ORIGINAL CONTRIBUTION

YALE JOURNAL OF BIOLOGY AND MEDICINE
YJBM
CELEBRATING 95 YEARS 1928–2023

Comparative Effectiveness Analysis of Lumpectomy and Mastectomy for Elderly Female Breast Cancer Patients: A Deep Learning-based Big Data Analysis

Jiping Wang[a], Shunqin Zhang[a,b], Huangdi Yi[c], and Shuangge Ma[a,*]

[a]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA; [b]School of Mathematical Sciences, University of Chinese Academy of Sciences; Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China; [c]Servier Pharmaceuticals, Boston, MA, USA

Objectives: To evaluate the comparative effectiveness of treatments, a randomized clinical trial remains the gold standard but can be challenged by a high cost, a limited sample size, an inability to fully reflect the real world, and feasibility concerns. The objective is to showcase a big data approach that takes advantage of large electronic medical record (EMR) data to emulate clinical trials. To overcome the limitations of regression analysis, a deep learning-based analysis pipeline was developed. **Study Design and Setting**: Lumpectomy (breast-conserving surgery) and mastectomy are the two most commonly used surgical procedures for early-stage female breast cancer patients. An emulation trial was designed using the Surveillance, Epidemiology, and End Results (SEER)-Medicare data to evaluate their relative effectiveness in overall survival. The analysis pipeline consisted of a propensity score step, a weighted survival analysis step, and a bootstrap inference step. **Results**: A total of 65,997 subjects were enrolled in the emulated trial, with 50,704 and 15,293 in the lumpectomy and mastectomy arms, respectively. The two surgery procedures had comparable effects in terms of overall survival (survival year change = 0.08, 95% confidence interval (CI): -0.08, 0.25) for the elderly SEER-Medicare early-stage female breast cancer patients. **Conclusion**: This study demonstrated the power of "mining large EMR data + deep learning-based analysis," and the proposed analysis strategy and technique can be potentially broadly applicable. It provided convincing evidence of the comparative effectiveness of lumpectomy and mastectomy.

*To whom all correspondence should be addressed: Shuangge Ma, Department of Biostatistics, Yale School of Public Health, New Haven, CT; Email: Shuangge.ma@yale.edu.

Abbreviations: EMR, electronic medical record; SEER, Surveillance, Epidemiology, and End Results; RCT, randomized clinical trial; DNN, deep neural network; ICD-O-3, International Classification of Diseases for Oncology, 3rd Edition; ICD-9/10-CM, International Classification of Diseases, Ninth/Tenth Revision, Clinical Modification; HCPCS, Healthcare Common procedure Coding System; CPT, Current Procedural Terminology; IPT, inverse probability treatment; AFT, accelerated failure time; ReLU, Rectified Linear Units; UOQ, upper outer quadrant; UIQ, upper inner quadrant; LOQ, lower outer quadrant; LIQ, lower inner quadrant; SMD, standardized mean difference.

Keywords: Electronic medical record, Deep learning, Emulation, Breast cancer, Lumpectomy, Mastectomy

Author Contributions: Conceptualization: JW, YH, and SM; methodology: JW, SZ, and SM; software: SZ and JW; validation: JW and SM; formal analysis: JW; investigation: JW, YH, and SM; data curation: JW; writing—original draft preparation, all; writing—review and editing, all; supervision: SM. All authors have read and agreed to the published version of the manuscript.

INTRODUCTION

To evaluate the comparative effectiveness of treatments (drugs, operation procedures, etc.), a rigorously designed and executed randomized clinical trial (RCT) remains the criterion standard. However, challenges have been well recognized. With a high cost, RCTs, although statistically sufficiently powered by design, often have limited sample sizes. It has been repeatedly observed that treatment effects seen in the real world may differ (sometimes significantly) from RCTs, which can be attributed to the higher-than-standard care, higher adherence, and other factors in RCTs. Additionally, for treatments that have been on the market and widely used, there are feasibility concerns for conducting new head-to-head comparison trials. The fast accumulation of data from registries, large electronic medical records (EMR) and insurance claims databases, such as data contained in the SEER (Surveillance, Epidemiology, and End Results) registries, Medicare, Medicaid, and Kaiser, as well as the development of data management techniques, have made it possible to mine a large amount of real-world data to complement RCTs. Comparatively, such analysis can be advantageous with a much larger sample size and hence significantly more power, a better reflection of real-world evidence, high cost-effectiveness, and limited feasibility hurdles. Among the available observational data analysis methods (that aim to complement RCT analysis), emulation has attracted special attention, with its trial-like design, potential for causal interpretations, and scalability [1]. Emulation analysis has been conducted on cancers [2-4], cardiovascular diseases [5,6], and many other diseases [7-11]. The first objective of this study is to present another showcase of the emulation analysis of large EMR data.

In RCT analysis, regression techniques (such as linear, logistic, and Cox regressions as well as t and logrank tests) remain the common practice. This is also true for emulation analysis [1]. As in the real world there is no randomized assignment, emulation analysis usually demands an additional propensity score step, which can be accomplished by logistic regression [12]. Regression analysis is easy to conduct and has lucid interpretation. On the other hand, it is often challenged by “suboptimal” estimation and prediction performance and stringent model assumptions (and hence an inability to accommodate, for example, unknown nonlinear effects) – such limitations can be overcome by deep learning, which has experienced an unparallel surge in the past few years. Deep learning has demonstrated great power in diverse fields such as engineering, business, and social science [13-15]. It has been applied in biomedical research to the analysis of epidemiological [16], omics [17], imaging [18,19], and other types of data [20,21]. Our literature review suggests that the application of deep learning in the emulation analysis of observational data has been very limited. Additionally, “classic” deep learning techniques have been criticized as “black boxes.” As such, the second objective of this study is to develop a deep learning-based emulation analysis pipeline, which may have broad applications far beyond this study. The proposed deep neural network (DNN) approaches inherit strengths from well-established statistical principles and “classic” DNNs and can be superior to both.

For early-stage breast cancer, the surgical treatment options include mastectomy (which removes the entire breast) and lumpectomy (which removes a tumor and some normal tissues surrounding it). In the past decades, studies have hinted at the equivalency of the two surgical procedures, and both have been extensively conducted in clinical practice. However, RCTs may potentially suffer from a lack of generalizability due to extensive exclusion criteria, and previous evidence and findings from RCTs were based on cohorts from the 1980s [22-25]. Since then, there have been significant developments in breast cancer detection and treatment, such as a better understanding of tumor biology and advancements in surgical techniques and adjuvant therapy. With the period effect, patient characteristics have also changed significantly over time. It is thus of interest to compare breast cancer outcomes for patients who underwent the two surgical procedures in a more recent era. Another critical limitation of the existing RCTs and some observational studies is that there was insufficient attention to the elderly patients – many of them excluded elderly women or only included a very limited number of elderly women. The undertreatment and underrepresentation of elderly patients in studies may be attributable to limited cosmetic needs, a shorter life expectancy, an increasing incidence of comorbidities, and other factors. Breast cancer disproportionately affects older women, with those over 65 having the highest age-specific probability of developing invasive diseases [26,27]. The third objective of this study is to evaluate the comparative effectiveness of mastectomy and lumpectomy for the overall survival of elderly female breast cancer patients diagnosed and treated more recently. This effort can complement the existing RCT and observational studies and directly inform breast cancer clinical practice.

METHODS

Data and Study Sample

The SEER-Medicare data was analyzed in this study. Surveillance, Epidemiology, and End Results (SEER) contains 22 cancer registries in the United States. Medicare is a federally-funded health insurance program in the US, and in 2010, approximately 94% of

