**SUPPLEMENTARY MATERIALS**

**Section A. Multiple Imputation of Missing Values, Derivation of Predicted Values, and Adjusting Standard Errors in Regression Models to Account for the Resulting Extra Sampling Variability**

Estimation of any two-stage residual inclusion (2SRI) model requires the computation of residuals from the first-stage regression, which then comprise the elements of a predictor variable (Residual) employed in the second-stage regression. Because Residual is derived as a function of the first-stage predictor variables, its values will reflect the sampling variability of those first-stage variables. Correspondingly, the standard errors of the odds ratios for all predictor variables in the second-stage, including Residual, should incorporate this extra sampling variability (1, 2). Similarly, when multiple imputation (MI) procedures are used to fill in missing data, the imputed variable values are derived as a function of observables that reflect sampling variability. Hence, the imputed values for each such variable are imbued with sampling variability, which should be accounted for in the standard error of the estimated odds ratios for the variable (3). Our base-case model for the determinants of an advanced stage diagnosis of breast cancer employs both 2SRI and MI. Thus, the model needs to be estimated in a way that permits both sources of sampling variability to be reflected in the standard errors. Below we briefly describe how these issues were addressed (with the key MI statements from Stata Version 12 in parentheses); greater detail, including the code employed, is available from the corresponding author.

1. 2SRI Model without Missing Value Imputation (e.g., model #8 in Table 4). We created 100 bootstrapped samples (that is, samples constructed by sampling with replacement) involving all the variables in the first-stage and the second-stage regression equations for the 2SRI model – except (of course) Residual, whose values were derived from the first-stage regression for each bootstrap sample. We then executed the bootstrapped analysis, wherein the standard error of the estimated odds ratio for each variable in the second-stage equation was equal to the standard deviation of the 100 odds ratios estimated for that variable. (To accomplish this, in a way that accounted for Residual, the authors wrote their own bootstrap program in Stata.)
2. Missing Value Imputation Only (e.g., models #5, 11, 12 in Table 4). The multiple imputation process here involves essentially three steps:
3. Imputation. In the base-case analysis, there were three variables with a significant number of missing observations: mole-subtype (28.1%), BMI (22.7%), and grade (9.1%). For each variable, patients with missing values were assigned imputed values via a regression model using all patients not missing that variable, with predictor variables chosen to yield statistically tight predictions. For the continuous variable BMI, a linear regression model was estimated by ordinary least squares. For mole-subtype and grade, each of which is a 4-level categorical variable, the regression models were multinomial logit. The predictors employed for all three variables were method of detection, stage at diagnosis, age category, comorbidity score, hypertension (yes/no), diabetes (yes/no), race/ethnicity status, educational status, marital status, urban/rural status, insurance, and histology status. (By implication, there is no requirement that predictors be limited to those implied by our conceptual model, nor is there a concern about model overfitting; rather, the aim is to arrive at highly accurate within-sample predictions for missing observations.)

The next step was to build in sampling variability to the set of predicted values for each variable, since they are in fact derived from an estimated regression. To accomplish this, random draws were made from the simulated error term distribution for the regression, and the *imputed value* for each missing value was the sum of the predicted value for the individual and a random error term draw.

(The key Stata commands for multiple imputation here are “mi set”; “mi register imputed”; “mi register passive”; “mi register regular”; and “mi impute chained”.)

1. Completed Data Set Analysis. Consistent with current practice, we performed 10 imputations (each as noted in a.), yielding 10 completed data sets. With each data set, we estimated the regression model of interest, e.g., single-equation logistic model for predicting advanced-stage diagnosis, resulting in 10 estimated models in parallel.
2. Pooling to Get Final Estimated Model. For each predictor variable in the model of interest, its odds ratio is the average of the 10 estimated odds ratios from step b.; and the corresponding standard error – which reflects the sampling variability imbedded in the imputed values – is a function of the estimated regression coefficients and standard errors for that variable across all 10 estimated models (3). (Key Stata command for b. and c. is "mi estimate, post: logit".)

An important (though generally not directly testable) assumption for MI to lead to consistent estimates of regression parameters and odds ratios is that the missing data are “missing at random” (MAR). This means that the probability a variable is missing for a given individual is not a function of the (unobserved) value of the variable, though it may be a function of other, observed variables. Indeed, those latter variables are the ones that should be employed in predicting (replacement) variable values in the imputation step of MI.

