Propensity Score Analysis to compare effects of radiation and surgery on survival time of lung cancer patients from National Cancer Registry (SEER)

Yan Wu

Advisor: Robert Pruzek

Epidemiology and Biostatistics,

School of Public Health

SUNY-Albany

December 2009

Data source: SEER

- Surveillance, Epidemiology, and End Results (SEER) registry from National Cancer Institute is a premier source for cancer statistics and an authoritative source of information on cancer incidence and survival in the United States. (http://seer.cancer.gov/)

- The SEER Program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data.

- SEER collects information on incidence, prevalence and survival from specific geographic areas representing 26 percent of the US population and compiles reports on all of these plus cancer mortality for the entire country and is intended for anyone interested in US cancer statistics or cancer surveillance methods.

- Updated annually and provided as a public service in print and electronic formats, SEER data are used by thousands of researchers, clinicians, public health officials, legislators, policymakers, community groups, and the public.

Background of the research

- Lung cancer is the second most diagnosed cancer in men and women, but it is the *number one cause of death* from cancer each year in both men and women.
- Surgical resection (cutting away) of the tumor generally is indicated for cancer that has not spread beyond the lung. It is the principal form of treatment for patients with stage 1 or stage 2 lung cancer
- Radiation therapy, or radiotherapy, delivers high-energy x-rays that can destroy rapidly dividing cancer cells. It can shrink the tumor(s) before surgery and eliminate most cancer cells that remain in the treated area after surgery.
- The SEER data that is analyzed in this report consists of 9474 cases pertaining to the effects of radiation and surgery on survival of patients with lung cancer.

Literature review of the research

- Cox regression analysis predicted long time survival after lung cancer surgery with early preoperative stage, age below 70 years and normal pulmonary function. ("Predictors of long time survival after lung cancer surgery: A retrospective cohort study" Kjetil Roth *et al*. *BMC Pulmonary Medicine* 2008, 8:22)

- Kaplan-Meier method was used to calculate 5-year postoperative survival in all categories of patient's characteristics and identified significant increase of survival in early stage patients. ("Surgery for non-small cell lung cancer: postoperative survival based on the revised tumor-node-metastasis classification and its time trend" Fumihiro Tanaka *et al*. European Journal of Cardio-thoracic Surgery 18 (2000) 147)

- Cox model and Kaplan-Meier curves can more accurately identify patients at risk for lung cancer death after surgery using histologic type and precise size specifications than using conventional tumor-node-metastasis staging, with SEER data. ("Survival after Surgery in Stage IA and IB Non–Small Cell Lung Cancer" David Ost *et al*. Am J Respir Crit Care Med Vol 177. pp 516–523, 2008)

Limitation of current studies:

- Research on the effects of surgery on survival focuses on building up survival models to identify significant variables in survival prediction.

- No research article on the effects of radiation treatment only on survival of lung cancer patients has been found.

- No studies seem to have comprehensively compared groups who did vs. did not have surgery or radiation with respect to survival for lung cancer. No comparison seems to have been made by controlling all available covariates!

Main purpose of this study:

- To answer the question: **Do lung cancer patients survive longer with treatment of surgery or radiation or both?**

- Since the observationally defined groups cannot be 'fairly' compared without adjustment, a wide variety of modern **PSA**-specific methods and graphics are used to compare groups after adjusting for selection bias.

Introduction to PSA

- **Definition of propensity scores**
- **Stage1- Estimation of Propensity Scores**
- **Stage2- Propensity Score Matching /Comparison methods**

PSA Definition

- **Definition**: The Propensity Score is the conditional probability that a unit *i* will receive a treatment (Z_i =1), given a set X of observed covariates:

$$e(x_i) = \text{pr}(Z_i = 1 \mid X_i = x_i)$$

(where it is assumed that, given the X's, the Z_i [0 or 1] are independent)

- **Key Properties of Propensity Scores:**

  - Given an appropriate (model) choice (for $e(x_i)$), treated and control subjects in the same stratum or matched set (having nearly the same PS) should have approximately the same distribution for each covariate, or combination of covariates.

  - Given an appropriate choice of covariates, treatment and control groups should differ from one another only by chance if their propensity scores are highly similar. This effect is similar to what happens when randomization is used to assign treatments; it is notable that it can also occur in well-designed observational studies.

Stage1-Estimate propensity scores

Method1: using logistic regression (LR) method

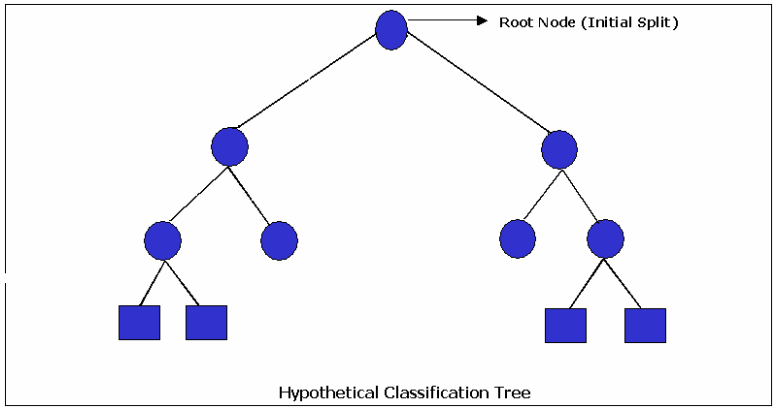
- Z_i = binary variable (1 = treatment, 0 = control)
- X_i = vector of Independent Variables (explanatory covariates). Propensity scores are estimated from the X_i; the estimation method should generally consider product functions, or interactions. Details later.
- Goal: use all covariates that appear to relate to treatment and outcome to improve prediction. (Not concerned about over-fitting in phase I of PSA, i.e. about external validity of PS estimation model.)
- If there is selection bias (for which adjustments are needed), will see (strong) discrimination reflected in ROC curve in the model. (But many of the best PSA studies show little discrimination.)
- Obtain fitted value estimate of $e(x_i)$, for each observation, based on covariates

$$\text{PScore= Prob(Zi=1) = e(x)} = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+\varepsilon_i)}}$$

Stage1-continued

Method2: classification tree

- Tree algorithm is not model-based; rather it is algorithmic
- Binary recursive partitioning used to form splits; bottom of tree identifies leaves
- Covariates and cut points are chosen to ensure the ‘best’ splits; interactions of covariates are therefore automatically detected
- Strata are formed naturally using leaves of tree; strata sizes are unconstrained & the # of terminal leaves/strata chosen based on prior experience with logistic regression and results of Rosenbaum and Rubin (1983).
- Propensity Scores are derived as # in treatment group / # of units, within each stratum; the number of strata equals the number of leaves (bottom of tree)
- All individuals are assigned the same estimated PScore in each stratum



Stage2-matching/comparison

- **Match each participant to one or more nonparticipants on propensity score:**
  - Nearest neighbor matching (only MM used here)
  - Caliper matching
  - Propensity score-based matching
- **Comparison**
  - Stratification-based comparison of outcomes across derived strata. Five (nearly equally sized) strata are generally sufficient to remove 90% of removable bias.
  - LOESS-based comparison of treatment to control groups for response (works best when Pscores are based on LR)
  - Paired dependent sample ANOVA with 1:1 matching

Nearest neighbor matching : selects the *m* comparison units whose propensity cores are “closest” to the treated unit in question.

- ❑ Matching with replacement
- Single nearest neighbor matching

- ❑ Matching without replacement (method used)
- Low-high, high-low or randomly ranked
- Highest ranked unit is matched first, the matched comparison unit is removed from further matching

Comparisons

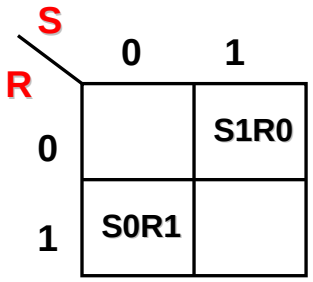
- ❖ Stratification-based comparison of outcomes across derived strata
- Comparison of means of treatment and control group within each stratum
- Calculate DAE-Direct Adjustment Estimator (weighted mean difference between control and treatment response across strata)

- ❖ Application of LOESS regression
- Local nonlinear regression curves obtained for treatment and control groups
- Requires choice of span argument to control smoothness
- The vertical distance between curves across ps range provides an index of the treatment effect weighted according to the density of the PScore distribution ignoring groups.

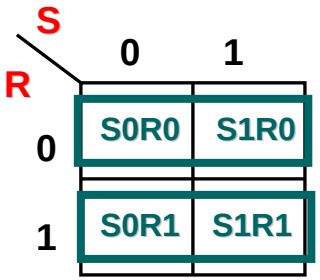
Purpose of the study

Overall question: **Do lung cancer patients survive longer with treatment of surgery or radiation or both?**

