Kidney Cancer Data Exploration

KL2 Aim 2

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

October 01, 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.

###### TOC

| **Note:** 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 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.  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 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 needed for our records.  Verbatim names of files, variables, or values are displayed in a special style, like this. Variable names are in addition linked to a glossary at the end of this document. This is where any relevant cleaning or tranformation steps will in the future be described for the respective variables. Tables, figures, and sections are also linked from text that references them. To follow a link in Word, please hold down the ‘control’ key and click on that link. | * [1 Overview](#sec:overview) * [2 Data preparation](#sec:dataprep) * [3 Plots of test data](#sec:descplots) * [4 Cohort characterization](#sec:cohorchar) * [5 Conclusion & next steps](#sec:nextsteps) * Appendices   + [A1. Stage/grade export sample](#sec:stage)   + [A2. TODO list](#sec:todo)   + [A3. Supplementary results](#sec:supp)   + [A4. Variable descriptions](#sec:vars)   + [A5. Audit trail](#sec:audit) |
| --- | --- |

# 1 Overview

A recent study of state death records[1](#ref-PinheiroHighcancermortality2017) reports that among US-born Texans of Hispanic ancestry (7.3 million, 27% of the State’s population), annual age-adjusted mortality rates for kidney cancer are 1.5-fold and 1.4-fold those of non-Hispanic whites for males and females respectively. My goal is to determine whether these findings can be replicated at UT Health (Aim 2) and Massachusetts General Hospital (Aim 3). If there is evidence for an ethnic disparity, I will look for possible *mediators* of this disparity among socioeconomic, lifestyle, and family history variables (Aim 2a). Otherwise the focus will shift to determining which of these same variables are the best *predictors* of mortality and recurrence.

At the Clinical Informatics Research Division (CIRD) we operate an i2b2[2](#ref-MurphyInstrumentinghealthcare2009) data warehouse containing deidentified data for over 1.3 million patients from the electronic medical record (EMR) systems of the UT Health faculty practice and the University Health System (UHS) county hospital. We use the HERON[3](#ref-AdagarlaSEINEMethodsElectronic2015) extract transform load (ETL) process to link data from multiple sources including copies of monthly reports that the Mays Cancer Center sends to the Texas Cancer Registry with detailed information on cancer cases including dates of diagnosis, surgery, and recurrence along with stage and grade at presentation. My first-pass eligibility query returned 2327 patients having one or more of the following in their records: an ICD9 code of 189.0 or any ICD10 code starting with C64; the NAACCR item [0400 Primary Site](http://datadictionary.naaccr.org/?c=10#400) having a value starting with C64 ([Kidney, NOS](#n_kcancer)); or the SEER Primary Site having a value of [Kidney and Renal Pelvis](#n_seer_kcancer).

My second pass criteria aimed at finding among these patients those for which NAACCR records also exist, defined as having a non-missing [0390 Date of Diagnosis](#n_ddiag) and at least one of [Kidney, NOS](#n_kcancer) or [Kidney and Renal Pelvis](#n_seer_kcancer). As can be seen from [table I](#tbl:cohortrectype) only 486 of the patient-set met these criteria and 1841 did not. In [Appendix 3.2.1](#sec:diag)-[Appendix 3.2.3](#sec:recur) I identified additional exclusion criteria which I will implement in the next major revision of this document.

In [sec. 2.1](#sec:linkagever) I summarize the evidence that NAACCR and EMR records are correctly matched with each other. In [sec. 2.2](#sec:reqelmnts) I summarize the minimum set of NAACCR data elements that is sufficient to replicate my analysis in an independent NAACCR data set. In [sec. 2.3](#sec:merging) I report the extent to which the completeness of NAACCR records can be improved by using EMR records of the same patients. In [sec. 3](#sec:descplots) is a technical demonstration of the data analysis scripts (on a small random sample). In [sec. 4](#sec:cohorchar) there is a characterization of the full (N=2327) patient cohort. Finally, in [sec. 5](#sec:nextsteps) I present my plans for overcoming the data issues I found, replicating the analysis on independent data, preparing additional variables, and starting work on Aim 1.

# 2 Data preparation

## 2.1 Verifying correct patient linkage

Since this is the first study at our site to make such extensive use of combined EMR and NAACCR data, it is important to first validate the data linkage done by our ETL.

