Kidney Cancer Data Exploration

KL2 Aim 2

Alex F. Bokov[[1]](#footnote-2)

September 14, 2018

Minimal necessary NAACCR variables chosen and process documented for preparing them for analysis, as well as supplementing some of them with additional data from EMR if available.

# 

|  |  |
| --- | --- |
| 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 because I value your perspective and perhaps my results might be useful to your own work.  So far, only de-identified data has been used to generate these results any dates or [patient num](#patient_num) values you see here are also de-identified (with size of time intervals preserved).  This portion of the study is under Dr. Michalek’s exempt project IRB number HSC20170563N. If you are a UT Health researcher who would like a copy of the data, please email me and I will get back to you with further instructions and any additional information I might need from you for our records.  Dr. Murphy, if you are interested in a copy of the data, I will talk to my local mentors and 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. I am doing these parts of Aim 2 ahead of Aim 1 to help me identify the need for additional data-transformations to incorporate into DataFinisher and will switch to the i2b2 plugin (Aim 1) once I hit a natural pausing-point on Aim 2. | * [Consistency-Checks](#consistency-checks) * [Cohort Characterization](#cohort-characterization) * [Testing/Interpreting Variables](#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. [Variable descriptions](#appendix-iv-variable-descriptions)   5. [Audit trail](#appendix-v-audit-trail) |

### Overview

### Questions for mentors and other domain experts:

* Question: What are the main problems with the NAACCR stage and grade information that I will need to clean up?
* Question: What is the typical time that elapses between diagnosis and surgery?
  + Answer (RR): 2-4 weeks, try to avoid more than 4
* Question: Is it possible for surgery to happen on the same day as the diagnosis? How common is that?
  + Answer (RR): Fairly common, if NAACCR diagnosis based on pathology rather than clinical examination, which is usually technically a renal mass, not a cancer. Might want to use imaging result date as the date of diagnosis if it isn’t already being used as such.
* Question: 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?
  + Answer (RR): No, there are a few local cases that took over a decade to get to surgery for various reasons (e.g. indolent tumor, or contact lost with patient).
* Question: What fraction of KC patients undergo surgery?
  + Answer (RR): Around 15%
* 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… (*need to reach out to Grace*)
  + …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) because [dates of birth](#how-well-do-birthdates-match-between-naaccr-and-the-emr), [sexes](#how-well-does-sex-match-up-between-the-emrs-and-naaccr), [races](#how-well-does-race-match-up-between-the-emrs-and-naaccr), and [Hispanic ethnicity](#how-well-does-hispanic-ethnicity-match-up-between-the-emrs-and-naaccr) do not exhibit a greater degree of mismatch between NAACCR records and EMR records than would be expected from routine data entry errors at the source. Furthermore, the mismatches do not seem to correlate with each other.
* 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: Cannot back-fill missing NAACCR values from EMR without chart review and interviewing registrar but within NAACCR the following have emerged as the main variables:
    1. [Diagnosis](#initial-diagnosis) = [n\_ddiag](#n_ddiag) , no others)
    2. [Surgery](#surgery-conclusion) = [n\_dsurg](#n_dsurg) surgery, no others so far but may incorporate information from additional variables after next data update)
    3. Recurrence and prior occurrence = [n\_drecur](#n_drecur)
    4. Death = [a\_tdeath](#a_tdeath)
* Question: Which records to exclude due to likely errors in the source data? E.g. surgery precedes diagnosis, recurrence precedes surgery (for some analysis) death precedes diagnosis or surgery
  + Answer: Currently excluding as incomplete any record lacking either an [n\_ddiag](#n_ddiag) event or both of [n\_kcancer](#n_kcancer) and [n\_seer\_kcancer](#n_seer_kcancer) events. May soon start excluding the few patients with V/Z or surgical history codes indicating missing kidney prior to first NAACCR diagnosis.

## Consistency-Checks

### How well do marital statuses match between NAACCR and the EMR?

Columns represent NAACCR, rows represent EMR. Whole dataset, not filtered for record completeness. Counts in bold are ones that agree between the two sources.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Divorced | Separated | Married | Domestic Partner | Single | Unknown | Widowed | NA | Sum |
| \*\*@\*\* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| **divorced** | **47** | 0 | 5 | 0 | 1 | 3 | 0 | **150** | 206 |
| **legally sepa** | 0 | **15** | 3 | 0 | 2 | 0 | 0 | 35 | 55 |
| **married** | 2 | 3 | **336** | 0 | 3 | 8 | 1 | 887 | 1240 |
| **other** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| **significant** | 0 | 0 | 0 | **0** | 0 | 0 | 0 | 2 | 2 |
| **single** | 5 | 1 | 13 | 0 | **119** | 32 | 1 | 423 | 594 |
| **unknown** | 2 | 0 | 5 | 0 | 0 | **22** | 0 | 66 | 95 |
| **widowed** | 0 | 0 | 7 | 0 | 1 | 1 | **35** | 89 | 133 |
| **Sum** | 56 | 19 | 369 | 0 | 126 | 66 | 37 | 1654 | 2327 |

### How well do birthdates match between NAACCR and the EMR?

There are 0 patients with complete NAACCR records by current criteria but no NAACCR birthdate [n\_dob](#n_dob). Interestingly there are a few [n\_dob](#n_dob) birthdates for patients who do *not* have an [n\_ddiag](#n_ddiag) (by informal inspection). There were a total of 24 patients with a mismatch between their NAACCR and EMR birthdates, and **of the patients with complete records by current criteria, 15 have a mismatch between their NAACCR and EMR birthdates** . Below is a summary of the distribution of their [birth\_date](#birth_date) variable minus [n\_dob](#n_dob):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| -12 | -6.5 | -3.162 | -3.186 | -0.7064 | 9.999 |

The 15 patients with otherwise complete records but mismatched birth dates vary by huge amounts from the EMR versions of their respective birth dates. However, as can be seen in [supplementary tables at the end of this document](#how-well-do-demographic-variables-match-up-for-just-the-patients-with) the 24 total patients with DOB mismatches are not particularly enriched for other mismatches I have tested so far which is more consistent with isolated errors in those respective variables rather than some subset of patients continuing to have their NAACCR and EMR records incorrectly linked.

### How well does sex match up between the EMRs and NAACCR?

Columns represent NAACCR, rows represent EMR. Whole dataset, not filtered for record completeness.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | f | NA | Sum |
| **m** | **428** | 9 | 937 | 1374 |
| **f** | 1 | **235** | 716 | 952 |
| **u** | 0 | 0 | 1 | 1 |
| **Sum** | 429 | 244 | 1654 | 2327 |

### How well does race match up between the EMRs and NAACCR?

Columns represent NAACCR, rows represent EMR. Whole dataset, not filtered for record completeness. Bolded values are those which agree between sources.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | White | Black | Asian | Pac Islander | Other | Unknown | NA | Sum |
| **White** | **591** | 1 | 0 | 0 | 1 | 13 | 1400 | 2006 |
| **Black** | 2 | **26** | 0 | 0 | 0 | 1 | 83 | 112 |
| **Asian** | 2 | 0 | **6** | 1 | 2 | 0 | 11 | 22 |
| **Pac Islander** | 0 | 0 | 0 | **0** | 0 | 0 | 1 | 1 |
| **Other** | 2 | 0 | 0 | 0 | **1** | 0 | 46 | 49 |
| **Unknown** | 23 | 0 | 0 | 0 | 0 | **1** | 113 | 137 |
| **Sum** | 620 | 27 | 6 | 1 | 4 | 15 | 1654 | 2327 |

