Joint Propensity Scores for the Analysis of Real-World Data with Biomarker Driven Treatment Selection

Debbie Jakubowski and Michael Crager

Genomic Health, Inc.

Bay Area Biotech-Pharma Statistics Workshop

7-8 November 2019

Setting for causal analysis for a treatment

- A cohort of patients with
  - A record of treatment received
  - Various other measures and assessments
  - An outcome of interest

Estimands in causal analysis

- Average treatment effect (ATE)
  - Reflects whole population
  - Expected treatment effect if population had been randomized to treatment

- Average treatment effect in the treated (ATT)
  - Reflects the population of patients selected for treatment
  - Expected treatment effect if the population selected for treatment had been randomized to treatment

Propensity score for treatment use

$$PS_T(\mathbf{z}_T) = \Pr(I_T = 1 | \mathbf{z}_T)$$

I_T = Indicator for treatment actually received

$\mathbf{z}_T$ = Vector of confounding covariates

- Is a balancing score
  - Weighted mean of each covariate is approximately equal across treatments

Setting for causal analysis for a treatment and a biomarker

- A cohort of patients with
  - A record of treatment received
  - An indication of whether a particular biomarker test was used and its result
    - Biomarker use influences treatment use
  - Various other measures and assessments
  - An outcome of interest

Propensity scores for biomarker test use

$$PS_B(\mathbf{z}_B) = \Pr(I_B = 1 | \mathbf{z}_B)$$

I_B = Indicator for biomarker test use

$\mathbf{z}_B$ = Vector of confounding covariates

- Is a balancing score
  - Weighted mean of each covariate is approximately equal for patients who used or did not use biomarker test

Conditional propensity score for treatment use given biomarker test use

$$PS_{T|B=1}\left(\mathbf{z}_T, \mathbf{z}_B\right) = \Pr\left(I_T = 1 \,|\, \mathbf{z}_T, I_B = 1\right)$$

- Is a balancing score
  - Weighted mean of each covariate is approximately equal across treatments given biomarker tes use

Joint propensity score for treatment and biomarker test use

$$PS_{T,B}\left(x,y;\mathbf{z}_T,\mathbf{z}_B\right)=\Pr\left(I_T=x,I_B=y\,|\,\mathbf{z}_T,\mathbf{z}_B\right) \qquad x,y\in\{0,1\}$$

$$=\Pr\left(I_T=x\,|\,\mathbf{z}_T,I_B=y\right)\Pr\left(I_B=y\,|\,\mathbf{z}_B\right)$$

$$PS_{T,B}\left(1,1;\mathbf{z}_T,\mathbf{z}_B\right)=PS_{T|B=1}\left(\mathbf{z}_T,\mathbf{z}_B\right)PS_B\left(\mathbf{z}_B\right)$$

- Is a balancing score
  - Weighted mean of each covariate is approximately equal across combinations of treatment and biomarker test use

Methods for using propensity scores in analysis

- Propensity score matching
  - Pair each treated patient with an untreated patient having a similar propensity score
  - May not be able to use all patients

- Estimates ATT not ATE

- Stratification by propensity score
  - Divide range of the propensity score into bins

- Use propensity score as a covariate

- Approximately unbiased estimates for
  - Linear models
- Biased estimates for
  - Cox regression of survival data
  - Logistic regression of categorical data

- Inverse probability of treatment weighting (IPTW)

- Approximately unbiased estimates for
  - Linear models
  - Cox regression of survival data
  - Logistic regression of categorical data

Ref: Austin (2014)

Inverse joint propensity weighting with focus on tested population

Weight for treated patients	Weight for untreated patients	Estimand
$1/\widehat{\Pr}(I_{Ti}=1, I_{Bi}=1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})$	$1/\{1-\widehat{\Pr}(I_{Ti}=1, I_{Bi}=1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})\}$	ATE as a function of test result in the tested
$1/\widehat{\Pr}(I_{Ti}=1, I_{Bi}=1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})$	$1/\widehat{\Pr}(I_{Ti}=0, I_{Bi}=1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})$	ATE as a function of test result in the whole population

$\mathbf{z}_{Ti}$ = covariate vector for treatment for patient i

$\mathbf{z}_{Bi}$ = covariate vector for biomarker for patient i

Stabilized inverse joint propensity weighting with focus on tested population

Weight for treated patients	Weight for untreated patients	Estimand
$\frac{\widehat{\Pr}(I_T = 1, I_B = 1)}{\widehat{\Pr}(I_T = 1, I_B = 1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})}$	$\frac{1 - \widehat{\Pr}(I_T = 1, I_B = 1)}{1 - \widehat{\Pr}(I_T = 1, I_B = 1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})}$	ATE as a function of test result in the tested
$\frac{\widehat{\Pr}(I_T = 1, I_B = 1)}{\widehat{\Pr}(I_T = 1, I_B = 1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})}$	$\frac{\widehat{\Pr}(I_T = 0, I_B = 1)}{\widehat{\Pr}(I_T = 0, I_B = 1 \mid \mathbf{z}_{Ti}, \mathbf{z}_{Bi})}$	ATE as a function of test result in the whole population

$\mathbf{z}_{Ti}$ = covariate vector for treatment for patient i

$\mathbf{z}_{Bi}$ = covariate vector for biomarker for patient i

Principles

- Develop and lock propensity score model without looking at outcome data
- Consider prior knowledge when selecting covariates for treatment or biomarker use
  - Do not rely solely on significance testing
- Check to see if propensity score model balances covariates
  - Absolute standardized difference between treatment groups (calculated using weights) < 10%

$$d = \frac{\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

Ref: Austin (2009)

Cohort from the Surveillance, Epidemiology and End Results (SEER) data base (National Cancer Institute)

- Oncotype DX Breast Cancer Recurrence Score® results provided to SEER registries using SEER methods (Petkov 2016)