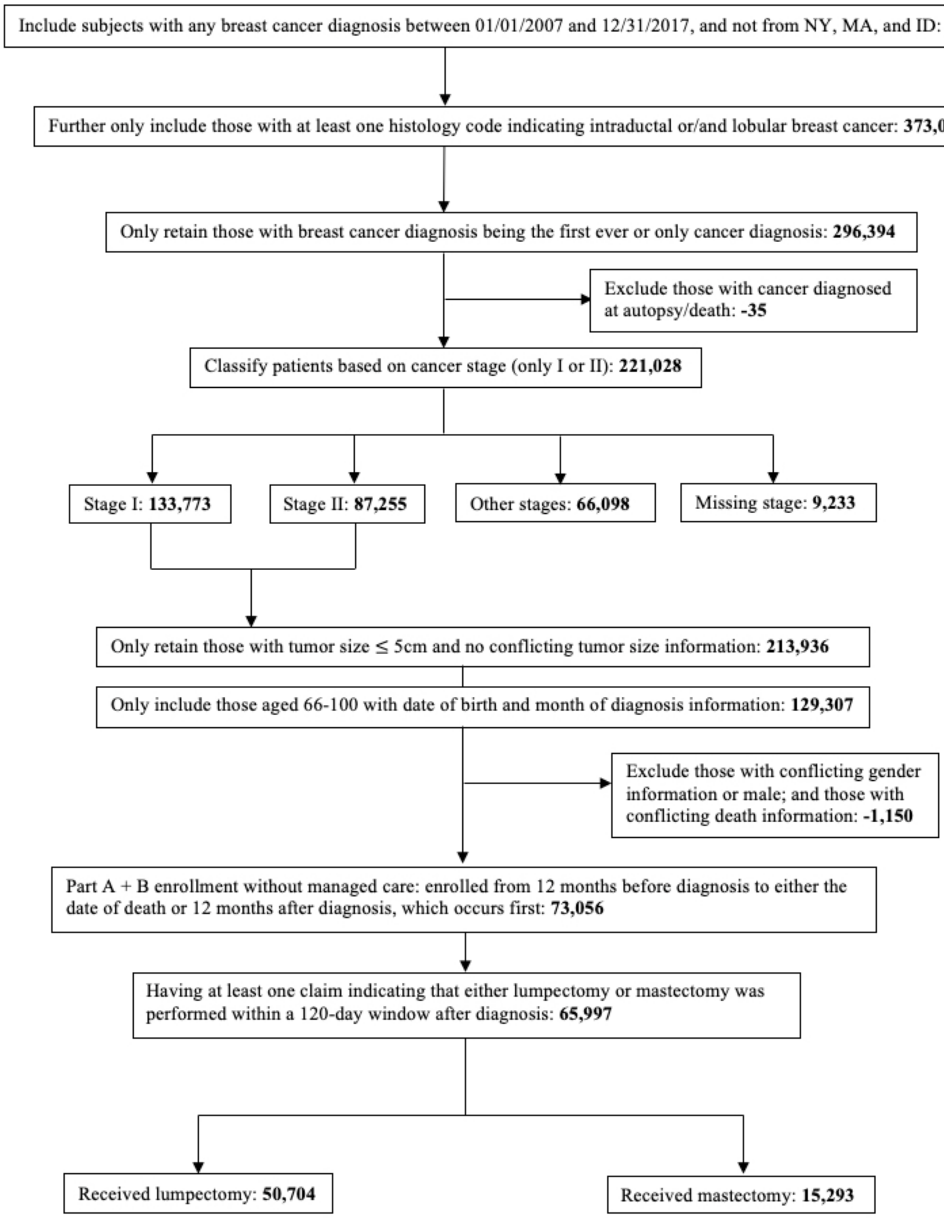
Include subjects with any breast cancer diagnosis between 01/01/2007 and 12/31/2017, and not from NY, MA, and ID: **472,238**
Further only include those with at least one histology code indicating intraductal or/and lobular breast cancer: **373,091**
Only retain those with breast cancer diagnosis being the first ever or only cancer diagnosis: **296,394**
Exclude those with cancer diagnosed at autopsy/death: **-35**
Classify patients based on cancer stage (only I or II): **221,028**
Stage I: **133,773**
Stage II: **87,255**
Other stages: **66,098**
Missing stage: **9,233**
Only retain those with tumor size ≤ 5cm and no conflicting tumor size information: **213,936**
Only include those aged 66-100 with date of birth and month of diagnosis information: **129,307**
Exclude those with conflicting gender information or male; and those with conflicting death information: **-1,150**
Part A + B enrollment without managed care: enrolled from 12 months before diagnosis to either the date of death or 12 months after diagnosis, which occurs first: **73,056**
Having at least one claim indicating that either lumpectomy or mastectomy was performed within a 120-day window after diagnosis: **65,997**
Received lumpectomy: **50,704**
Received mastectomy: **15,293**

Figure 1. **Flowchart of sample selection**.

the US population age 65 years or older was enrolled in Medicare. The SEER-Medicare data provides an ideal opportunity to study the US elderly cancer population. It has been extensively analyzed, including in emulation analysis [3,28]. Here, it is noted that the adopted mining observational data for emulation analysis strategy and deep learning-based analysis pipeline can be directly applied to other databases. Data analyzed in this study was obtained through a Data Use Agreement (DUA) by the National Cancer Institute (NCI). The study design was motivated by relevant RCTs [22-25] and observational analysis [29-32], while taking into consideration data availability in the SEER-Medicare database. The target and emulated trial designs are described in Appendix Table S1. The flowchart of emulated trial patient selection is summarized in Figure 1. Briefly, the emulated trial enrolled Medicare beneficiaries with first primary invasive intraductal or/and lobular breast cancer (International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3): 8500, 8520, and 8522) who were: 1) female and 66-100 years old at diagnosis, 2) diagnosed between 1/1/07 and 12/31/17, 3) stage I-II, 4) with tumor size ≤ 5cm, 5) received lumpectomy or mastectomy within 120 days after diagnosis, and 6) had continuous enrollment in Medicare Part A and Part B, with no health maintenance organization (HMO) enrollment from one year before to

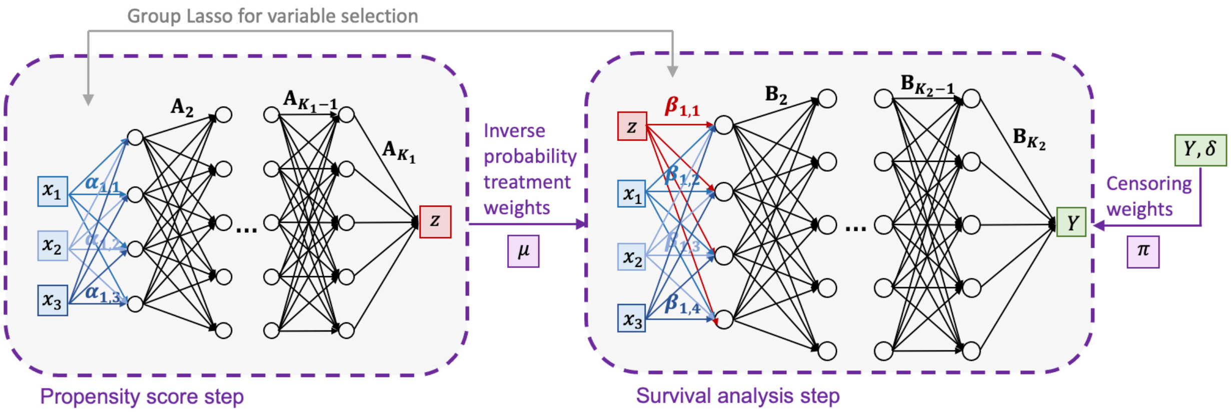

Figure 2. **Scheme of the proposed DNN-based analysis**.

one year after the cancer diagnosis or death, whichever occurred first. A subject was excluded if she 1) was reported from registries in New York (NY), Massachusetts (MA), and Idaho (ID) because of missing cancer-related information (eg, stage), 2) had missing information on month of diagnosis, gender, stage or tumor size, and 3) was reported from death certificate or autopsy.

Variable Information

After reviewing the relevant observational studies, and also considering data availability, we included the following variables as potential confounders, which may have an impact on treatment initiation (choice of surgery) and outcome (overall survival): 1) baseline demographics: age, race, and marital status; 2) cancer-related variables: tumor size, morphology, stage, primary site, laterality, grade, HR status; and 3) comorbidities: Elixhauser comorbid conditions index. It is noted that human epidermal growth factor receptor 2 (HER2) was not measured for patients diagnosed prior to 2010, and, as such, it was not included in the main analysis. In addition, as only a few (<11) patients had unknown laterality or paired site/midline tumor, they were excluded from the analysis. A patient was enrolled into the lumpectomy arm if she received lumpectomy as the first surgery procedure after being diagnosed and would be censored if a subsequent mastectomy was performed. A patient was enrolled into the mastectomy arm if she received mastectomy as the first surgery procedure after being diagnosed regardless of subsequent surgical procedures. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), ICD-10-CM, and the Healthcare Common Procedure Coding System (HCPCS)/Current Procedural Terminology (CPT) codes appeared in Medicare inpatient and outpatient records were adopted to define surgeries. Details are provided in Appendix A: Table S2. Time zero was set as the time of surgery. Immortal time bias, if any, is expected to be limited, as the two treatment groups were comparable and the mortality rate right after diagnosis was very low compared to the long follow-up. The overall survival information was obtained from the Medicare records. All subjects were followed to the end of 2019 or death, whichever happened first.

A DNN-based Analysis Pipeline

Under the regression framework, the emulation technique has been well developed. We refer to [1,33] for methodological developments and [2,34] for representative case studies. Additionally, some computer codes are also publicly available, for example at [35]. Briefly, to emulate the estimation of an intention-to-treat (ITT) effect, the observed treatments in medical records are used as treatment assignments. Special considerations may be needed if a patient received multiple treatments. In a well-executed RCT, different arms can be sufficiently balanced. As such, there may be no need for accounting for confounding – although it is still commonly done out of caution. In the emulation analysis of observational data, since randomization did not really happen, there is a risk of imbalance. To tackle this problem, the propensity score and inverse probability treatment (IPT) weighting techniques are usually adopted to achieve baseline balance as in an RCT. The propensity score calculation is usually built on logistic regression that includes the treatment indicator as response and baseline covariates as covariates. Usually, this analysis only includes linear effects. Based on the logistic regression results, the IPT weight can be calculated as the inverse of the propensity score for one treatment group and the inverse of one minus the propensity score for the other group. Stabilized weights with a truncation at the upper 99.5% percentile (that is, if a weight is above the 99.5% percentile, it is set as the 99.5% percentile) are usually adopted for retaining the same group size ratio and avoiding putting extreme weights on certain subjects. With such weights, a pseudo-population can be created, for which there is no (or very weak) association between the baseline confounders and received treatment. Then, survival analysis, for example, Cox regression, can be conducted using the weighted pseudo-population to estimate the causal effect under the assumption of no unmeasured confounders. Similar to logistic regression, Cox and other survival analysis models also rely on strong assumptions and linear effects. In practice, it is not uncommon to observe violations of such assumptions.

Our overall strategy is to follow the above framework but replace regression-based analysis with DNN-based analysis, which relies on weaker model assumptions and can "automatically" accommodate unspecified nonlinear relationships. The architecture of the DNNs for both propensity score and survival analysis steps is sketched in Figure 2 and consists of a sparse layer, multiple hidden layers, and an output layer. The key building components of a DNN include its loss function and estimation approach, which are presented below. Denote n as the number of independent subjects. For subject i, we observe p baseline covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$, a binary treatment indicator Z_i, and right censored survival outcome $\{Y_i = \min(T_i, C_i), \delta_i = \mathrm{I}(T_i \leq C_i)\}$. Here, T_i, C_i are the event and censoring times, respectively, and $\mathrm{I}(\cdot)$ is the indicator function.