### How well does Hispanic ethnicity match up between the EMRs and NAACCR?

This time columns represent EMR and rows represent NAACCR. Whole dataset, not filtered for record completeness.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Non\_Hispanic | Hispanic | Sum |
| **Non\_Hispanic** | **304** | 15 | 319 |
| **Hispanic** | 56 | **298** | 354 |
| NA | 983 | 671 | 1654 |
| **Sum** | 1343 | 984 | 2327 |

More detailed ethnicity breakdown…

Again columns represent EMR and rows represent NAACCR. Whole dataset, not filtered for record completeness.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Non\_Hispanic | Hispanic | Sum |
| **Non\_Hispanic** | **291** | 11 | 302 |
| **Unknown** | 13 | **4** | 17 |
| **Hispanic\_NOS** | 44 | **256** | 300 |
| **Mexican** | 9 | **39** | 48 |
| **Spanish\_Surname** | 2 | **2** | 4 |
| **Cuban** | 1 | **0** | 1 |
| **S\_Ctr\_America** | 0 | **1** | 1 |
| NA | 983 | 671 | 1654 |
| **Sum** | 1343 | 984 | 2327 |

## 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. No KC in NAACCR means there is an EMR diagnosis of kidney cancer and there may in some cases also be a *record* for that patient in NAACCR but that record does not show the patient’s site of occurence being kidney.