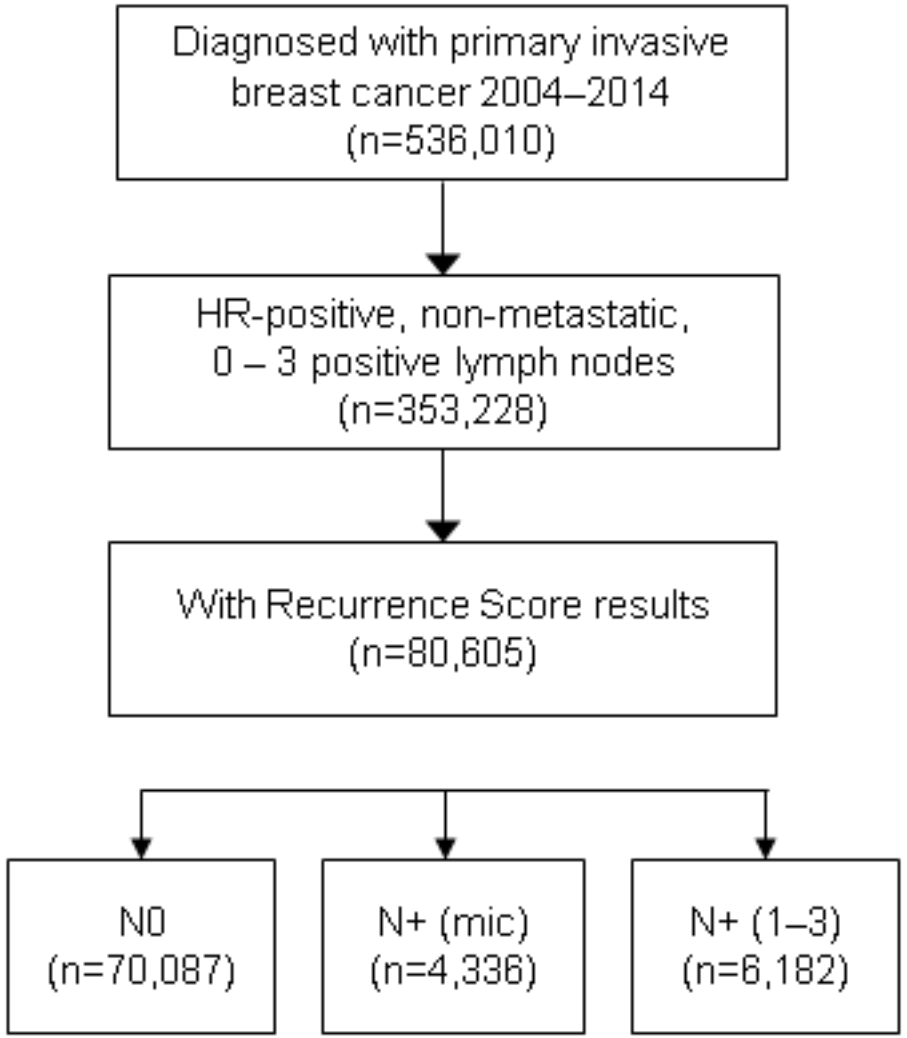
- Eligibility requirements:
  - Breast cancer diagnosis Jan 2004 - Dec 2014
  - Node-negative (N0), micromets (N1mic) or 1-3 positive nodes (N1-3), HR+, HER2-negative
  - No prior malignancy or multiple tumors

- Endpoint: breast cancer mortality
  - Follow-up through Dec 2015

- Chemotherapy (CT) use reported as yes vs. no/unknown

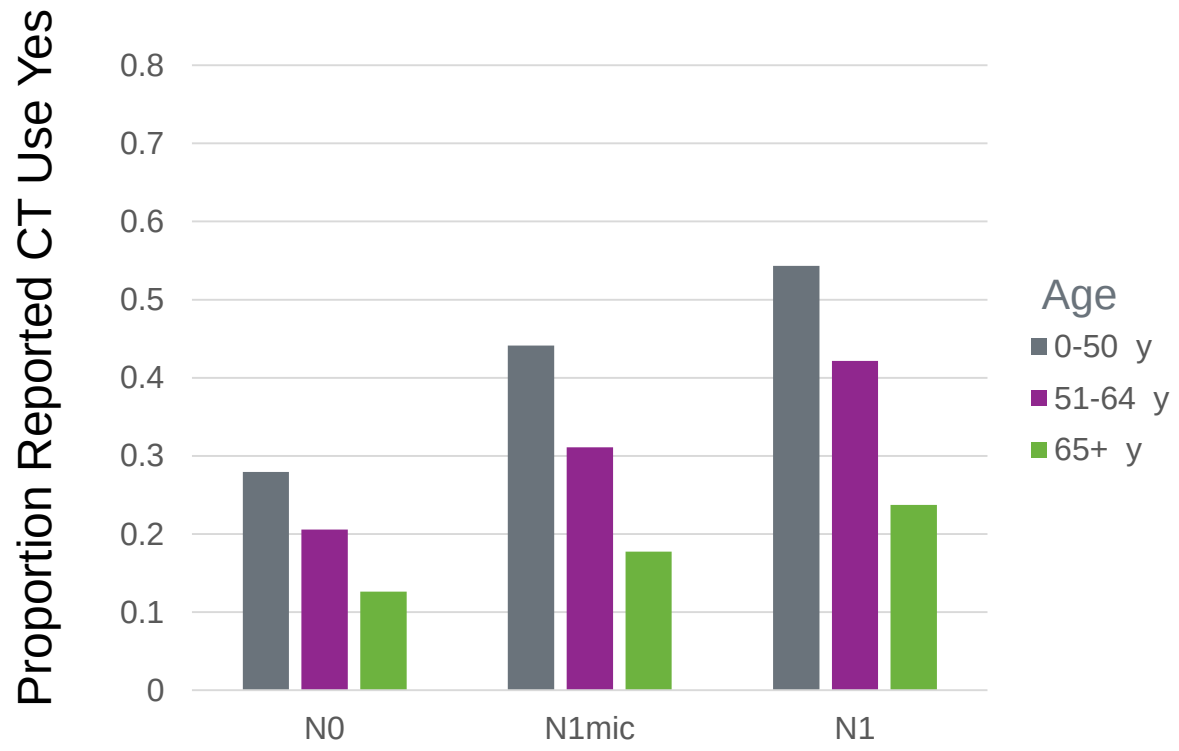
Ref: Hortobagyi et al (2018)

