# Predicting **7-year Prostate Cancer** Survivability (cont.)
# Survival Analysis Techniques

*IDS.506 - Robert Minh Duc Bui - mbui7@uic.edu*

# Executive Summary

- After experimenting with survival analysis, we can build a **Cox proportional-hazard model**. This model has middle-of-the-pack performance. As of current stage, model does not improve the results of our previous models.

- Final model at-a-glance:
  - **Cox Proportional-Hazard model.**
  - Text parsing + dummy variables + low variance filter + **derivative features**
  - Quick Performance Metric compared to naive null model:
    - Accuracy: **acc = 0.6583** (**+28.76%** lift)

# Train-Test **Splitting**

- Split proportion:
  - 80% training
  - 20% testing

# **Preprocessing** Steps #1

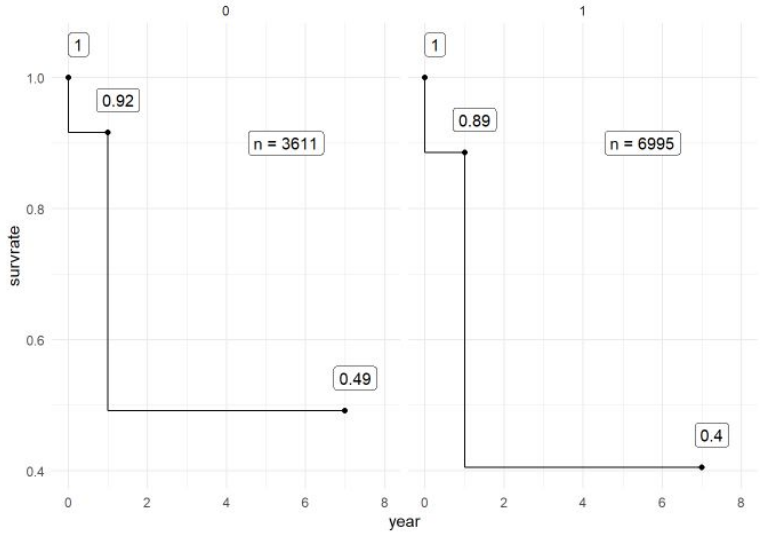
| posi | Definition | varname | NA | rate.rest | rate.NAs | NA_method | NA_method_justification |
|---|---|---|---|---|---|---|---|
| 18 | Size of primary tumor 6 months after diagnosis, in mm | tumor_6_months | 0.6540 | 0.4862 | 0.4798 | **remove col** | too much missing. |
| 21 | Level of prostate-specific antigen in blood 6 months after diagnosis, in ng/mL | psa_6_months | 0.6179 | 0.4859 | 0.4796 | **remove col** | too much missing. |
| 12 | count of family members who have been diagnosed with prostate cancer | family_history | 0.1034 | 0.4810 | 0.4905 | **remove col** | not used in modeling |
| 13 | count of brothers and fathers of the patient who have been diagnosed with prostate cancer | first_degree_history | 0.1034 | 0.4810 | 0.4905 | **remove col** | not used in modeling |
| 14 | flag indicating whether the patient has ever been diagnosed with any cancer previously | previous_cancer | 0.1034 | 0.4810 | 0.4905 | force to **0** | predominant subgroup + maybe nothing to record? = 0 |
| 15 | flag indicating whether the patient describes himself as a smoker | smoker | 0.1034 | 0.4810 | 0.4905 | force to **0** | predominant subgroup + maybe nothing to record? = 0 |
| 23 | How many times the patient reports drinking tea per week | tea | 0.1034 | 0.4810 | 0.4905 | **remove col** | not used in modeling |
| 20 | Level of prostate-specific antigen in blood at time of diagnosis, in ng/mL | psa_diagnosis | 0.0909 | 0.4816 | 0.4856 | **remove** | cannot make meaningful prediction without diagnosis info |
| 10 | Height of patient at time of diagnosis | height | 0.0894 | 0.4831 | 0.4704 | **remove** | no default value, cannot impute. |
| 11 | Weight of patient at time of diagnosis | weight | 0.0856 | 0.4846 | 0.4539 | **remove** | no default value, cannot impute. |