The following data elements exist in both NAACCR and the EMR, respectively: date of birth ([0240 Date of Birth](#n_dob) and [birth\_date](#birth_date)), marital status ([0150 Marital Status at DX](#n_marital) and [Marital Status](#e_marital)), sex ([0220 Sex](#n_sex) and [sex\_cd](#sex_cd)), race ([Race (NAACCR 0160-0164)](#a_n_race) and [race\_cd](#race_cd)), and Hispanic ethnicity ([0190 Spanish/Hispanic Origin](#n_hisp) and [Hispanic or Latino](#e_hisp)). The agreement between NAACCR and the EMR is never going to be 100% for these variables because of errors in the original source data as well as variation in how a patient chooses to self-report from one visit to another (race, Hispanic ancestry, and marital status are expected to be especially variable). Nonetheless, if patient counts for NAACCR and EMR are tabulated against each variable, then *most* of the values should agree.

I have confirmed that this *is* the case for marital status ([Appendix 3.1.1](#sec:xcmarital)), sex ([Appendix 3.1.2](#sec:xcsex)), race ([Appendix 3.1.3](#sec:xcrace)), and Hispanic ancestry ([Appendix 3.1.4](#sec:xchisp)). Furthermore, there are 0 eligible patients lacking a [0240 Date of Birth](#n_dob) and only 15 with a mismatch between [0240 Date of Birth](#n_dob) and [birth\_date](#birth_date). Independent evidence for correct linkage is that EMR ICD9/10 codes for primary kidney cancer rarely precede [0390 Date of Diagnosis](#n_ddiag) ([fig. 5](#fig:diag_plot)), EMR surgical history of nephrectomy and ICD9/10 codes for acquired absence of a kidney rarely precede [1200 RX Date--Surgery](#n_dsurg) or [3170 RX Date--Most Defin Surg](#n_rx3170) ([fig. 6](#fig:surg0_plot0)), and death dates from non-NAACCR sources ([Death, i2b2](#e_death), [Deceased per SSA](#s_death) , and [Expired[7,579 facts; 7,544 patients]](#e_dscdeath)) rarely precede [1760 Vital Status](#n_vtstat) ([fig. 10](#fig:death_plot)).

## 2.2 Required NAACCR data elements.

The primary outcome variables I need are date of initial diagnosis, date of surgery (if any), date of recurrence (if any), and date of death (if any). The primary predictor variable is whether or not a patient is Hispanic. There are many covariates of interest, but these five values are the scaffolding on which the rest of the analysis will be built.

**I found the following NAACCR fields sufficient for deriving all the above analytic variables:** [**0190 Spanish/Hispanic Origin**](#n_hisp)**,** [**1880 Recurrence Type--1st**](#n_rectype)**,** [**3170 RX Date--Most Defin Surg**](#n_rx3170)**,** [**1340 Reason for No Surgery**](#n_surgreason)**,** [**0390 Date of Diagnosis**](#n_ddiag)**,** [**1200 RX Date--Surgery**](#n_dsurg)**,** [**1750 Date of Last Contact**](#n_lc)**,** [**1760 Vital Status**](#n_vtstat)**,** [**1770 Cancer Status**](#n_cstatus)**,** [**1860 Recurrence Date--1st**](#n_drecur)**,** [**Kidney and Renal Pelvis**](#n_seer_kcancer)**, and** [**Kidney, NOS**](#n_kcancer)**.** More details about how these were selected can be found in the [“Which EMR and NAACCR variables are reliable event indicators?”] section. In addition the following will almost certainly be needed for the covariates: [0220 Sex](#n_sex), [0240 Date of Birth](#n_dob), [0150 Marital Status at DX](#n_marital) , [0250 Birthplace](#n_brthplc) and any field whose name contains Race , Comorbid/Complication, Derived AJCC, or TNM. For crosschecking purposes it may also be useful to have [2850 CS Mets at DX](#n_mets), [0580 Date of 1st Contact](#n_fc) , and [0446 Multiplicity Counter](#n_mult). Additional items are likely to be needed as this project evolves, but **I believe the elements listed so far will be sufficient to replicate my analysis de-identified State or National NAACCR data**.