SEER study population



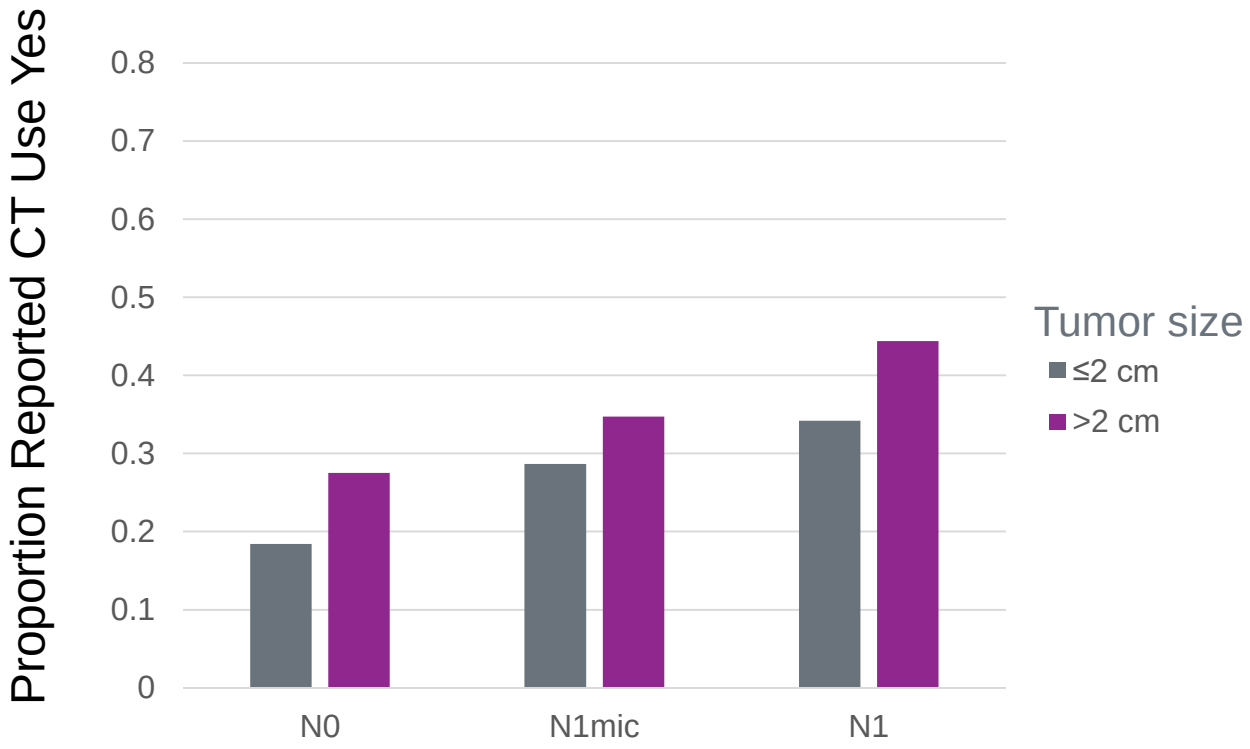
Ref: Hortobagyi et al (2018)

CT use in Breast Recurrence Score®-tested patients by nodal status and age (N=80,605)



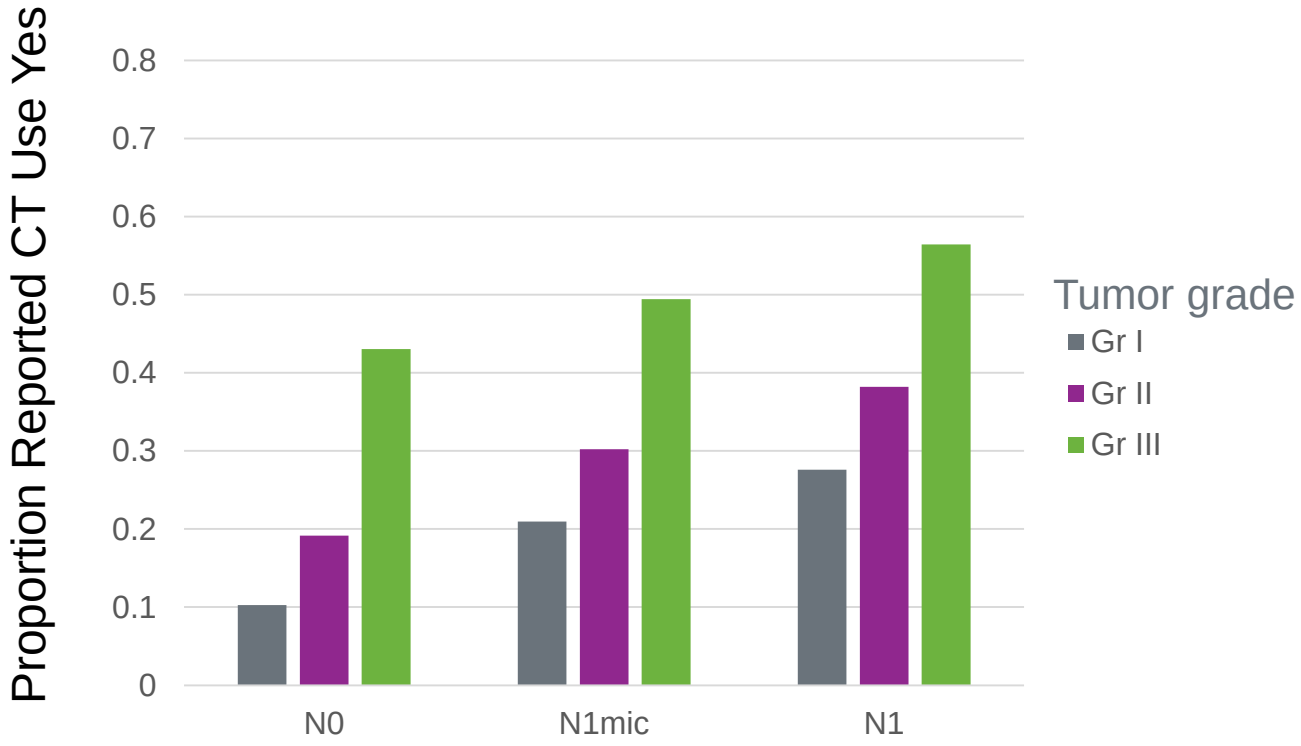
Ref: Hortobagyi et al (2018)

