Imputing Missing Estrogen Receptor Status from Population-based SEER Cancer Registries

Nadia Howlader
Mathematical Statistician, MS
Surveillance Research Program, NCI

NAACCR Cancer Surveillance Webinar Series

NIH NATIONAL CANCER INSTITUTE

May 20, 2016

Overview

- Part I
  - Background & motivation
  - Data source & ER status missing patterns
  - Imputation method
  - Results
  - Discussions
  - Brief description of imputation of missing HER2 status
- Part II
  - Demonstrate how to use imputed dataset in SEER*Stat

Background

- Epidemiologic studies examining trends of tumor subtypes are important, e.g. estrogen receptor (ER), progesterone receptor (PR) status
- Tumor markers are prone to missing data. Why?
- Therefore, it is important to understand extent of missing information and impact of missing tumor markers when assessing trends

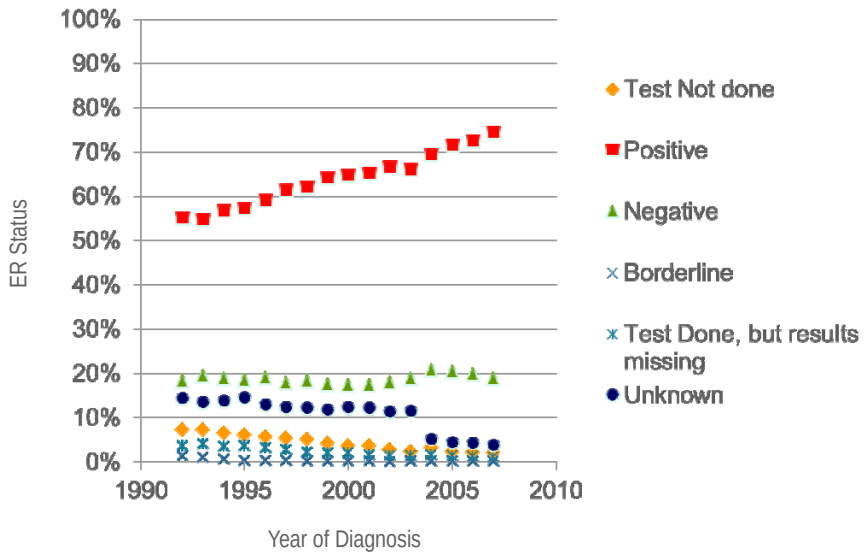
Objective

- Describe missing pattern with ER status (main variable of interest) and explore other related variables
- Impute missing ER status
- Present breast cancer incidence trends by original (ignoring missing ER) and imputed ER status

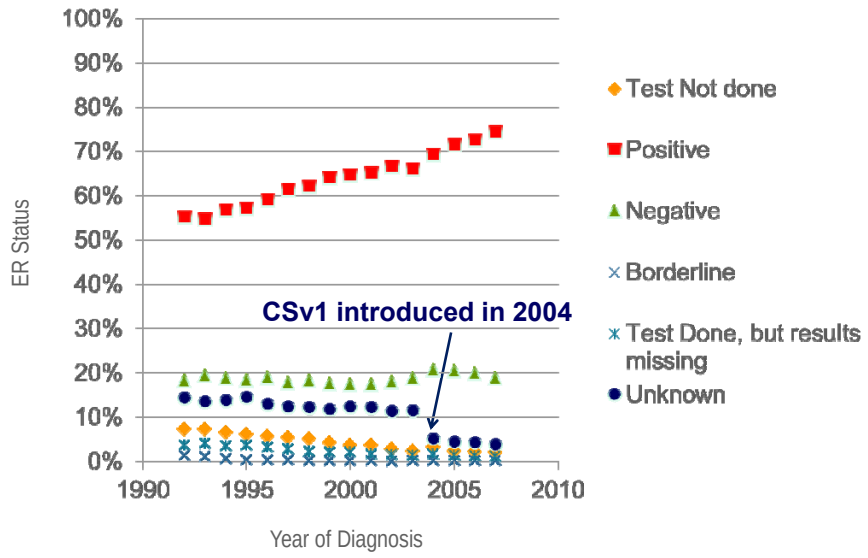
Study Cohort

- SEER-13 cancer registries, representing ~14% of total US population
- Female breast cancer patients diagnosed between 1992-2007 (malignant cases only)
- N = 401,741