## 2.3 Merging NAACCR and EMR variables

EMR records can not only enrich the data with additional elements not available in NAACCR alone, but might also make it possible to fill in missing [0390 Date of Diagnosis](#n_ddiag), [3170 RX Date--Most Defin Surg](#n_rx3170) / [1200 RX Date--Surgery](#n_dsurg), [1860 Recurrence Date--1st](#n_drecur), and [1750 Date of Last Contact](#n_lc) values. It may even be possible to reconstruct entire records for the 1841 kidney cancer patients in the EMR lacking NAACCR records. This depends on how much the EMR and NAACCR versions of a variable agree when neither is missing.

**Data elements representing date of death and Hispanic ethnicity are in sufficient agreement (** [**sec. Appendix 3.1.4**](#sec:xchisp) **and** [**sec. Appendix 3.2.4**](#sec:death) **) to justify merging information from the EMR and NAACCR.** The process for combining them is described in the [Death](#a_tdeath), [Hispanic (strict)](#a_hsp_strict), and [Hispanic (broad)](#a_hsp_broad) sections of [Appendix 4](#sec:vars) respectively. At this time I cannot merge diagnosis, surgery, or recurrence– where data from both sources is available I found that EMR dates lag considerably behind NAACCR dates ( secs. [Appendix 3.2.1](#sec:diag)-[Appendix 3.2.3](#sec:recur) ) and their variability is probably larger than the effect size. The surgery and recurrence lags might be because those actual visits are not yet available in the data warehouse and I am only seeing them as reflected in the patient history at visits long after the fact. The diagnosis lag may be due to the decision to proceed with surgery often being made based on imaging data,[4](#ref-pcRodriguez2018) with definitive pathology results only available after surgery ([sec. Appendix 3.2.2](#sec:surg)). Attempting to merge these elements would bias the data and obscure the actual differences. However there are several ways forward that I will discuss in [sec. 5](#sec:nextsteps) below.

EMR data can still be used to flag records for exclusion or verification via chart review if, despite the lag, EMR codes for kidney cancer or secondary tumors precede [Diagnosis](#a_tdiag) or [Recurrence](#a_trecur) respectively. This can also apply to nephrectomy EMR codes and [Surgery](#a_tsurg) but I will need to distinguish between the prior nephrectomy being due cancer versus other indications.

For now I am analyzing the data as if I only have access to NAACCR except mortality where I do it both with ( [fig. 3](#fig:naaccrdeath_survfit) ) and without ( [fig. 4](#fig:alldeath_survfit) ) the EMR.

# 3 Plots of test data

The point of this section is **solely** to confirm that my scripts succeeded in turning the data into a time-to-event (TTE) format to which Kaplan-Meier curves can be fit without numeric errors or grossly implausible results. All the results below are from a small random subset of the data– N=127, 82 Hispanic and NA non-Hispanic. There were 5 excluded. This is further reduced in some cases as described in the figure captions. These sample sizes doe not have sufficient power to detect clinically significant differences and **this is not the goal yet**. Again, the intent is only to insure that my software performs correctly without viewing the hold-out data on which the hypothesis testing will ultimately be done.

Furthermore, these survival curves are not yet adjusted for covariates such as age and stage at diagnosis. Finally, there are upcoming refinements to [Hispanic (NAACCR)](#a_hsp_naaccr) and the exclusion criteria which I discuss below in [sec. 5](#sec:nextsteps).

In all the plots below, the time is expressed in weeks and + signs denote censored events (the last follow-up of patients for whom the respective outcomes were never observed). The lightly-shaded regions around each line are 95% confidence intervals. In all cases a 3-year follow-up period is represented on the x-axis meaning patients for whom the outcomes occurred beyond that are censored at 3 years. The vertical scales for the plots vary and are automatically determined by the data.

###### blank

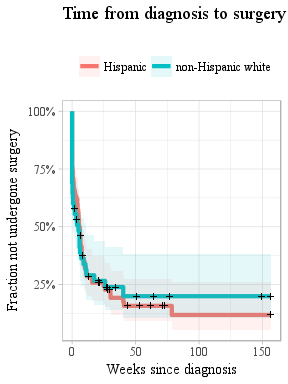


Figure 1: Number of weeks elapsed from [Diagnosis](#a_tdiag) (time 0) to [Surgery](#a_tsurg) for 82 Hispanic and 45 non-Hispanic white patients.