CT use in Breast Recurrence Score-tested patients by nodal status and tumor size (N=80,605)



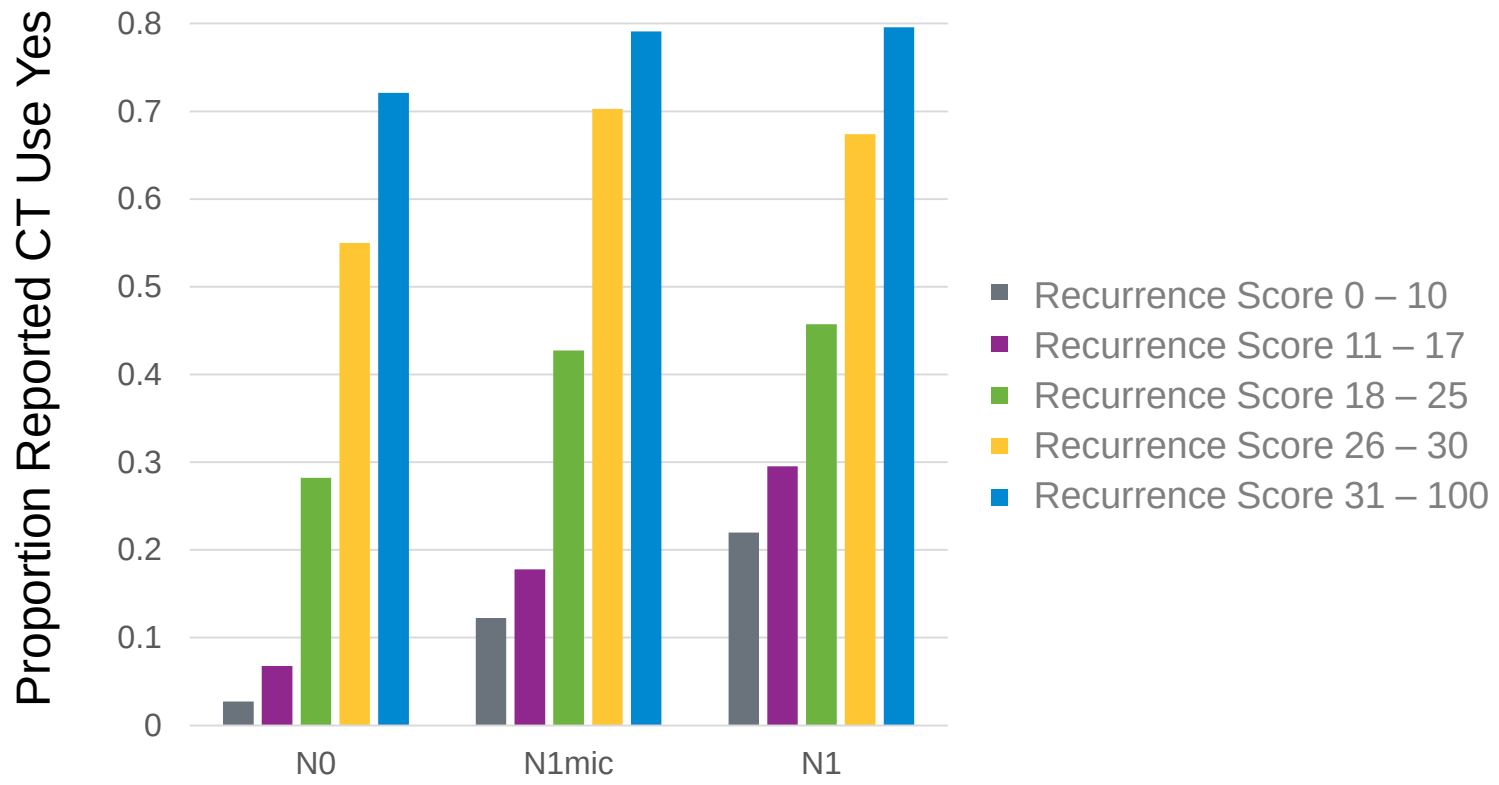
Ref: Hortobagyi et al (2018)

CT use in Breast Recurrence Score-tested patients by nodal status and tumor grade (N=80,605)



Ref: Hortobagyi et al (2018)

CT use in Breast Recurrence Score-tested patients by nodal status and Recurrence Score group (N=80,605)



Ref: Hortobagyi et al (2018)