Propensity score step. A feed-forward DNN is constructed with input being $\boldsymbol{X}_i$'s after standardization for continuous variables and one-hot encoding for categorical variables and output being Z_i,'s. EMR databases contain rich information, and collecting data from such databases is almost free. As such, to avoid losing important information, "more than necessary" covariates may be included. In addition, in regression-based emulation analysis, it is also commonly observed that some covariates are not significant. For the proposed DNN, we first propose imposing group Lasso penalization to the input layer weights. In particular, the weights corresponding to one input variable are treated as a group. Group Lasso has been extensively developed under the regression framework and has the "group in or group out" selection property [36]. With this penalty, we are able to distinguish important input variables from noises, which can lead to a simpler DNN structure, fewer variables and hence more stable estimation, and more lucid interpretations. Additionally, we also propose applying ridge penalization [37] to the hidden layer weights. Here, the goal is to regularize estimation and avoid extreme values. In principle, it is possible to also apply group Lasso (or its individual counterpart Lasso). However, this may lead to an overly sparse DNN and unstable optimization.

Use $\boldsymbol{\alpha}_{1,j}$, a vector, where $j = 1, \ldots, p$, to denote the input layer weights corresponding to the jth input variable, and $\mathbf{A}_{k1}$, a matrix, where $k_1 = 1, \ldots, K_1$, to denote the weights of the k_1th hidden layer. Additionally, use A to collectively denote all weights. For subject j denote $p_i = \frac{1}{1+\exp(-h_{\mathrm{A}}(\boldsymbol{X}_i))}$ as the probability output (that is, a Sigmoid activation function is adopted for the output layer). Intuitively, $h_{\mathrm{A}}(\boldsymbol{X}_i)$ corresponds to the covariate effect in logistic regression. For the input layer and hidden layers, respectively, we adopt the Tanh and Hardtanh activation functions. We adopt the logarithmic loss function (binary cross-entropy), which has connections with the logistic likelihood function, and has been popularly adopted. This has been partly motivated by the fact that, in the literature, logistic regression has been popularly adopted in propensity score estimation and has demonstrated sensible performance. The overall penalized loss function is:

$$l_1 = -\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(Z_i \times h_{\mathrm{A}}(\boldsymbol{X}_i) - \log\left(1+\exp\left(h_{\mathrm{A}}(\boldsymbol{X}_i)\right)\right)\right) + \lambda_1\sum\nolimits_{j=1}^{p}\|\boldsymbol{\alpha}_{1,j}\|_2 + \lambda_2\sum\nolimits_{k_1=1}^{K_1}\|\mathrm{A}_k\|_F^2$$

where λ_1 and λ_2 are tuning parameters, $\|\cdot\|_2$ denotes the l_2-norm, and $\|\cdot\|_F$ denotes the Frobenius norm. Denote the estimate of A as $\hat{\mathrm{A}}$ and $\hat{p}_i = \frac{1}{1+\exp(-h_{\hat{\mathrm{A}}}(\boldsymbol{X}_i))}$. Following regression-based analysis, we calculate the inverse probability treatment weight: if subject i is in the treatment arm, $\hat{\mu}_i = \frac{1}{\hat{p}_i}$; otherwise, $\hat{\mu}_i = \frac{1}{1-\hat{p}_i}$.

Survival analysis step. In regression analysis, the Cox model has been widely used because of clear interpretation and easy implementation, although alternative models have also been extensively adopted. In deep learning, a loss function that corresponds to the negative partial likelihood function of the Cox model has been developed [38]. However, it is not as popular as its counterpart under regression. Here, we develop an alternative loss function.

For subject $i = 1, \ldots, n$, consider weight $\pi_i = \frac{\delta_i}{S(Y_i)}$, where $S(t)$ is the survival function for the event time of interest. In practice, we estimate $S(t)$ using the Kaplan-Meier estimator $\hat{S}(t)$ and denote the corresponding weight as $\hat{\pi}_i$. We construct a feed-forward weighted DNN with $(Z_i, \boldsymbol{X}_i)$'s as input and Y_i's as output. For subject i, $\hat{\pi}_i$ is imposed as weight. Note that the weight calculation does not depend on the DNN construction. In addition, censored subjects have zero weights, and as such, are virtually removed from the DNN analysis. For the input layer, the Tanh activation function is adopted. For the hidden and output layers, the Rectified Linear Units (ReLU) activation function is adopted. For the jth input variable, use $\boldsymbol{\beta}_{1,j,}$ a vector, to denote its input layer weights. For the k_2th $(= 1, \ldots, K_2)$ hidden layer, use $\mathbf{B}_{k2}$, a matrix, to denote its collection of weights. And use B to collectively denote all network weights. For subject i denote $f_{\mathrm{B}}(Z_i, \boldsymbol{X}_i)$ the network output. Following the same penalization strategy as above, we propose the loss function:

$$l_2 = \frac{1}{\sum_i \mathrm{I}\{\delta_i \neq 0\} \times \hat{\mu}_i \times \hat{\pi}_i}\sum\nolimits_{\{i:\delta_i \neq 0\}}\hat{\mu}_i \times \hat{\pi}_i \times \left(Y_i - f_{\mathrm{B}}(Z_i, \boldsymbol{X}_i)\right)^2 + \lambda_3\sum\nolimits_{j=1}^{p+1}\|\boldsymbol{\beta}_{1,j}\|_2 + \lambda_4\sum\nolimits_{k_2=1}^{K_2}\|\mathrm{B}_{k_2}\|_F^2,$$

where λ_3 and λ_4 are tuning parameters. The first term has a least squares form, which makes computation simple.

The proposed loss has been motivated by the accelerated failure time (AFT) model [39]. Compared to the Cox model, the AFT model can sometimes be preferred because of its more lucid interpretation and simpler computation. Although it has been widely adopted in regression, its DNN development has not been well pursued. This study can fill this important knowledge gap. There can be multiple estimation approaches under the AFT modeling. We adopt the weighted least squares technique [40,41], which has the lowest computational cost and can be easily combined with the IPT weighting.

Inference on treatment effect. In RCT analysis, quantification of the significance of the treatment effect (using p-value, confidence interval, etc.) can be as important as estimation. In most of the existing DNN studies, there has been a lack of attention to the effect of an individual input variable and the quantification of its significance. To fill this knowledge gap, for the proposed analysis, we propose estimating treatment effects through a permutation of the treatment indicator Z_i for each subject. That is, the population treatment effect is defined as the average increase/decrease in predicted survival time if switching $Z_i = 1$ to $Z_i = 0$ (and vice versa) for the whole cohort. For inference, we propose a weighted bootstrap technique, which has been motivated by the weighted bootstrap for regression analysis [42] and consists of the following steps:

1) Randomly generate $\eta_1, \ldots, \eta_\eta$ from exp(1), the exponential distribution with rate = 1.

2) Conduct the weighted penalized estimation. In particular, the first term in the survival analysis objective function is revised as

$$\frac{1}{\sum_i \mathrm{I}\{\delta_i \neq 0\}\, \eta_i \times \hat{\mu}_i \times \hat{\pi}_i} \sum_{\{i:\delta_i \neq 0\}} \eta_i \times \hat{\mu}_i \times \hat{\pi}_i \times \left(Y_i - f_{\mathrm{B}}\left(Z_i, \boldsymbol{X}_i\right)\right)^2$$

The penalty terms remain unchanged.

3) Repeat Steps 1) - 2), eg, 500 times. The resulted estimates are used to construct a confidence interval (CI) for the treatment effect.

Computation. The bootstrap analysis can be conducted using the same approach as without the exponential weights. For both the propensity score and survival analysis steps, we adopt state-of-the-art techniques, including input standardization, Adaptive Moment Estimation for the gradient descent algorithm, Nesterov momentum, and learning rate scheduling with the exponential decay technique. We perform a Random hyperparameter optimization search and tune network depth and size, learning rate, penalization parameters, dropout rate, exponential learning rate decay constant, and momentum, with the assistance of the Python package Optunity. For tuning parameter selection, we conduct a grid search and minimize the cross entropy in the propensity score step and the weighted mean squared error in the survival analysis step. As the sample size is much larger than the number of parameters, cross-validation is not adopted but can be easily revised. The open-source Python module PyTorch is adopted.

Sensitivity analysis. In the main analysis, HER2 was not included because of complete missingness for patients diagnosed prior to 2010. It is recognized that HER2 is an important confounder and may impact both surgical choices and overall survival. In Appendix A: S1, analysis was conducted on a revised cohort with available HER2 information (that is, patients diagnosed in or after 2010), and HER2 was included as a confounder. Additionally, it was recognized that whether to perform radiotherapy was still being debated. Also motivated by the observed high proportion of patients receiving adjuvant radiotherapy after lumpectomy, we conducted the second sensitivity analysis in Appendix A: S2 and compared the overall survival of patients receiving mastectomy without adjuvant radiation therapy versus lumpectomy with radiation therapy. The ICD-9-CM, ICD-10-CM, HCPCS/CPT codes for identifying radiation therapy are provided in Appendix Table S2. We further fit a standard Cox regression model that included all the confounders and compared the results with those using the deep learning approach.

RESULTS

As shown in Figure 1, a total of 65,997 patients were enrolled in the emulated trial, with 50,704 in the lumpectomy arm and 15,293 in the mastectomy arm. The waiting time from diagnosis to surgery was observed to be short – over 95% of the surgeries were performed within 120 days after diagnosis. The baseline characteristics are summarized in Table 1. It was observed that before the IPT weighting, the patients treated with lumpectomy were slightly younger, more likely to be stage I and with smaller tumor size and had lower comorbidity scores.

The unadjusted Kaplan-Meier survival curves by treatment are shown in Figure 3. For the lumpectomy arm, the overall mortality rate and median survival were 22.6% and 6.0 years, respectively. For the mastectomy arm, they were 35.5% and 5.8 years, respectively. The patients who received lumpectomy were observed to have longer overall survival (p-value < 0.001 by logrank test). The proportional hazards assumption was tested, and the global Chi-squared test returned a p-value < 0.001, which justifies the adoption of a more flexible modeling technique.