Note: the below variables are subject to change as the validity criteria and creation of analytic variables from multiple columns of raw data evolve.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Never disease-free | Disease-free | Recurred | Unknown if recurred or was ever gone | Not in NAACCR |
| **n** | 211 | 160 | 95 | 20 | 1841 |
| **Age at Last Contact (mean (sd))** | 63.43 (13.76) | 54.32 (20.42) | 62.51 (15.23) | 55.59 (23.01) | 61.34 (14.18) |
| **a\_hsp\_broad (%)** |  |  |  |  |  |
| Hispanic | 116 (55.0) | 106 ( 66.2) | 50 (52.6) | 8 ( 40.0) | 857 (46.6) |
| non-Hispanic | 92 (43.6) | 50 ( 31.2) | 45 (47.4) | 11 ( 55.0) | 620 (33.7) |
| Unknown | 3 ( 1.4) | 4 ( 2.5) | 0 | 1 ( 5.0) | 364 (19.8) |
| **a\_hsp\_strict (%)** |  |  |  |  |  |
| Hispanic | 68 (32.2) | 62 ( 38.8) | 27 (28.4) | 6 ( 30.0) | 562 (30.5) |
| non-Hispanic | 76 (36.0) | 33 ( 20.6) | 37 (38.9) | 10 ( 50.0) | 577 (31.3) |
| Unknown | 67 (31.8) | 65 ( 40.6) | 31 (32.6) | 4 ( 20.0) | 702 (38.1) |
| **a\_tdeath (mean (sd))** | 48.92 (169.24) | -0.99 (0.08) | 19.42 (107.60) | -1.00 (0.00) | 11.17 (89.54) |
| **a\_tdiag (mean (sd))** | 1208.61 (692.54) | 1457.04 (689.39) | 2254.03 (1343.83) | 1291.75 (696.24) | -1.00 (0.00) |
| **a\_trecur (mean (sd))** | 2.58 (52.05) | -1.00 (0.00) | 1130.04 (805.58) | -1.00 (0.00) | 27.66 (216.71) |
| **a\_tsurg (mean (sd))** | 710.35 (844.78) | 1353.01 (709.37) | 2186.35 (1372.01) | 794.45 (716.09) | 98.67 (443.79) |
| **BMI (mean (sd))** | 27.77 (7.26) | 31.19 (8.34) | 29.32 (7.11) | 29.66 (9.92) | 30.63 (9.31) |
| **Deceased, EMR = TRUE (%)** | 90 (42.7) | 7 ( 4.4) | 22 (23.2) | 3 ( 15.0) | 298 (16.2) |
| **Deceased, Registry = TRUE (%)** | 71 (33.6) | 1 ( 0.6) | 18 (18.9) | 3 ( 15.0) | 43 ( 2.3) |
| **Deceased, SSN = TRUE (%)** | 12 ( 5.7) | 1 ( 0.6) | 5 ( 5.3) | 0 ( 0.0) | 89 ( 4.8) |
| **Diabetes, i2b2 = TRUE (%)** | 54 (25.6) | 56 ( 35.0) | 27 (28.4) | 1 ( 5.0) | 585 (31.8) |
| **Diabetes, Registry = TRUE (%)** | 26 (12.3) | 31 ( 19.4) | 8 ( 8.4) | 0 ( 0.0) | 26 ( 1.4) |
| **Hispanic, i2b2 = TRUE (%)** | 96 (45.5) | 92 ( 57.5) | 43 (45.3) | 7 ( 35.0) | 746 (40.5) |
| **Hispanic, Registry (%)** |  |  |  |  |  |
| Non\_Hispanic | 92 (43.6) | 54 ( 33.8) | 47 (49.5) | 11 ( 55.0) | 98 ( 5.3) |
| Unknown | 5 ( 2.4) | 6 ( 3.8) | 2 ( 2.1) | 1 ( 5.0) | 3 ( 0.2) |
| Hispanic\_NOS | 96 (45.5) | 86 ( 53.8) | 43 (45.3) | 8 ( 40.0) | 67 ( 3.6) |
| Mexican | 17 ( 8.1) | 13 ( 8.1) | 1 ( 1.1) | 0 ( 0.0) | 17 ( 0.9) |
| Spanish\_Surname | 1 ( 0.5) | 0 ( 0.0) | 1 ( 1.1) | 0 ( 0.0) | 2 ( 0.1) |
| Cuban | 0 | 1 ( 0.6) | 0 | 0 ( 0.0) | 0 |
| S\_Ctr\_America | 0 | 0 ( 0.0) | 1 ( 1.1) | 0 ( 0.0) | 0 |
| NA | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1654 (89.8) |
| **Insurance, Registry (%)** |  |  |  |  |  |
| Not Insured | 21 (10.0) | 17 ( 10.6) | 7 ( 7.4) | 2 ( 10.0) | 17 ( 0.9) |
| Self-Pay | 21 (10.0) | 22 ( 13.8) | 15 (15.8) | 0 ( 0.0) | 14 ( 0.8) |
| Insurance NOS | 5 ( 2.4) | 1 ( 0.6) | 0 | 0 ( 0.0) | 1 ( 0.1) |
| Managed Care, HMO, PPO | 53 (25.1) | 56 ( 35.0) | 28 (29.5) | 10 ( 50.0) | 40 ( 2.2) |
| Private Fee-for-Service | 1 ( 0.5) | 0 ( 0.0) | 0 | 0 ( 0.0) | 0 |
| Medicaid | 14 ( 6.6) | 10 ( 6.2) | 1 ( 1.1) | 0 ( 0.0) | 10 ( 0.5) |
| Medicaid Managed Care Plan | 6 ( 2.8) | 14 ( 8.8) | 6 ( 6.3) | 3 ( 15.0) | 10 ( 0.5) |
| Medicare/Medicaid NOS | 30 (14.2) | 13 ( 8.1) | 12 (12.6) | 1 ( 5.0) | 36 ( 2.0) |
| Medicare w Supplement NOS | 2 ( 0.9) | 3 ( 1.9) | 2 ( 2.1) | 0 ( 0.0) | 6 ( 0.3) |
| Medicare Managed Care Plan | 16 ( 7.6) | 9 ( 5.6) | 7 ( 7.4) | 3 ( 15.0) | 13 ( 0.7) |
| Medicare w Private Supplement | 22 (10.4) | 5 ( 3.1) | 9 ( 9.5) | 0 ( 0.0) | 20 ( 1.1) |
| Medicare w Medicaid | 5 ( 2.4) | 3 ( 1.9) | 2 ( 2.1) | 0 ( 0.0) | 7 ( 0.4) |
| TriCare | 1 ( 0.5) | 3 ( 1.9) | 0 | 0 ( 0.0) | 4 ( 0.2) |
| VA | 7 ( 3.3) | 1 ( 0.6) | 1 ( 1.1) | 0 ( 0.0) | 3 ( 0.2) |
| Unknown | 7 ( 3.3) | 3 ( 1.9) | 5 ( 5.3) | 1 ( 5.0) | 6 ( 0.3) |
| NA | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1654 (89.8) |
| **Kidney Cancer, i2b2 = TRUE (%)** | 193 (91.5) | 152 ( 95.0) | 85 (89.5) | 17 ( 85.0) | 1729 (93.9) |
| **Kidney Cancer, Registry = TRUE (%)** | 204 (96.7) | 156 ( 97.5) | 87 (91.6) | 19 ( 95.0) | 20 ( 1.1) |
| **Language, i2b2 (%)** |  |  |  |  |  |
| English | 173 (82.0) | 128 ( 80.0) | 84 (88.4) | 19 ( 95.0) | 1588 (86.3) |
| Spanish | 29 (13.7) | 31 ( 19.4) | 7 ( 7.4) | 1 ( 5.0) | 213 (11.6) |
| Other | 3 ( 1.4) | 0 ( 0.0) | 0 | 0 ( 0.0) | 4 ( 0.2) |
| Unknown | 6 ( 2.8) | 1 ( 0.6) | 4 ( 4.2) | 0 ( 0.0) | 36 ( 2.0) |
| **Marital Status, Registry (%)** |  |  |  |  |  |
| Divorced | 16 ( 7.6) | 13 ( 8.1) | 11 (11.6) | 0 ( 0.0) | 16 ( 0.9) |
| Separated | 2 ( 0.9) | 8 ( 5.0) | 1 ( 1.1) | 2 ( 10.0) | 6 ( 0.3) |
| Married | 125 (59.2) | 79 ( 49.4) | 56 (58.9) | 7 ( 35.0) | 102 ( 5.5) |
| Domestic Partner | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 0 |
| Single | 30 (14.2) | 39 ( 24.4) | 16 (16.8) | 9 ( 45.0) | 32 ( 1.7) |
| Unknown | 24 (11.4) | 15 ( 9.4) | 8 ( 8.4) | 2 ( 10.0) | 17 ( 0.9) |
| Widowed | 14 ( 6.6) | 6 ( 3.8) | 3 ( 3.2) | 0 ( 0.0) | 14 ( 0.8) |
| NA | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1654 (89.8) |
| **n\_cstatus (%)** |  |  |  |  |  |
| Tumor\_Free | 1 ( 0.5) | 160 (100.0) | 7 ( 7.4) | 0 ( 0.0) | 58 ( 3.2) |
| Tumor | 210 (99.5) | 0 ( 0.0) | 81 (85.3) | 0 ( 0.0) | 114 ( 6.2) |
| Unknown | 0 | 0 ( 0.0) | 7 ( 7.4) | 20 (100.0) | 15 ( 0.8) |
| NA | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1654 (89.8) |
| **Race, i2b2 (%)** |  |  |  |  |  |
| White | 185 (87.7) | 149 ( 93.1) | 87 (91.6) | 19 ( 95.0) | 1566 (85.1) |
| Black | 10 ( 4.7) | 3 ( 1.9) | 3 ( 3.2) | 1 ( 5.0) | 95 ( 5.2) |
| Asian | 6 ( 2.8) | 3 ( 1.9) | 0 | 0 ( 0.0) | 13 ( 0.7) |
| Pac Islander | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1 ( 0.1) |
| Other | 3 ( 1.4) | 0 ( 0.0) | 0 | 0 ( 0.0) | 46 ( 2.5) |
| Unknown | 7 ( 3.3) | 5 ( 3.1) | 5 ( 5.3) | 0 ( 0.0) | 120 ( 6.5) |
| **Race, Registry (%)** |  |  |  |  |  |
| White | 188 (89.1) | 153 ( 95.6) | 91 (95.8) | 18 ( 90.0) | 170 ( 9.2) |
| Black | 10 ( 4.7) | 3 ( 1.9) | 2 ( 2.1) | 1 ( 5.0) | 11 ( 0.6) |
| Asian | 3 ( 1.4) | 1 ( 0.6) | 0 | 0 ( 0.0) | 2 ( 0.1) |
| Pac Islander | 1 ( 0.5) | 0 ( 0.0) | 0 | 0 ( 0.0) | 0 |
| Other | 4 ( 1.9) | 0 ( 0.0) | 0 | 0 ( 0.0) | 0 |
| Unknown | 5 ( 2.4) | 3 ( 1.9) | 2 ( 2.1) | 1 ( 5.0) | 4 ( 0.2) |
| NA | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1654 (89.8) |
| **Sex, i2b2 (%)** |  |  |  |  |  |
| m | 151 (71.6) | 100 ( 62.5) | 63 (66.3) | 13 ( 65.0) | 1047 (56.9) |
| f | 60 (28.4) | 60 ( 37.5) | 32 (33.7) | 7 ( 35.0) | 793 (43.1) |
| u | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1 ( 0.1) |
| **Sex, Registry (%)** |  |  |  |  |  |
| m | 149 (70.6) | 98 ( 61.3) | 63 (66.3) | 13 ( 65.0) | 106 ( 5.8) |
| f | 62 (29.4) | 62 ( 38.8) | 32 (33.7) | 7 ( 35.0) | 81 ( 4.4) |
| NA | 0 | 0 ( 0.0) | 0 | 0 ( 0.0) | 1654 (89.8) |

## 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](#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 dat3.

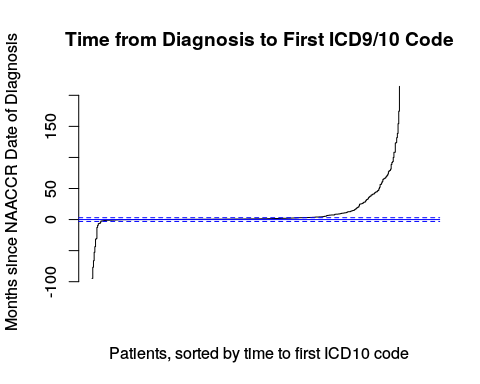
### Initial diagnosis

The c\_kcdiag group of columns in dct0.

* NAACCR: [n\_ddiag](#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](#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 0 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](#n_ddiag) **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](#n_ddiag). Unfortunately, as can be seen from the below table, for the 486 patients who have non-missing [n\_ddiag](#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](#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.
  + Chart review of the 60-100 patients with ICD9/10 codes for kidney cancer that seemingly pre-date their [n\_ddiag](#n_ddiag).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | before | +/- 2 weeks | after | NA | Sum |
| **before** | 29 | 2 | 15 | 1 | 47 |
| **+/- 2 weeks** | 0 | 38 | 34 | 1 | 73 |
| **after** | 0 | 1 | 316 | 3 | 320 |
| NA | 0 | 0 | 7 | 39 | 46 |
| **Sum** | 29 | 41 | 372 | 44 | 486 |

Here is a plot centered on [n\_ddiag](#n_ddiag) (blue horizontal line at 0) with black lines indicating ICD10 codes for primary kidney cancer from the EMR and dashed red lines indicating ICD9 codes. The dashed horizontal blue lines indicate +- 3 months from [n\_ddiag](#n_ddiag).



