Kidney Cancer Data Exploration (KL2 Aim 2)

Alex F. Bokov

08/09/2018

## Overview

This is not (yet) a manuscript. We are still at the data cleaning/alignment stage and it is far too early to draw conclusions. Rather, this is a regularly updated progress report that I am sharing with you to keep you in the loop on my work and/or because you are also working on NAACCR, i2b2, Epic, or Sunrise and this might be useful to you or you might wish to offer advice.

Currently only de-identified data is being used, under Dr. Michalek’s exempt project IRB number HSC20170563N. Dr. Michalek has given me a set of guidelines under which we can share the de-identified data with UTHSCSA collaborators. If you would like a copy of the data, please email me and indicate which of the following versions you would like, confirm that you are up to date on your HIPAA training and will comply with it, and that if somebody else wants a copy of the data you will forward them to this team (my mentors and myself) rather than directly giving them a copy:

* Raw: the data as it literally exists when I input it into my scripts.
* Lightly curated: the main dataset as it is after my scripts are done processing it
* Moderately curated: the dataset pared down to just the columns and rows currently being used for analysis

Dr. Murphy, if you are interested in a copy of the data, I’ll talk to IRB about the best way to do that. It’s probably time we start talking about what approvals in general will be necessary for the full project. In case you are wondering, I am doing Aim 2 ahead of Aim 1 because it will help me identify the need for any additional recurring data-transformation rules that I can then incorporate into DataFinisher all at once. I will switch to Aim 1, the i2b2 plugin, once I hit a natural pausing-point on Aim 2.

### Questions for mentors and other domain experts:

* What are the main problems with the NAACCR stage and grade information that I will need to clean up?
* What is the typical time that elapses between diagnosis and surgery?
* Is it possible for surgery to happen on the same day as the diagnosis? How common is that?
* What would be the threshold on the lag to surgery until we must conclude that there is an error in that record? E.g. is four years too long?
* What fraction of KC patients undergo surgery?
* How would one distinguish the chart of a patient who is was diagnosed for the first time with a kidney tumor from that of a patient experiencing a relapse…
  + …in Epic?
  + …in Sunrise?
* Where in the chart would one positively establish the date of the patient’s first nephrectomy…
  + …in Epic?
  + …in Sunrise?
* Is there some additional data source that the UTHealth NAACCR registrar consults?

### Questions to answer empirically:

* Question: Are NAACCR-EMR linkages now correct?
  + Motivation: For Sub-Aim 2a, I will be looking for possible mediators of disparity, many of which will come from data outside NAACCR, linked via i2b2. For this reason I need to establish that NAACCR patients are linked to the correct records in the rest of i2b2.
  + Answer: [Yes](#consistency-checks)
* Question: Which elements in the raw data to use as our highest priority analytic variables (dates of diagnosis, surgery, recurrence, and death as well as ethnicity)
  + Motivation: For the main Aim 2, I am trying to determine whether there is an outcomes disparity associated with Hispanic ethnicity. There needs to be a way to quickly validate it against data independent of UTHealth. With the local data we cannot conclude anything at all about prevalence or incidence in the general population because we lack a comparator group for that and this is not part of my project. Instead, I am testing the existence in outcome disparities among patients already diagnosed with kidney cancer and those who have undergone surgery for kidney cancer. Here local results *can* be compared to de-identified NAACCR regional and national data. Therefore I need to establish the minimum set of NAACCR-only variables needed to replicate this analysis. If possible it would also be good to find corresponding EMR data elements so that incomplete NAACCR records can be back-filled with EMR data from i2b2.
  + Answer: So far looking like:
    1. [Diagnosis](#initial-diagnosis) = n\_ddiag (NAACCR date of diagnosis, no others)
    2. [Surgery](#surgery-conclusion) = n\_dsurg (NAACCR date of surgery, no others)
    3. Recurrence and prior occurrence: *in progress*
* Q: Which records to exclude due to likely errors in the source data? I.e. surgery precedes diagnosis, recurrence precedes surgery (for some analysis) death precedes diagnosis or surgery

