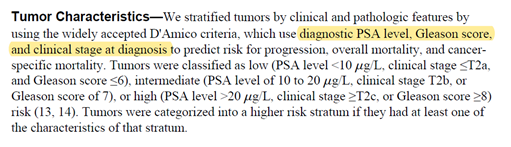
1. for PSA and stage variables - are the missing values (~10%) coming from individuals from certain sites? or are they a small number from all sites?

For PSA, 192 (/340) missing values are from Georgia. For stage, 171 (/436) missing values are from Georgia too. I created tabs to show missing in each registry. **Do we need to worry about this missing percentage?**

1. For aggressiveness, definition clarify if the criterion is (PSA or Clinical or Gleason) vs. (PSA and Clinical and Gleason)? That is, does an individual have to have all these factors or just one to be included? Also, would be good to see the correlation between the various outcomes

I used the definition from the paper Ann sent to us and I believe for high and intermediate-risk, they were using “or”. As shown below:



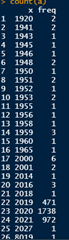
PSA and Gleason’s scores are numeric, but Stage and Aggressiveness are categorical. **Do we need to code them to 1,2, and 3 to get the correlation? Or we don’t need them to be in the correlation table?**

1. Can the comorbidity score be simplified into a “high vs. low”? has this been done in the literature? If so, maybe we can include an additional variable?

Yes, it always was categorized to “No comorbid conditions”, “1 Comorbid condition”, “2 comorbid conditions” and “3 or more comorbid conditions”. I categorized it and used categorized ones to create table 1. **Do we still need numeric comorbidity score to be in the table 1?**

1. for the data of birth through age variable: some additional cleaning is needed - some values are too large/small

The variable of “Current year” means the year of finishing the survey. I cleaned some extreme values. But there are other weird values, such as 1941, 1943, 1945 and 1946.



One thing to note is that there were some subjects with an age of 0. Since the birthday was from the registry, which is reliable (0% missing). I believe these people wrongly wrote down their birthday in the current date question. So, I used 2021 as their “filling out year”.

1. Drinking variable: I would not use the term “alcoholic”. Looks like you dichotomized to make “Heavy” drinking vs “Other” - You could also make this “Heavy”, “Modest” and “Light/None”

Made “drink everyday” and “5-6 times” to “Heavy”, “3-4 times” and “1-2 times” to “Modest” and “1-2 times” and “do not drink” to “Light or None”.

1. Table 1 PSA/: change labels to more informative (“High” “Med” and “Low” for PSA; actual Stage values for Stage, Gleason Score criteria for Gleason: you can do this by creating factors

Already done.

1. Table 1: can we have one for aggressiveness?

Added table 1 for aggressiveness.

1. Also, for the Table 1, can you include or for the final Table 1 a test of if the values differ by categories? i.e. an overall ANOVA test for differences by category and a ANOA test for trend?

I created a function to perform the chi-square tests and one-way ANOVA tests. Most of them are not significant.