Distribution of ER Status, 1992-2007



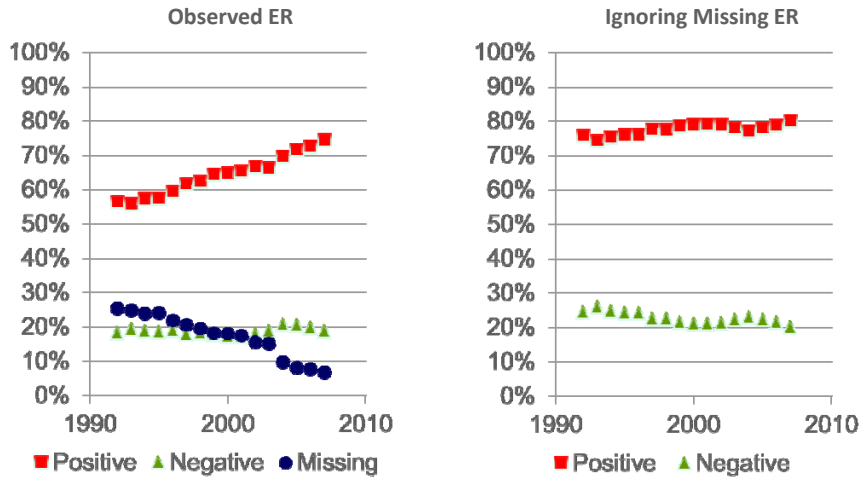
Distribution of ER Status, 1992-2007



Define ER Status for Analysis

- Positive ER Status
  - positive + borderline
- Negative ER Status
  - negative
- Missing ER Status
  - test not done
  - test done, but results are not interpretable
  - unknown

Distribution of ER Status, 1992-2007



SEER Breast Cancer Missing Data*

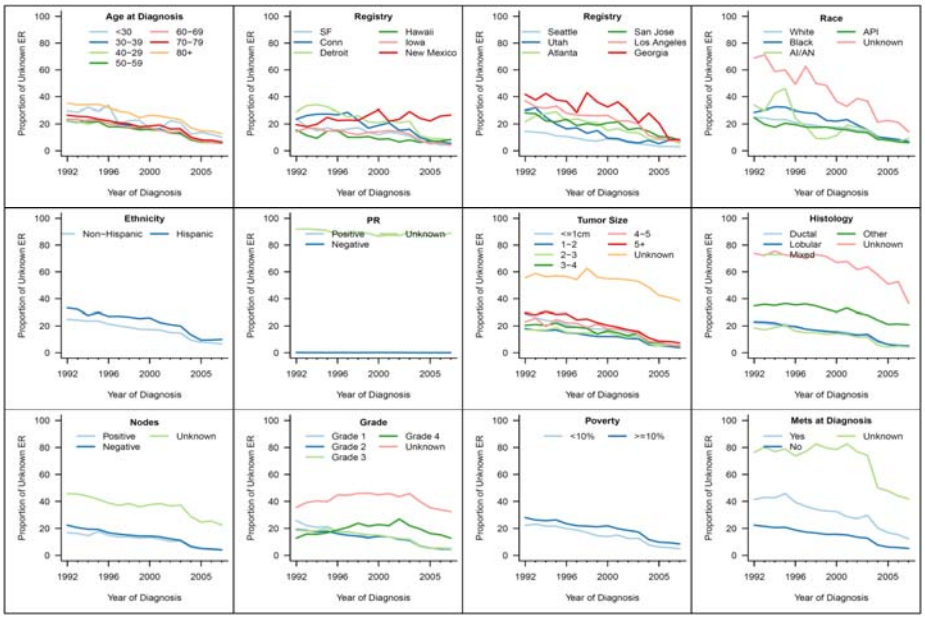
Variables	% Missing
ER status	17%
PR status	19%
Tumor size	8%
Histology	2%
Node positive status	14%
Grade	14%
Presence of metastasis	4%

*Age at diagnosis and county level poverty were minimally missing (< 0.5% of cases); Registry, year of diagnosis, Hispanic ethnicity had no missing information.