###### blank

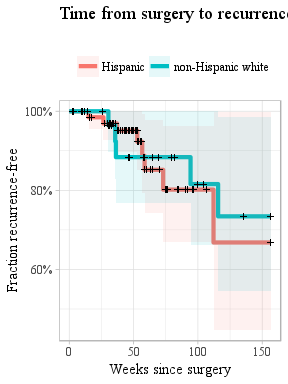


Figure 2: Number of weeks elapsed from [Surgery](#a_tsurg) (time 0) to [Recurrence](#a_trecur) for 67 Hispanic and 34 non-Hispanic white patients. The numbers are lower than for [fig. 1](#fig:surg_survfit) because patients not undergoing surgery are excluded.

###### blank

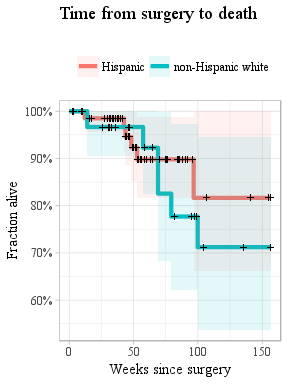


Figure 3: Like [fig. 2](#fig:recur_survfit) except now the outcome is [1760 Vital Status](#n_vtstat) for 67 Hispanic and 34 non-Hispanic white patients.

###### blank

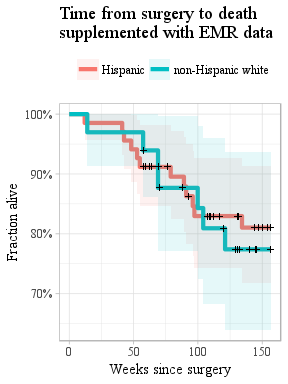


Figure 4: Like [fig. 3](#fig:naaccrdeath_survfit) but now supplemented EMR information to see how much of a difference it makes. For the predictor [Hispanic (broad)](#a_hsp_broad) is used instead of [Hispanic (NAACCR)](#a_hsp_naaccr) and for the outcome [Death](#a_tdeath) is used instead of [1760 Vital Status](#n_vtstat) . There were 68 Hispanic and 33 non-Hispanic white patients. There were 7 fewer censored events than in [fig. 3](#fig:naaccrdeath_survfit) which may improve sensitivity in the actual analysis.

# 4 Cohort Characterization

The below variables are subject to change as the data validation and preparation processes evolve.

Table I: Summary of all the variables in the combined i2b2/NAACCR set broken up by [Recurrence Status](#a_n_recur). Disease-free and Never disease-free have the same meanings as codes 00 and 70 in the [NAACCR definition](http://datadictionary.naaccr.org/?c=10#1880) for [1880 Recurrence Type--1st](#n_rectype). Recurred is any code other than (00, 70, or 99), and Unknown if recurred or was ever gone is 99. Not 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 it does not indicate kidney as the site