1. 2SRI Model with Missing Value Imputation (e.g., #7 (Base-case), 9, 10 in Table 4). In this case, we have two sources of sampling variability to reflect appropriately in the regression model standard errors. In response, we combined the MI and bootstrap procedures in the following way. First, we conducted imputation, as described in 2.a., and now it is useful to characterize what emerged as one large data set (call it A), consisting of all the variables in the 2SRI model that were not imputed *and* the implied 30 variables here that were imputed in the 10 successive imputations: namely,

mole-subtype1,…..,mole-subtype10, BMI1,…, BMI10, grade1,….,grade10. Data set A can be so described since all the non-imputed variables take on precisely the same values across the 10 imputed data sets. Then, we applied the bootstrap method to A in the estimation of the 2SRI model of interest in order to adjust the SEs in the second stage. Specifically, we drew 100 bootstrap samples from A, and estimated the 2SRI model for each sample, which also involved deriving values for Residual for use in the second-stage regression. In the final estimated 2SRI model, the standard error of the odds ratio for each variable was the standard deviation of the 100 estimated odds ratios for the variable. In this way, the standard errors ultimately reflect the sampling variability from both the MI process and the fact that Residual is computed. (Key Stata commands are "program"; "mi estimate, saving( ): logit"; "mi estimate, post: logit"; and "bootstrap:".)

 The computer programs used to carry out our bootstrapping analyses and imputation of missing variable values are available from the corresponding author upon request. The reader should be aware, however, that the codes developed here (implemented in Stata) are highly specific to models employed in this paper.

**Section B. First-Stage Regression Analysis in Support of the Base-Case Two-Stage Residual Inclusion (2SRI) Instrumental Variables Model for Breast Cancer Stage of Diagnosis**

The estimated model below is specified expressly as prescribed in the 2SRI instrumental variables (IV) approach to estimation in the face of potential selection bias. Such bias is a threat if method of detection is an important predictor of stage at diagnosis and if stage and method are both influenced by certain common covariates, some of which are unobservable in the current data set. Under these hypothesized circumstances, the first-stage regression is a binary logistic formulation in which the log odds of detection by mammography (versus some other method) is a function of (1) all the other predictors in the second-stage regression for stage at diagnosis and (2) the selected IV’s – state, histology, and mammography-capacity – which play a pivotal role in the 2SRI model’s effort to detect and reduce selection bias. A “strong” IV in the current context will be highly significantly related to method of detection (the “intervention”) while not being significantly related to stage at diagnosis (the “outcome”) *except* through its influence on method of detection.

While mammography-capacity is a compelling measure of the woman’s physical access to screening, there will still be individual-specific variations in actual distance and time to screening facilities and that may influence the propensity to be tested. Elkin et al. (4) found that distance or time to appointment may or may not influence screening behavior, depending on a complex of factors.

Although the model below is thus ancillary to the main task of predicting stage, the results are arguably of interest in their own right:

**Section B Table: First-Stage Logistic Regression Results for 2SRI Model in the Base Case: Determinants of Detection of Breast Cancer by *Mammograph*y vs. *Other* Methods**