| 22 | Level of prostate-specific antigen in blood 1 year after diagnosis, in ng/mL | psa_1_year | 0.0675 | 0.4820 | 0.4823 | \=6/**diag**, fallback 0 | impute as no change from diagnosis |
| 8 | Age of patient at time of diagnosis | age | 0.0478 | 0.4823 | 0.4765 | **remove** | no default value, cannot impute. |
| 19 | Size of primary tumor 1 year after diagnosis, in mm | tumor_1_year | 0.0389 | 0.4818 | 0.4860 | \=6/**diag**, fallback 0 | impute as no change from diagnosis |
| 24 | A list of codes indicating the presence of various symptoms. Meaning has been removed. | symptoms | 0.0270 | 0.4800 | 0.5550 | parsed as **empty** | nothing to record = 0 |
| 3 | A measurement of how abnormal the cancer cells look compared to normal cells | gleason_score | 0.0207 | 0.4813 | 0.5140 | **remove** | small proportion + cannot predict w/o diagnosis |
| 17 | Size of primary tumor at time of diagnosis, in mm | tumor_diagnosis | 0.0197 | 0.4823 | 0.4669 | **remove** | small proportion + cannot predict w/o diagnosis |
| 9 | Race of patient | race | 0.0112 | 0.4821 | 0.4740 | **remove col** | not used in modeling |
| 29 | brachytherapy used | brch_thrpy | 0.0000 | 0.4820 | 0.0000 | | |
| 27 | chemotherapy used | chm_thrpy | 0.0000 | 0.4820 | 0.0000 | | |
| 28 | crypotherapy used | cry_thrpy | 0.0000 | 0.4820 | 0.0000 | | |
| 2 | the month and year of diagnosis | diagnosis_date | 0.0000 | 0.4820 | 0.0000 | **remove col** | no observable relationship + "00" dates + irregular data shape |
| 26 | hormone therapy used | h_thrpy | 0.0000 | 0.4820 | 0.0000 | | |
| 1 | An identifier used for scoring dataset | id | 0.0000 | 0.4820 | 0.0000 | **remove col** | not used in modeling |
| 6 | Describes whether or not the cancer has spread to distant parts of the body | m_score | 0.0000 | 0.4820 | 0.0000 | | |
| 31 | multiple therapies used in conjunction | multi_thrpy | 0.0000 | 0.4820 | 0.0000 | **remove col** | not used in modeling |
| 5 | Describes whether or not the cancer has spread to the lymph nodes | n_score | 0.0000 | 0.4820 | 0.0000 | | |
| 30 | prostate surgically removed | rad_rem | 0.0000 | 0.4820 | 0.0000 | | |
| 25 | external beam radiotherapy used | rd_thrpy | 0.0000 | 0.4820 | 0.0000 | | |
| 16 | What side of the prostate the cancer has been found in | side | 0.0000 | 0.4820 | 0.0000 | **remove col** | not used in modeling |
| 7 | Stage of cancer | stage | 0.0000 | 0.4820 | 0.0000 | | |
| 32 | survived 1 year from diagnosis flag | survival_1_year | 0.0000 | 0.4820 | 0.0000 | | |
| 4 | Describes local extent of prostate tumor | t_score | 0.0000 | 0.4820 | 0.0000 | | |