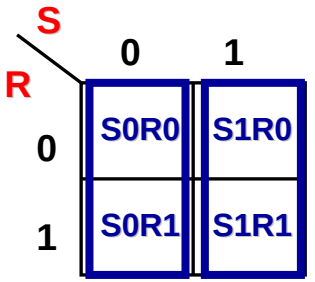
I: surgery/radiation



II: surgery



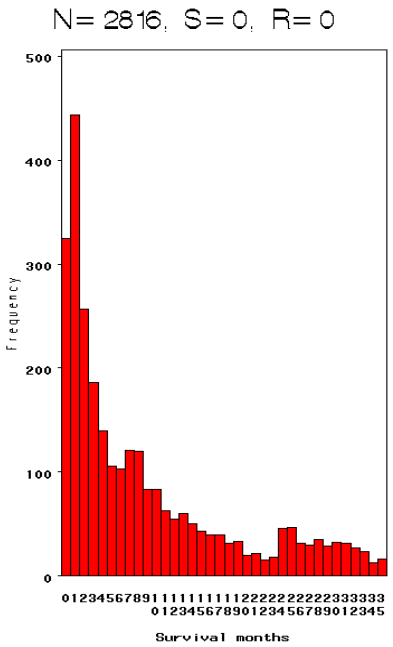
III: radiation

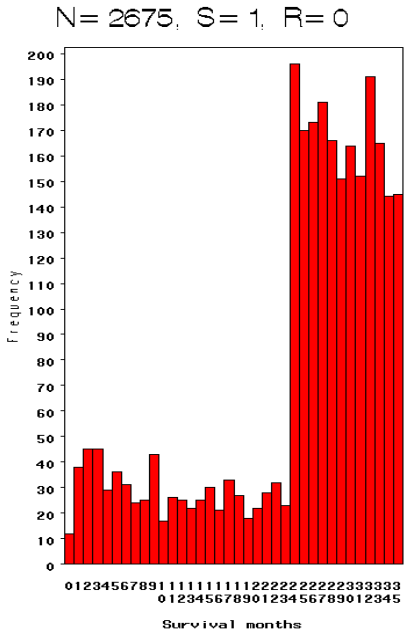


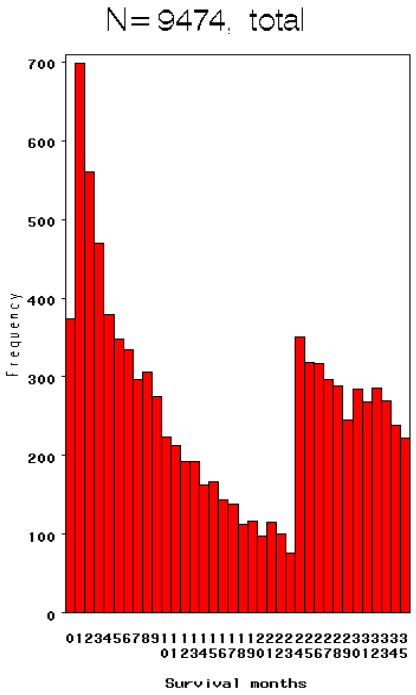
Data dictionary

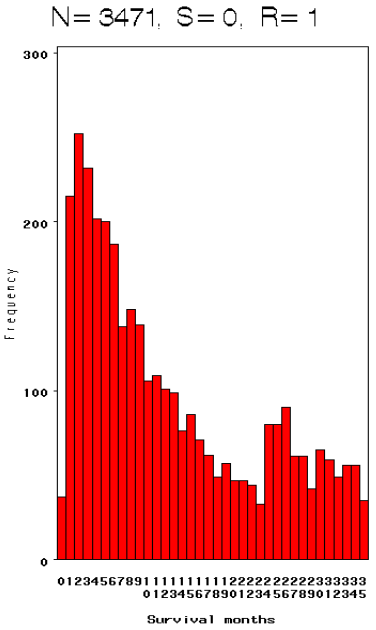
Variable	Description
survival	# of months 0…35 (All cases diagnosed in 2004 and cutoff is in 2006. If survive at the end of 3 years, survival=35)
surgery	Surgery is performed or not
radiation	Radiation is performed or not
laterality	Which side the tumor originated
size	The diameter of the primary tumor recorded in millimeters
marital	marital status
race	patient race
sex	patient gender
age	patient age
grade	Grading and differentiation codes
histg	Grouped histological type
diag_confirm	The method used to confirm the presence of cancer (histology or cytology)
node	Exact number of regional lymph nodes containing metastases
exten	Contiguous growth of the primary tumor
stage	Describes the extent of the cancer
site	the site in which the primary tumor originated
mets	the distant site(s) of metastatic involvement
malignant	First malignant primary indicator
number_pri	the actual number of primaries
vitalr	Whether patient dies from the cancer
lymph	the regional lymph nodes involved with cancer

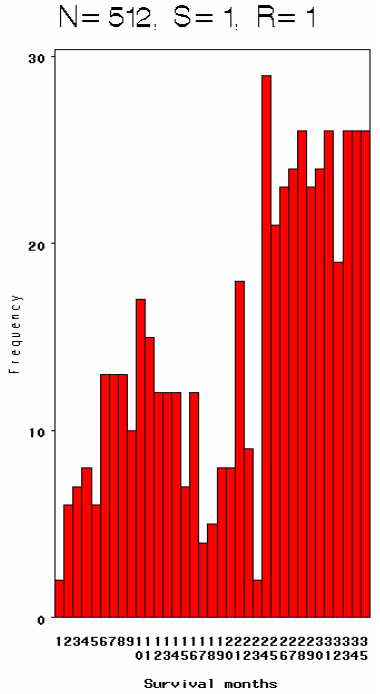
Survival time distribution







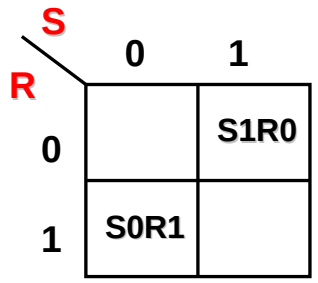


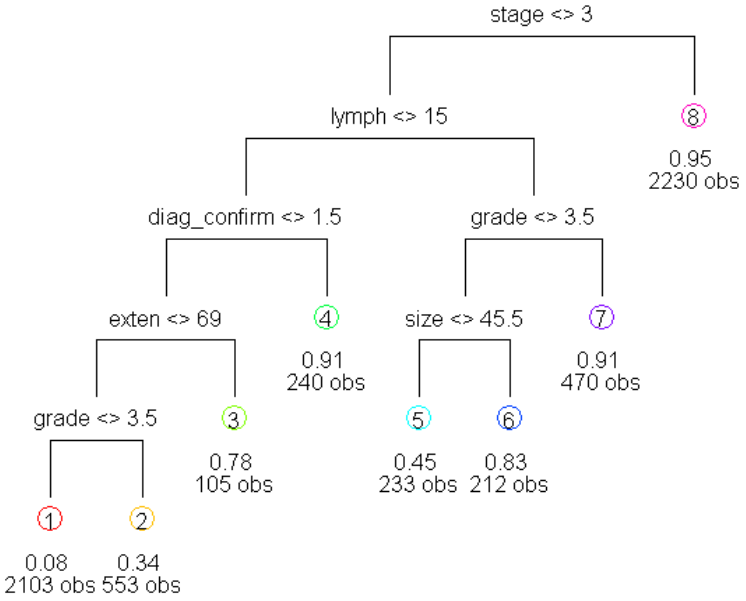


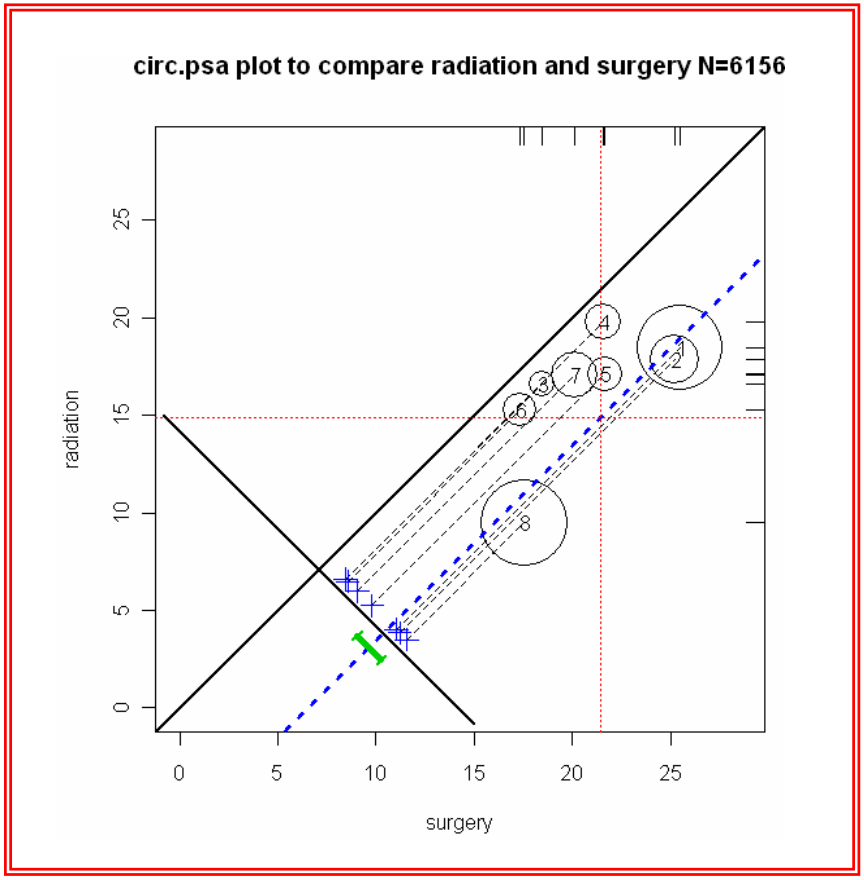
These histograms suggest several questions about survival for certain groups that would be worthy of deeper analysis, but they require detailed knowledge of archive and so are not considered here.

Classification tree stratification

Surgery treatment only vs. radiation treatment only (R=1, S=0) vs. (R=0, S=1)

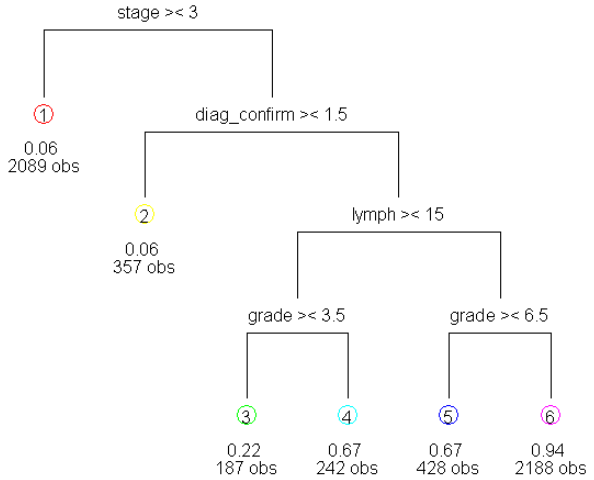


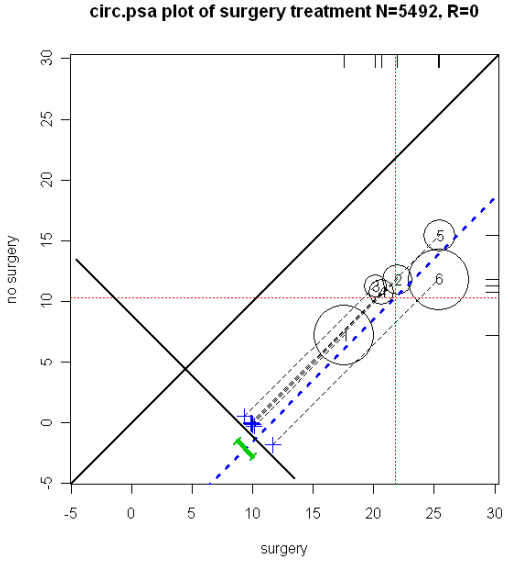




Classification tree stratification (surgery effects)

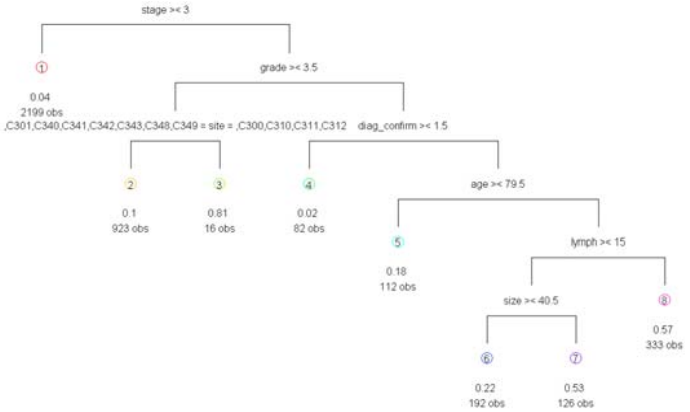
Surgery treatment only (R=0, S=0) vs. (R=0, S=1)