Propensity model covariates for CT use and for Recurrence Score use

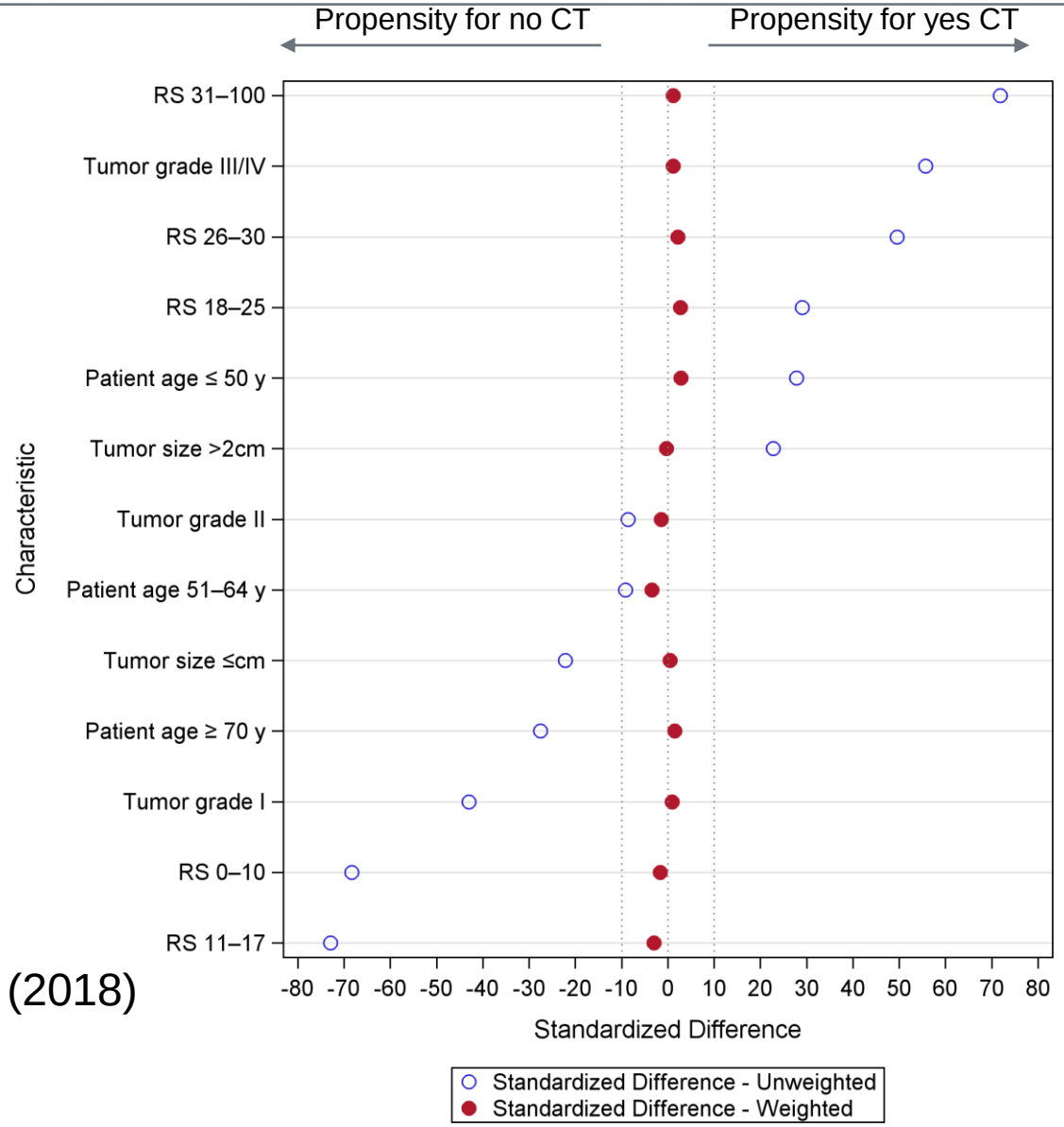
- Tumor size
- Tumor grade
- Race and ethnicity
- Type of surgery
- Histologic subtype
- State of residence
- Socioeconomic status
- Patient age
- Year of diagnosis

Interactions of other variable with these

- RS (propensity for CT only)

Model separately by nodal status => relationship of covariates with Recurrence Score test use and CT use may differ among nodal status groups

Propensity model adjustment for imbalances in baseline covariates N=70,087 patients with N0 disease



Ref: Hortobagyi et al (2018)

Analysis Methods

- Cox proportional hazards regression
  - Inverse joint propensity score weighting
  - Variance estimation using robust method of Lin and Wei (1989)

- Weighted Kaplan-Meier curves

Truncating weights

- Goal: avoid variance inflation due to a few patients with extreme weights

- Truncation of stabilized weights
  - Set weights < 5^{th} percentile to 5^{th} percentile
  - Set weights > 95^{th} percentile to 95^{th} percentile

Ref: Lee, Lessler and Stuart (2011), Austin and Stuart (2015)

Significance tests for interaction with chemotherapy treatment N=70,087 patients with N0 disease