### Outline

* [Consistency-Checks](#consistency-checks)
* [Cohort Characterization](#cohort-characterization)
* [Which EMR and NAACCR variables are reliable event indicators?](#which-emr-and-naaccr-variables-are-reliable-event-indicators)
* [Descriptive Plots (Preliminary)](#descriptive-plots-preliminary)
* Appendices
  1. [Example of stage/grade data](#appendix-i-example-of-stagegrade-data)
  2. [Next steps](#appendix-ii-next-steps)
  3. [Supplementary tables](#appendix-iii-supplementary-tables)
  4. [Audit trail](#appendix-iv-audit-trail)

## Consistency-Checks

## Cohort Characterization

Summary of all the variables in the combined i2b2/NAACCR set. Tumor\_Free means no recurrence, Tumor means recurrence, and Unknown means unknown. Not in NAACCR means there is an EMR diagnosis of kidney cancer but no record for that patient in NAACCR.

Note: the below variables have not yet all been updated with the conclusions about what will be used as the final date of diagnosis and date of surgery nor have the criteria for dropping invalid records been finalized, so these numbers are subject to change.

## Which EMR and NAACCR variables are reliable event indicators?

We need the following variables for starters. For most or all of these events, both data sources have multiple variables some or all of which could be indicators. We will likely need to merge groups of synonymous variables into one analytic variable each for NAACCR and for the EMR. This is to mitigate for missing data. We can then do the same analysis on the same patients using NAACCR-only variables and EMR-only variables confirm that they agree. If we can either show agreement or find and resolve the causes of discrepancy this will permit other sites, which have not necessarily merged NAACCR and EMR data, to replicate our analysis. It will also allow us to compare our results to national or Texas NAACCR data-sets which of course are not linked to EMR data.

However, there will be even fewer missing observations and a richer choice of predictor variables if we work on a combined NAACCR and EMR dataset. Therefore for each of the below we will also need a third analytic variable combining NAACCR and EMR information.

Our standard way of indexing time in this study is age\_at\_visit\_days. The main table dat1 will be collapsed into one row per patient, and the value for each of the above columns will be replaced with the age in days when that event was recorded (if any, otherwise NA). This table will be called xdat1.

#### Initial diagnosis

The c\_kcdiag group of columns in dct0.

* NAACCR: n\_ddiag. The other two– the date accompanying the SEER site and the date accompanying the NAACCR primary site– are not date fields in NAACCR, so whatever start\_date they are getting assigned must be from our ETL process, not NAACCR and that is the code I will need to review. There is data element 443, [Date Conclusive DX](http://datadictionary.naaccr.org/default.aspx?c=10" \l "443) but that is never recorded in our NAACCR. All other NAACCR data elements containing the word ‘date’ seem to be retired or related to later events, not initial diagnosis. Whatever the case, there are only 19 patients with a missing date of diagnosis but non-missing dates for the SEER site variable, so within the range of reasonable error at the NAACCR end. **Therefore** n\_ddiag **(date of initial diagnosis) is the only NAACCR variable on which we can rely for onset.**
* EMR: First occurence of any ICD9/10 code for kidney cancer. Naively, I had hoped that the first ICD9/10 code for kidney cancer would closely track the date for the n\_ddiag. Unfortunately, as can be seen from the below table, for the 616 patients who have non-missing n\_ddiag values, the first ICD9 and first ICD10 code most often occurs after initial diagnosis, sometimes before the date of diagnosis, and coinciding with the date of diagnosis rarest of all. By inspection I found that several of the ICD9/10 first observed dates lead or trail the n\_ddiag by multiple years! **Therefore, one or both of the following steps are needed before EMR data can be relied on at all for establishing date of onset** :
  + Meeting with CTRC NAACCR registrar to see how she obtains her dates of onset
  + Chart review of a sample of NAACCR patients to understand what information visible in Epic sets them apart from non kidney cancer patients.