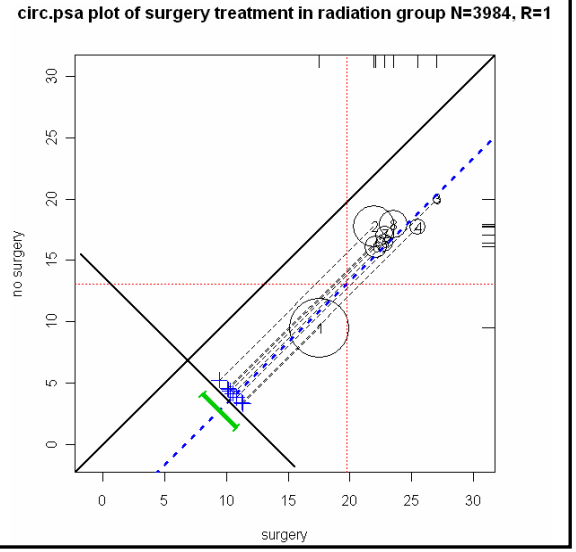




S / R	0	1
0	S0R0	S1R0
1	S0R1	S1R1

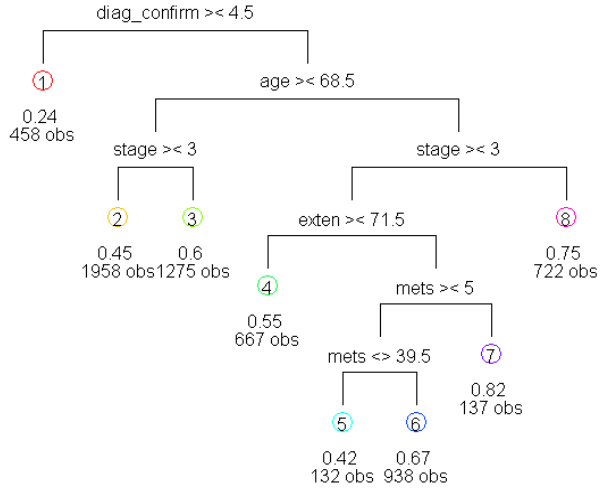
Surgery treatment in radiation group (R=1, S=0) vs. (R=1, S=1)

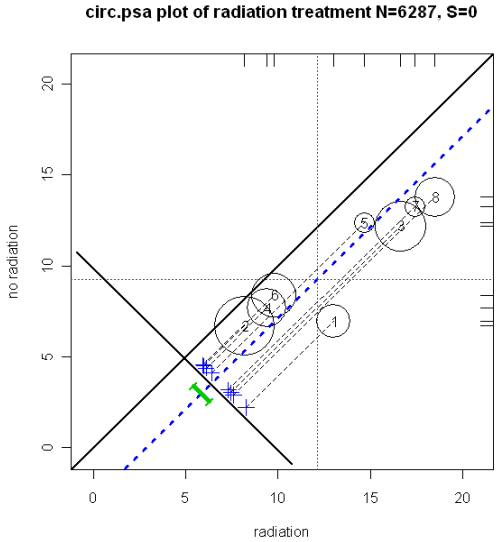


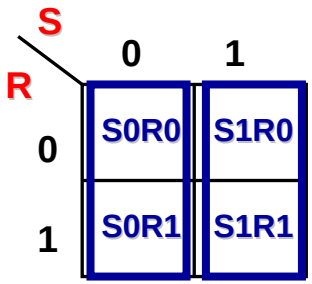


Classification tree stratification (radiation effects)

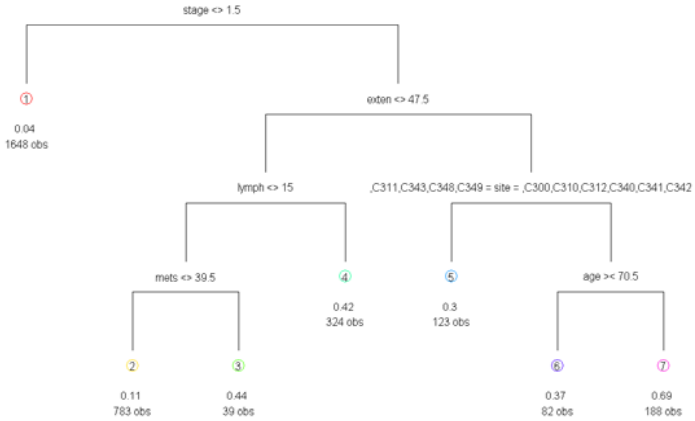
Radiation treatment only (S=0, R=1) vs. (S=0, R=0)

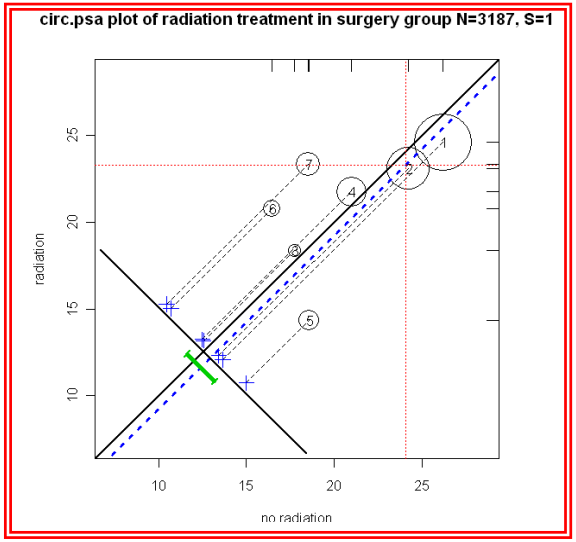






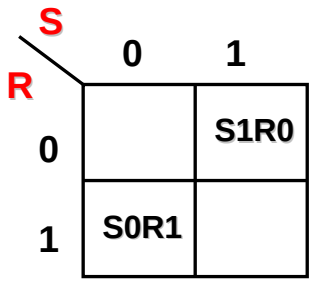
Radiation treatment in surgery group (R=0, S=1) vs. (R=1, S=1)



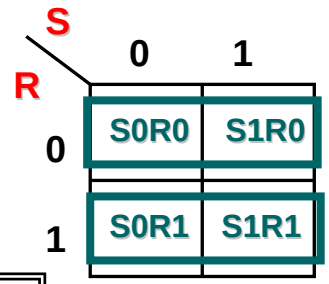


PSA stage 1: LR to estimate PS of surgery alone/radiation alone

N=6123	
Effect	**P value**
marital	0.4989
race	**0.0187**
sex	**0.0766**
age	**<.0001**
site	**<.0001**
laterality	0.2940
grade	**<.0001**
diag_confirm	**<.0001**
size	0.0971
exten	**<.0001**
lymph	**<.0001**
mets	**<.0001**
stage	**<.0001**
number_pri	**0.0009**
malignant	0.2335



PSA stage 1: LR to estimate PS of surgery



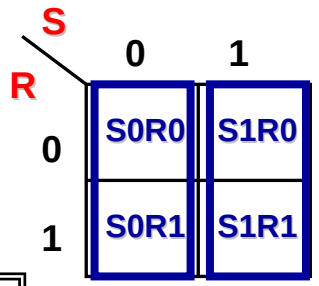
R=0, N=5484

Effect	P value
marital	<.0001
race	0.0941
sex	0.1588
age	<.0001
site	0.0003
laterality	0.0363
grade	<.0001
diag_confirm	<.0001
size	0.1293
exten	<.0001
lymph	<.0001
mets	<.0001
stage	<.0001
number_pri	0.0020
malignant	0.0904

R=1, N=3983

Effect	P value
marital	0.2718
race	0.0229
sex	0.8959
age	<.0001
site	0.1708
laterality	0.0529
grade	<.0001
diag_confirm	<.0001
size	0.0188
exten	0.0132
lymph	<.0001
mets	<.0001
stage	<.0001
number_pri	0.0005
malignant	0.0100

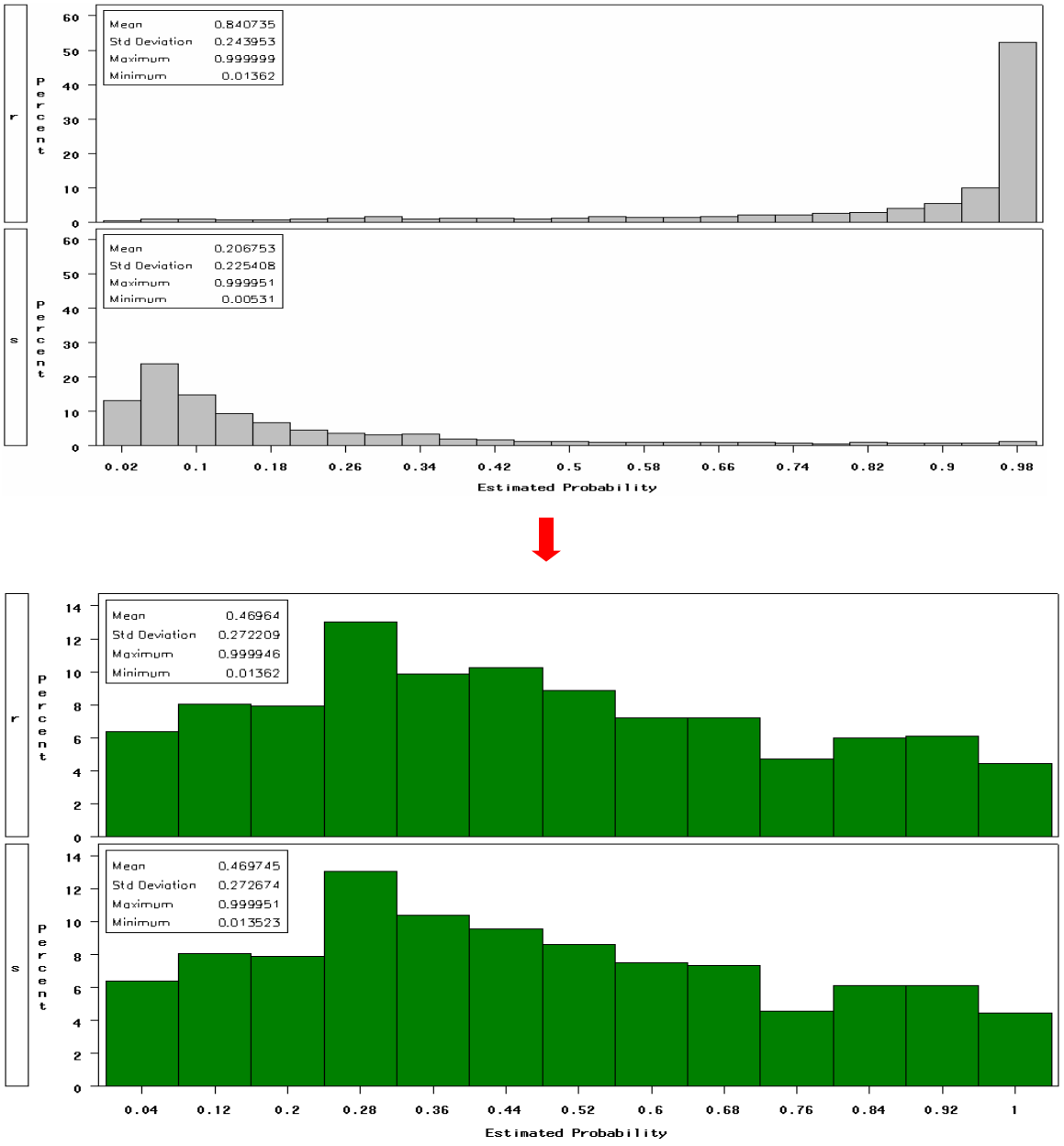
PSA stage 1: LR to estimate PS of radiation