Cox proportional hazards regression with inverse joint propensity weighting

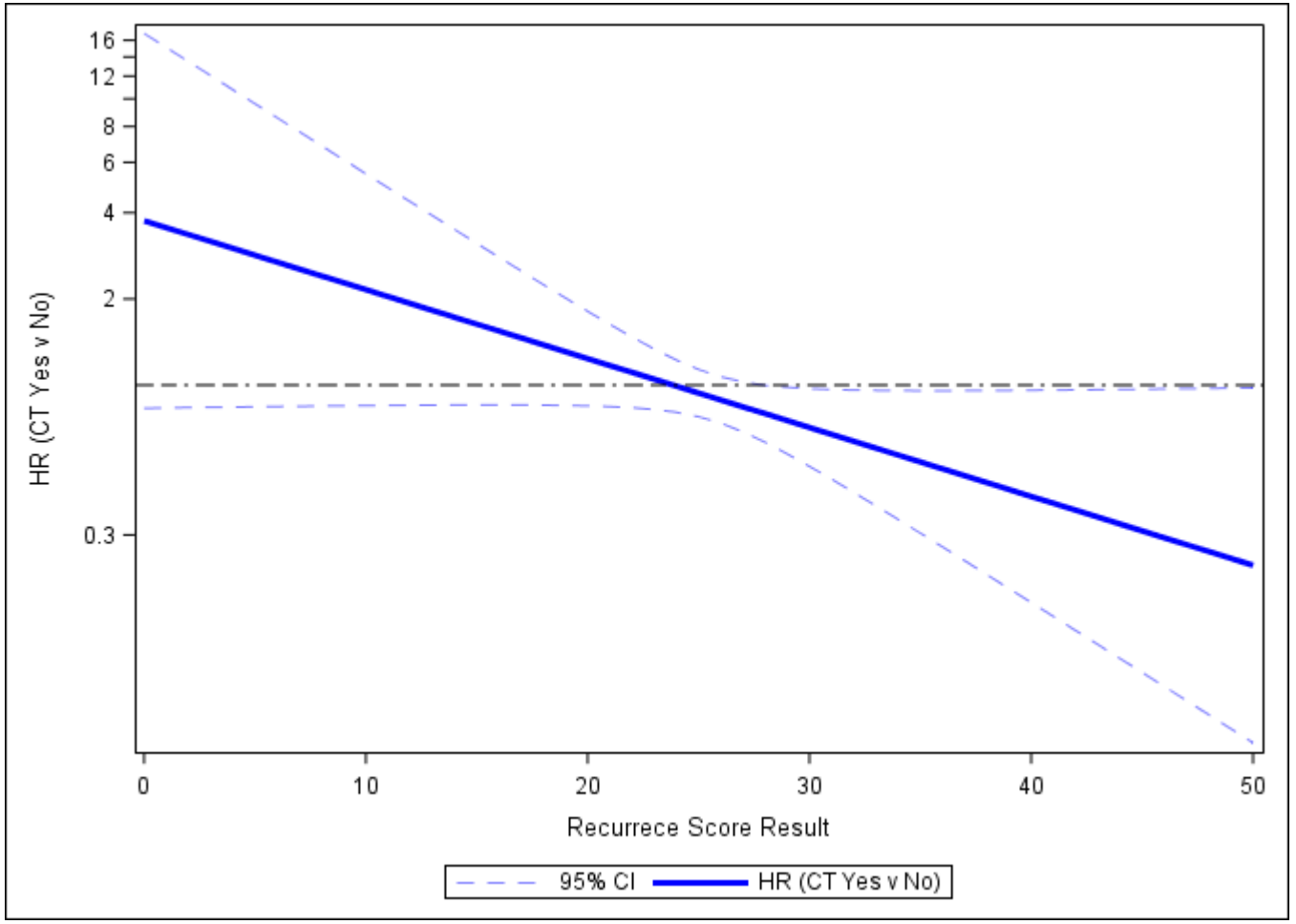
Interaction	HR (95% CI)	p-value
Age (≤ 50 y vs > 50 years) with CT	0.581 (0.303,1.116)	.103
Tumor Size (≤ 2 cm vs > 2 cm) with CT	1.579 (0.821,3.040)	.171
Tumor Grade (II vs I, III vs I) with CT	0.299 (0.085,1.053)	.083
	0.252 (0.074,0.858)	
RS (RS 26-100 vs. RS 0-25) with CT	0.432 (0.229,0.812)	.009

Each interaction added separately to model adjusting for Recurrence Score group (RS 26-100 vs. RS 0-25), tumor size (≤2 cm vs >2 cm), age (≤50 y vs >50 y), tumor grade (II vs I, III vs I), and chemotherapy use (yes vs. no/unknown).

Ref: Hortobagyi et al (2018)

Hazard ratio for chemotherapy benefit as a function of Recurrence Score result

Cox proportional hazards regression with inverse joint propensity weighting (N=70,087)



1.0 = No CT Benefit

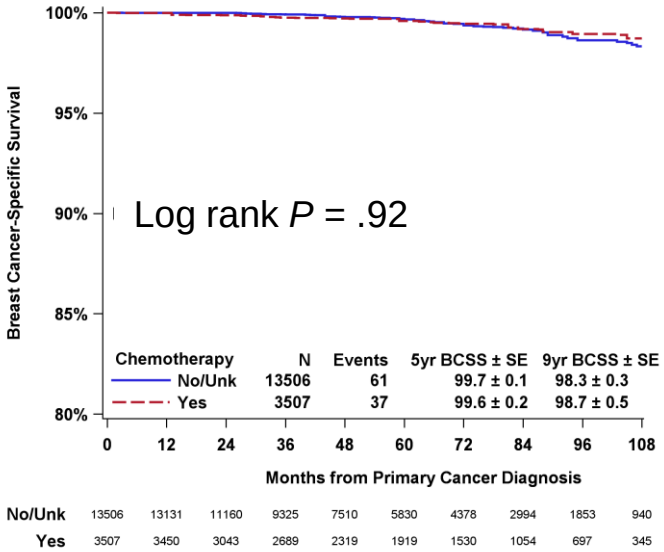
Increasing CT Benefit

Ref: Hortobagyi et al (2018)

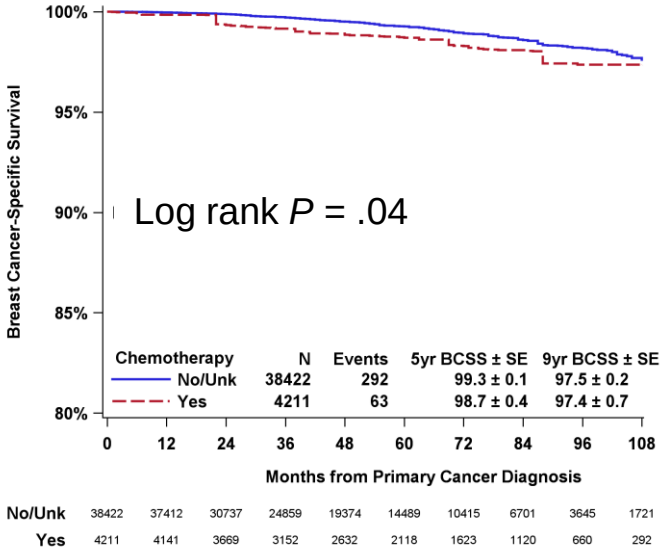
Breast cancer-specific survival in N0 disease, Kaplan-Meier estimates with inverse joint propensity weighting (N=70,087)