| ***Variable*** | *Odds ratio* |  | *p|* | *95% CI* |
| --- | --- | --- | --- | --- |
| ***Age group*** |  |  |  |  |  |
|  *50-64* | Ref |  |  |  |  |
|  *Under 40* | 0.21 |  | < 0.001 | 0.16 | 0.27 |
|  *40\_49* | 0.62 |  | < 0.001 | 0.55 | 0.71 |
|  *65\_69* | 1.38 |  | 0.001 | 1.15 | 1.66 |
|  *70\_79* | 1.04 |  | 0.61 | 0.89 | 1.23 |
|  *80+* | 0.73 |  | 0.003 | 0.59 | 0.90 |
| ***Mole-Subtype*** |  |  |  |  |  |
|  *Luminal A* | Ref |  |  |  |  |
|  *Triple Negative*  | 0.63 |  | < 0.001 | 0.53 | 0.75 |
|  *HER2 Type* | 1.17 |  | 0.08 | 0.99 | 1.38 |
|  *Luminal B*  | 0.80 |  | 0.07 | 0.63 | 1.02 |
| ***Grade*** |  |  |  |  |  |
|  *Well differentiated* | Ref |  |  |  |  |
|  *Moderately differentiated* | 0.80 |  | 0.002 | 0.69 | 0.92 |
|  *Poorly or undifferentiated* | 0.51 |  | < 0.001 | 0.44 | 0.60 |
| ***Comorbidity status*** |  |  |  |  |  |
|  *None* | Ref |  |  |  |  |
|  *Mild*  | 1.05 |  | 0.39 | 0.94 | 1.18 |
|  *Moderate*  | 0.85 |  | 0.10 | 0.71 | 1.03 |
|  *Severe*  | 0.60 |  | < 0.001 | 0.46 | 0.79 |
| ***Race/ethnicity*** |  |  |  |  |  |
|  *White* | Ref |  |  |  |  |
|  *Black* | 0.93 |  | 0.29 | 0.82 | 1.06 |
|  *API* | 0.79 |  | 0.05 | 0.63 | 1.00 |
|  *AI\_AN* | 0.93 |  | 0.81 | 0.53 | 1.65 |
|  *Hispanic* | 0.72 |  | 0.001 | 0.59 | 0.88 |
| ***Education status*** |  |  |  |  |  |
|  *<25% w/o HS educ* | Ref |  |  |  |  |
|  *≥ 25% w/o HS educ* | 0.92 |  | 0.21 | 0.81 | 1.05 |
| ***Marital status*** |  |  |  |  |  |
|  *Married* | Ref |  |  |  |  |
|  *Single* | 0.84 |  | 0.02 | 0.72 | 0.97 |
|  *Separated / widowed /* *divorced* | 0.85 |  | 0.006 | 0.75 | 0.95 |
| ***Poverty status*** |  |  |  |  |  |
|  *< 20% below federal level* | Ref |  |  |  |  |
|  *≥ 20% below federal level* | 1.01 |  | 0.94 | 0.87 | 1.16 |
| ***Urban/rural status*** |  |  |  |  |  |
|  *Urban* | Ref |  |  |  |  |
|  *Rural* | 0.92 |  | 0.34 | 0.78 | 1.09 |
|  *Urban-rural mix* | 1.07 |  | 0.23 | 0.96 | 1.21 |
| ***Insurance*** |  |  |  |  |  |
|  *Private*  | Ref |  |  |  |  |
|  *Uninsured* | 0.56 |  | < 0.001 | 0.42 | 0.76 |
|  *Medicaid* | 0.71 |  | < 0.001 | 0.61 | 0.83 |
|  *Medicare* | 0.90 |  | 0.15 | 0.78 | 1.04 |
| ***BMI*** |  |  |  |  |  |
|  *< 25* | Ref |  |  |  |  |
|  *25 - 29.9*  | 1.25 |  | 0.001 | 1.10 | 1.43 |
|  *30 - 39.9* | 1.32 |  | < 0.001 | 1.14 | 1.52 |
|  *≥ 40* | 1.30 |  | 0.03 | 1.03 | 1.65 |
| ***State*** |  |  |  |  |  |
|  *LA* | Ref |  |  |  |  |
|  *WI* | 1.23 |  | 0.03 | 1.02 | 1.47 |
|  *NC* | 0.97 |  | 0.74 | 0.81 | 1.16 |
|  *CA* | 1.13 |  | 0.19 | 0.94 | 1.35 |
|  *KY* | 0.90 |  | 0.34 | 0.72 | 1.12 |
|  *GA* | 1.12 |  | 0.12 | 0.97 | 1.29 |
| ***Histology*** |  |  |  |  |  |
|  *Lobular* | Ref |  |  |  |  |
|  *Ductal or*  *Ductal/lobular*  | 2.07 |  | < 0.001 | 1.67 | 2.56 |
|  *Other* | 2.79 |  | < 0.001 | 2.21 | 3.50 |
| ***Mammography-Capacity*** | 1.11 |  | 0.02 | 1.02 | 1.21 |

**Section C. Propensity Score Analysis as an Approach to Bias Reduction in Estimating the Determinants of Stage**

 An alternative, and frequently used, statistical approach to reduce bias when estimating treatment-outcome relationships with non-experimental data is propensity score (PS) modeling (1), which can take several specific forms: PS matching of patients across treatment arms; the separate estimation of treatment impacts within fractiles (typically quintiles) of patients who have “sufficiently similar” propensity scores; and the use of patient-specific PS weights in full-sample regression analyses (5-7). Regardless of the specific PS technique selected, a successfully executed analysis will result in a satisfactory balance of predictor variable values across treatment arms, just as one expects if the patients had instead been randomized successfully. A necessary condition for PS to achieve its bias-reducing potential is that there be no substantively and statistical important *unobservable* variables that influence the treatment-outcome relationship. Indeed, a recent simulation analysis found that PS adjustments may actually exacerbate the bias if the unobservables are sufficiently important relative to the observable variables in the model (8). Because unobservables may be influential in the method-of-detection/stage-at-diagnosis relationship, we selected an instrumental variables approach (2SRI) rather than PS for the base-case and most sensitivity analyses.