#### Surgery

The c\_nephx group of columns

* NAACCR: in addition to 1200 RX Date--Surgery (in this script shortened to n\_dsurg) and 3180 RX Date--Surgical Disch the following possibly relevant fields are available in our local NAACCR and will be evaluated after the next data-pull:
  + 1260 Date of Initial RX--SEER
  + 1270 Date of 1st Crs RX--CoC
  + 3170 RX Date--Most Defin Surg
* EMR: First occurrence of any ICD9/10 code for acquired absence of kidney; or first occurence of surgical history of nephrectomy

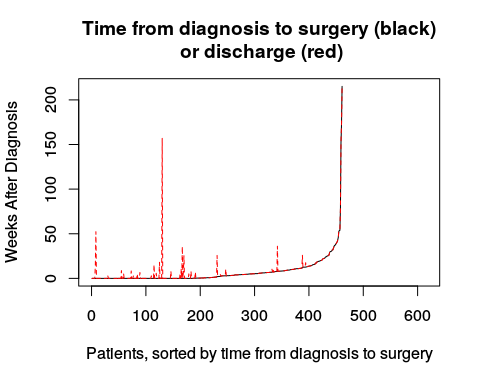
As can be seen in the table below, the variables v088\_hx\_nphrctm, v102\_acqrd\_absnc, v109\_acqrd\_absnc *sometimes* precede n\_ddiag by many weeks. However, they *usually* follow n\_ddiag by more weeks than the two NAACCR variables n\_dsdisc and n\_dsurg. Those two NAACCR variables never occur before n\_ddiag and usually occur within 2-8 weeks.

As can be seen from the NA's column, the inactive ICD9/10 V/Z codes for acquired absence of kidney are disqualified because they are very rare in addition to being even more divergent from the n\_ddiag than the non-inactive codes.

It’s understandable if the Epic EMR lags behind NAACCR (because as an outpatient system, it’s probably recording just the visits after the original surgery, and perhaps we are not yet importing the actual surgery events from Sunrise EMR). But for the V or Z or surgical history codes that precede n\_ddiag, it could mean that those NAACCR cases are not first-time occurrences. How big of a problem is this?

Not too bad. Though we cannot trust the ICD9/10 codes as replacements for missing surgery dates, there are few enough of them preceding diagnosis that we can remove them as source data errors without ruining the sample size.

Now, as far as the two NAACCR variables go, does n\_dsdisc (date of discharge contribute anything more than n\_dsurg? There are 0 non-missing values of n\_dsdisc when n\_dsurg is missing. As can be seen from the plot below where n\_dsdisc are the red dashed lines and n\_dsurg are the black lines, both relative to date of diagnosis, n\_dsdisc either coincides with n\_dsurg or lags by multiple weeks, as might be expected of a discharge date (what is the plausible threshold on time from surgery to discharge?).



##### Surgery Conclusion

The sole variable on which we can rely for date of surgery is n\_dsurg, though this might get supplemented by additional NAACCR variables in the next data-pull. However, we can rely on v088\_hx\_nphrctm, v102\_acqrd\_absnc, v109\_acqrd\_absnc for excluding possibly invalid records if any of them occur prior to n\_ddiag.

#### Re-ocurrence

#### Death

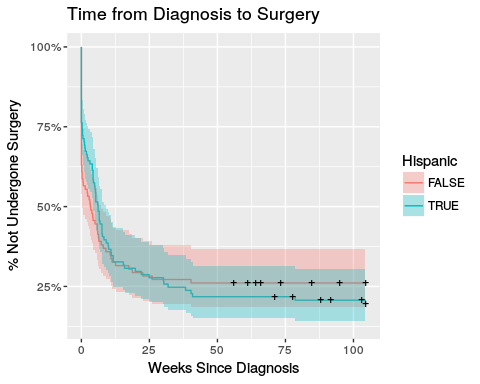
#### Whether or not the patient is Hispanic

A similar process needs to be done for Hispanic ethnicity, but as an ordinary static variable rather than time-to-event. I think I’ll do two variables: one that is true if we are very sure the patient is Hispanic, and the other one that is true if we aren’t certain the patient is *not* Hispanic. In both cases, there will also be an Unknown bins for where all variables are unanimous on the patient’s Hispanic status being unknown.