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Disease-free | Never disease-free | Recurred | Unknown if recurred or was ever gone | Not in NAACCR |
| **n** | 160 | 211 | 95 | 20 | 1841 |
| **Age at Last Contact, combined (mean (sd))** | 54.32 (20.42) | 63.43 (13.76) | 62.51 (15.23) | 55.59 (23.01) | 61.34 (14.18) |
| **a\_hsp\_broad (%)** |  |  |  |  |  |
| Hispanic | 106 ( 66.2) | 116 ( 55.0) | 50 ( 52.6) | 8 ( 40.0) | 857 (46.6) |
| non-Hispanic white | 47 ( 29.4) | 75 ( 35.5) | 42 ( 44.2) | 10 ( 50.0) | 525 (28.5) |
| Other | 3 ( 1.9) | 17 ( 8.1) | 3 ( 3.2) | 1 ( 5.0) | 13 ( 0.7) |
| Unknown | 4 ( 2.5) | 3 ( 1.4) | 0 | 1 ( 5.0) | 364 (19.8) |
| NA | 0 | 0 | 0 | 0 | 82 ( 4.5) |
| **a\_hsp\_naaccr (%)** |  |  |  |  |  |
| Hispanic | 100 ( 62.5) | 114 ( 54.0) | 46 ( 48.4) | 8 ( 40.0) | 86 ( 4.7) |
| non-Hispanic white | 50 ( 31.2) | 74 ( 35.1) | 45 ( 47.4) | 10 ( 50.0) | 84 ( 4.6) |
| Other | 4 ( 2.5) | 18 ( 8.5) | 2 ( 2.1) | 1 ( 5.0) | 14 ( 0.8) |
| Unknown | 6 ( 3.8) | 5 ( 2.4) | 2 ( 2.1) | 1 ( 5.0) | 3 ( 0.2) |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |
| **a\_hsp\_strict (%)** |  |  |  |  |  |
| Hispanic | 62 ( 38.8) | 68 ( 32.2) | 27 ( 28.4) | 6 ( 30.0) | 562 (30.5) |
| non-Hispanic white | 29 ( 18.1) | 64 ( 30.3) | 35 ( 36.8) | 9 ( 45.0) | 53 ( 2.9) |
| Other | 4 ( 2.5) | 12 ( 5.7) | 2 ( 2.1) | 1 ( 5.0) | 84 ( 4.6) |
| Unknown | 65 ( 40.6) | 67 ( 31.8) | 31 ( 32.6) | 4 ( 20.0) | 702 (38.1) |
| NA | 0 | 0 | 0 | 0 | 440 (23.9) |
| **a\_tdeath (%)** | 8 ( 5.0) | 99 ( 46.9) | 30 ( 31.6) | 3 ( 15.0) | 305 (16.6) |
| **a\_tdiag (%)** | 160 (100.0) | 211 (100.0) | 95 (100.0) | 20 (100.0) | 0 |
| **a\_trecur (%)** | 0 | 1 ( 0.5) | 83 ( 87.4) | 0 | 41 ( 2.2) |
| **a\_tsurg (%)** | 157 ( 98.1) | 113 ( 53.6) | 94 ( 98.9) | 13 ( 65.0) | 113 ( 6.1) |
| **BMI (mean (sd))** | 31.19 (8.34) | 27.77 (7.26) | 29.32 (7.11) | 29.66 (9.92) | 30.63 (9.31) |
| **Deceased, EMR (%)** | 7 ( 4.4) | 90 ( 42.7) | 22 ( 23.2) | 3 ( 15.0) | 298 (16.2) |
| **Deceased, Registry (%)** | 1 ( 0.6) | 71 ( 33.6) | 18 ( 18.9) | 3 ( 15.0) | 43 ( 2.3) |
| **Deceased, SSN (%)** | 1 ( 0.6) | 12 ( 5.7) | 5 ( 5.3) | 0 | 89 ( 4.8) |
| **Diabetes, i2b2 (%)** | 56 ( 35.0) | 54 ( 25.6) | 27 ( 28.4) | 1 ( 5.0) | 585 (31.8) |
| **Diabetes, Registry (%)** | 31 ( 19.4) | 26 ( 12.3) | 8 ( 8.4) | 0 | 26 ( 1.4) |
| **Hispanic, i2b2 (%)** | 92 ( 57.5) | 96 ( 45.5) | 43 ( 45.3) | 7 ( 35.0) | 746 (40.5) |
| **Hispanic, Registry (%)** |  |  |  |  |  |
| Non\_Hispanic | 54 ( 33.8) | 92 ( 43.6) | 47 ( 49.5) | 11 ( 55.0) | 98 ( 5.3) |
| Unknown | 6 ( 3.8) | 5 ( 2.4) | 2 ( 2.1) | 1 ( 5.0) | 3 ( 0.2) |
| Hispanic\_NOS | 86 ( 53.8) | 96 ( 45.5) | 43 ( 45.3) | 8 ( 40.0) | 67 ( 3.6) |
| Mexican | 13 ( 8.1) | 17 ( 8.1) | 1 ( 1.1) | 0 | 17 ( 0.9) |
| Spanish\_Surname | 0 | 1 ( 0.5) | 1 ( 1.1) | 0 | 2 ( 0.1) |
| Cuban | 1 ( 0.6) | 0 | 0 | 0 | 0 |
| S\_Ctr\_America | 0 | 0 | 1 ( 1.1) | 0 | 0 |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |
| **Insurance, Registry (%)** |  |  |  |  |  |
| Not Insured | 17 ( 10.6) | 21 ( 10.0) | 7 ( 7.4) | 2 ( 10.0) | 17 ( 0.9) |
| Self-Pay | 22 ( 13.8) | 21 ( 10.0) | 15 ( 15.8) | 0 | 14 ( 0.8) |
| Insurance NOS | 1 ( 0.6) | 5 ( 2.4) | 0 | 0 | 1 ( 0.1) |
| Managed Care HMO / PPO | 56 ( 35.0) | 53 ( 25.1) | 28 ( 29.5) | 10 ( 50.0) | 40 ( 2.2) |
| Private Fee-for-Svc | 0 | 1 ( 0.5) | 0 | 0 | 0 |