From this we can conclude that for most patients (291), the first EMR code is recorded within 3 months of first diagnosis as recorded by NAACCR. Of those with a larger time difference, the majority (143) have their first EMR code occur *after* first NAACCR diagnosis. Only 13 patients have ICD9/10 diagnoses that precede their NAACCR diagnoses by more than 3 months. And additional 54 patients have first EMR diagnoses that precede NAACCR diagnosis by less than three months. These might need to be eliminated from the sample on the grounds of not being first occurrences of kidney cancer. However, we cannot back-fill missing NAACCR records or NAACCR records lacking a diagnosis date because there is too frequently a difference between the the two sources, and the EMR records are currently biased toward later dates.

### Surgery

* NAACCR:
  + In addition to [n\_dsurg](#n_dsurg)) the following possibly relevant fields are available in our local NAACCR:
    - [n\_rx1260](#n_rx1260)
    - [n\_rx1270](#n_rx1270)
    - [n\_rx3170](#n_rx3170)
  + Here are the questions raised:
    - Do they agree with [n\_dsurg](#n_dsurg) sufficiently that missing [n\_dsurg](#n_dsurg) can be backfilled from some or all of them?
    - Under what circumstances can they be interpreted as surgery dates rather dates for something else?
    - How accurate is [n\_surgreason](#n_surgreason) in distinguishing surgical cases from non-surgical cases as per EMR records?
* EMR: First occurrence of any ICD9/10 code for acquired absence of kidney; or first occurence of surgical history of nephrectomy. How much do they agree with NAACCR?

As can be seen in the table below, the variables v080\_acqrd\_absnc, v094\_hx\_nphrctm, v122\_acqrd\_absnc, e\_surgonc *sometimes* precede [n\_ddiag](#n_ddiag) by many weeks. However, they *usually* follow [n\_ddiag](#n_ddiag) by more weeks than the two NAACCR variables [n\_dsdisc](#n_dsdisc) and [n\_dsurg](#n_dsurg). Those two NAACCR variables never occur before [n\_ddiag](#n_ddiag) and usually occur within 2-8 weeks after it.

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](#n_ddiag) than the non-inactive codes.

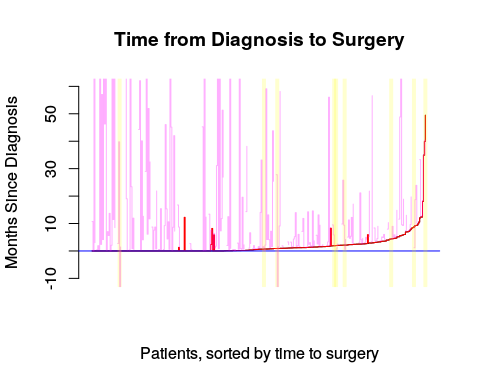
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA’s |
| **v080\_acqrd\_absnc** | -361.1 | 8.143 | 31.43 | 69.5 | 82.71 | 957.4 | 261 |
| **v080\_acqrd\_absnc\_inactive** | 2.571 | 83.29 | 205.3 | 233.5 | 375.4 | 708.3 | 477 |
| **v094\_hx\_nphrctm** | -91.86 | 10.11 | 37.07 | 77.85 | 93.96 | 758.1 | 318 |
| **v122\_acqrd\_absnc** | -20.14 | 9.607 | 37.86 | 85.12 | 111.2 | 957.4 | 226 |
| **v122\_acqrd\_absnc\_inactive** | 2.571 | 72.68 | 155.1 | 237 | 375.5 | 708.3 | 478 |
| **n\_rx3170** | 0 | 0 | 3 | 8.461 | 9.643 | 215.1 | 119 |
| **n\_rx1270** | 0 | 0 | 2.929 | 6.431 | 6.964 | 318.3 | 28 |
| **n\_rx1260** | 0 | 0 | 3.857 | 8.213 | 8.571 | 270.9 | 198 |
| **n\_dsurg** | 0 | 0 | 2.857 | 7.83 | 9 | 215.1 | 109 |
| **e\_surgonc** | -194.9 | 0.2143 | 4.714 | 23.58 | 46 | 236.6 | 455 |
| **n\_drecur** | 0 | 40.04 | 73.71 | 137.2 | 205.3 | 935.9 | 402 |

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](#n_ddiag), it could mean that those NAACCR cases are not first-time occurrences. How big of a problem is this?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | before | same-day | after | NA |
| **v080\_acqrd\_absnc** | 3 | 0 | 222 | 261 |
| **v080\_acqrd\_absnc\_inactive** | 0 | 0 | 9 | 477 |
| **v094\_hx\_nphrctm** | 3 | 2 | 163 | 318 |
| **v122\_acqrd\_absnc** | 1 | 0 | 259 | 226 |
| **v122\_acqrd\_absnc\_inactive** | 0 | 0 | 8 | 478 |
| **n\_rx3170** | 0 | 138 | 229 | 119 |
| **n\_rx1270** | 0 | 149 | 309 | 28 |
| **n\_rx1260** | 0 | 83 | 205 | 198 |
| **n\_dsurg** | 0 | 146 | 231 | 109 |
| **e\_surgonc** | 7 | 1 | 23 | 455 |

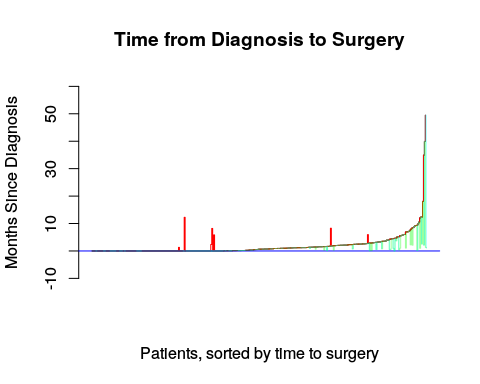
Not too bad. Though we cannot trust the ICD9/10 codes as replacements for missing surgery dates, there are few enough of them preceding diagnosis that we can remove them as source data errors without ruining the sample size.

Below is a plot of all patients sorted by [n\_dsurg](#n_dsurg) (black line). On the same axis is [n\_rx3170](#n_rx3170) (red line) which is almost identical to [n\_dsurg](#n_dsurg) except for a small number of cases where it occurs later than [n\_dsurg](#n_dsurg). Never earlier. The purple lines indicate for each patient the earliest EMR code implying that a surgery had taken place (acquired absence of kidney ICD V/Z codes or surgical history of nephrectomy).

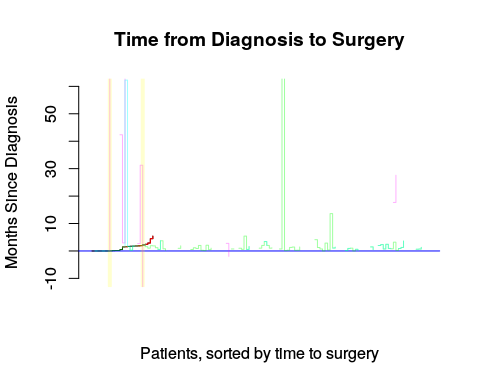


In the above plot the 9 patients for which one or more EMR codes are recorded prior to [n\_dsurg](#n_dsurg) are highlighted in yellow.