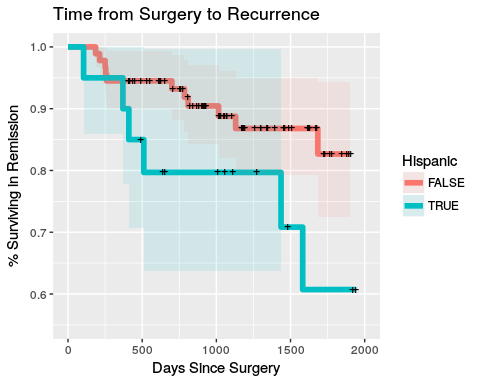
Basically two variables because there are the two ends of the spectrum for resolving disagreement about a binary variable between multiple sources.

## Descriptive Plots (Preliminary)

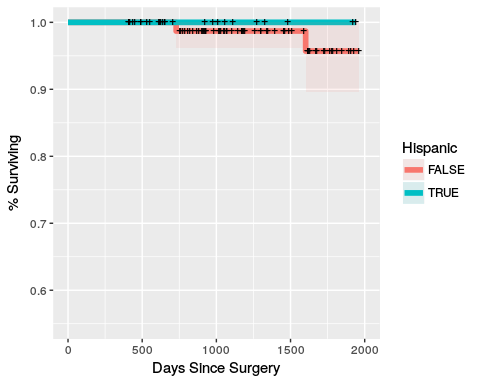
To avoid bias/overfitting all descriptive data and visualizations that relate the predictor variable to the outcome are done using a randomly selected subset of the records (N=758).



So far it seems there is no great difference in the raw lag between Hispanic and non-Hispanic patients but with the major caveat that I have not yet finalized the Hispanic variable and this does not account for other covariates like age and stage.



Does recurrence-free survival after surgery differ between hispanic and non hispanic patients?



Does survival after surgery (insofar that it is reliably represented in the records) differ between hispanic and non-hispanic patients?