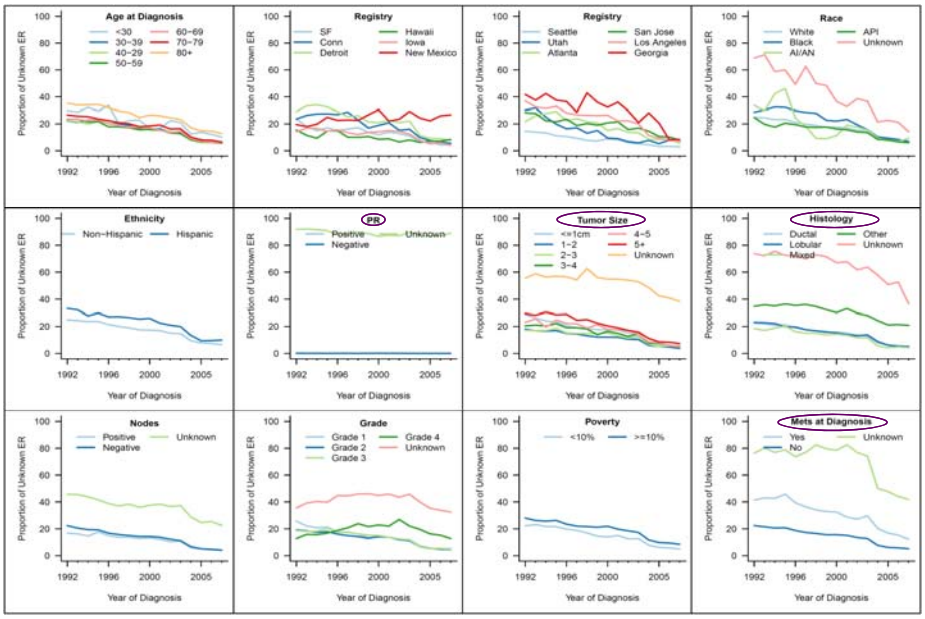
How Does Missing ER Status Vary over Time by Variables of Interest?

- Age at diagnosis
- Race
- Ethnicity
- Stage
- Registry
- Tumor size
- Socioeconomic status

Distribution of missing ER status over time by variables



Distribution of missing ER status over time by variables



Multiple Imputation of ER Status

- Imputed missing ER status under MAR assumption
- Basic idea behind imputation:
  - Fit regression model among observed cases, use to predict response for individuals with missing cases; add a random error term to account for uncertainty
- Specially, imputation of missing ER status, we used sequential regression multiple imputation (SRMI)
  - Impute each variable one at a time
  - Tailor the imputation to that specific variable (e.g., binary, continuous)

Multiple Imputation of ER Status (Cont'd)

- Variables: age (continuous), race (categorical with 3 levels), ER status (binary)

Id	Age	Race	ER
1	65	W	.
2	40	.	0
3	77	W	1
4	80	B	.
5	.	W	.

Steps in SRMI:

1. Do a single imputation to fill in missing values for all 3 variables
2. Using cases with observed age, fit normal regression model for age ~ race + ER; predict missing values of age

Multiple Imputation of ER Status (Cont'd)

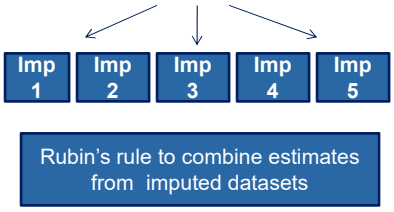
3. Using cases with observed race, fit multinomial logistic regression model for race ~ age + ER; predict missing values of race
4. Using cases with observed ER, fit logistic regression model for ER ~ age + race; predict missing values of ER
5. Iterate steps 2 through 4
6. Repeat step 5 to get multiple imputations

Id	Age	Race	ER
1	65	W	.
2	40	.	0
3	77	W	1
4	80	B	.
5	.	W	.