Recurrence Score 0 – 25

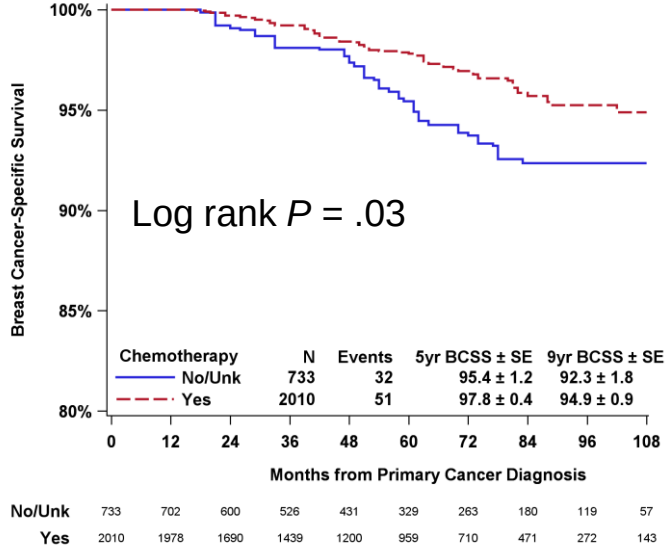
Age ≤50y

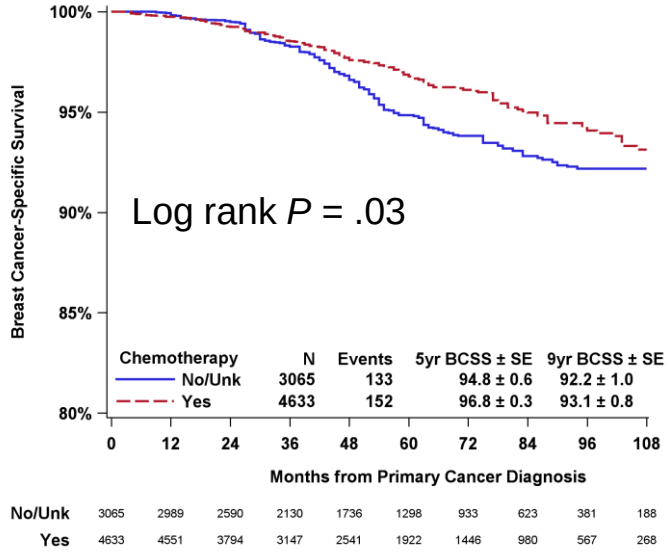


Age >50y



Recurrence Score 26 – 100

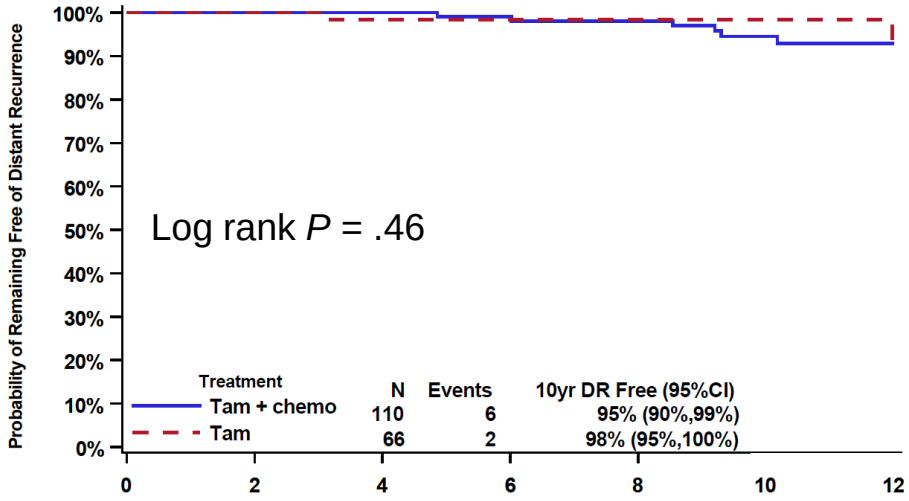




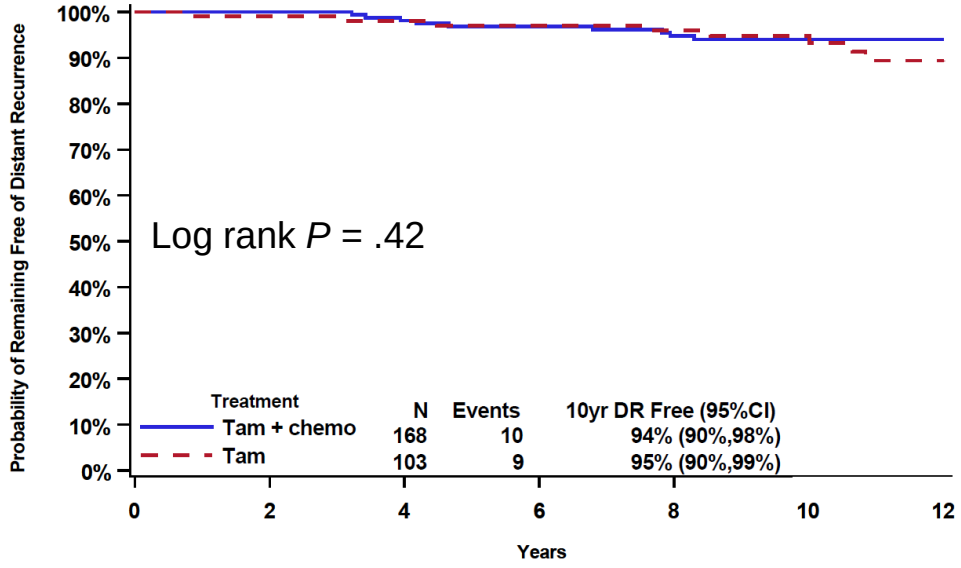
Ref: Hortobagyi et al (2018)