 Nonetheless, given the extensive application of propensity score methods in outcomes research now and also that the *degree* of bias reduction through IV models is ultimately an empirical matter, we elected to include PS models for the likelihood of advanced stage in our sensitivity analyses. In a similar vein, several recent studies of cancer treatment effects have employed both PS and the 2SRI modeling (9, 10).

We used two alternative PS weighting schemes (Table S5) and then examined whether covariate values were adequately balanced between the mammography and other methods arms under either/both weighting schemes (Table S6). The schemes, as employed in a recent analysis of cancer therapy selection and impact (9), are the “inverse probability of treatment weighting” (IPTW) approach and the “standardized mortality ratio weighting” (SMRW) approach (7).

In the present context, IPTW assumes that women in the sample are sufficiently similar that mammography could be regarded as a clinically and behaviorally plausible method of detection, regardless of the woman’s actual method. The SMRW approach assumes, instead, that the observed “treatment effect” (resulting from detection by mammography) applies specifically to women whose characteristics are aligned with the portion of the sample actually detected by mammography. As shown in Table S6, both the IPTW and SMRW approaches led to a successful rebalancing of the distribution of covariate values between the mammography and the other methods of detection “arms.” Prior to PS weighting, the null of no imbalance in covariate values could be rejected for 9 of the 11 predictor variables; after reweighting the data under either approach, the resulting high p values on the null for all 11 variables indicate balance was achieved.

**Section D. Sensitivity Analyses Examining Robustness of Findings on Determinants of Advanced-Stage Breast Cancer Diagnosis to Variations on the Base-Case Model**

In each of Tables S1-S5 below, some particular operating assumption made in the base-case models (Table 3) is altered, and we examine the impact on the direction and strength of individual predictors and overall model performance. Table S6 reports analyses undertaken to appraise the statistical performance of the propensity-score models presented in Table S5.

The first five tables here are formatted and organized identically to Table 3, so that the primary element requiring additional explanation is the reported sample size for model estimation, compared with N=7,503 in Table 3. In Table S1, the net impact of eliminating the “tumor biology variables” (mole-subtype, grade, BMI) is to increase the sample to 7,610. Even though these 3 variables are imputed in the base case, not all of the cases receiving imputed values are included in the Table 3 estimates, since a small number will have missing values for one or more of the other predictor variables; removing the 3 variables from the analysis serves to restore just over a hundred patients to the sample. When there is no imputation of missing values for these 3 variables (Table S2), the available sample (that is, patients with no missing variable values whatsoever) shrinks to 4,078. Re-estimating the base-case models using the sample weights (Table S3) has no effect on sample size relative to the base case. When the statistically important instrumental variable, mammography-capacity, is removed from the first-stage regression of the 2SRI model, so that observations from the state of Minnesota are added, the resulting net uptick in sample makes N = 8,385. When propensity score weighting is employed (rather than instrumental variables) to adjust for possible selection bias (Table S5), the sample is identical to that used in the base-case (N = 7,503).













**Section E. Breast Cancer Method of Detection and Stage at Diagnosis: Some Interpretative Issues**

In this paper there are several defined modalities for detecting breast cancer: screening mammography (call it now M), clinical breast exam (CBE), breast self-exam (BSE), and symptom-driven discovery. The latter three comprise the “other” category for our dichotomous method of detection variable. Note that the frequency and time-spacing with which each – or any – of M, CBE, BSE, and symptom-driven discovery are being executed by the at-risk woman cannot be observed in our data set; what we know is the recorded Method around the time of diagnosis. Note also that each of M, CBE, BSE, and symptom-driven discovery can generate true and false positives and true and false negatives, and at different rates that vary not only by the underlying “technology” of the test but also the physiological make-up of the woman, including (for example) tumor growth rates, mammographic breast density, and the size and location of the breast tumor if there is one to be detected. In sum, in the observational setting of our study, there are multiple factors potentially at work prior to the moment when the breast cancer is diagnostically confirmed and a method of detection is recorded.