In the below plot the [n\_rx1270](#n_rx1270) (green) and [n\_rx1260](#n_rx1260) (cyan) events are superimposed over time until [n\_dsurg](#n_dsurg) from above (but EMR codes for nephrectomy are omitted on this one). The [n\_rx1270](#n_rx1270) and [n\_rx1260](#n_rx1260) variables trend toward occurring earlier than [n\_dsurg](#n_dsurg).



Furthermore, it can be seen from an equivalent plot but for patients who do *not* have a [n\_surgreason](#n_surgreason) code equal to Surgery Performed there are many [n\_rx1270](#n_rx1270) and [n\_rx1260](#n_rx1260) events, but only a small number of [n\_dsurg](#n_dsurg) (black) and [n\_rx3170](#n_rx3170) (red). The [n\_dsurg](#n_dsurg) and [n\_rx3170](#n_rx3170) that do occur track each other perfectly. Together with NAACCR data dictionary’s description this suggests that [n\_rx3170](#n_rx3170) is the legitimate principal surgery date in close agreement with [n\_dsurg](#n_dsurg), so perhaps missing [n\_rx3170](#n_rx3170) values can be filled in from [n\_dsurg](#n_dsurg). However [n\_rx1270](#n_rx1270) and [n\_rx1260](#n_rx1260) seem like non-primary surgeries or other events



Here is a table of every NAACCR surgery event variable versus the [n\_surgreason](#n_surgreason) variable:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | n\_rx3170=FALSE | n\_rx3170=TRUE | n\_rx1270=FALSE | n\_rx1270=TRUE | n\_rx1260=FALSE | n\_rx1260=TRUE | n\_dsurg=FALSE | n\_dsurg=TRUE |
| **Surgery Performed** | 15 | 457 | 13 | 459 | 170 | 302 | 14 | 458 |
| **Surgery Not First Course** | 136 | 10 | 20 | 126 | 82 | 64 | 122 | 24 |
| **No Surgery, Contra Indicated** | 17 | 1 | 3 | 15 | 10 | 8 | 16 | 2 |
| **No Surgery, Deceased** | 4 | 0 | 1 | 3 | 2 | 2 | 4 | 0 |
| **No Surgery, No Reason Given** | 5 | 0 | 2 | 3 | 2 | 3 | 5 | 0 |
| **No Surgery, Refused** | 5 | 3 | 2 | 6 | 4 | 4 | 4 | 4 |
| **Unknown Whether Surgery Done** | 16 | 1 | 11 | 6 | 13 | 4 | 15 | 2 |
| **Unknown Whether Surgery Recommended or Done** | 3 | 0 | 2 | 1 | 2 | 1 | 3 | 0 |

##### Surgery Conclusion

As of now the sole variables on which I can rely for date of surgery are [n\_rx3170](#n_rx3170) supplemented by [n\_dsurg](#n_dsurg), and the small number of cases where EMR codes imply surgery prior to diagnosis will be excluded. For the purposes of determining whether there is a difference in the time from diagnosis to surgery I could also create an alternative ‘naive’ variable that is simply the earliest of all possible surgery events for each patient. For the time elapsed from surgery to death or recurrence, I will use the first ([n\_rx3170](#n_rx3170) and [n\_dsurg](#n_dsurg)) variable as above with the additional criterion that only cases where the [n\_surgreason](#n_surgreason) is Surgery Performed be included.

TODO: Might need to rework t\_priorcond

### Re-occurrence

The current available variables are: [n\_cstatus](#n_cstatus) which corresponds to [1770 Cancer Status](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1770) ~~hopefully with~~ [start\_date](#start_date) ~~set by the ETL to~~ [1772 Date of Last Cancer Status](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1772) ~~(need to double-check that it is)~~ and [n\_drecur](#n_drecur), [1860 Recurrence Date--1st](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1860). UPDATE: Our site is on NAACCR v16, not v18, and we do not have [1772 Date of Last Cancer Status](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1772). According to the v16 standard, instead the [1750 Date of Last Contact](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1750) should be used.

~~It looks like it would also be useful in the next data pull to include~~ [1880 Recurrence Type--1st](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1880) ~~which our NAACCR does use.~~ Done.

Now we can reconcile the [n\_cstatus](#n_cstatus) and [n\_rectype](#n_rectype) variables. We can see below that almost all [n\_cstatus](#n_cstatus) Tumor\_Free patients also have a Disease-free in their [n\_rectype](#n_rectype) column, the Tumor ones have a variety of values, and the Unknown ones are also mostly Unknown if recurred or was ever gone.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tumor\_Free | Tumor | Unknown |
| **Disease-free** | 201 | 0 | 0 |
| **In situ invasive** | 0 | 2 | 0 |
| **In situ original** | 0 | 3 | 0 |
| **Local, insufficient info** | 1 | 8 | 0 |
| **Local invasive** | 2 | 15 | 0 |
| **Regional, insufficient info** | 0 | 3 | 1 |
| **Invasive adjacent tissue only** | 0 | 3 | 0 |
| **Invasive regional lymph nodes only** | 0 | 3 | 0 |
| **Invasive adjacent tissue and regional lymph nodes** | 0 | 2 | 0 |
| **Regional in situ, NOS** | 0 | 1 | 0 |
| **Multiple true for invasive tumor** | 0 | 2 | 0 |
| **Distant, insufficient info** | 1 | 16 | 0 |
| **Distant invasive lung only** | 1 | 22 | 1 |
| **Distant invasive pleura only** | 0 | 1 | 0 |
| **Distant invasive liver only** | 0 | 3 | 0 |
| **Distant invasive bone only** | 1 | 7 | 0 |
| **Distant invasive CNS only** | 0 | 5 | 0 |
| **Distant invasive lymph node only** | 0 | 3 | 0 |
| **Distant invasive single site and local/trocar/regional** | 0 | 4 | 0 |
| **Distant invasive multiple sites** | 1 | 4 | 0 |
| **Never disease-free** | 0 | 242 | 0 |
| **Recurred but no other info** | 0 | 2 | 0 |
| **Unknown if recurred or was ever gone** | 0 | 2 | 31 |
| **Ambig\_6070** | 0 | 1 | 0 |
| **Ambig\_7000** | 0 | 3 | 0 |

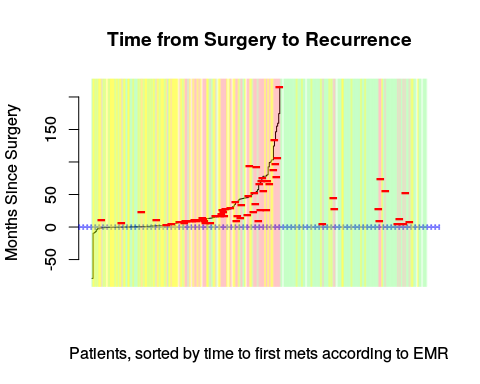
This suggest the following rules for binning them:

* [n\_rectype](#n_rectype) is Disease-free (disease free)
* [n\_rectype](#n_rectype) is Never disease-free (never disease free)
* [n\_rectype](#n_rectype) raw code includes 70 then assume never diease free
* [n\_rectype](#n_rectype) is Unknown if recurred or was ever gone (unknown)
* Otherwise, (recurred)