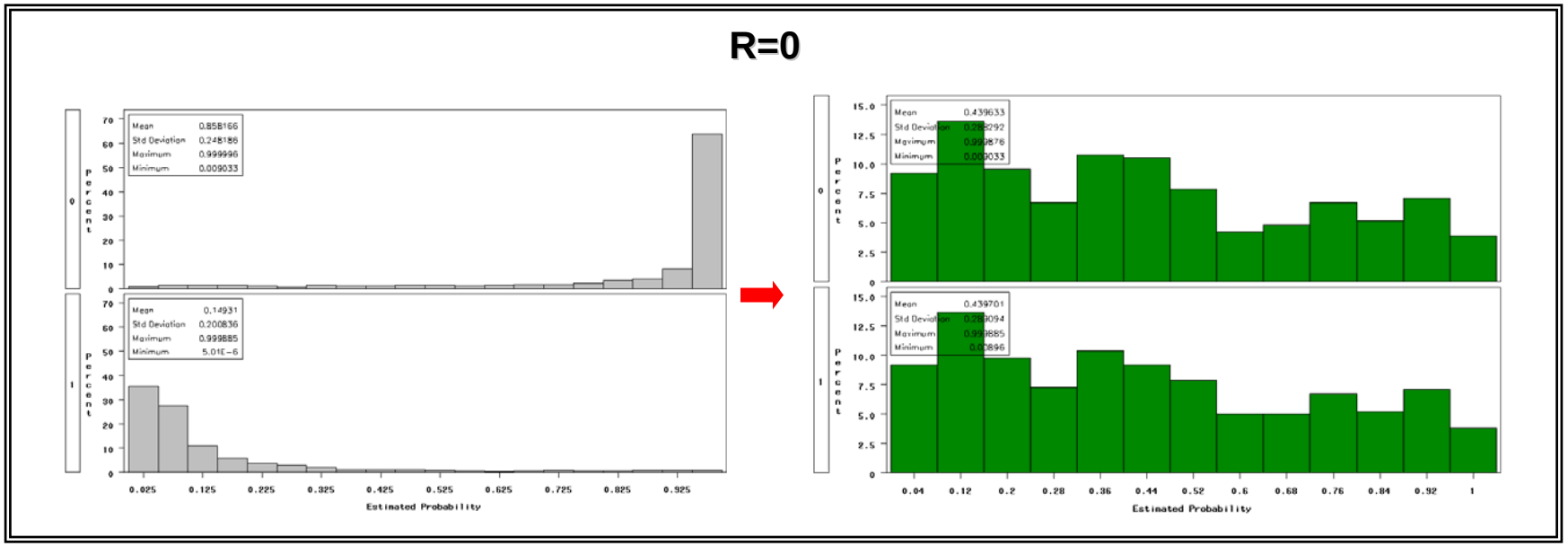
S=0, N=6287	
Effect	**P value**
marital	**0.0079**
race	0.9893
sex	0.1557
age	**<.0001**
site	**0.0006**
laterality	0.2164
grade	**0.0097**
diag_confirm	**<.0001**
size	0.1293
exten	**0.0020**
lymph	**0.0027**
mets	**0.0035**
stage	**<.0001**
number_pri	0.3926
malignant	**0.0884**

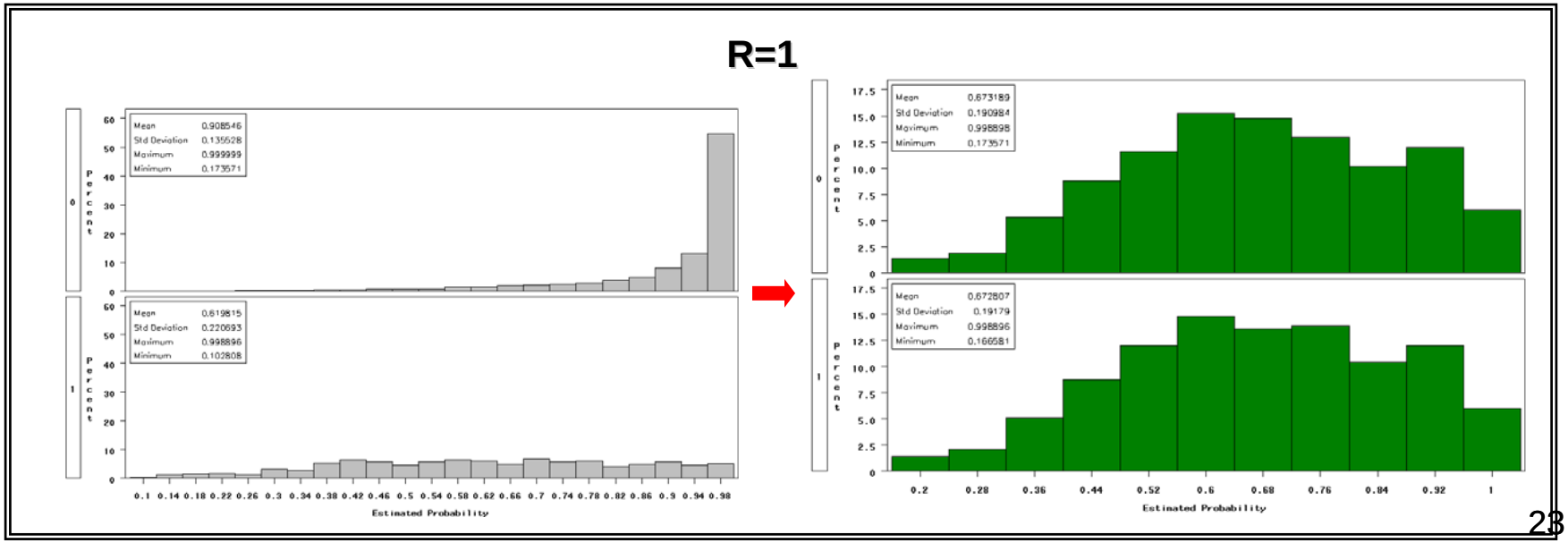
S=1, N=3187	
Effect	**P value**
marital	**0.0791**
race	0.9025
sex	0.6478
age	**<.0001**
site	**<.0001**
laterality	0.2112
grade	**<.0001**
diag_confirm	**0.0763**
size	0.7740
exten	**<.0001**
lymph	**<.0001**
mets	**0.0422**
stage	**<.0001**
number_pri	0.4891
malignant	0.113

PSA stage 2A: 1:1 NN matching of surgery only/radiation only



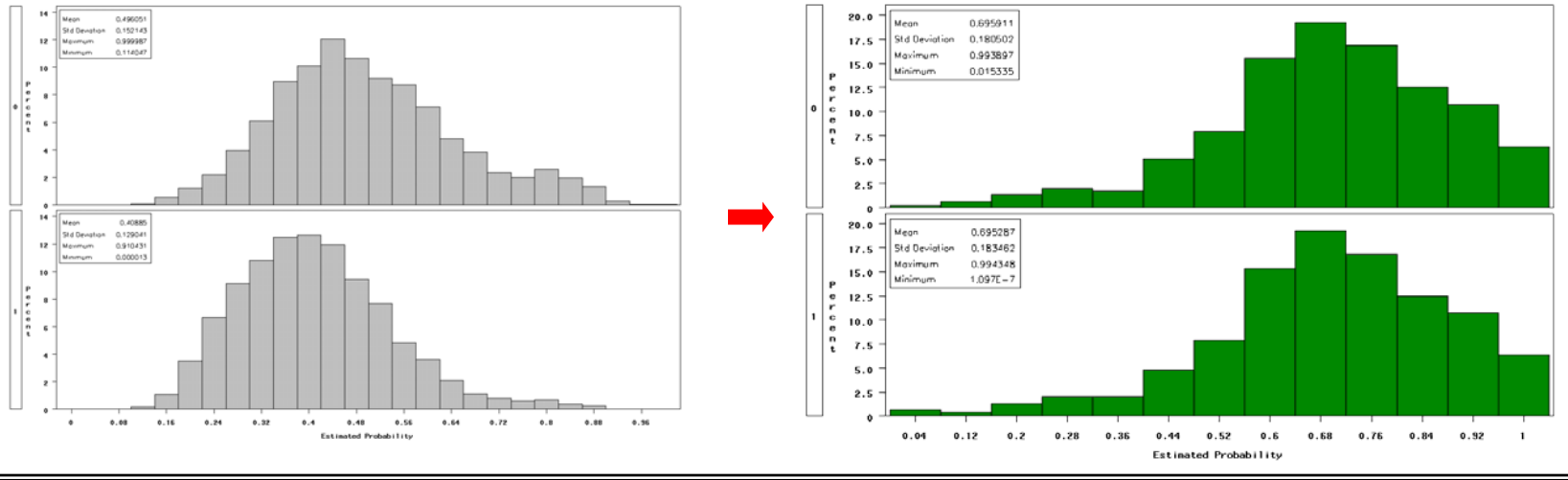
PSA stage 2A: 1:1 NN matching of surgery, w/ and w/o radiation