In the propensity score analysis, we evaluated the impact of tuning parameters by visualizing parameter paths, which is a common technique for penalized estimation [43]. The parameter paths as a function of tuning parameter λ_1 are shown in Figure 4a. It was observed that tumor size was the only variable that was included when λ_1 was large, suggesting its higher importance. As the

Table 1. Patients' Characteristics by Treatment

	Mastectomy (n = 15,293)	Lumpectomy (n = 50,704)	Overall (n = 65,997)
Demographics			
Age at diagnosis	76.1 (6.85)	75.3 (6.55)	75.5 (6.63)
Race (Non-Hispanic White)	12474 (81.6%)	43038 (84.9%)	55512 (84.1%)
Marital Status (married)	6715 (43.9%)	25048 (49.4%)	31763 (48.1%)
Year of diagnosis			
2007	1571 (10.3%)	4586 (9.0%)	6157 (9.3%)
2008	1540 (10.1%)	4535 (8.9%)	6075 (9.2%)
2009	1493 (9.8%)	4504 (8.9%)	5997 (9.1%)
2010	1456 (9.5%)	4468 (8.8%)	5924 (9.0%)
2011	1472 (9.6%)	4575 (9.0%)	6047 (9.2%)
2012	1512 (9.9%)	4508 (8.9%)	6020 (9.1%)
2013	1475 (9.6%)	4493 (8.9%)	5968 (9.0%)
2014	1373 (9.0%)	4561 (9.0%)	5934 (9.0%)
2015	1244 (8.1%)	4721 (9.3%)	5965 (9.0%)
2016	1126 (7.4%)	4902 (9.7%)	6028 (9.1%)
2017	1031 (6.7%)	4851 (9.6%)	5882 (8.9%)
Tumor characteristics			
Morphology			
Intraductal	12194 (79.7%)	42100 (83.0%)	54294 (82.3%)
Lobular	2046 (13.4%)	5498 (10.8%)	7544 (11.4%)
Intraductal and Lobular	1053 (6.9%)	3106 (6.1%)	4159 (6.3%)
Tumor size (mm)	20.6 (11.1)	14.6 (8.74)	16.0 (9.68)
Stage (II vs ref: I)	8235 (53.8%)	14236 (28.1%)	22471 (34.0%)
Primary site			
Nipple	74 (0.5%)	186 (0.4%)	260 (0.4%)
Central	1193 (7.8%)	2168 (4.3%)	3361 (5.1%)
UIQ	1682 (11.0%)	6942 (13.7%)	8624 (13.1%)
LIQ	878 (5.7%)	3203 (6.3%)	4081 (6.2%)
UOQ	4671 (30.5%)	18704 (36.9%)	23375 (35.4%)
LOQ	1133 (7.4%)	3717 (7.3%)	4850 (7.3%)
Axillary tail	38 (0.2%)	207 (0.4%)	245 (0.4%)
Overlapping lesion	3417 (22.3%)	11850 (23.4%)	15267 (23.1%)
Breast, NOS	2207 (14.4%)	3727 (7.4%)	5934 (9.0%)
Laterality			
Right: origin of primary	7454 (48.7%)	25053 (49.4%)	32507 (49.3%)
Left: origin of primary	>7828 (>51.2%)	>25640 (>50.6%)	>33479 (>50.7%)
Other or unspecified	<11	<11	<11
Grade			
I	3266 (21.4%)	14862 (29.3%)	18128 (27.5%)
II	7281 (47.6%)	24097 (47.5%)	31378 (47.5%)
III	4366 (28.5%)	10376 (20.5%)	14742 (22.3%)

IV	43 (0.3%)	103 (0.2%)	146 (0.2%)
Unknown	337 (2.2%)	1266 (2.5%)	1603 (2.4%)
HR status			
Positive or borderline	12742 (83.3%)	44922 (88.6%)	57664 (87.4%)
Negative	2111 (13.8%)	4856 (9.6%)	6967 (10.6%)
Unknown	440 (2.9%)	926 (1.8%)	1366 (2.1%)
Elixhauser Comorbidities			
Comorbidity index	9.62 (8.28)	7.47 (6.79)	7.97 (7.22)

*For a categorical variable, count (percent). For a continuous variable, mean (standard deviation).

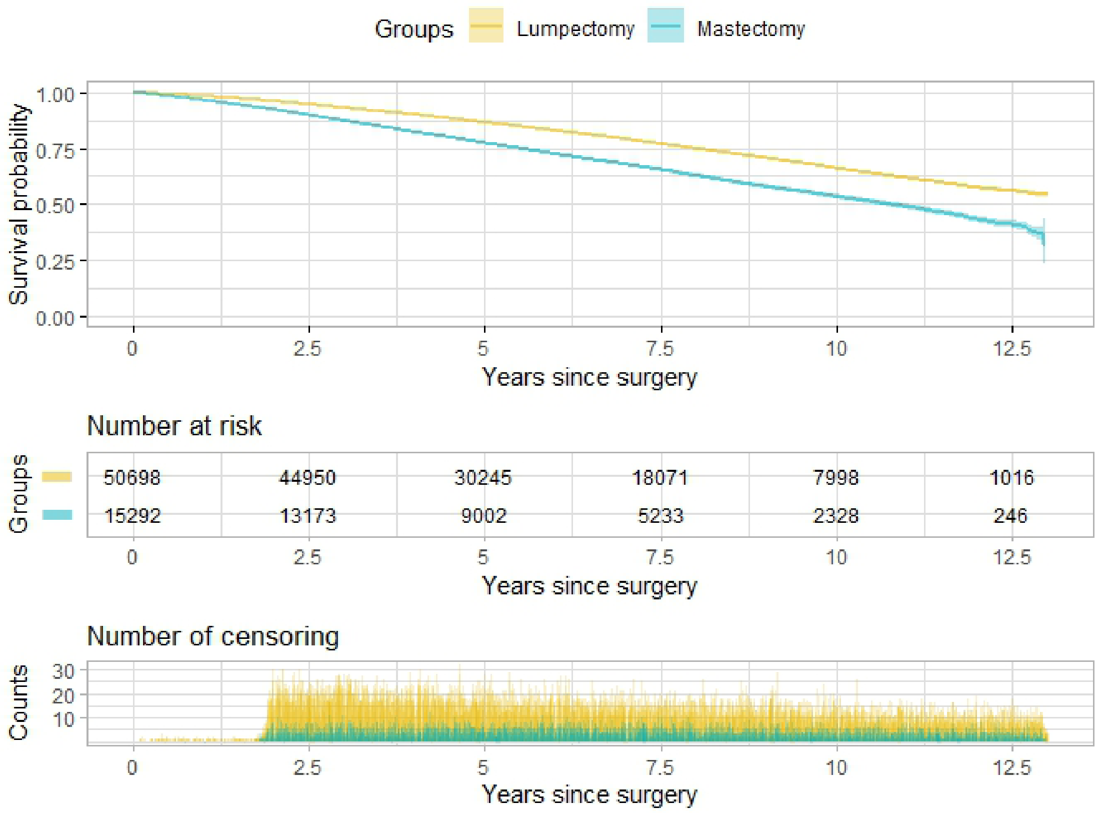

Figure 3. **Unadjusted survival analysis**.

penalty level decreases, all variables can be included into the propensity score model. For example, race and HR status positive also have relatively larger group norms, which suggests that those variables are important factors and may be associated with the decision-making between lumpectomy and mastectomy. Figure 5a and b show the distributions of the estimated propensity scores and weights. The two treatment groups have visibly different propensities towards surgical options. As shown in Figure 5c, prior to the weighting, some of the absolute standardized mean difference (SMD) values were above the 0.1 threshold, especially for tumor size and stage. Both variables are important factors to be considered when making a choice between lumpectomy and mastectomy. All the absolute SMDs significantly reduced to be below 0.1 and even nearly 0 with the weighting, with which the two arms could be considered as balanced. For comparison, we also considered the standard logistic regression approach. The Chi-squared test for deviance returned a p-value < 0.001, which suggested unsatisfactory model fitting and hence justified the adoption of DNN.

In the survival analysis step, we also examined the parameter paths to assess the relative importance of input variables. As shown in Figure 4b, overall survival was found to be highly related to Elixhauser comorbidities and age, which remained in the model even with larger tun-

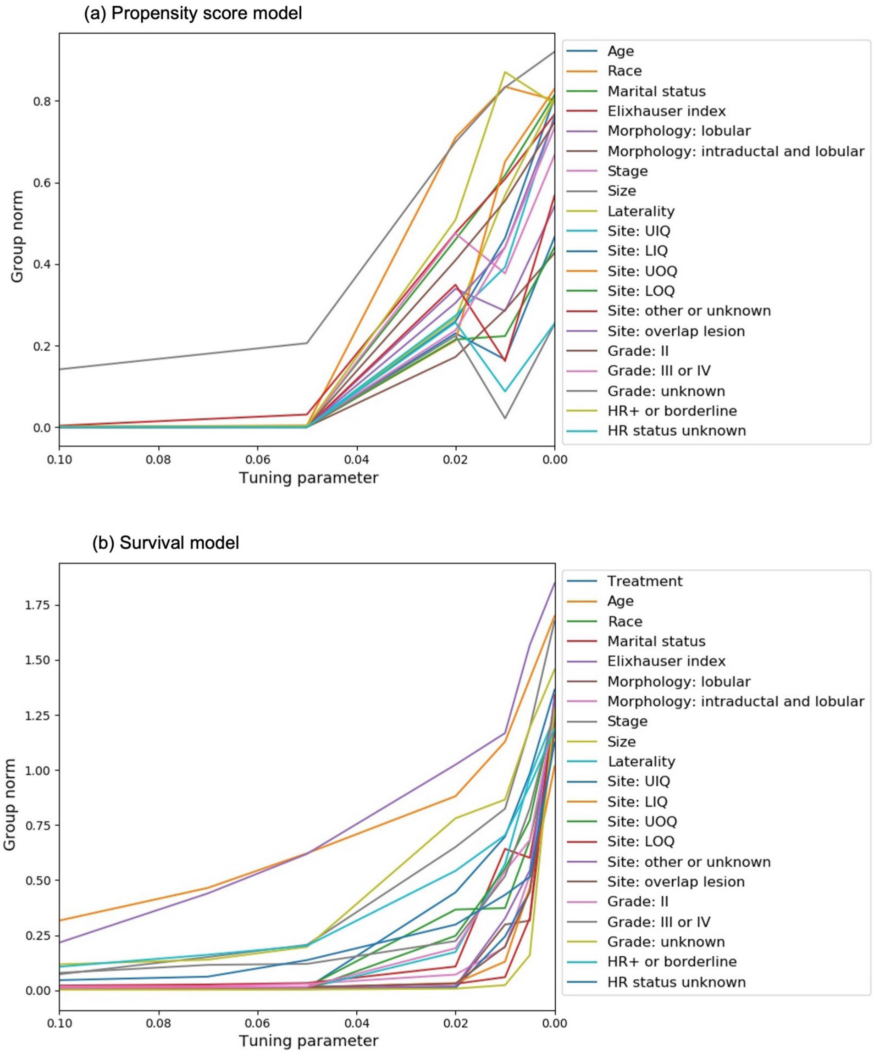

Figure 4. **Parameter paths for the propensity score analysis (upper) and survival analysis (lower)**.