Here is the condensed version after having followed the above rules. Looks like the only ones who have a [n\_drecur](#n_drecur) (recurrence date) are the ones which also have a Recurred status for [a\_n\_recur](#a_n_recur) (with 19 missing an [n\_drecur](#n_drecur)). The only exception is 1 Never diease-free patient that had an [n\_drecur](#n_drecur).

|  |  |  |
| --- | --- | --- |
|  | Recur Date=FALSE | Recur Date=TRUE |
|  | 1654 | 0 |
| **Disease-free** | **215** | 0 |
| **Never disease-free** | **281** | 1 |
| **Recurred** | 19 | **124** |
| **Unknown if recurred or was ever gone** | **33** | 0 |

This explains why [n\_drecur](#n_drecur) values are relatively rare in the data– they are specific to actual recurrences which are not a majority of the cases. This is a good from the standpoint of data consistency. Now we need to see to what extent the EMR codes agree with this. In the below plot, the black line represents months elapsed between surgery and the first occurence of an EMR code for secondary tumors, if any. The horizontal red line segments indicate individual NAACCR dates of recurrence, [n\_drecur](#n_drecur). The blue horizontal line is the date of surgery. Patients whose status ([n\_rectype](#n_rectype)) is Disease-free are highlighted in green, Never disease-free in yellow, and Recurred in red.

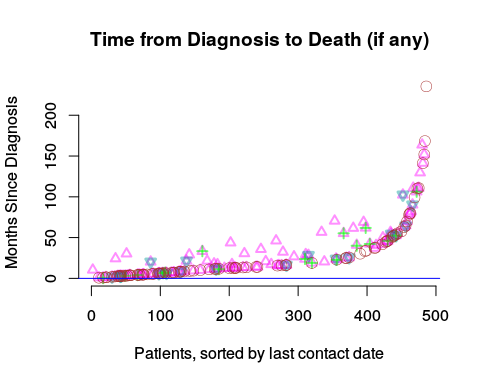


The green highlights are *mostly* where one would expect, but why are there 38 patients on the left side of the plot that have EMR codes for secondary tumors? Also, there are 32 patients with metastatic tumor codes earlier than [n\_dsurg](#n_dsurg) and of those 5 occur more than 3 months prior to [n\_dsurg](#n_dsurg). Did they present with secondary tumors to begin with but remained disease free after surgery? Removing the \_inactive versions of the secondary tumor codes does not make the left-side green patients go away.

### Death

Below are plotted times of death (for patients that have them) relative to date of diagnosis [n\_ddiag](#n_ddiag) (horizontal blue line). The four data sources are [e\_death](#e_death) the EMR death date ($\tiny\color{magenta}\triangle$), [s\_death](#s_death) the social security death date ($\color{blue}\triangledown$), [e\_dscdeath](#e_dscdeath) the EMR hospital discharge death date ($\color{green}+$), and n\_vtstat the NAACCR death date ($\tiny\color{brown}\bigcirc$).

When more than one source has a death date, they are in agreement. To be fair, the agreement between [e\_death](#e_death), [e\_dscdeath](#e_dscdeath), and [s\_death](#s_death) is probably due to our i2b2 ETL already merging [e\_dscdeath](#e_dscdeath) and [s\_death](#s_death) into [e\_death](#e_death). But it is also encouraging that none of them seem (by visual inspection) to occur prior to the date of last contact in NAACCR. That suggests I can simply take the mininum of available death dates to fill in data for patients that NAACCR is not aware are deceased. It also means that the ETL’s coverage of vital status can be further improved by using the NAACCR vital status and last contact variables in combination.



### 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 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.

Here are the variables to process:

* [language\_cd](#language_cd) is an i2b2 PATIENT\_DIMENSION variable that is simplified by data.R and levels\_map.csv
  + Hispanic : Spanish
  + non-Hispanic: Other
  + Unknown: English or Unknown or NA
* [e\_lng](#e_lng) is an i2b2 OBSERVATION\_FACT variable currently in the raw form that DataFinisher uses for complex variables lacking a specific rule. Below are regexp patterns for a non case-sensitive match.
  + Hispanic: ^.\*spanish.\*$ ELSE
  + Unknown: ^.\*(english|sign language|unknown).\*$ or NA ELSE
  + non-Hispanic: anything not caught by the above two filters
* [n\_hisp](#n_hisp) is the [0190 Spanish/Hispanic Origin](http://datadictionary.naaccr.org/default.aspx?c=10" \l "190) variable from NAACCR. Slightly processed by data.R and levels\_map.csv
  + non-Hispanic: Non\_Hispanic
  + Unknown: Unknown
  + Hispanic: any other value
* [e\_hisp](#e_hisp) is the indicator variable for Hispanic ethnicity from i2b2 OBSERVATION\_FACT.
  + Hispanic: TRUE
  + Unknown: FALSE
* [e\_eth](#e_eth) is the whole ethnicity variable from i2b2 OBSERVATION\_FACT and suprprisingly it is not in full agreement with [e\_hisp](#e_hisp)
  + Hispanic: hispanic
  + Unknown: other,unknown,unknown/othe,i choose not,@
  + non-Hispanic: arab-amer,non-hispanic

The strict Hispanic variable.

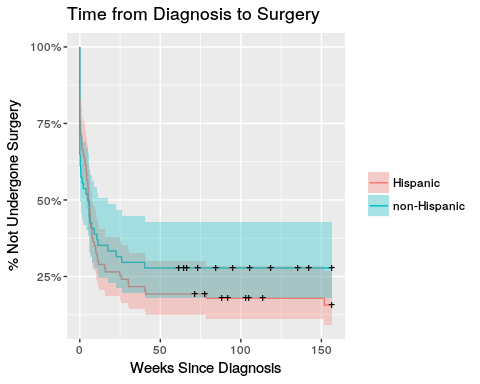
* Hispanic if ALL non-missing values of [n\_hisp](#n_hisp), [e\_hisp](#e_hisp), and [e\_eth](#e_eth) are unanimous for Hispanic
* non-Hispanic if ALL non-missing values of [n\_hisp](#n_hisp) and [e\_eth](#e_eth) are unanimous for non-Hispanic (note that [e\_hisp](#e_hisp) not included here) and neither [e\_lng](#e_lng) nor [language\_cd](#language_cd) vote for Hispanic
* Unknown if any other result.

The lenient Hispanic variable.

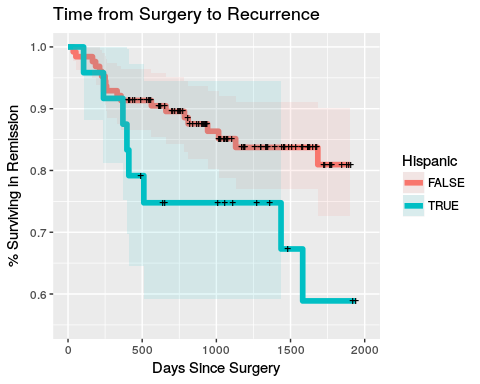
* Hispanic if ANY non-missing values of [language\_cd](#language_cd), [e\_lng](#e_lng), [n\_hisp](#n_hisp), [e\_hisp](#e_hisp), and [e\_eth](#e_eth) have value Hispanic
* Unknown if ALL non-missing values of [language\_cd](#language_cd), [e\_lng](#e_lng), [n\_hisp](#n_hisp), [e\_hisp](#e_hisp), and [e\_eth](#e_eth) are unanimous for Unknown
* non-Hispanic if any other result

## Descriptive Plots (Preliminary)

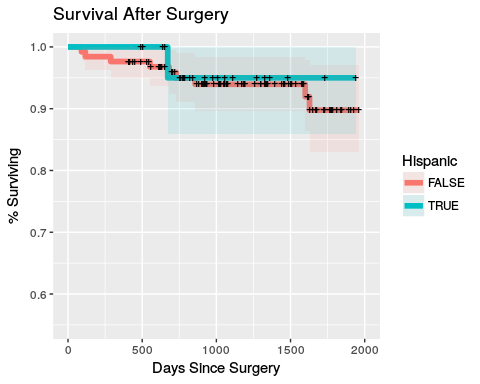
To avoid bias/overfitting all descriptive data and visualizations below that relate the predictor variable to the outcome are done using a randomly selected subset of the records (N=758).



