of imputing those covariates with accuracy.

References

1. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons;1987.

2. Berglund PA. An Introduction to Multiple Imputation of Complex Sample Data using SAS®. SAS Global Forum 2010. Paper 265-2010.