ing parameter values (and hence higher penalty). Subsequently, tumor size and tumor grade III or IV were found to contribute to survival. Based on the obtained DNN, the predicted subject-level survival would, on average, slightly increase, if a patient would have switched from lumpectomy to mastectomy. The result is summarized in Figure 6a. With the predicted survival time for each patient (without censoring), the Kaplan-Meier survival curves were constructed and shown in Figure 6b and c. In Figure 6b, the treatment assignment was created to be random. The survival curves of the predicted survival times for the two arms almost completely overlap (p-value = 0.2 by logrank test). In this assessment, the two surgical groups are balanced for treatment assignments, mimicking the randomization process in a well-executed RCT. In comparison, in Figure 6c, a beneficial effect of lumpectomy (p-value < 0.001 by logrank test) was observed, if the treatment assignment was based on what was observed and without addressing potential confounding. This assessment mimics real-world observations where there are always some inherent differences between those receiving different surgical procedures. This may also explain why some observational studies concluded that lumpectomy was associated with improved survival [29-32]. Overall, we concluded that lumpectomy and mastectomy had similar effects on overall survival. For all subjects in the studied cohort, the average estimated increase in overall survival is 0.05 years if a patient received lumpectomy in comparison to mastectomy. With the weighted bootstrap inference, the average mean survival year change (across bootstraps) was found to be 0.08, with a standard deviation of 0.08 and 95% CI [-0.08, 0.25].

In the sensitivity analysis with respect to HER2, a similar conclusion was made. In particular, the average estimated increase in overall survival is 0.094 years if a received lumpectomy in comparison to mastectomy. With

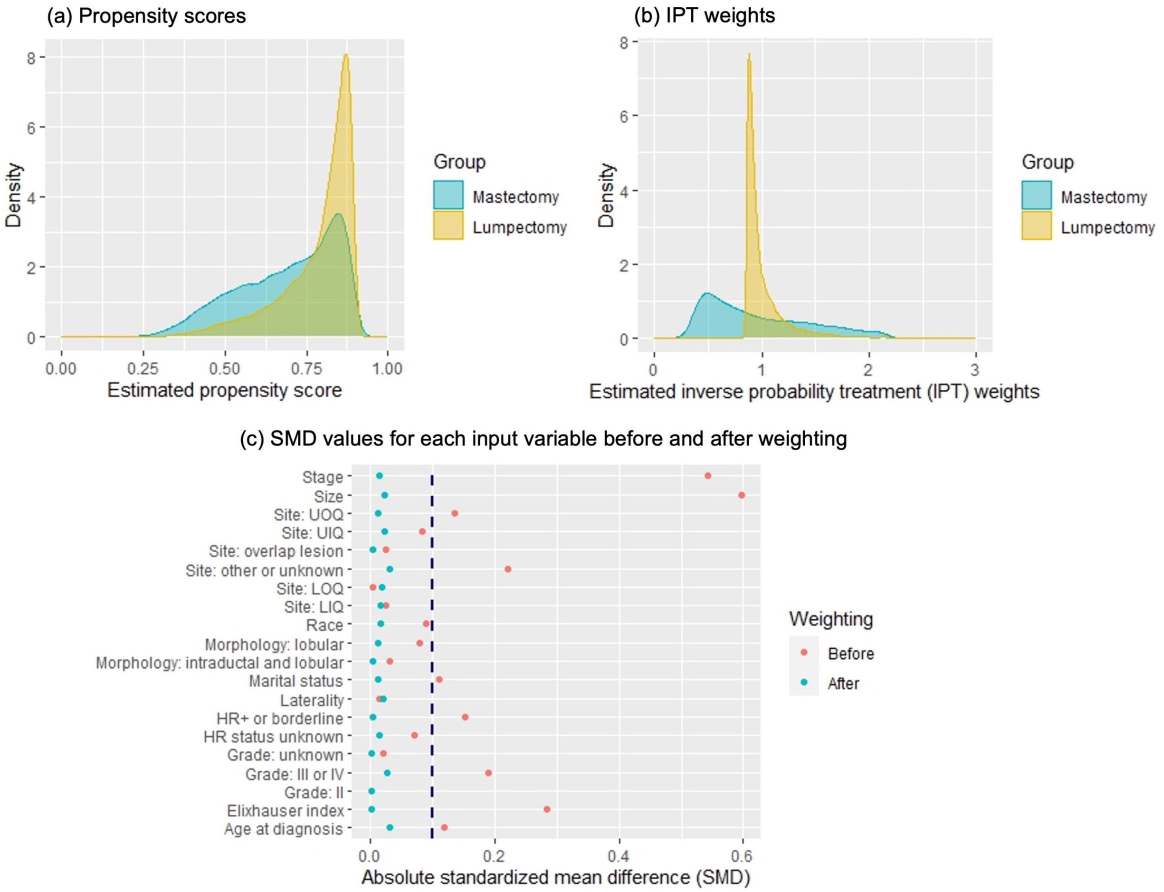

Figure 5. Distributions of estimated (a) propensity score and (b) inverse probability treatment weights; and (c) absolute standard mean difference of all confounders before and after weighting.

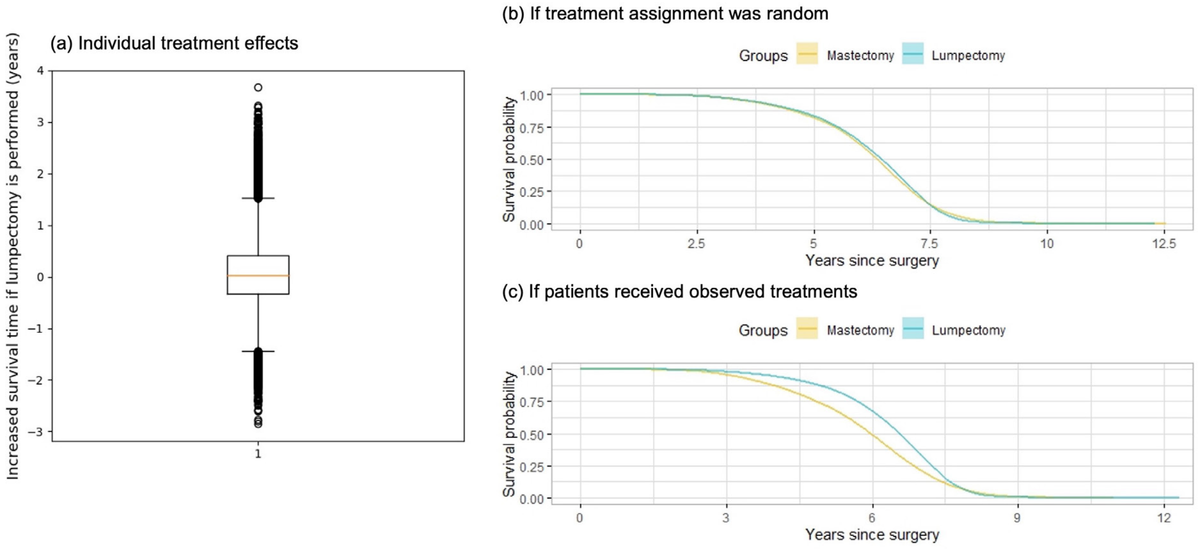

Figure 6. Analysis of predicted survival: (a) predicted subject-level improvement in survival if a patient would switch treatment; (b) survival probabilities based on the predicted survival times if the treatment assignment was random; and (c) survival probabilities with patients under the original observed treatments.

the weighted bootstrap inference, the average mean survival year change was found to be 0.14, with a standard deviation of 0.08 and 95% CI [-0.02, 0.29]. In the sensitivity analysis with respect to radiotherapy, it was concluded that lumpectomy with radiotherapy was superior. The average estimated increase in overall survival is 0.30 years if a patient received lumpectomy in comparison to mastectomy. With the weighted bootstrap inference, the average mean survival year change was found to be 0.29, with a standard deviation of 0.09 and 95% CI [0.13, 0.49]. It is noted that the definition of time zero changed and that the results could be subject to immortal time bias (which may result in a misleading conclusion on the beneficial treatment effect of lumpectomy with radiation therapy). As such, this conclusion should be taken with caution. With the standard Cox regression analysis, it was found that lumpectomy was slightly superior, and the hazard ratio was 0.96, with 95% CI [0.92, 0.99]. As the Cox model failed model diagnostics, the conflict with the deep learning-based analysis is not surprising. Additionally, it is noted that the estimated treatment under the Cox model is in fact very small.