Multiple Imputation of ER Status (Cont'd)

- Imputation was repeated 5 times to account for imputation uncertainty
- Each imputed dataset was analyzed separately to obtain an estimate
- Rubin's rule is used for getting a final estimate combining across each dataset

Id	Age	Race	ER
1	65	W	1
2	40	B	0
3	77	W	1
4	80	B	0
5	79	W	0



Rubin's Rule

- Overall Estimate:

$$\overline{Q}_j = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$$

Number of imputed datasets (m = 5)

- Overall Variance: within and between-imputation

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

Within imputation variance

Between imputation variance

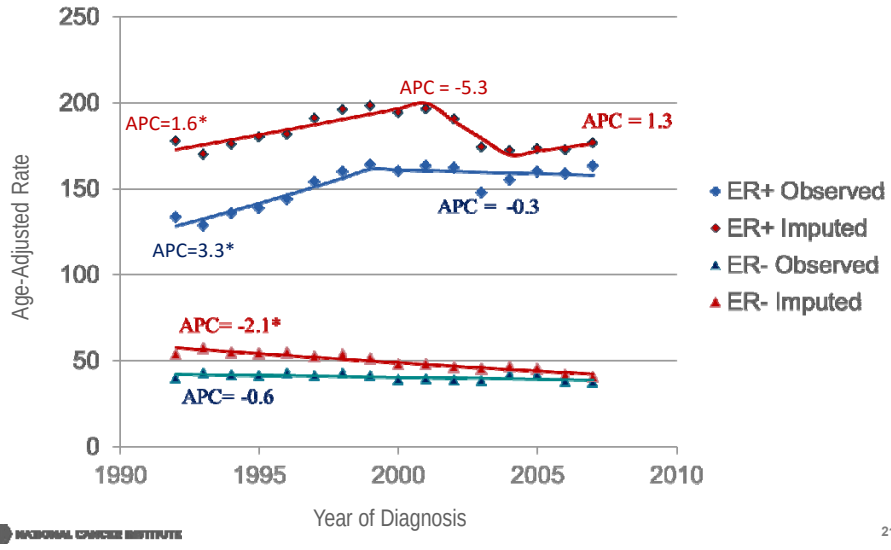
Multiple Imputation* of ER Status (Cont'd)

Demographic Variables	Clinical Variables
Age at diagnosis	Node positive status
Year of diagnosis	Metastasis at diagnosis
Registry	PR Status
Race	Histology
Ethnicity	Tumor Grade
County level poverty	Tumor Size

* Iveware (v 0.2) in SAS used for multiple imputation

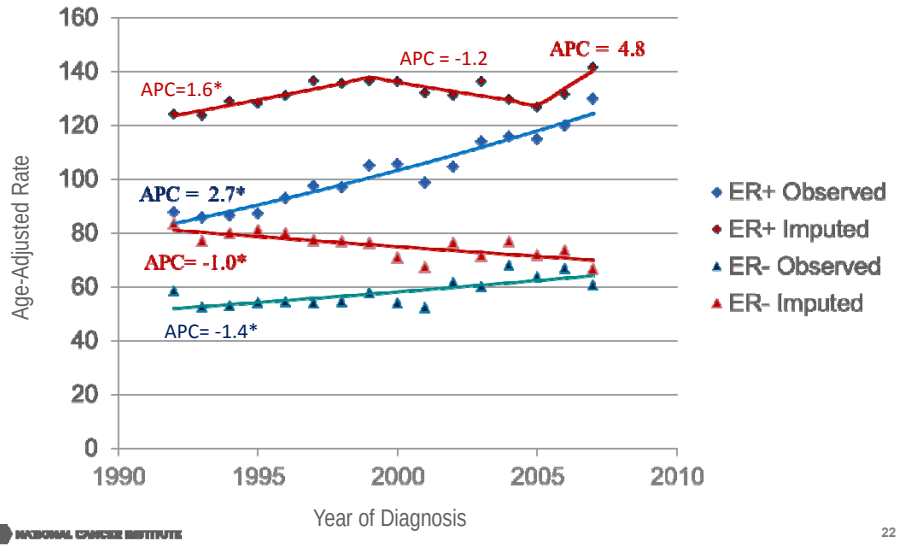
How Do Breast Cancer Incidence Trends Compare Before and After the Imputation?