## Appendix I: Example of stage/grade data

## Appendix II: Next steps

* TODO: Prior to doing the above tte() put in a safeguard to make sure all the c\_tte variables are TRUE/FALSE only. They are right now as it happens, but nothing enforces that.
* TODO: Create combined (if applicable) variables for each of the following:
  + ~~Initial diagnosis~~
  + ~~Surgery~~ *pending additional variables from next data pull*
  + Re-ocurrence
  + *Last follow-up ?*
  + Death
  + Strict Hispanic designator
  + Lenient Hispanic designator
* TODO: Clean up TNM variables, in consultation with domain expert (Peter?)
* TODO: Create unified Hispanic indicators
* TODO: Create access/quality variables including: number of visits per year, number of lab tests and imaging orders per visit, time spent with provider per visit
* TODO: Resume effort to link Mays Center historic trial records from IDEAS to get information about enrollment in adjuvant trials
* TODO: Start validating and using additional 2a variables already in current data
  + [CN101] OPIOID ANALGESICS (EMR)
  + [CN103] NON-OPIOID ANALGESICS (EMR)
  + 0250 Birthplace (NAACCR possibly EMR)
  + Language (NAACCR and EMR)
  + smoking and alcohol (EMR)
  + Diabetes (NAACCR and EMR)
  + Family history (EMR)
  + Labs (EMR) including: hemoglobin A1c, HDL, VLDL
  + Vitals (EMR) including: systolic and diastolic blood pressure, BMI
* TODO: In next re-run of query…
  + Follow up re additional patient linkages, more recent NAACCR data
  + Miperamine, other anti-depressants
  + 1260 Date of Initial RX--SEER
  + 1270 Date of 1st Crs RX--CoC
  + 3170 RX Date--Most Defin Surg
  + income and education
* DONE: ~~tableOne~~
* DONE: ~~Create time-since-first-diagnosis variable~~
* DONE: ~~Create a special TTE variable from the main i2b2 age at death~~
* DONE: ~~Matrices of pairwise differences between all TTE variables~~
* DONE: ~~Create TTE variable for death (several raw variables)~~
* DONE: ~~Create TTE variable for recurrence~~
* DONE: ~~Create TTE variable for surgery date~~
* DONE: ~~Plot time from diagnosis to surgery, hisp vs non~~
  + ~~First need to confirm interpretation of outcome variable~~
* DONE: ~~Apply the~~ tte() ~~function to all variable in~~ c\_tte
* DONE: ~~Create censoring variable for surgery~~
* DONE: ~~Create censoring variable for recurrence/death~~
* DONE: ~~Map cancer status variable (didn’t turn out to be useful)~~
* DONE: ~~Create unified comorbidity variable for:~~
  + DONE ~~Diabetes~~
* DONE: ~~Mappings for other numcode variables~~
* DONE: ~~Re-run query with additional variables (~~*~~query completed~~*~~):~~
  + ~~EMR codes for secondary tumors~~
  + ~~median household income, 2016 and 2013~~
  + ~~HbA1c~~
  + ~~Family history of diabetes and cancer~~

## Appendix III: Supplementary tables

### What is the coverage of valid records in each data source.

How many patients are in NAACCR, the EMR, both, neither, or have a diagnosis prior to first available record?

*This has been temporarily moved from the main section pending finalization of the recurrence variables. For now, the only ones we can be sure of* [*as indicators of a pre-existing condition*](#surgery-conclusion) *as exclusion criteria for possibly invalid records are v088\_hx\_nphrctm, v102\_acqrd\_absnc, v109\_acqrd\_absnc if they occur prior to* n\_ddiag *and those will exclude far fewer records than suggested by this table* .