DISCUSSION

This article has provided another showcase of mining large observational data, extracting valuable information, complementing RCTs, and informing clinical practice. In our data analysis, the sample size is dramatically larger than regular RCTs, making the analysis much more powerful. It is usually not easy to significantly increase sample sizes in RCTs. However, with the accumulation of data, the sample size of this study and other observational data analyses can easily keep increasing. The analyzed real-world data can be closer to real clinical practice than RCTs. With a much broader coverage and more diverse population, the findings can be more generalizable than RCTs. As acknowledged in the literature [34,44], emulation and other big data analysis of observational data can be the most useful for assessing the effectiveness of existing treatments (that have not been directly compared in RCTs) and updating comparisons (that have been done years ago on patients with different characteristics). It is noted that the proposed analysis can also be done with local data, for example, generated by the Yale New Haven Healthcare System, to better reflect local clinical practice and patient characteristics. With the promising performance of deep learning on other biomedical problems (for example, imaging and omics data analysis), it is a natural step to develop deep learning-based emulation analysis. The proposed analysis pipeline has combined the strengths of both statistical analysis and “regular” DNNs. Specifically, key building blocks of the “standard” DNN architecture have been retained. As such, the proposed analysis, similar to other DNNs, can have better estimation/prediction performance and more flexibly accommodate unspecified nonlinear relationships. It can be especially useful when the logistic, Cox, and other model assumptions are not satisfied, as in this data analysis. Here, it is noted that some recent studies have pointed out that DNNs may not be entirely model-free [45,46]. Our literature review suggested that there is still a lack of model diagnostics tools for DNNs – it will be of interest to diagnose the proposed DNNs when such tools become available. On the other hand, the proposed analysis has inherited the lucid framework of emulation analysis, with the clear propensity score and survival analysis steps. Additionally, it has incorporated penalized estimation, examination of parameter paths, and bootstrap-based inference, which are routine in regression analysis but still have not been well developed in deep learning. It is noted that deep learning is not “free.” Different from regression analysis, there is a lack of simple models. The resulting DNNs are available from the authors, however, do not have easily interpretable forms. Additionally, they do not deliver simple statistics such as hazard ratio. The proposed assessment based on the predicted survival time can largely alleviate this problem.

In data analysis, our main finding is that lumpectomy and mastectomy have comparable long-term overall survival for early-stage elderly female breast cancer patients. It is noted that the studied population differs from many of the existing studies. The conclusion is consistent with most of the existing RCTs and can complement the existing literature. For example, the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-04 study [25] and the study at the National Cancer Institute in Milan [47] both suggested no difference in 20-year overall survival after the two surgical procedures. More recent observational studies, interestingly, reached different conclusions. For example, Agarwal et al. [29], Hwang et al. [30], and Wrubel et al. [32] concluded that lumpectomy was superior in terms of breast cancer-specific and overall survival compared to mastectomy. The difference in findings can be at least partly attributed to the differences in study populations (for example, different age groups and years of diagnosis). Mogal et al. [31] specially analyzed elderly patients and also concluded that lumpectomy was significantly superior in terms of overall survival. It is noted that all of the above studies adopted regression-based association analysis techniques. Most of them did not report any model diagnostics, and there could be a risk of model misspecification. Additionally, some adopted the Kaplan-Meier technique without addressing possible confounding bias.

Limitations of the proposed analysis are fully recognized. SEER-Medicare and other EMR and insurance claims databases may not have comprehensive and accurate information as RCTs. For example, to capture most of the observed treatments of interest, we had an additional inclusion criterion in our emulated trial (compared to the target trial) as having at least one year or until death enrollment in Medicare part A or B. This may limit the sample size and introduce a selection bias to our analysis. For the main data analysis, unmeasured and possibly relevant confounders may include HER2 status, lymphovascular invasion, extracapsular invasion, and size of nodal metastases, which may have an impact on both surgical choices and prognosis. This is the practice paid for collecting data for a large population without specific targets as in RCTs. To partially address this limitation, we conducted a sensitivity analysis with respect to HER2 status. The proposed analysis may suffer the same limitations as the existing emulation techniques. For example, it demands a proper definition of treatment assignment and time zero. Otherwise, it may lead to falsely defined treatment effects and succumb to immortal time bias. In the main analysis, we omitted the subsequent chemotherapy, radiotherapy, or endocrine therapy before and after surgery, as our main interest was in the two surgical procedures. As discussed in [29], the indication for systemic chemotherapy is comparable for both surgical groups, thereby minimizing a disproportionate impact of chemotherapy. To partially mitigate this limitation, we additionally conducted a sensitivity analysis with respect to adjuvant radiation therapy performed after surgery. However, including neoadjuvant and adjuvant therapies in the analysis may pose a challenge to the proper definition of treatment and choice of time zero. Future research will be needed to address the corresponding immortal time bias – otherwise, we may not be able to perform a valid emulation analysis (but more of an observation analysis informing association). In the analysis, data from a time window much wider than regular RCTs was pooled. The analysis can be repeated for shorter time windows to examine potential temporal variations. It is also recognized that adopting DNN (and other complex modelings) may have the risk of overfitting. In this study, this risk can be largely mitigated via penalization, the relatively simple structure of the DNNs, and the large sample size.

CONCLUSION

This study has further advanced big data analysis for biomedicine by developing a deep learning-based emulation analysis pipeline for mining large observational data and complementing RCTs. Equally importantly, the analysis of SEER-Medicare data has led to the conclusion that lumpectomy and mastectomy have similar effects on overall survival for elderly SEER-Medicare early-stage female breast cancer patients. This finding can directly inform breast cancer clinical practice.

Acknowledgments: The collection of the California cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract N01-PC-35136 awarded to the Northern California Cancer Center, contract N01-PC-35139 awarded to the University of Southern California, and contract N02-PC-15105 awarded to the Public Health Institute; and the Centers for Disease Control and Prevention's National Program of Cancer Registries, under agreement #U55/CCR921930-02 awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the author(s) and endorsement by the State of California, Department of Public Health the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors is not intended nor should be inferred. The authors acknowledge the efforts of the Applied Research Program, NCI; the Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database. The interpretation and reporting of the SEER-Medicare data are the sole responsibility of the authors.

Funding: This study was partly supported by NIH HL161691 and CA276790.

REFERENCE

1. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr;183(8):758–64.
2. Maringe C, Benitez Majano S, Exarchakou A, Smith M, Rachet B, Belot A, et al. Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. Int J Epidemiol. 2020 Oct;49(5):1719–29.
3. Petito LC, García-Albéniz X, Logan RW, Howlader N, Mariotto AB, Dahabreh IJ, et al. Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)-Medicare Linked Database. JAMA Netw Open. 2020 Mar;3(3):e200452.
4. García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol. 2017 Jun;32(6):495–500.
5. Mei H, Wang J, Ma S. An emulated target trial analysis based on Medicare data suggested non-inferiority of Dabigatran versus Rivaroxaban. J Clin Epidemiol. 2021 Nov;139:28–37.
6. Danaei G, García Rodríguez LA, Cantero OF, Logan RW, Hernán MA. Electronic medical records can be used to emulate target trials of sustained treatment strategies. J

Clin Epidemiol. 2018 Apr;96:12–22.

7. Atkinson A, Zwahlen M, Barger D, d'Arminio Monforte A, De Wit S, Ghosn J, et al. Withholding primary PcP prophylaxis in virologically suppressed HIV patients: an emulation of a pragmatic trial in COHERE. Clin Infect Dis. 2020;•••: https://doi.org/10.1093/cid/ciaa615.
8. Young J, Scherrer AU, Calmy A, Tarr PE, Bernasconi E, Cavassini M, et al.; Swiss HIV Cohort Study. The comparative effectiveness of NRTI-sparing dual regimens in emulated trials using observational data from the Swiss HIV Cohort Study. Antivir Ther. 2019;24(5):343–53.
9. Caniglia EC, Robins JM, Cain LE, Sabin C, Logan R, Abgrall S, et al. Emulating a trial of joint dynamic strategies: an application to monitoring and treatment of HIV-positive individuals. Stat Med. 2019 Jun;38(13):2428–46.
10. Lyu B, Chan MR, Yevzlin AS, Gardezi A, Astor BC. Arteriovenous Access Type and Risk of Mortality, Hospitalization, and Sepsis Among Elderly Hemodialysis Patients: A Target Trial Emulation Approach. Am J Kidney Dis. 2021: https://doi.org/10.1053/j.ajkd.2021.03.030.
11. Hoffman KL, Schenck EJ, Satlin MJ, Whalen W, Pan D, Williams N, et al. Comparison of a target trial emulation framework vs Cox regression to estimate the association of corticosteroids with COVID-19 mortality. JAMA Network Open. 2022;5(10):e2234425-e.
12. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998 Oct;17(19):2265–81.
13. Deng L, Yu D. Deep learning: methods and applications. Foundations and trends® in signal processing. 2014;7(3–4):197-387. https://doi.org/10.1561/9781601988157.
14. Yu D, Deng L. Deep learning and its applications to signal and information processing [exploratory dsp]. IEEE Signal Process Mag. 2010;28(1):145–54.
15. Heaton JB, Polson NG, Witte JH. Deep learning for finance: deep portfolios. Appl Stochastic Models Bus Ind. 2017;33(1):3–12.
16. Weichenthal S, Hatzopoulou M, Brauer M. A picture tells a thousand…exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environ Int. 2019 Jan;122:3–10.
17. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver CancerUsing deep learning to predict liver cancer prognosis. Clin Cancer Res. 2018 Mar;24(6):1248–59.
18. Zhou K, Greenspan H, Shen D. Deep learning for medical image analysis. Academic Press; 2017.
19. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access. 2017;6:9375–89.
20. Mei H, Xu Y, Wang J, Ma S. Evaluation of Survival Outcomes of Endovascular Versus Open Aortic Repair for Abdominal Aortic Aneurysms with a Big Data Approach. Entropy (Basel). 2020 Nov;22(12):1349.
21. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods. 2019 Jan;16(1):67–70.
22. Sarrazin D, Lê MG, Arriagada R, Contesso G, Fontaine F, Spielmann M, et al. Ten-year results of a randomized trial comparing a conservative treatment to mastectomy in early breast cancer. Radiother Oncol. 1989 Mar;14(3):177–84.
23. Veronesi U, Saccozzi R, Del Vecchio M, Banfi A, Clemente C, De Lena M, et al. Comparing radical mastectomy with quadrantectomy, axillary dissection, and radiotherapy in patients with small cancers of the breast. N Engl J Med. 1981 Jul;305(1):6–11.
24. Lichter AS, Lippman ME, Danforth DN Jr, d'Angelo T, Steinberg SM, deMoss E, et al. Mastectomy versus breast-conserving therapy in the treatment of stage I and II carcinoma of the breast: a randomized trial at the National Cancer Institute. J Clin Oncol. 1992 Jun;10(6):976–83.
25. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. N Engl J Med. 2002 Oct;347(16):1233–41.
26. Smith BD, Jiang J, McLaughlin SS, Hurria A, Smith GL, Giordano SH, et al. Improvement in breast cancer outcomes over time: are older women missing out? J Clin Oncol. 2011 Dec;29(35):4647–53.
27. Schonberg MA, Marcantonio ER, Li D, Silliman RA, Ngo L, McCarthy EP. Breast cancer among the oldest old: tumor characteristics, treatment choices, and survival. J Clin Oncol. 2010 Apr;28(12):2038–45.
28. Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining bias in studies of statin treatment and survival in patients with cancer. JAMA Oncol. 2018 Jan;4(1):63–70.
29. Agarwal S, Pappas L, Neumayer L, Kokeny K, Agarwal J. Effect of breast conservation therapy vs mastectomy on disease-specific survival for early-stage breast cancer. JAMA Surg. 2014 Mar;149(3):267–74.
30. Hwang ES, Lichtensztajn DY, Gomez SL, Fowble B, Clarke CA. Survival after lumpectomy and mastectomy for early stage invasive breast cancer: the effect of age and hormone receptor status. Cancer. 2013 Apr;119(7):1402–11.
31. Mogal HD, Clark C, Dodson R, Fino NF, Howard-McNatt M. Outcomes after mastectomy and lumpectomy in elderly patients with early-stage breast cancer. Ann Surg Oncol. 2017 Jan;24(1):100–7.
32. Wrubel E, Natwick R, Wright GP. Breast-conserving therapy is associated with improved survival compared with mastectomy for early-stage breast cancer: a propensity score matched comparison using the national cancer database. Ann Surg Oncol. 2021 Feb;28(2):914–9.
33. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
34. Danaei G, Rodríguez LA, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. Stat Methods Med Res. 2013 Feb;22(1):70–96.
35. Stuart E. Propensity Score Software [cited 2023 May 23]. Available from: https://www.elizabethstuart.org/psoftware/
36. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. J R Stat Soc Series B Stat Methodol. 2008;70(1):53–71.
37. Hastie T. Ridge regularization: an essential concept in data