Results are consistent with randomized clinical trial results
Example: National Surgical Adjuvant Breast and Bowel Project (NSABP) Trial B-20

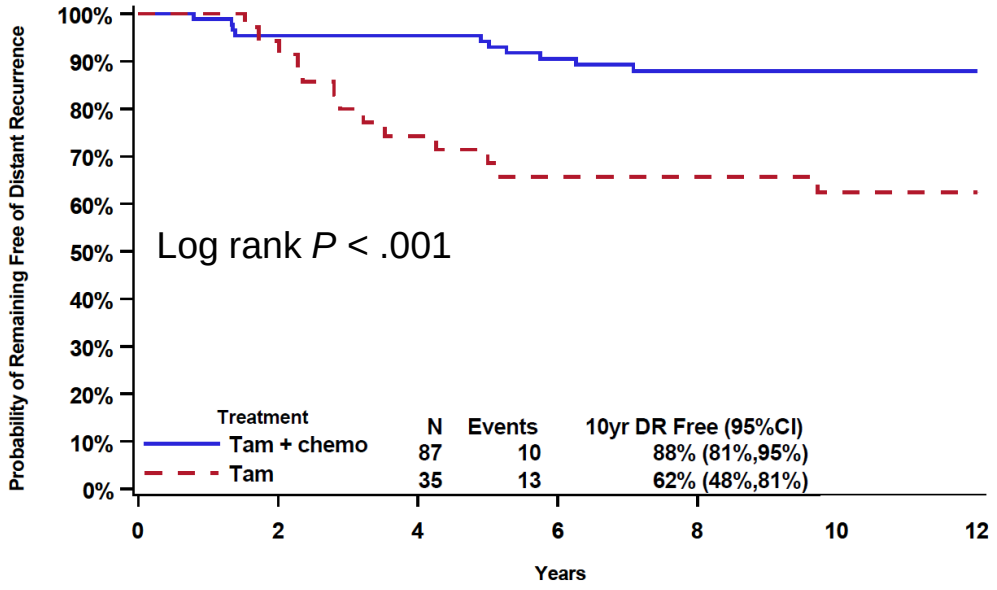
Recurrence Score < 11



Recurrence Score 11 – 25



Recurrence Score > 25



Test for interaction between Recurrence Score group and treatment: p=.014

Ref: Geyer et al. 2018

Discussion

- Causal analysis using propensity scores
  - Reduces bias due to non-random use of treatment and biomarkers
  - Increases variability of treatment main effect and interaction estimates

- All causal analyses assume no unmeasured confounders
  - Results should be interpreted with caution

Discussion

From Karim and Booth (2019):

“[Real-world data (RWD) comparative effectiveness] studies are best suited for settings in which there is existing evidence that a given treatment is efficacious . . . In settings where RCTs do not exist or may not be feasible, RWD can be informative; however, these studies should be interpreted with caution.”

Discussion

- Real-world evidence can help supplement evidence from randomized controlled trials
  - Evaluate treatment effects or interactions in actual use populations
  - Address effectiveness in patient populations under-represented in clinical trials
    - Example: young and old patients

- Analysis of real-word data related to biomarker-directed treatment should account for joint propensity for biomarker and treatment use

Acknowledgements

John Bennett (Senior Statistician, Clinical Biostatistics, Genomic Health, Inc.)
Gabriel Hortobagyi, M.D. (University of Texas M.D. Anderson Cancer Center)
Steven Shak, M.D. (Co-Founder, Chief Scientific Officer, Genomic Health, Inc.)
Jim Whitmore, Ph.D. (Vice President, Biostatistics and Data Management, Genomic Health, Inc.)

References

Austin PC (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* **28**: 3083–3107.

Austin PC (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine* **33:**1242–1258.

Austin PC, Stuart EA. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* **34**:3661–3679.

Geyer CE, Tang G, Mamounas EP, Rastogi P, Paik S, Shak S, Baehner FL, Crager M, Wickerham DL, Costantino JP, Wolmark N (2018). 21-Gene assay as predictor of chemotherapy benefit in HER2-negative breast cancer. *npj Breast Cancer* **4**:37. DOI 10.1038/s41523-018-0090-6

Hortobagyi GN, Shak S, Sledge, Jr. GW, Winer EP, Albain KS, Mamounas EP, Jakubowski DM, Petkov VI, Wolmark N (2018). Breast cancer-specific mortality (BCSM) in patients (pts) with node-negative (N0) and node-positive (N+) breast cancer (BC) guided by the 21-gene assay: A SEER-genomic population-based study. Poster presentation at the 2018 San Antonio Breast Cancer Symposium.

Karim S, Booth CM (2019). Effectiveness in the absence of efficacy: cautionary tales from real-world evidence. *Journal of Clinical Oncology* **37**:1047–1050. DOI 10.1200/JCO.18.02105

Lee BK, Lessler J, Stuart EA (2011). Weight trimming and propensity score weighting. *PLoS One* **6**:e18174. DOI 10.1371/journal.pone.0018174.

Petkov VI, Miller DP, Howlader N, Gliner N, Howe W, Schussler N, Cronin K, Baehner FL, Cress R, Deapen D, Glaser SL, Hernandez BY, Lynch CF, Mueller L, Schwartz AG, Schwartz SM, Stroup A, Sweeney C, Tucker TC, Ward KC, Wiggins C, Wu X-C, Penberthy L, Shak S (2016). Breast-cancer-specific mortality in patients treated based on the 21-gene assay: a SEER population-based study. *npj Breast Cancer* **2**:16017. DOI 10.1038/npjbcancer.2016.17