| Medicaid | 10 ( 6.2) | 14 ( 6.6) | 1 ( 1.1) | 0 | 10 ( 0.5) |
| Medicaid Mgd. Care Pln. | 14 ( 8.8) | 6 ( 2.8) | 6 ( 6.3) | 3 ( 15.0) | 10 ( 0.5) |
| Medicare/Medicaid NOS | 13 ( 8.1) | 30 ( 14.2) | 12 ( 12.6) | 1 ( 5.0) | 36 ( 2.0) |
| Medicare w Suppl. NOS | 3 ( 1.9) | 2 ( 0.9) | 2 ( 2.1) | 0 | 6 ( 0.3) |
| Medicare Mgd. Care Pln. | 9 ( 5.6) | 16 ( 7.6) | 7 ( 7.4) | 3 ( 15.0) | 13 ( 0.7) |
| Medicare w Private Suppl. | 5 ( 3.1) | 22 ( 10.4) | 9 ( 9.5) | 0 | 20 ( 1.1) |
| Medicare w Medicaid | 3 ( 1.9) | 5 ( 2.4) | 2 ( 2.1) | 0 | 7 ( 0.4) |
| TriCare | 3 ( 1.9) | 1 ( 0.5) | 0 | 0 | 4 ( 0.2) |
| VA | 1 ( 0.6) | 7 ( 3.3) | 1 ( 1.1) | 0 | 3 ( 0.2) |
| Unknown | 3 ( 1.9) | 7 ( 3.3) | 5 ( 5.3) | 1 ( 5.0) | 6 ( 0.3) |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |
| **Kidney Cancer, i2b2 (%)** | 152 ( 95.0) | 193 ( 91.5) | 85 ( 89.5) | 17 ( 85.0) | 1729 (93.9) |
| **Kidney Cancer, Registry (%)** | 156 ( 97.5) | 204 ( 96.7) | 87 ( 91.6) | 19 ( 95.0) | 20 ( 1.1) |
| **Language, i2b2 (%)** |  |  |  |  |  |
| English | 128 ( 80.0) | 173 ( 82.0) | 84 ( 88.4) | 19 ( 95.0) | 1588 (86.3) |
| Spanish | 31 ( 19.4) | 29 ( 13.7) | 7 ( 7.4) | 1 ( 5.0) | 213 (11.6) |
| Other | 0 | 3 ( 1.4) | 0 | 0 | 4 ( 0.2) |
| Unknown | 1 ( 0.6) | 6 ( 2.8) | 4 ( 4.2) | 0 | 36 ( 2.0) |
| **Marital Status, Registry (%)** |  |  |  |  |  |
| Divorced | 13 ( 8.1) | 16 ( 7.6) | 11 ( 11.6) | 0 | 16 ( 0.9) |
| Separated | 8 ( 5.0) | 2 ( 0.9) | 1 ( 1.1) | 2 ( 10.0) | 6 ( 0.3) |
| Married | 79 ( 49.4) | 125 ( 59.2) | 56 ( 58.9) | 7 ( 35.0) | 102 ( 5.5) |
| Domestic Partner | 0 | 0 | 0 | 0 | 0 |
| Single | 39 ( 24.4) | 30 ( 14.2) | 16 ( 16.8) | 9 ( 45.0) | 32 ( 1.7) |
| Unknown | 15 ( 9.4) | 24 ( 11.4) | 8 ( 8.4) | 2 ( 10.0) | 17 ( 0.9) |
| Widowed | 6 ( 3.8) | 14 ( 6.6) | 3 ( 3.2) | 0 | 14 ( 0.8) |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |
| **n\_cstatus (%)** |  |  |  |  |  |
| Tumor\_Free | 160 (100.0) | 1 ( 0.5) | 7 ( 7.4) | 0 | 58 ( 3.2) |
| Tumor | 0 | 210 ( 99.5) | 81 ( 85.3) | 0 | 114 ( 6.2) |
| Unknown | 0 | 0 | 7 ( 7.4) | 20 (100.0) | 15 ( 0.8) |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |
| **Race, i2b2 (%)** |  |  |  |  |  |
| White | 149 ( 93.1) | 185 ( 87.7) | 87 ( 91.6) | 19 ( 95.0) | 1566 (85.1) |
| Black | 3 ( 1.9) | 10 ( 4.7) | 3 ( 3.2) | 1 ( 5.0) | 95 ( 5.2) |
| Asian | 3 ( 1.9) | 6 ( 2.8) | 0 | 0 | 13 ( 0.7) |
| Pac Islander | 0 | 0 | 0 | 0 | 1 ( 0.1) |
| Other | 0 | 3 ( 1.4) | 0 | 0 | 46 ( 2.5) |
| Unknown | 5 ( 3.1) | 7 ( 3.3) | 5 ( 5.3) | 0 | 120 ( 6.5) |
| **Race, Registry (%)** |  |  |  |  |  |
| White | 153 ( 95.6) | 188 ( 89.1) | 91 ( 95.8) | 18 ( 90.0) | 170 ( 9.2) |
| Black | 3 ( 1.9) | 10 ( 4.7) | 2 ( 2.1) | 1 ( 5.0) | 11 ( 0.6) |
| Asian | 1 ( 0.6) | 3 ( 1.4) | 0 | 0 | 2 ( 0.1) |
| Pac Islander | 0 | 1 ( 0.5) | 0 | 0 | 0 |
| Other | 0 | 4 ( 1.9) | 0 | 0 | 0 |
| Unknown | 3 ( 1.9) | 5 ( 2.4) | 2 ( 2.1) | 1 ( 5.0) | 4 ( 0.2) |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |
| **Sex, i2b2 (%)** |  |  |  |  |  |
| m | 100 ( 62.5) | 151 ( 71.6) | 63 ( 66.3) | 13 ( 65.0) | 1047 (56.9) |
| f | 60 ( 37.5) | 60 ( 28.4) | 32 ( 33.7) | 7 ( 35.0) | 793 (43.1) |
| u | 0 | 0 | 0 | 0 | 1 ( 0.1) |
| **Sex, Registry (%)** |  |  |  |  |  |
| m | 98 ( 61.3) | 149 ( 70.6) | 63 ( 66.3) | 13 ( 65.0) | 106 ( 5.8) |
| f | 62 ( 38.8) | 62 ( 29.4) | 32 ( 33.7) | 7 ( 35.0) | 81 ( 4.4) |
| NA | 0 | 0 | 0 | 0 | 1654 (89.8) |