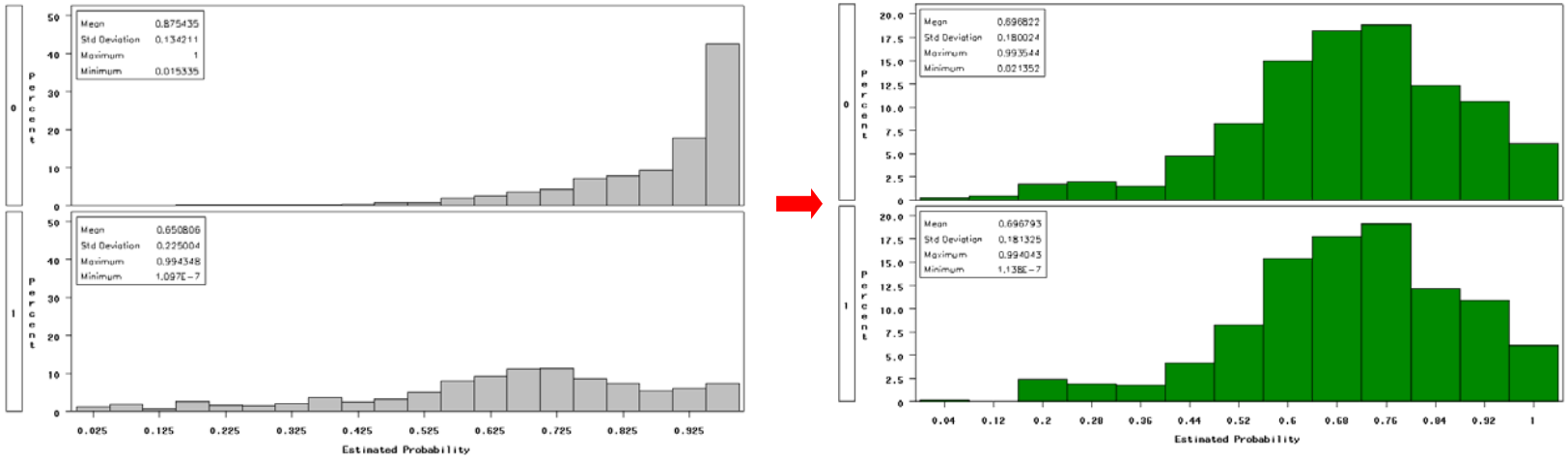


PSA stage 2A: 1:1 NN matching of radiation, w/ or w/o surgery

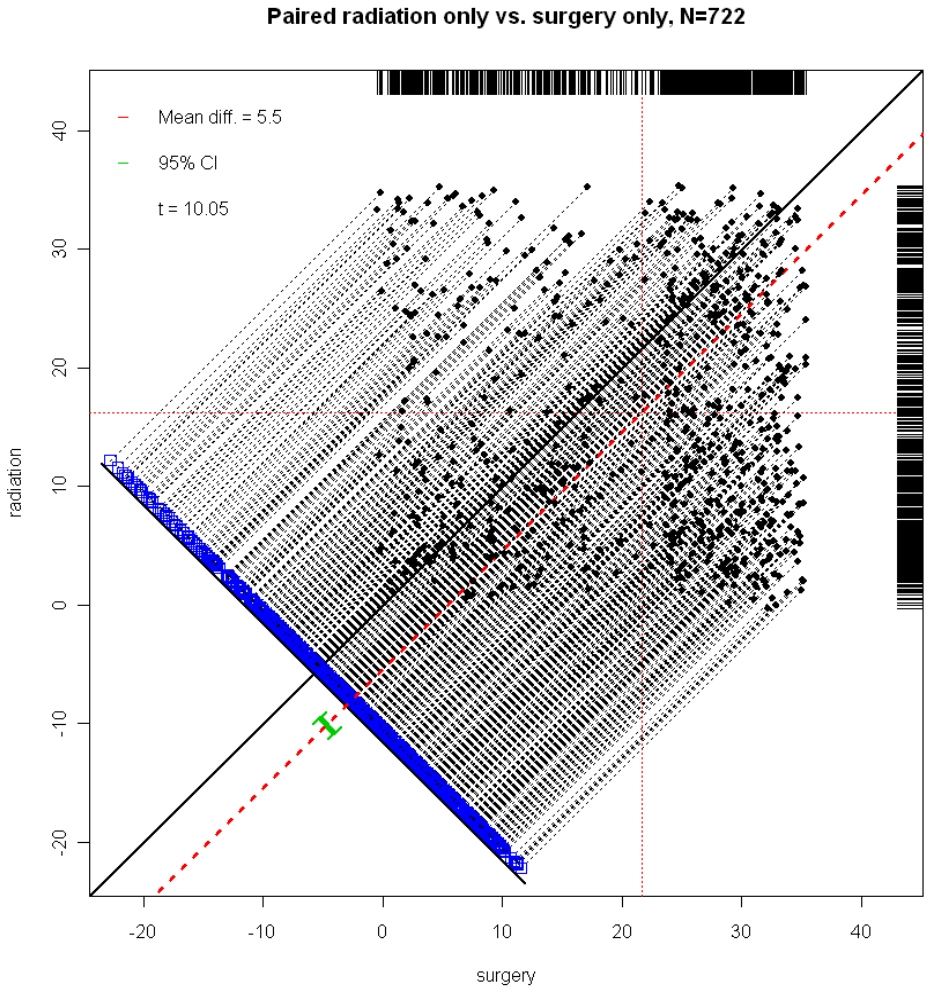
S=0



S=1

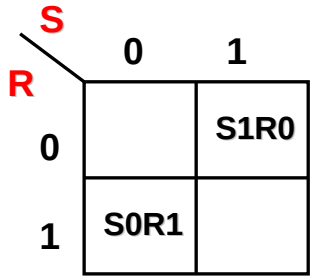


PSA stage 2B: `granova.ds` comparison of surgery only/radiation only



Summary Stats

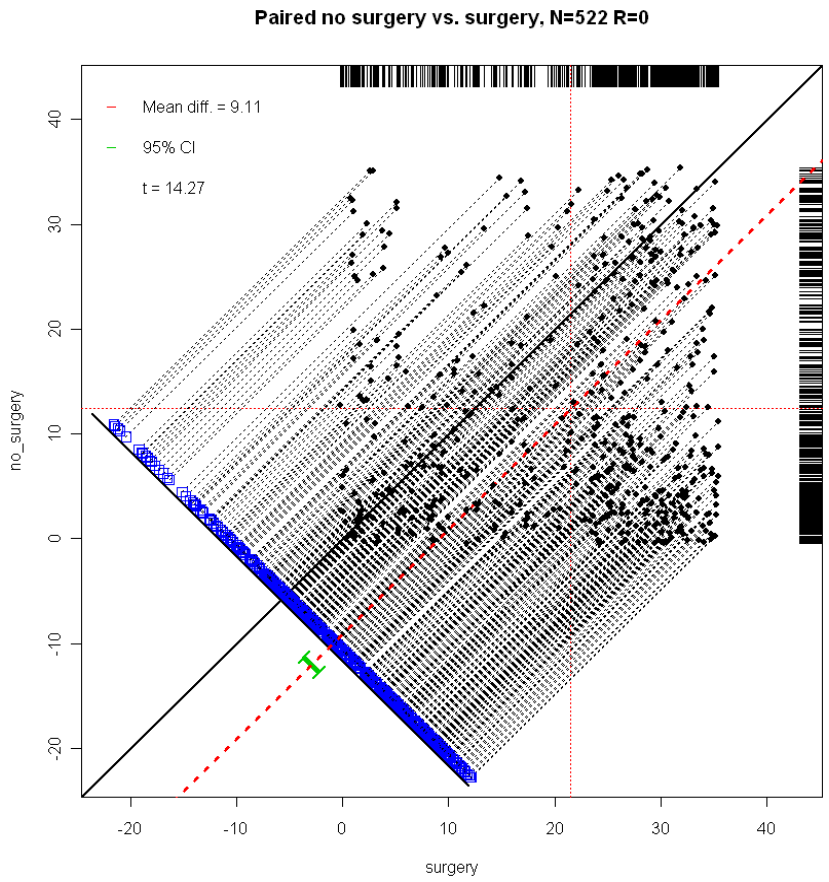
n	722
mean(x)	21.695
mean(y)	16.183
mean(D=x-y)	5.512
SD(D)	14.652
ES(D)	0.376
r(x,y)	-0.003
r(x+y,d)	0.029
LL 95%CI	**4.441**
UL 95%CI	**6.583**
t(D-bar)	10.101
df.t	721.000
pval.t	0.000

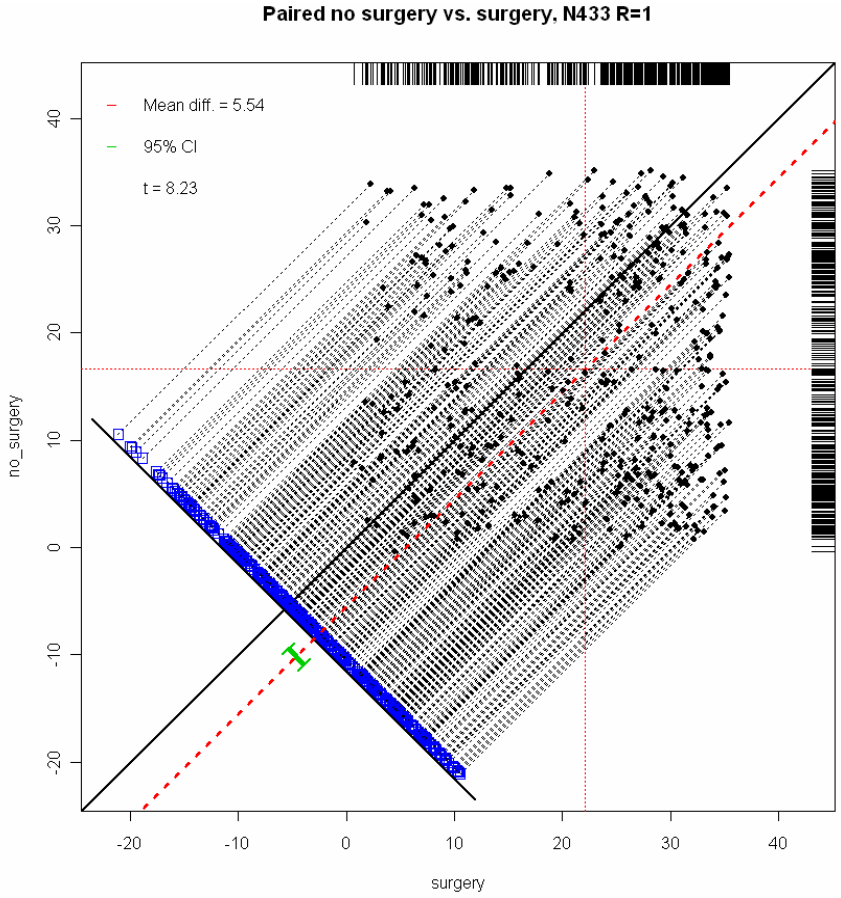


PSA stage 2B:

`granova.ds` comparisons of surgery treatment

S / R	0	1
0	S0R0	S1R0
1	S0R1	S1R1

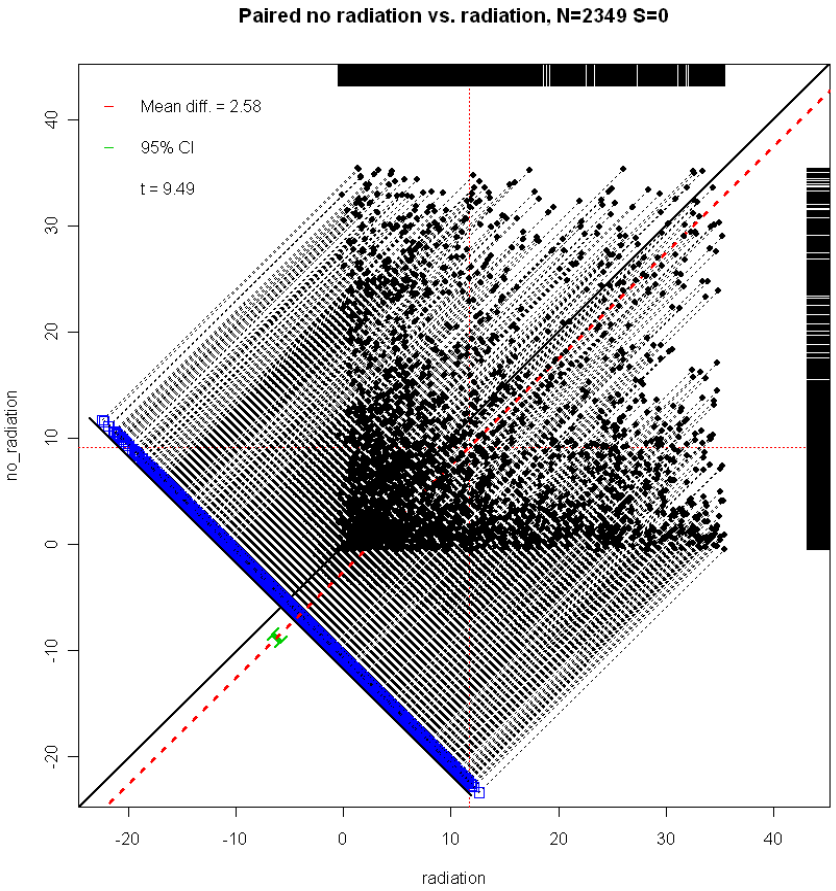


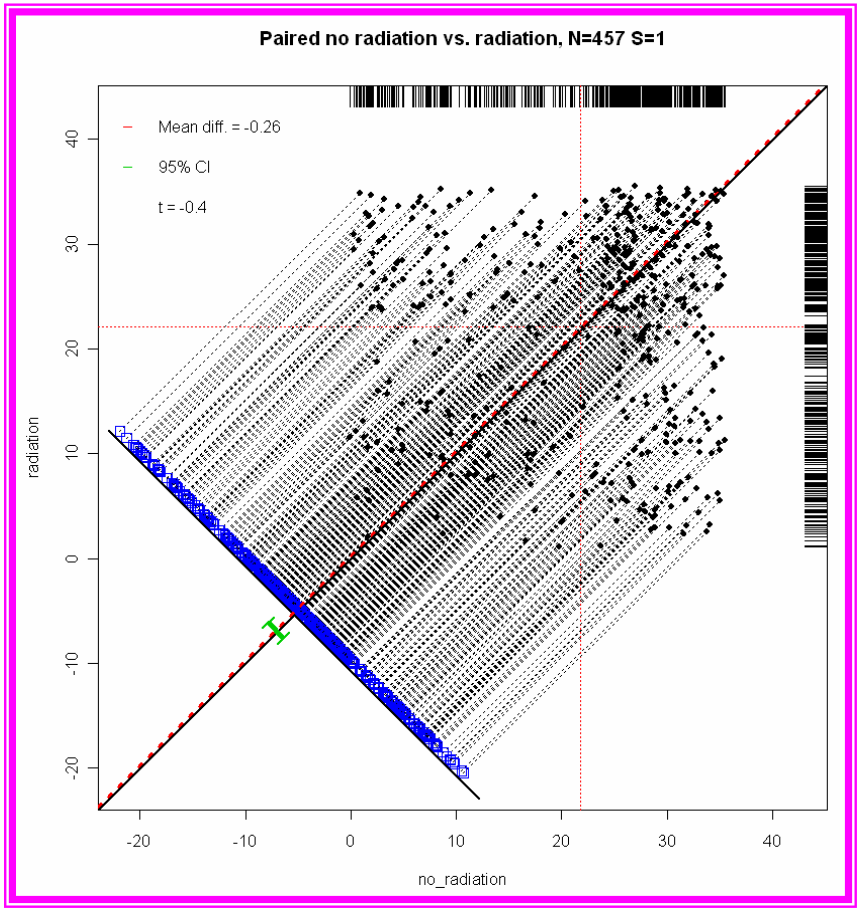


PSA stage 2B:

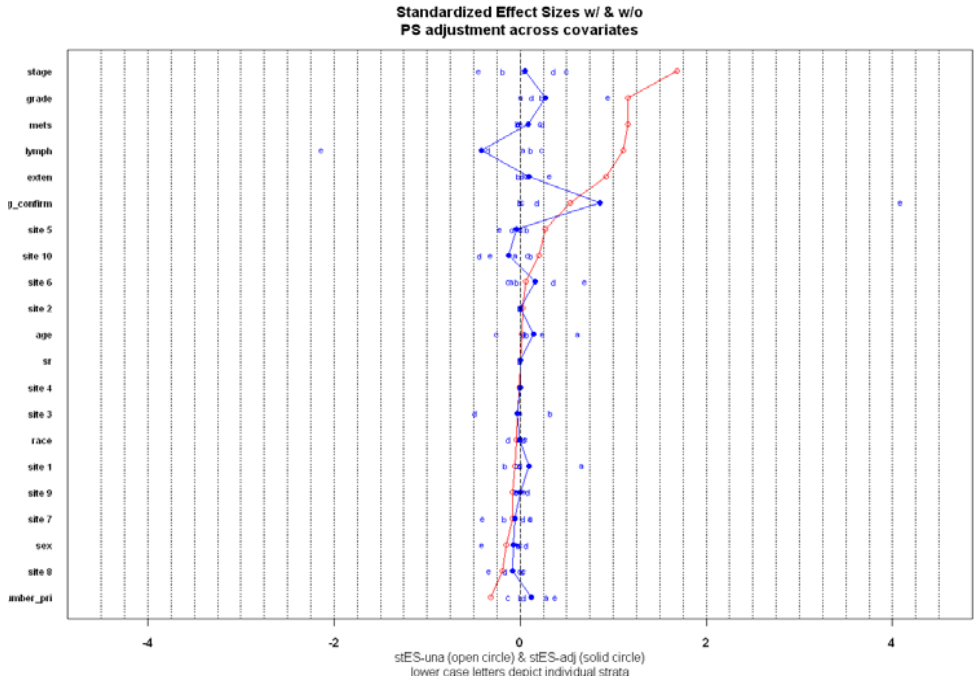
`granova.ds` comparisons of radiation treatment

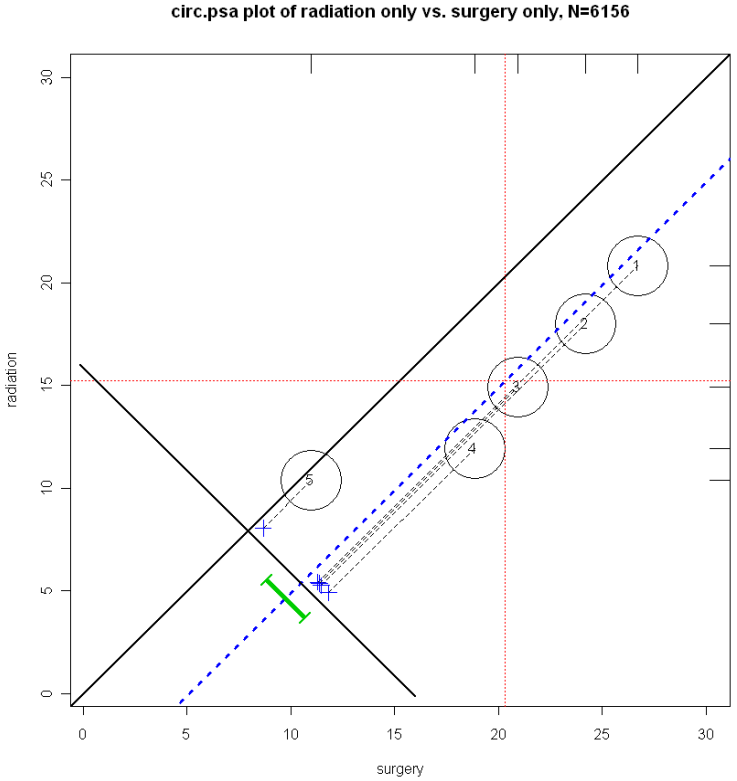
S \ R	0	1
0	S0R0	S1R0
1	S0R1	S1R1





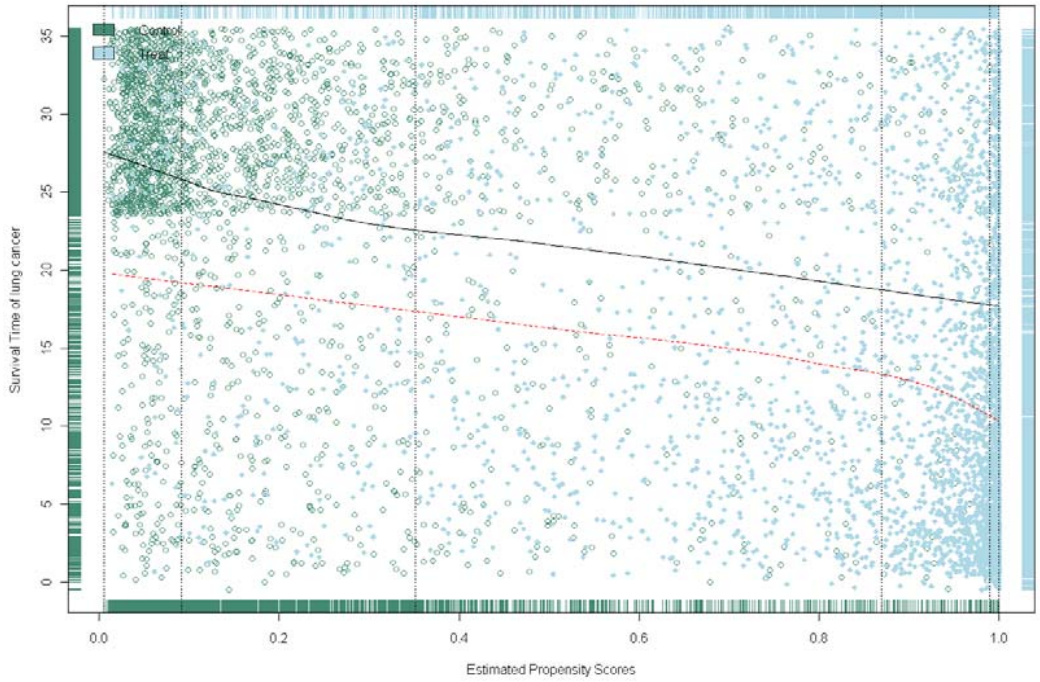
PSA: Comparison of radiation vs. surgery across range of PS

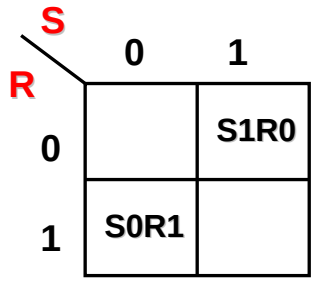




S / R	0	1
0		S1R0
1	S0R1	

PSA: Comparison of surgery vs. radiation across range of PS

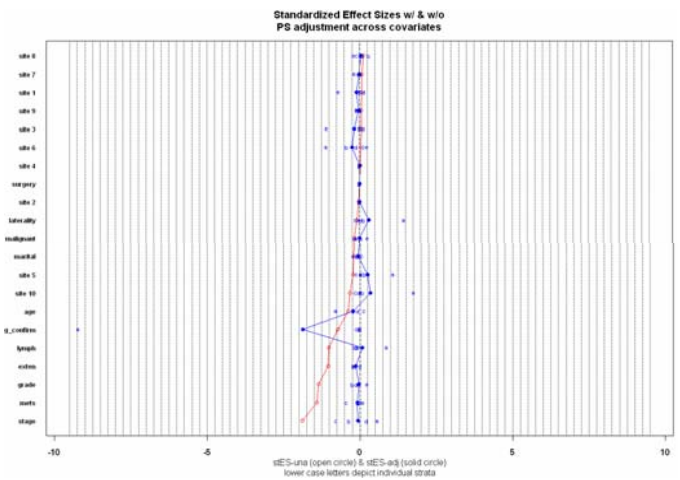


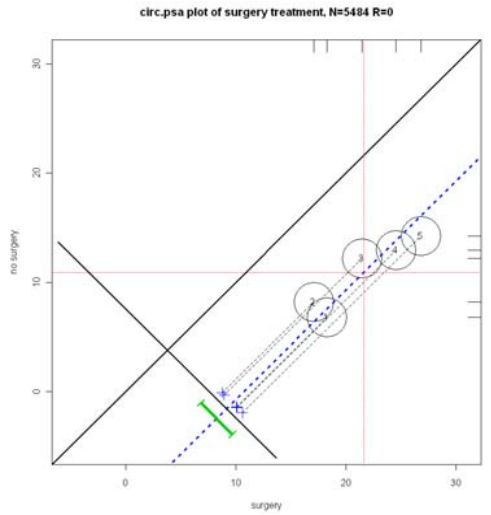



```
$dae: -6.65
$se.wtd: 0.893
$CI95: -8.43 -4.86
$summary.strata
  counts.0 counts.1 means.0 means.0 diff.means
1     1165       64    26.6    19.4      -7.22
2     1019      209    24.3    18.2      -6.16
3      408      821    21.2    15.1      -6.04
4       73     1155    18.2    11.8      -6.48
5        8     1221    17.8    10.4      -7.34
```