# **Preprocessing** Steps #2

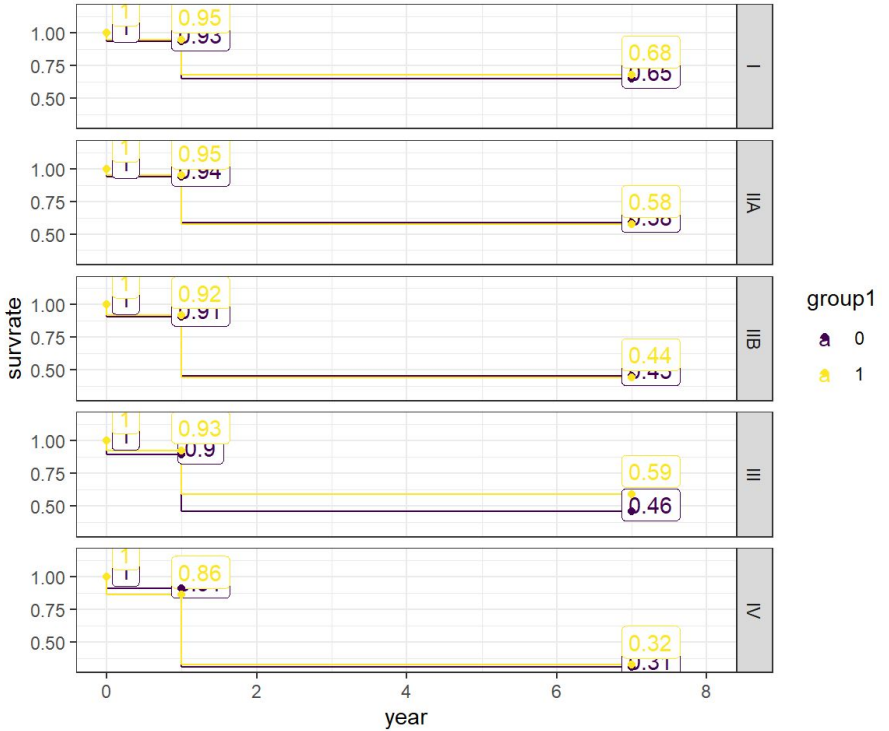
- Additional Feature Engineering carried over:
  - Calculating BMI & binning into **weight classes** (under/normal /overweight/obese)
  - Calculating **PSA** & **tumor** size change: **1-year ratio minus 1**
  - Aggregating treatments: **summing** up all **treatment** cols
  - Deriving "**Treatment Effectiveness Index**"

# **Appendix**: the Problem with **Treatments**

- Feat. **Simpson's Paradox**




`thrpy_chm`



# **Appendix**: Treatment Effectiveness Index

- Goal of this derived index:
  - Compensate for the fact that less severe (i.e. more survivable) cases are less likely to receive treatments to begin with and **skew** survivability for no-treatment group.
  - Gauge patient case **recurrence/remission (1)**, as an effect of and adjusting for…
  - The amount of **treatments (2)**.
- Components:
  - **(1)**: tumor/psa ratio.
  - **(2)**: number of treatment columns ticked.

# **Appendix**: Treatment Effectiveness Index

- Intuition behind variables:
  - **Smaller** tumor/PSA growth ratio is **better**.
  - **Smaller** number of treatments is **better**.
- Since the two elements interact in the **same direction**, the index is the **product of the two terms**.
- Viability:
  - **Pros: Convenient** abstraction of many treatment columns & clinical measurements
  - **Cons:** Abstraction renders the index **without intrinsic meaning** and thus **reduces** model **explainability**.

# **Appendix**: Treatment Effectiveness Index

- Formula Pseudocode:
  - `tumor_index = (tumor_t / tumor_dx) * count(thrpy_* == 1)`
  - `psa_index   = (psa_t   / psa_dx  ) * count(thrpy_* == 1)`