science. Technometrics. 2020;62(4):426–33.

38. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018 Feb;18(1):24.
39. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med. 1992;11(14-15):1871–9.
40. Stute W. Nonlinear censored regression. Stat Sin. 1999:1089–102.
41. Stute W. Consistent estimation under random censorship when covariables are present. J Multivariate Anal. 1993;45(1):89–103.
42. Ma S, Kosorok MR. Robust semiparametric M-estimation and the weighted bootstrap. J Multivariate Anal. 2005;96(1):190–217.
43. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7.
44. Fiks AG, Grundmeier RW, Margolis B, Bell LM, Steffes J, Massey J, et al. Comparative effectiveness research using the electronic medical record: an emerging area of investigation in pediatric primary care. J Pediatr. 2012 May;160(5):719–24.
45. Pearce T, Brintrup A, Zaki M, Neely A, editors. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. International conference on machine learning; 2018: PMLR.
46. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst. 2017:30.
47. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, et al. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. N Engl J Med. 2002 Oct;347(16):1227–32.

Appendix A

Table S1. Comparison of the target trial and the emulated trial.

Component	Target trial	Emulated trial
Aim	To test the comparative effectiveness of Mastectomy and Lumpectomy for overall survival in early-stage breast cancer patients of age 66 and older	Same
Inclusion Criteria	Subjects are female aged 66 or older and meet the following criteria: a) were diagnosed with the first primary invasive intraductal or/and lobular breast cancer, b) were stage I or II at the time of diagnosis, c) tumor size ≤ 5cm at diagnosis. Subjects were able and willing to provide informed consent.	Other than informed consent, the eligibility in the target trial should all be met. Subjects also a) had continuous enrollment in Medicare Part A and Part B, with no health maintenance organization (HMO) enrollment from one year before to one year after cancer diagnosis or death, whichever occurred first, and b) received lumpectomy or mastectomy within 120 days after diagnosis.
Exclusion Criteria	1) Enrolment in a conflicting clinical trial, 2) Any disease other than breast cancer associated with a likelihood of survival of less than one year, and 3) unlikeness to complete the trial and follow-up activities.	1) missing key confounder information: age, gender, cancer stage, and tumor size at the time of diagnosis. 2) missing essential records for medical history and treatment initiation, for example, did not have continuous enrollment in Medicare before and after one year of diagnosis or death.
Treatment	Individuals were randomly assigned to a strategy in an unblinded fashion. Either received Mastectomy or Lumpectomy.	Same. In the emulated trial, set the treatment assignment window as one year after diagnosis.
Follow-up	Time of randomization to treatment was considered as the starting point of the follow-up.	Time of randomization was not directly observable. Time of surgery was used as time zero for both groups.
Outcome	All-cause mortality	Same
Causal Contrast	Intention-to-treat effect, i.e., effect of being assigned to Mastectomy versus Lumpectomy at baseline.	Same
Statistical Analysis	Logrank test and Cox regression as the primary analysis techniques. Possible confounders (e.g., study site and co-morbidity score) would be adjusted.	A deep learning technique based on the AFT model. Propensity score-type weighting used to improve covariance balance. Variance estimation/inference obtained via the weighted bootstrap technique.

Table S2. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), ICD-10-CM, and Healthcare Common Procedure Coding System (HCPCS)/Current Procedural Terminology (CPT) codes for identifying treatments.

Procedure	ICD-9	ICD-10	CPT/HCPCS
Mastectomy *,**	Procedure: 85.33–85.36, 85.4, 85.41-85.48,	Procedure: 0HRT075-0HRT079, 0HRT07Z, 0HRT0JZ, 0HRT0KZ, 0HRT37Z, 0HRT3JZ, 0HRT3KZ, 0HRU075-0HRU079, 0HRU07Z, 0HRU0JZ, 0HRU0KZ, 0HRU37Z, 0HRU3JZ, 0HRU3KZ, 0HRV075-0HRV079, 0HRV07Z, 0HRV0JZ, 0HRV0KZ, 0HRV37Z, 0HRV3JZ, 0HRV3KZ, 0HTT0ZZ, 0HTU0ZZ, 0HTV0ZZ	19300, 19303–19307
Lumpectomy *,**	Procedure: 85.20–85.23	Procedure: 0HBT0ZZ, 0HBT3ZZ, 0HBT7ZZ, 0HBT8ZZ, 0HBU0ZZ, 0HBU3ZZ, 0HBU7ZZ, 0HBU8ZZ, 0HBV0ZZ, 0HBV3ZZ, 0HBV7ZZ, 0HBV8ZZ	19301, 19302, 19120, 19125, 19126
Radiation therapy	Diagnosis: V580, V661, V671 Procedure: 9221-9229	Diagnosis: Z510 Procedure: See DM-Radiation, DM0-Beam Radiation***	77401-77499, 77750-77799, 77520, 77523, G0256, G0261

* Only included cases that had surgery performed in less than 120 days after the initial date of diagnosis (with the consideration that these cases might have been treated with neoadjuvant chemotherapy).

** Patients were categorized as having lumpectomy if they had lumpectomy as the first surgical procedure after breast cancer diagnosis. Patients were categorized as having mastectomy if they had mastectomy as the first surgical procedure after diagnosis. If lumpectomy was performed first followed by mastectomy conducted within a one-year window after diagnosis, the patient would be censored at the time of mastectomy. If mastectomy was performed first, further lumpectomy was ignored. If receiving both mastectomy and lumpectomy on the same day, a patient was classified as receiving mastectomy.

*** Analysis was focused on radiotherapy conducted on breast only. Codes include: DM000ZZ, DM001ZZ, DM002ZZ ,DM003Z0, DM003ZZ, DM004ZZ, DM005ZZ, DM006ZZ, DM010ZZ, DM011ZZ, DM012ZZ, DM013Z0, DM013ZZ, DM014ZZ, DM015ZZ, DM016ZZ, DM1097Z, DM1098Z, DM1099Z, DM109BZ, DM109CZ, DM109YZ, DM10B6Z, DM10B7Z, DM10B8Z, DM10B9Z, DM10BB1, DM10BBZ, DM10BCZ, DM10BYZ, DM1197Z, DM1198Z, DM1199Z, DM119BZ, DM119CZ, DM119YZ, DM11B6Z, DM11B7Z, DM11B8Z, DM11B9Z, DM11BB1, DM11BBZ, DM11BCZ, DM11BYZ, DM20DZZ, DM20HZZ, DM20JZZ, DM21DZZ, DM21HZZ, DM21JZZ, DMY07ZZ, DMY08ZZ, DMY0FZZ, DMY17ZZ, DMY18ZZ, DMY1FZZ.

Section S1: Sensitivity analysis with respect to HER2 status

In the main analysis, HER2 status was not included, with the consideration that its information was missing for patients diagnosed before 2010. To examine the potential confounding effect of HER2 status, we further limited the cohort for analysis to those diagnosed on or after 2010 and included HER2 as a confounder. The patient characteristics are summarized in Table S3.

Table S3. Patients' characteristics by treatment for those diagnosed on or after 2010.