PSA: Comparison of surgery treatment across range of PS

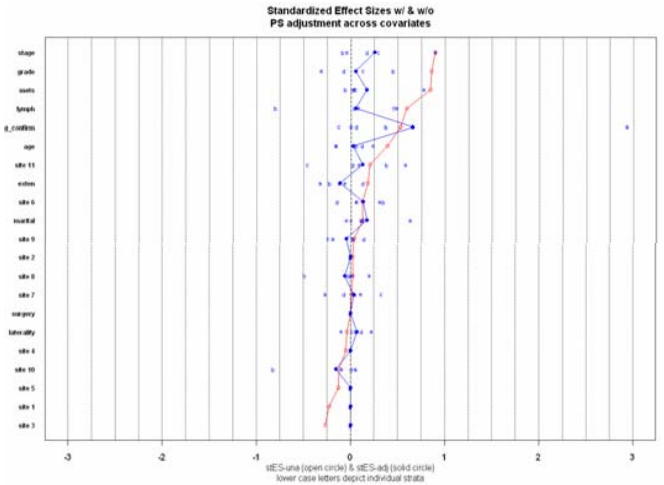
R=0

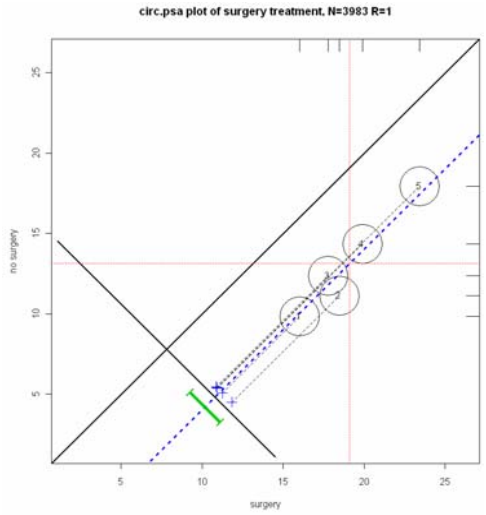




S / R	0	1
0	S0R0	S1R0
1	S0R1	S1R1

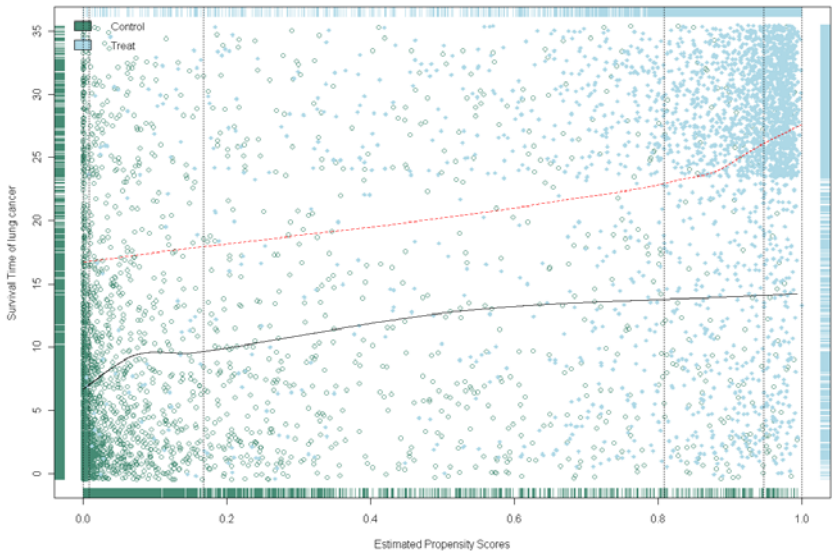
R=1



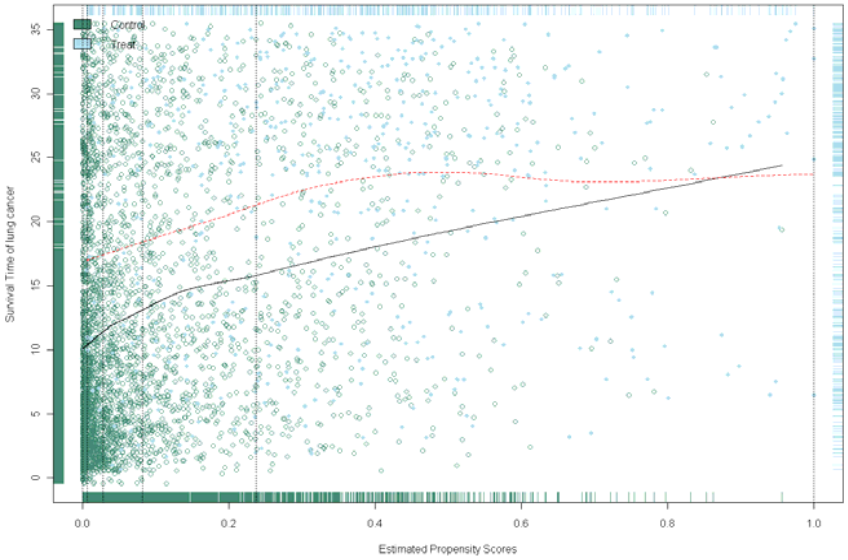


PSA: Comparison of surgery treatment across range of PS

Loess regression, N=5484 R=0

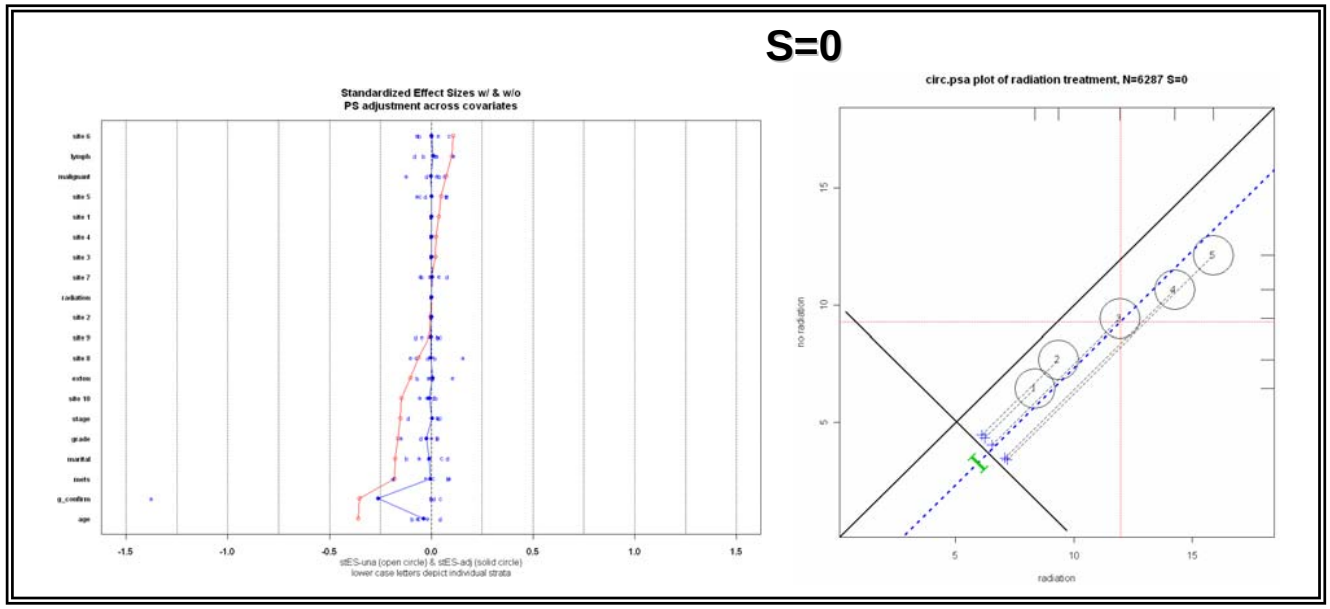


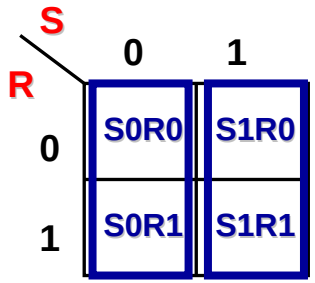
Loess regression, N=3983 R=1

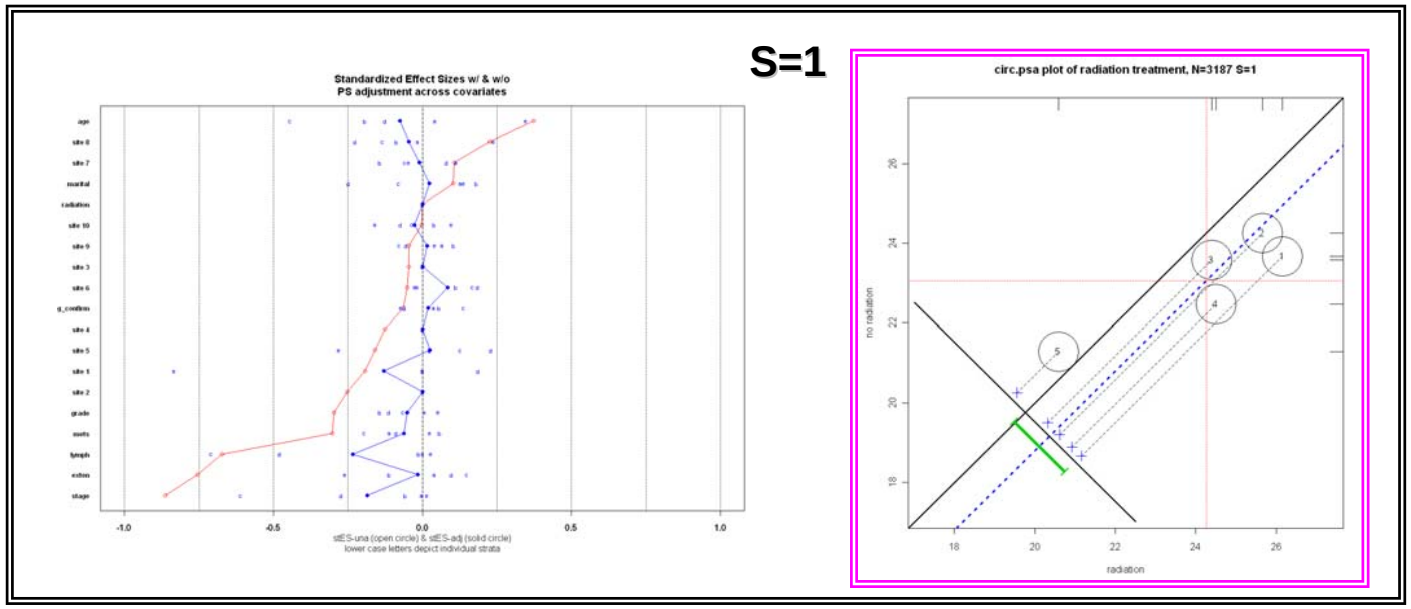


R \ S	0	1
0	S0R0	S1R0
1	S0R1	S1R1

Comparison of radiation treatment across range of PS



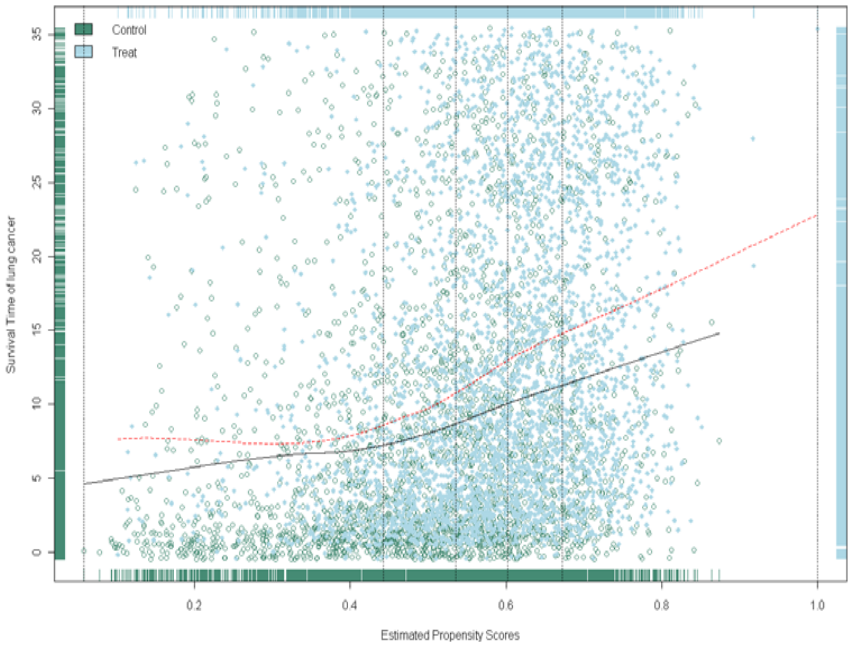




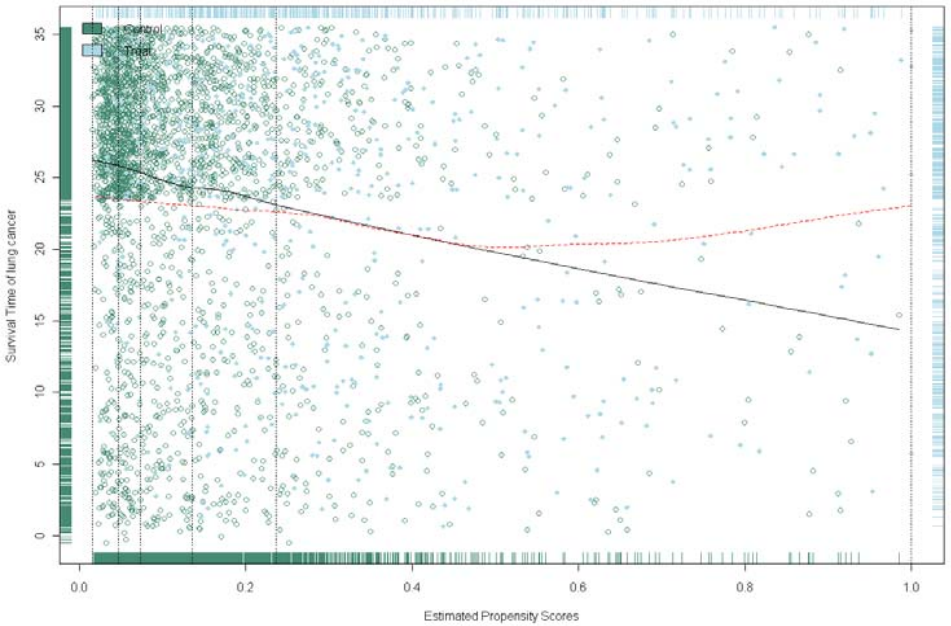
Comparison of radiation treatment across range of PS

S / R	0	1
0	S0R0	S1R0
1	S0R1	S1R1

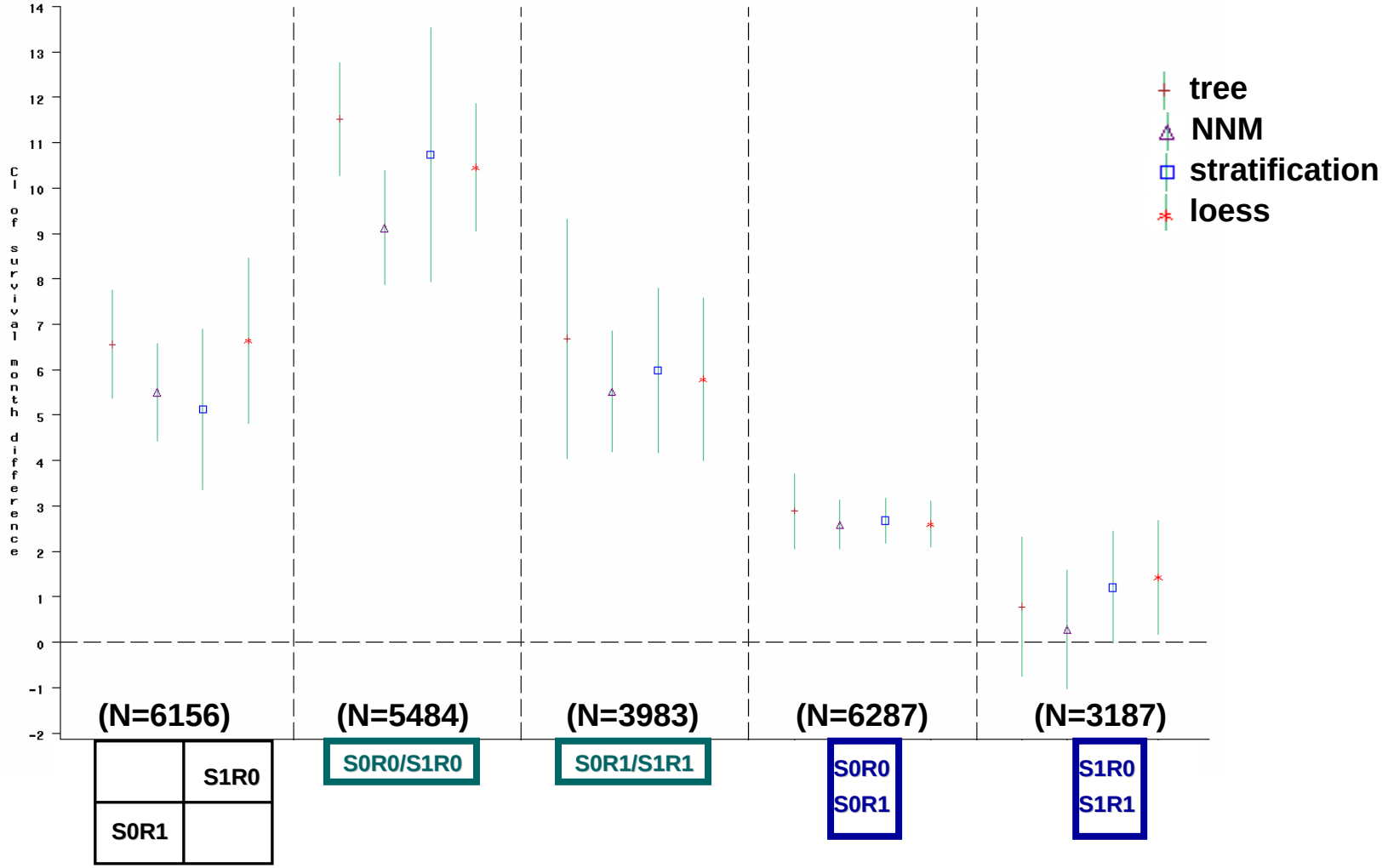
Loess regression, N=6287 S=0

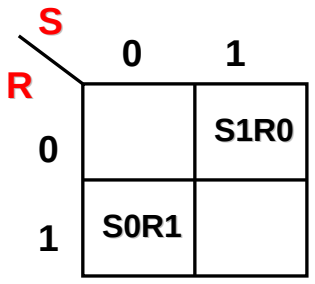


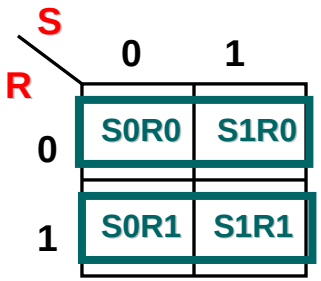
Loess regression, N=3187 S=1

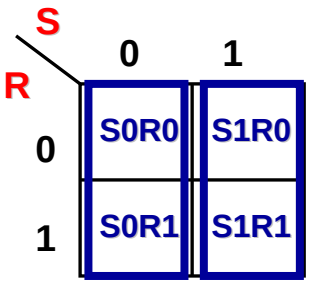


Summary of CI results for Lung Cancer Survival









Summary:

- Lung cancer patients treated with surgery alone had longer survival times than those treated with radiation alone (averaging about ½ year longer).

- For patients who received no radiation, surgery yielded large mean gains in survival times (almost 1 year longer); among patients who had radiation, the addition of surgery had smaller average effects (roughly ½ year difference).

- Radiation tended to improve survival of lung cancer patients who have not had surgery (averaging 2-3 months); but radiation tended to have little affect following surgery (adding just over 1 month on average).

Contribution of the study:

This is a relatively thorough comparison of lung cancer patients' survival times following radiation or surgery treatments after adjusting for all available covariates in SEER. It has been based on modern methods of covariate adjustment using propensity scores so that the interpretations require fewer qualifications than are needed when covariates adjustments are not used; and it has demonstrated use of different PSA methods and several graphics (using R). This study has helped quantify mean differences (expressed as months of survival) between various treatments, after adjusting for covariate differences .

Acknowledgements

Advisor: ***Robert Pruzek***

Committee: **Stratton, Howard**
Zurbenko, Igor
Yucel, Recai M.
DiRienzo, Gregory

TA advisor: Gensburg, Lenore

Nikki Malachowski
Judith L. Moran

Friends and family…

Thank you for attending!