$$Index_t = \frac{x_t}{x_0} \sum_{j=1}^{ncol} T_{tj}$$

# **Survival** Analysis: Model Construction

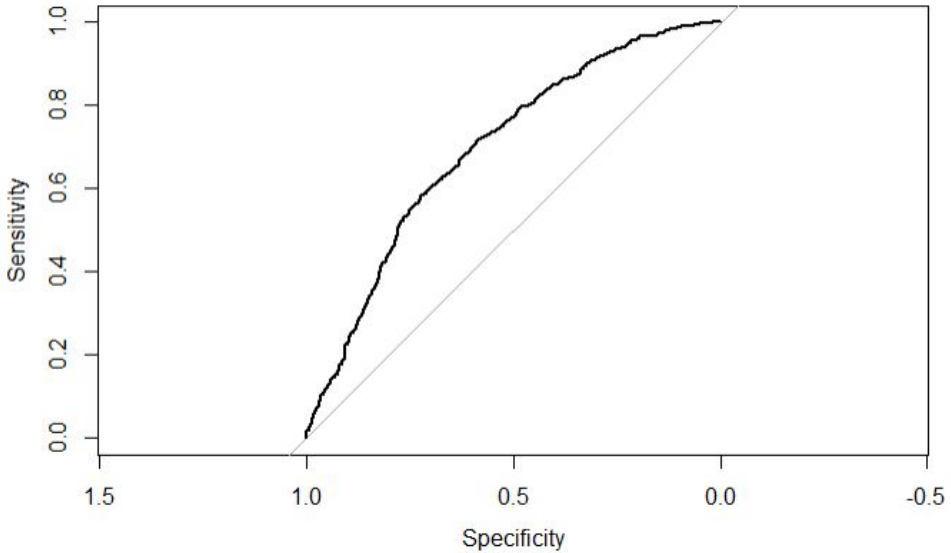
- Package: `R/survival` and `R/survminer` and `survival::coxph`.
- Predictors included in formula:

```
res.cox <- coxph(Surv(time,status) ~ gleason_score + t_score + n_score + m_score + stage + age + previous_cancer + smoker +
                   wgt_under + wgt_norml + wgt_overw +
                   tf_symptoms_o01 + tf_symptoms_o08 + tf_symptoms_o09 + tf_symptoms_o10 + tf_symptoms_o11 +
                   tf_symptoms_p01 + tf_symptoms_p02 + tf_symptoms_p03 + tf_symptoms_s04 + tf_symptoms_s07 +
                   tf_symptoms_s10 + tf_symptoms_u01 + tf_symptoms_u02 + tf_symptoms_u03 + tf_symptoms_u05 +
                   tf_symptoms_u06 + tumor_index + psa_index,
                 data = d.train, x = T)
```