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

(proof of feasibility)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| patient\_num | 3430 Derived AJCC-7 Stage Grp | 3422 Derived AJCC-7 M Descript | 3420 Derived AJCC-7 M | 3412 Derived AJCC-7 N Descript | 3410 Derived AJCC-7 N | 3402 Derived AJCC-7 T Descript | 3400 Derived AJCC-7 T |
| 114314 | 500 | c | 000 | p | 000 | p | 320 |
| 274467 | 888 | N | 888 | N | 888 | N | 888 |
| 317889 | 500 | c | 000 | p | 000 | p | 320 |
| 337717 | 500 | c | 000 | c | 000 | p | 310 |
| 387599 | 700 | p | 100 | c,p | 000 | p | 310,320 |
| 401774 | 700 | p | 100 | p | 000 | p | 310 |
| 444345 | 888 | N | 888 | N | 888 | N | 888 |
| 692996 | 010 | c | 000 | c | 000 | c | 010 |
| 731060 | 700 | c | 100 | p | 000 | p | 320 |
| 800320 | 100 | c | 000 | c | 000 | p | 120 |
| 857476 | 500 | c | 000 | p | 000 | p | 300 |
| 1003998 | 888 | N | 888 | N | 888 | N | 888 |
| 1158986 | 100 | c | 000 | c | 000 | p | 150 |
| 1231407 | 888 | N | 888 | N | 888 | N | 888 |
| 1270762 | 700 | c | 100 | c | 100 | p | 310 |

## Appendix II: Next steps

* TODO: Update and clean up the plots, including labels.
* TODO: Update and streamline the narrative.
* 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: Clean up TNM variables, in consultation with domain expert (Peter?)
* 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
  + income (Census)
  + Miperamine, other anti-depressants
  + DONE: ~~Should use~~ [0580 Date of 1st Contact](http://datadictionary.naaccr.org/default.aspx?c=10" \l "580) ~~as the diagnosis date if earlier than~~ [n\_ddiag](#n_ddiag)~~!~~ *Actually, evidence that it’s neither a diagnosis date nor a first contact. Not known what it is.*
  + DONE: ~~Surgery fields:~~
    - [1260 Date of Initial RX--SEER](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1260)
    - [1270 Date of 1st Crs RX--CoC](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1270)
    - [3170 RX Date--Most Defin Surg](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3170)
  + DONE: ~~Recurrence:~~ [1880 Recurrence Type--1st](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1880)
* TODO: In a future re-run of query…
  + Follow up re additional patient linkages, more recent NAACCR data
  + education (Census, not ready, ETL needs fixing)
* DONE: ~~Create combined (if applicable) variables for each of the following:~~
  + ~~Initial diagnosis~~ [a\_tdiag](#a_tdiag), [a\_cdiag][a\_cdiag]
  + ~~Surgery~~ [a\_tsurg](#a_tsurg), [a\_csurg][a\_csurg]
  + ~~Re-ocurrence~~ [a\_trecur](#a_trecur), [a\_crecur][a\_crecur]
  + *~~Last follow-up ?~~*
  + ~~Death~~ [a\_tdeath](#a_tdeath), [a\_cdeath][a\_cdeath]
  + ~~Strict Hispanic designator~~ [a\_hsp\_strict](#a_hsp_strict)
  + ~~Lenient Hispanic designator~~ [a\_hsp\_broad](#a_hsp_broad)
  + ~~NAACCR-only Hispanic designator~~ [a\_hsp\_naaccr](#a_hsp_naaccr)
* DONE: ~~Verify that the [ETL](http://www.hostedredmine.com/issues/719444" \l "note-11) gets~~ [start\_date](#start_date) ~~for~~ 1770 Cancer Status ~~from~~ [1772 Date of Last Cancer Status](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1770) *in NAACCR v16 it cannot/doesn’t need to*
* 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

### How well do demographic variables match up for just the patients with mismatched birthdates?

#### Sex

Columns represent NAACCR, rows represent EMR. Only DOB mismatched patients.

|  |  |  |  |
| --- | --- | --- | --- |
|  | m | f | Sum |
| **m** | **12** | **0** | 12 |
| **f** | **1** | **11** | 12 |
| **u** | 0 | 0 | 0 |
| **Sum** | 13 | 11 | 24 |

#### Race

Columns represent NAACCR, rows represent EMR. Only DOB mismatched patients.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | White | Black | Asian | Pac Islander | Other | Unknown | Sum |
| **White** | **19** | **0** | **0** | **0** | 0 | 2 | 21 |
| **Black** | **0** | **1** | **0** | **0** | 0 | 0 | 1 |
| **Asian** | **0** | **0** | **0** | **0** | 0 | 0 | 0 |
| **Pac Islander** | **0** | **0** | **0** | **0** | 0 | 0 | 0 |
| **Other** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Unknown** | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| **Sum** | 21 | 1 | 0 | 0 | 0 | 2 | 24 |

#### Hispanic ethnicity

This time columns represent EMR and rows represent NAACCR. Only DOB mismatched patients.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Non\_Hispanic | Hispanic | Sum |
| **Non\_Hispanic** | **14** | **1** | 15 |
| **Hispanic** | **0** | **9** | 9 |
| **Sum** | 14 | 10 | 24 |

#### Nephrectomy according to EMR preceding diagnosis according to NAACCR

Only complete NAACCR records with mismatched DOBs.

Looks like the 15 DOB-mismatched patients otherwise meeting completeness criteria for kidney cancer records do not coincide with the set of patients seeming to have nephrectomies prior to their NAACCR diagnoses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | before | same-day | after | NA |
| **v080\_acqrd\_absnc** | 0 | 0 | 7 | 8 |
| **v080\_acqrd\_absnc\_inactive** | 0 | 0 | 0 | 15 |
| **v094\_hx\_nphrctm** | 0 | 0 | 4 | 11 |
| **v122\_acqrd\_absnc** | 0 | 0 | 7 | 8 |
| **v122\_acqrd\_absnc\_inactive** | 0 | 0 | 0 | 15 |
| **n\_rx3170** | 0 | 4 | 6 | 5 |
| **n\_rx1270** | 0 | 4 | 10 | 1 |
| **n\_rx1260** | 0 | 2 | 6 | 7 |
| **n\_dsurg** | 0 | 4 | 6 | 5 |
| **e\_surgonc** | 0 | 0 | 0 | 15 |

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

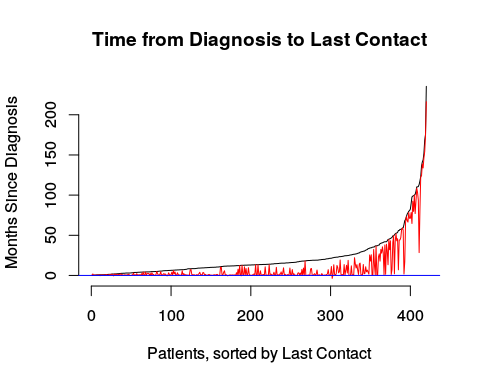
How many patients are in NAACCR, the EMR, both, neither, or have a diagnosis prior to first available record?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NAACCR | EMR | PreExisting | N | N Cumulative |
| FALSE | FALSE | TRUE | 109 | 2327 |
| FALSE | TRUE | TRUE | 360 | 2218 |
| TRUE | TRUE | TRUE | 3 | 1858 |
| FALSE | FALSE | FALSE | 3 | 1855 |
| TRUE | FALSE | FALSE | 39 | 1852 |
| FALSE | TRUE | FALSE | 1369 | 1813 |
| TRUE | TRUE | FALSE | 444 | 444 |

*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 v080\_acqrd\_absnc, v094\_hx\_nphrctm, v122\_acqrd\_absnc, e\_surgonc if they occur prior to* [n\_ddiag](#n_ddiag) *and those will exclude far fewer records than suggested by this table* .