As one of the reviewers of this paper emphasized, there lurks here that matter of the “implicit counterfactual.” For a woman diagnosed at advanced stage (IIIB, IIIC, or IV) whose recorded method of detection is other, one should not infer that *if* the woman had been undergoing M, she would necessarily have been detected at an Earlier Stage (0, I, IIA, IIB, or IIIA). There are several reasons to be cautious here: (1) the woman may have just started doing M, at a time when her tumor was already at an advanced stage (this is referred to as the prevalent screen, and in both randomized trials and evaluations of service screening, it always has yielded a mix of early and advanced tumors); (2) even for those doing M more frequently, the test could yield a false negative (especially for women with very mammographically dense breasts or with lobular histology); and (3) when the M was done, the tumor was very small and missed, but it happened to be a very aggressive tumor that moved to advanced stage before the next scheduled M (heightening the likelihood of a “interval” detection). Consequently, we have not claimed (nor does anyone else today) that if a woman gets regular mammograms, her cancer always will be detected by mammography at an early stage.

There are kindred variations on the theme: For a woman diagnosed at an Earlier Stage by M, we must be cautious in inferring that she would have been diagnosed at an Advanced Stage, especially as defined here, without M. Some women whose cancers are self-detected are diagnosed at an early stage simply because the tumor is palpable but not aggressive, i.e., an “underachiever.” In the randomized controlled trials for mammography screening, interval cancers diagnosed in the invited arm tended to have more favorable stage at diagnosis simply because women had a heightened sense of awareness and responded to symptoms sooner compared with the control groups. More specifically, if a woman also was receiving CBE, it is possible that *that* test could have triggered an Earlier Stage diagnosis, notwithstanding its inferior sensitivity relative to M. Indeed, for a woman with a tumor not yet exhibiting symptoms and who is doing both M and CBE, the cancer could conceivably be detected by either test, depending on the timing of their execution relative to the emergence of a sufficiently discernible mass that she herself would notice via BSE. The test that “got there first” will be given credit for the diagnosis in the woman’s clinical notes (that’s how we identified the Method for each patient in this study), but that does not mean the other test(s) could not also have gotten the job done successfully.

Analyses of the association between invitation and exposure to screening in both randomized controlled trials and observational studies have focused mostly on mortality. Only a few observational studies, to our knowledge, have brought both theory and data to bear to investigate the effectiveness of screening mammography (or CBE) in reducing the likelihood of an advanced stage breast cancer. Indeed, our paper does not lay claim to that ambition either. *Rather, the role of the method of detection variable in our analyses is to account for what might be termed the “net overall impact” of these unobserved (to us) screening behaviors on the likelihood of an advanced stage diagnosis.* And we do so in a way that allows for the possibility that some certain observable and unobservable variables influence both method and stage (e.g., breast density), although possibly in different ways; this is the rationale for the instrumental variables analysis. At the same time, this modeling approach permits us to examine the impact of other important variables, such as race/ethnicity and insurance status, on stage while adjusting for method of detection (and while also accounting for the role those variables play in influencing method).

**References**

1. Pizer SD. An intuitive review of methods for observational studies of comparative effectiveness. *Health Serv Outcomes Res Method* 2009;9:54-68.
2. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J Health Econ* 2008;27:531-543.
3. Allison PD. *Missing Data*. Thousand Oaks, CA: Sage, 2001.
4. Elkin EB, Ishill NM, Snow JG, Panageas KS, Bach PB, Liberman L, et al. Geographic access and the use of screening mammography. *Med Care* 2010;48:349-356.
5. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757-763.
6. D’Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of treatment to non-randomized control group (Tutorial in Biostatistics). *Statist Med* 1998;17:2265-2281.
7. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006;163:262-270.
8. Brooks JM, Ohsfeldt RL. Squeezing the balloon: propensity scores and unmeasured covariate balance. *HSR: Health Services Research* 2013;48(4):1487-1507.
9. Hadley J, Yabroff KR, Barrett MJ, Penson DF, Saigal CS, Potosky AL. Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data. *J Natl Cancer Inst* 2010;102(23):1780-1793.
10. Wright JD, Ananth CV, Tsui J, Glied SA, Burke WM, Lu Y-S, et al. Comparative effectiveness of upfront treatment strategies in elderly women with ovarian cancer. *Cancer* 2014;120:1246-1254.