# **Survival** Analysis: Model Terms

| | coef | exp(coef) | se(coef) | z | p |
|---|---|---|---|---|---|
| gleason_score | 0.0644050 | 1.0665242 | 0.0088557 | 7.273 | 3.52e-13 |
| t_scoreT1b | -0.0906434 | 0.9133433 | 0.0871124 | -1.041 | 0.298091 |
| t_scoreT1c | 0.0381373 | 1.0388738 | 0.0846448 | 0.451 | 0.652309 |
| t_scoreT2a | -0.0421482 | 0.9587277 | 0.0805307 | -0.523 | 0.600710 |
| t_scoreT2b | -0.0143635 | 0.9857391 | 0.0808944 | -0.178 | 0.859069 |
| t_scoreT2c | -0.0556717 | 0.9458496 | 0.0827279 | -0.673 | 0.500980 |
| t_scoreT3a | 0.0159729 | 1.0161012 | 0.1046946 | 0.153 | 0.878740 |
| t_scoreT3b | -0.0538012 | 0.9476205 | 0.1051441 | -0.512 | 0.608868 |
| t_scoreT3c | 0.0177987 | 1.0179580 | 0.1050491 | 0.169 | 0.865457 |
| t_scoreT4 | -0.0810280 | 0.9221679 | 0.0971568 | -0.834 | 0.404286 |
| n_scoreN1 | 0.5006640 | 1.6498164 | 0.0531185 | 9.425 | < 2e-16 |
| n_scoreNX | 0.0146739 | 1.0147821 | 0.0547251 | 0.268 | 0.788593 |
| m_scoreM1a | -0.0675651 | 0.9346669 | 0.1767294 | -0.382 | 0.702233 |
| m_scoreM1b | -0.0369236 | 0.9637497 | 0.1861721 | -0.198 | 0.842786 |
| m_scoreM1c | -0.0522534 | 0.9490884 | 0.1773354 | -0.295 | 0.768255 |
| stageIIA | 0.0506789 | 1.0519851 | 0.1098655 | 0.461 | 0.644596 |
| stageIIB | 0.2068032 | 1.2297405 | 0.1090913 | 1.896 | 0.058001 |
| stageIII | 0.0272964 | 1.0276723 | 0.1351565 | 0.202 | 0.839947 |
| stageIV | 0.2974153 | 1.3463743 | 0.1319696 | 2.254 | 0.024217 |
| age | -0.0001344 | 0.9998656 | 0.0015572 | -0.086 | 0.931219 |
| previous_cancer | 0.0606083 | 1.0624827 | 0.0586724 | 1.033 | 0.301606 |
| smoker | -0.0211739 | 0.9790487 | 0.0675214 | -0.314 | 0.753834 |
| wgt_under | -0.6412160 | 0.5266516 | 1.0031313 | -0.639 | 0.522683 |
| wgt_norml | -0.2604874 | 0.7706759 | 0.0668338 | -3.898 | 9.72e-05 |
| wgt_overw | -0.1887178 | 0.8280201 | 0.0640556 | -2.946 | 0.003218 |
| tf_symptoms_o01 | 0.0065759 | 1.0065975 | 0.0965802 | 0.068 | 0.945716 |
| tf_symptoms_o08 | 0.3653418 | 1.4410065 | 0.1046698 | 3.490 | 0.000482 |
| tf_symptoms_o09 | 0.4045386 | 1.4986109 | 0.1144693 | 3.534 | 0.000409 |
| tf_symptoms_o10 | 0.0780225 | 1.0811470 | 0.1321820 | 0.590 | 0.555012 |
| tf_symptoms_o11 | 0.0066296 | 1.0066516 | 0.0397247 | 0.167 | 0.867458 |
| tf_symptoms_p01 | 0.3301131 | 1.3911254 | 0.1708527 | 1.932 | 0.053341 |
| tf_symptoms_p02 | 0.4208928 | 1.5233210 | 0.1749528 | 2.406 | 0.016139 |
| tf_symptoms_p03 | 0.4259891 | 1.5311041 | 0.2074875 | 2.053 | 0.040065 |
| tf_symptoms_s04 | -0.0384694 | 0.9622611 | 0.0337887 | -1.139 | 0.254899 |
| tf_symptoms_s07 | -0.0341737 | 0.9664036 | 0.0296086 | -1.154 | 0.248427 |
| tf_symptoms_s10 | 0.3282958 | 1.3885997 | 0.0564022 | 5.821 | 5.86e-09 |
| tf_symptoms_u01 | 0.0173062 | 1.0174568 | 0.0296489 | 0.584 | 0.559419 |
| tf_symptoms_u02 | -0.0260264 | 0.9743093 | 0.0291096 | -0.894 | 0.371277 |
| tf_symptoms_u03 | 0.0264984 | 1.0268526 | 0.0302954 | 0.875 | 0.381756 |
| tf_symptoms_u05 | 0.2324570 | 1.2616962 | 0.0447970 | 5.189 | 2.11e-07 |
| tf_symptoms_u06 | 0.0134157 | 1.0135061 | 0.0357289 | 0.375 | 0.707299 |
| tumor_index | 0.1160608 | 1.1230642 | 0.0160708 | 7.222 | 5.13e-13 |
| psa_index | 0.0328662 | 1.0334123 | 0.0183021 | 1.796 | 0.072532 |

# **Survival** Analysis: Test Set Performance




```
m.test %>% accuracy(truth = y, estimate = y_hat)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.658
```