	Mastectomy (n = 10,689)	Lumpectomy (n = 37,079)	Overall (n = 47,768)
Demographics			
Age at diagnosis	75.9 (6.82)	75.2 (6.49)	75.3 (6.57)
Race (Non-Hispanic White)	8632 (80.8%)	31282 (84.4%)	39914 (83.6%)
Marital Status (married)	4787 (44.8%)	18636 (50.3%)	23423 (49.0%)
Year of diagnosis			
2010	1456 (13.6%)	4468 (12.1%)	5924 (12.4%)
2011	1472 (13.8%)	4575 (12.3%)	6047 (12.7%)
2012	1512 (14.1%)	4508 (12.2%)	6020 (12.6%)
2013	1475 (13.8%)	4493 (12.1%)	5968 (12.5%)
2014	1373 (12.8%)	4561 (12.3%)	5934 (12.4%)
2015	1244 (11.6%)	4721 (12.7%)	5965 (12.5%)
2016	1126 (10.5%)	4902 (13.2%)	6028 (12.6%)
2017	1031 (9.6%)	4851 (13.1%)	5882 (12.3%)
Tumor characteristics			
Morphology			
Intraductal	8434 (78.9%)	30778 (83.0%)	39212 (82.1%)
Lobular	1503 (14.1%)	4089 (11.0%)	5592 (11.7%)
Intraductal and Lobular	752 (7.0%)	2212 (6.0%)	2964 (6.2%)
Tumor size (mm)	20.6 (11.1)	14.6 (8.73)	15.9 (9.64)
Stage (II vs ref: I)	5749 (53.8%)	10177 (27.4%)	15926 (33.3%)
Primary site			
Nipple	53 (0.5%)	123 (0.3%)	176 (0.4%)
Central	812 (7.6%)	1509 (4.1%)	2321 (4.9%)
UIQ	1210 (11.3%)	5239 (14.1%)	6449 (13.5%)
LIQ	608 (5.7%)	2365 (6.4%)	2973 (6.2%)
UOQ	3300 (30.9%)	13707 (37.0%)	17007 (35.6%)
LOQ	797 (7.5%)	2784 (7.5%)	3581 (7.5%)
Axillary tail	25 (0.2%)	130 (0.4%)	155 (0.3%)
Overlapping lesion	2400 (22.5%)	8775 (23.7%)	11175 (23.4%)
Breast, NOS	1484 (13.9%)	2447 (6.6%)	3931 (8.2%)
Laterality			
Right: origin of primary	5223 (48.9%)	18280 (49.3%)	23503 (49.2%)
Left: origin of primary	>5455 (>51.0%)	>18788 (>50.7%)	>24254 (>50.8%)
Other or unspecified	<11	<11	<11
Grade			
I	2291 (21.4%)	11120 (30.0%)	13411 (28.1%)
II	5226 (48.9%)	17693 (47.7%)	22919 (48.0%)
III	2945 (27.6%)	7384 (19.9%)	10329 (21.6%)
IV	17 (0.2%)	46 (0.1%)	63 (0.1%)

Unknown	210 (2.0%)	836 (2.3%)	1046 (2.2%)
HR status			
Positive or borderline	9134 (85.5%)	33312 (89.8%)	42446 (88.9%)
Negative	1366 (12.8%)	3317 (8.9%)	4683 (9.8%)
Unknown	189 (1.8%)	450 (1.2%)	639 (1.3%)
HER2 status			
Positive or borderline	1552 (14.5%)	3627 (9.8%)	5179 (10.8%)
Negative	8800 (82.3%)	32537 (87.8%)	41337 (86.5%)
Unknown	337 (3.2%)	915 (2.5%)	1252 (2.6%)
Elixhauser Comorbidities			
Comorbidity index	9.48 (8.31)	7.36 (6.79)	7.83 (7.21)

* For a categorical variable, count (percent). For a continuous variable, mean (standard deviation).

In this analysis, it was concluded that lumpectomy and mastectomy had similar effects on overall survival. For all subjects in the studied cohort, the average estimated increase in overall survival was 0.094 years, if a patient received lumpectomy in comparison to mastectomy. With the weighted bootstrap inference, the average mean survival year change was found to be 0.14, with a standard deviation of 0.08 and 95% CI [-0.02, 0.29].

Section S2: Sensitivity analysis with respect to radiotherapy

In the studied cohort, most of the patients who received lumpectomy had following adjuvant radiation therapy performed. However, this was not the case for the patients with mastectomy. The observed patterns of receiving radiotherapy are summarized in Table S4. As radiotherapy can have a strong association with overall survival, we redefined the treatments of the target trial as Mastectomy without radiotherapy and Lumpectomy with radiotherapy. The patient characteristics are summarized in Table S5.

Table S4. Patterns of receiving radiotherapy.

Pattern	**Lumpectomy** (n = 50,704)	**Mastectomy** (n = 15,293)	**Overall** (n = 65,997)
Radiotherapy performed before the first surgery	698 (1.4%)	77 (0.5%)	775 (1.17%)
Radiotherapy performed on the same day of the first surgery	775 (1.5%)	18 (0.1%)	793 (1.2%)
Radiation therapy performed after the first surgery but before death or censoring	36,699 (72.4%)	2,266 (14.8%)	38,965 (59.0%)

Table S5. Patients' characteristics by treatment.

	Mastectomy w/o radiotherapy (n = 13,027)	**Lumpectomy w/ radiotherapy** (n = 36,699)	**Overall** (n = 49,726)
Demographics			
Age at diagnosis	76.4 (6.93)	74.1 (5.74)	74.7 (6.16)
Race (Non-Hispanic White)	10636 (81.6%)	31201 (85.0%)	41837 (84.1%)
Marital Status (married)	5639 (43.3%)	19495 (53.1%)	25134 (50.5%)
Year of diagnosis			
2007	1371 (10.5%)	3268 (8.9%)	4639 (9.3%)
2008	1325 (10.2%)	3264 (8.9%)	4589 (9.2%)
2009	1287 (9.9%)	3304 (9.0%)	4591 (9.2%)
2010	1247 (9.6%)	3242 (8.8%)	4489 (9.0%)
2011	1265 (9.7%)	3381 (9.2%)	4646 (9.3%)
2012	1298 (10.0%)	3239 (8.8%)	4537 (9.1%)
2013	1251 (9.6%)	3317 (9.0%)	4568 (9.2%)
2014	1154 (8.9%)	3337 (9.1%)	4491 (9.0%)
2015	1034 (7.9%)	3421 (9.3%)	4455 (9.0%)
2016	947 (7.3%)	3505 (9.6%)	4452 (9.0%)
2017	848 (6.5%)	3421 (9.3%)	4269 (8.6%)
Tumor characteristics			
Morphology			
Intraductal	10464 (80.3%)	30531 (83.2%)	40995 (82.4%)
Lobular	1683 (12.9%)	3919 (10.7%)	5602 (11.3%)
Intraductal and Lobular	880 (6.8%)	2249 (6.1%)	3129 (6.3%)
Tumor size (mm)	20.0 (10.8)	14.4 (8.35)	15.9 (9.37)
Stage (II vs ref: I)	6546 (50.2%)	10228 (27.9%)	16774 (33.7%)
Primary site			
Nipple	62 (0.5%)	108 (0.3%)	170 (0.3%)
Central	1020 (7.8%)	1438 (3.9%)	2458 (4.9%)
UIQ	1458 (11.2%)	5115 (13.9%)	6573 (13.2%)
LIQ	750 (5.8%)	2380 (6.5%)	3130 (6.3%)
UOQ	3974 (30.5%)	13954 (38.0%)	17928 (36.1%)
LOQ	954 (7.3%)	2756 (7.5%)	3710 (7.5%)
Axillary tail	31 (0.2%)	152 (0.4%)	183 (0.4%)
Overlapping lesion	2892 (22.2%)	8550 (23.3%)	11442 (23.0%)
Breast, NOS	1886 (14.5%)	2246 (6.1%)	4132 (8.3%)
Laterality			
Right: origin of primary	6301 (48.4%)	18236 (49.7%)	24537 (49.3%)
Left: origin of primary	>6715 (>51.5%)	>18452 (>50.3%)	>25178 (>50.6%)
Other or unspecified	<11	<11	<11
Grade			
I	2897 (22.2%)	10484 (28.6%)	13381 (26.9%)
II	6169 (47.4%)	17553 (47.8%)	23722 (47.7%)
III	3648 (28.0%)	7776 (21.2%)	11424 (23.0%)
IV	34 (0.3%)	70 (0.2%)	104 (0.2%)
Unknown	279 (2.1%)	816 (2.2%)	1095 (2.2%)
HR status			
Positive or borderline	32483 (88.5%)	10840 (83.2%)	5474 (11.0%)
Negative	3695 (10.1%)	1779 (13.7%)	43323 (87.1%)
Unknown	521 (1.4%)	408 (3.1%)	929 (1.9%)
Elixhauser Comorbidities			
Comorbidity index	9.21 (8.09)	7.03 (6.33)	7.60 (6.90)

* For a categorical variable, count (percent). For a continuous variable, mean (standard deviation).

It was concluded that lumpectomy with adjuvant radiation therapy outperformed mastectomy without adjuvant radiation therapy for overall survival. For all the subjects in the studied cohort, the average estimated increase in overall survival was 0.30 years if a patient received lumpectomy in comparison to mastectomy. With the weighted bootstrap inference, the average mean survival year change was found to be 0.29, with a standard deviation of 0.09 and 95% CI [0.13, 0.49]. One limitation of this sensitivity analysis is the potential vulnerability to the immortal time bias. Specifically, immortal time bias may arise because those who first received mastectomy might not receive subsequent radiotherapy due to worse health conditions and shorter survival after mastectomy, and because those who first received lumpectomy had to survive long enough to have the choice of receiving subsequent radiotherapy. The immortal time bias may distort the estimation of the treatment effects and lead to the misleading conclusion that lumpectomy with radiation therapy leads to longer survival.