### Which variables are near-synonymous?

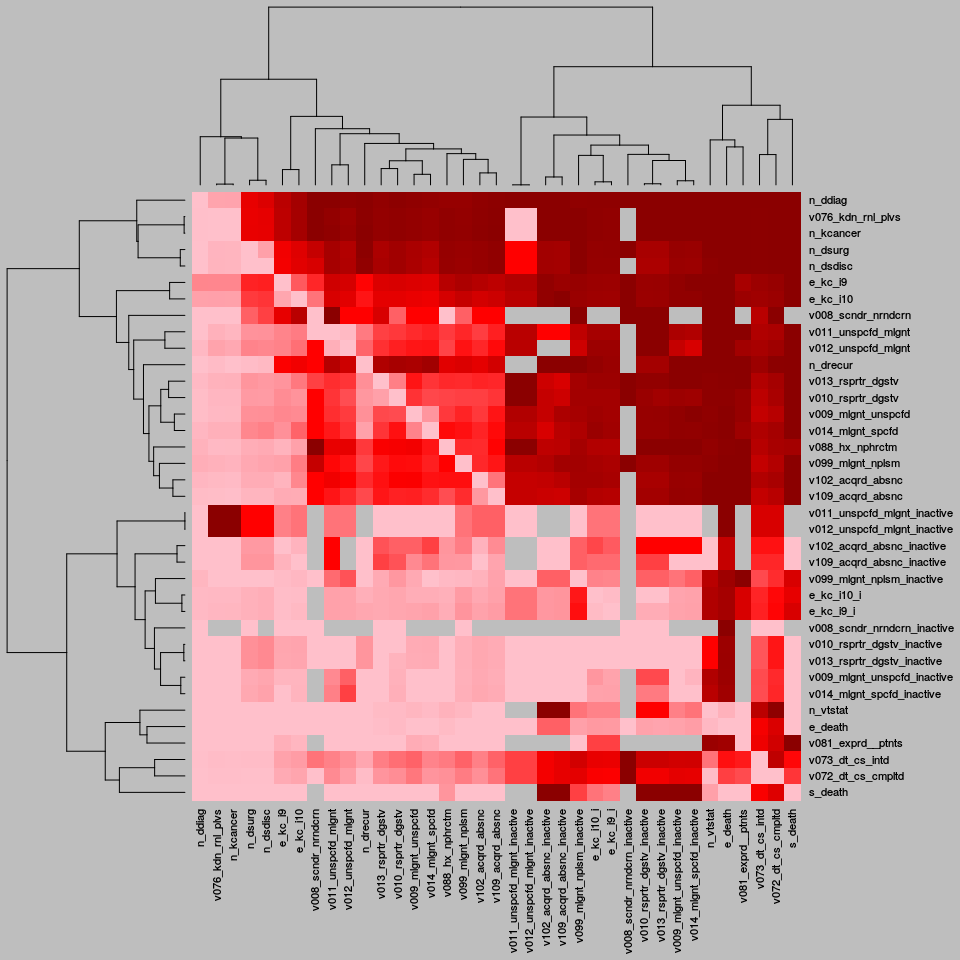
Some variables will, despite what they sound like will be clearly unrelated to each other. Others will be in high pairwise agreement when both are non-missing. The ones in between need to be investigated further to determine whether they are more informative than no information at all, whether they can be cleaned up, and whether there is a bias (i.e. one variable will consistently lag another variable).

In the data.R script we will convert all the event variables to a time to event (tte) form. The above variables plus a few that are dates which aren’t currently known to correlate with any of the events of interest, but doesn’t hurt to check. The overall approach will be:

1. Take for each patient the first visit where the variable is TRUE, non-missing, or in some cases meets some other criteria.
2. Center the age\_at\_visit\_days variable on that visit, so for that patient it is 0 on the visit, a negative integer prior to the visit, and a positive integer after. It will be seen later that this will help make survival analysis easier when we get to it. For patients where an event is never observed, these numbers will be shifted to that the value at the last visit is -1, *not* 0. This is so that we can easily distinguish patients where the event never occurred.

Then we will be ready to probe the degree of agreement and size of lags between these variables.

We will then obtain a diagonal matrix of median differences between each pair of variables. Not only the ones believed to reflect the same event, but all of them. This is so that we can do an overall sanity check on the relationships between groups of variables. For example, if the supposed dates of surgery are in good agreement with each other, but they often happen after the supposed date of reoccurence, then that would be a problem we need to resolve before proceeding further.



*RED indicates row-event occurred after column-event and BLUE indicates that row-event occurred before column-event.*

A lot to unpack here! We can already see that some variables are in close agreement (but these are just medians, this needs to be confirmed on just those groups of variables by checking whether they EVER differ when when both are present… if they never differ, we can treat them as synonymous in casese where one or the other is missing assuming these conclusions are on a reasonably large sample size). Another early conclusion from this is that it isn’t looking good for EMR events lining up with NAACCR events out of the box… they seem to lag behind NAACCR dates, especially diagnoses and (not surprisingly) surgical history. Might need to see if there is something in the EMR that captures date of surgery (especially in Sunrise) and chart review to see why the KC diagnosis codes lag behind NAACCR diagnosis date.

Closer visualization of individual groups of variables can be accomplished by subsetting from this master table.

In addition to medians, we might also generate tables of the 5th and 95th percentiles of the differences as well as medians of the absolute values of the differences. The former are for identifying directional trends and the latter are to distinguish variables that track each other from variables that are uncorrelated but their difference is unbiased in one direction versus another.

## Appendix IV: Audit trail