### What is going on with the first contact variable?

Wierd observation– the date of first contact [n\_fc](#n_fc) (red) is almost always between last contact [n\_lc](#n_lc) (black) and diagnosis [n\_ddiag](#n_ddiag) (blue), though diagnosis is usually on a biopsy sample and that’s why it’s dated as during or after surgery we thought. If first contact is some kind of event after first diagnosis, what is it?



Surgery [n\_dsurg](#n_dsurg) seems to happen in significant amounts both before and after first contact [n\_fc](#n_fc).

### 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](#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 diagonal matrices of various pairwise comparisons of the timing of events. 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. The below heatmap indicates the fraction of the column events that occurred before or at the same time as the row events.



A lot to unpack here! We can already see that some variables are in close agreement. Another early conclusion from this is that it isn’t looking good for EMR events lining up with NAACCR events… they seem to lag behind NAACCR dates, especially diagnoses and 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.

However, most of this shotgun approach is now superseded by the more focused investigation in the [initial diagnosis](#initial-diagnosis) and [surgery](#surgery-conclusion) sections in the main document above. This is just for historic reference.

## Appendix IV: Variable descriptions

Here are descriptions of the variables referenced in this document.

###### start\_date

start\_date

###### birth\_date

birth\_date

###### language\_cd

language\_cd; Language, i2b2

###### age\_at\_visit\_days

age\_at\_visit\_days; Age at Last Contact

###### n\_rectype

1880 Recurrence Type–1st

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1880](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1880)

###### n\_rx3170

3170 RX Date–Most Defin Surg; Deat of most definitive surgery.

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3170](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3170)

###### n\_rx1270

1270 Date of 1st Crs RX–CoC; Date of initiation of the first therapy for the cancer being reported, using the CoC definition of first course. The date of first treatment includes the date a decision was made not to treat the patient.

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1270](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1270)

###### n\_rx1260

1260 Date of Initial RX–SEER; Date of initiation of the first course therapy for the tumor being reported, using the SEER definition of first course. See also Date 1st Crs RX CoC [1270].

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1260](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1260)

###### n\_fc

0580 Date of 1st Contact; Can also be date of clinical (as opposed to path) diagnosis

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#580](http://datadictionary.naaccr.org/default.aspx?c=10" \l "580)

###### n\_dsdisc

3180 RX Date–Surgical Disch

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3180](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3180)

###### n\_surgreason

1340 Reason for No Surgery

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1340](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1340)

###### n\_hisp

0190 Spanish/Hispanic Origin; Hispanic Origin, Registry

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#190](http://datadictionary.naaccr.org/default.aspx?c=10" \l "190)

###### n\_dob

0240 Date of Birth

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#240](http://datadictionary.naaccr.org/default.aspx?c=10" \l "240)

###### n\_ddiag

0390 Date of Diagnosis

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#390](http://datadictionary.naaccr.org/default.aspx?c=10" \l "390)

###### n\_dsurg

1200 RX Date–Surgery

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1200](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1200)

###### n\_lc

1750 Date of Last Contact; Last Contact

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1750](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1750)

###### n\_cstatus

1770 Cancer Status; Cancer Status, Registry

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1770](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1770)

###### n\_drecur

1860 Recurrence Date–1st

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#1860](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1860)

###### s\_death

Deceased per SSA; Death, SSN

###### e\_hisp

Hispanic or Latino; Hispanic Origin, i2b2

###### e\_dscdeath

Expired[7,579 facts; 7,544 patients]; Discharge Disposition

###### e\_eth

Ethnicity; EMR demographics

###### n\_seer\_kcancer

Kidney and Renal Pelvis; SEER site

###### n\_kcancer

Kidney, NOS; KC, Registry

###### e\_lng

Language

###### patient\_num

Patient Number (anonymized); Patient

###### e\_death

Death, i2b2; Death according to the combined i2b2 records from all sources

###### a\_n\_recur

Recurrence Status; *This is the main analytic variable for recurrence.* This is based on [n\_rectype](#n_rectype) but with all values that signify recurrence binned together leaving Unknown if recurred or was ever gone,Never disease-free,Disease-free, and Recurred.

###### a\_hsp\_broad

Hispanic; Code patients as Hispanic if there is even the slightest evidence they are, otherwise assume they re non-Hispanic, and only if there is really zero evidence either way return Unknown

###### a\_hsp\_strict

Hispanic (strict); Code patients as Hispanic or non-Hispanic only if all available evidence is unanimous, otherwise err on the side of Unknown

###### a\_hsp\_naaccr

Hispanic (NAACCR); The [n\_hisp](#n_hisp) variable binned to Hispanic, non-Hispanic, and Unknown

###### a\_tdiag

Diagnosis Date

###### a\_trecur

Recurrence Date

###### a\_tsurg

Surgery Date

###### a\_tdeath

Death Date

## Appendix V: Audit trail

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sequence | time | type | name | hash |
| 0001 | 2018-09-14 19:00:23 | info | sessionInfo | - |
| 0002 | 2018-09-14 19:00:23 | this\_script | exploration.spin.Rmd | 9165cca |
| 0003 | 2018-09-14 19:00:40 | rdata | .depdata = “data.R.rdata” | 2755e1fd7ece31348b75fad478f04397 |
| 0003.0001 | 2018-09-14 17:35:04 | info | sessionInfo | - |
| 0003.0002 | 2018-09-14 17:35:04 | this\_script | data.R | fac8ce1 |
| 0003.0003 | 2018-09-14 17:35:06 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0003.0004 | 2018-09-14 17:35:17 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0003.0005 | 2018-09-14 17:35:17 | file | rawdct = “local/in/meta\_HSC20170563N\_kc\_v200.int.csv” | 77226290495672d030798e64327fe10a |
| 0003.0006 | 2018-09-14 17:35:17 | file | tpldct = “datadictionary\_static.csv” | b150f32ecdc9a3ed746282ea3012ca01 |
| 0003.0007 | 2018-09-14 17:35:19 | file | levels\_map\_file = “levels\_map.csv” | 146f04f7f9d207d35649867c7eb7da4d |
| 0003.0008 | 2018-09-14 17:35:50 | seed | project\_seed | - |
| 0003.0009 | 2018-09-14 17:36:31 | info | sessionInfo | - |
| 0003.0010 | 2018-09-14 17:36:31 | save | save | - |

1. UT Health, Department of Epidemiology and Biostatistics [↑](#footnote-ref-2)