Breast Cancer Incidence Trends Among White Women by ER Status, SEER-13



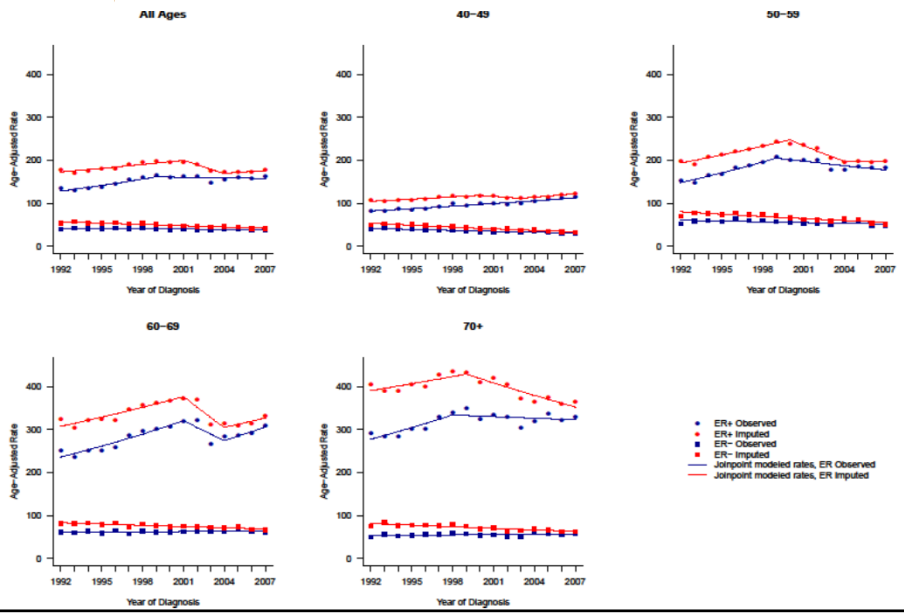
21

Breast Cancer Incidence Trends Among Black Women by ER Status, SEER-13

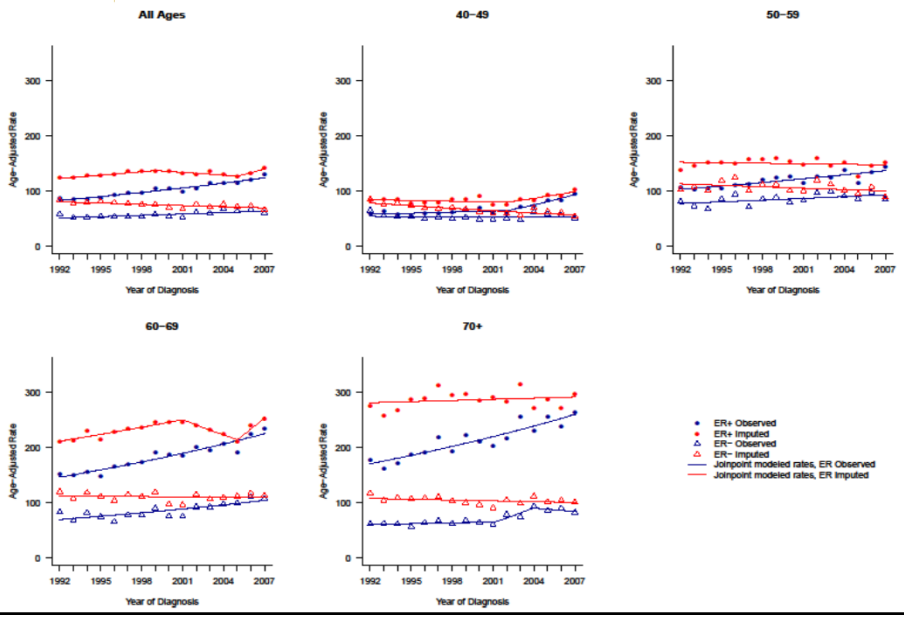


22

Breast Cancer Incidence Trends Among White Women, SEER-13



Breast Cancer Incidence Trends Among Black Women, SEER-13



Discussions

- ER status in SEER becoming more complete over time (25% in 1992 to 7% in 2007)
- Imputation method appears to be a reasonable approach to correct for missing ER status and to present trends more accurately
- Important to address missing ER status as we saw trends differ based on original vs imputed ER status

Discussions (Cont'd)

- Key assumption behind imputation is ER status is missing at random (MAR)
- What if ER status missingness were not at random (MNAR)? (Rebecca's talk)

How to Access the Imputed Dataset:

- Imputed dataset available through SEER*Stat for SEER-13 registries for 1992-2012 year of diagnosis

Contact:

Nadia Howlader

Email: howladern@mail.nih.gov

American Journal of EPIDEMIOLOGY

Use of Imputed Population-based Cancer Registry Data as a Method of Accounting for Missing Information: Application to Estrogen Receptor Status for Breast Cancer

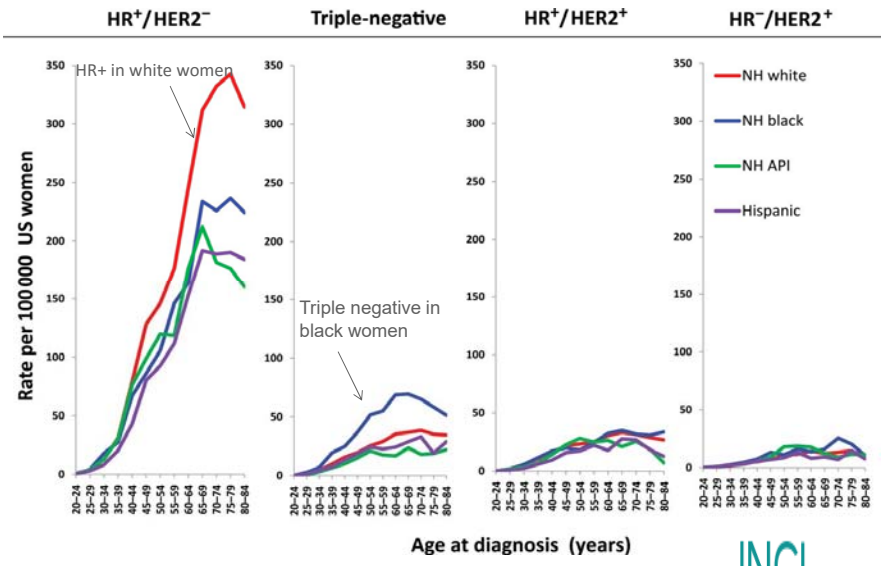
Nadia Howlader*, Anne-Michelle Noone, Mandi Yu and Kathleen A. Cronin

Imputed HER2 Status

Data Collection for HER2 Status

- Beginning with 2010 breast cancer cases
  - All registries from the SEER program for the first time collected HER2 receptor status
- ER and PR status were collected by SEER registries since the beginning of 1990
- The major molecular subtypes of breast cancer are approximated by the joint expression of these 3 tumor markers
  - With the availability of HER2/ER/PR, demographic & clinical assessment of major breast cancer subtypes for ~28% of US female population

Breast Cancer Incidence by Molecular Subtypes 2010, SEER



Nadia Howlader et al. JNCI J Natl Cancer Inst 2014;jnci.dju055

JNCI

Breast Cancer Molecular Subtypes

- Beginning with 2011 breast cancer cases, most US cancer registries started collecting HER2 status routinely
- Therefore, we expand the analysis to include data from 42 states plus the District of Columbia
  - Covering ~84% of the US female population

Breast Cancer Molecular Subtypes (Cont'd)

- To report breast cancer subtype by age group, race/ethnicity, area-based poverty status, and state
- However, one major challenge in reporting subtypes was that HER2 status was missing
  - ~10% of all breast cancer cases

JNCI Journal of the National Cancer Institute

Annual Report to the Nation on the Status of Cancer, 1975-2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State

Betsy A. Kohler, Recinda L. Sherman, Nadia Howlader, Ahmedin Jemal, A. Blythe Ryerson, Kevin A. Henry, Francis P. Boscoe, Kathleen A. Cronin, Andrew Lake, Anne-Michelle Noone, S. Jane Henley, Christie R. Eheman, Robert N. Anderson and Lynne Penberthy

How to Access Imputed HER2 Status:

- Available in SEER*Stat on request:
- CINA file:

  Recinda Sherman: rsherman@naaccr.org
- SEER file:

  Nadia Howlader: howladern@mail.nih.gov

Ongoing Work

- Imputed HER2 status is available for one year only (2011 breast cancer cases)
- Updating HER2 status for more recent years (2010-2013)
- Performing sensitivity analyses under MNAR assumption (using methods developed by Rebecca)


NIH
NATIONAL CANCER INSTITUTE
www.cancer.gov
www.cancer.gov/espanol