# 5 Conclusion and next steps

This detailed investigation of the available data elements and development of analysis scripts opens the following directions: more data, *external* data, more covariates, and improved pre-processing at the i2b2 end (Aim 1).

More data can be acquired by reclaiming values that are currently inconsistent or missing. There are various ad-hoc consistency checks described in sections. I need to gather these checks in one place and systematically run them on every patient to get a total count of records that need manual chart review (Dr. Rodriguez’s protocol) and for each record a list of issues to resolve.

To reclaim missing values I will need to solve the problem of lag and disagreement between the EMR and NAACCR ([sec. 2.3](#sec:merging)). I will meet with the MCC NAACCR registrar and learn where exactly in the EMR and other sources she looks to abstract [1880 Recurrence Type--1st](#n_rectype), [3170 RX Date--Most Defin Surg](#n_rx3170), [1340 Reason for No Surgery](#n_surgreason), [0390 Date of Diagnosis](#n_ddiag), [1200 RX Date--Surgery](#n_dsurg), [1750 Date of Last Contact](#n_lc), [1760 Vital Status](#n_vtstat), [1770 Cancer Status](#n_cstatus), [1860 Recurrence Date--1st](#n_drecur), [Kidney and Renal Pelvis](#n_seer_kcancer), and [Kidney, NOS](#n_kcancer). I will also meet with personnel experienced in Urology chart review to learn their methods. This may lead to revisions of the CIRD process for loading UHS, UT Health, and NAACCR data into i2b2 (IRB protocol HSC20150212HR). As per Dr. Rodriguez I already plan on adding all ICD codes for ‘renal mass’ to my i2b2 query ([sec. Appendix 3.2.1](#sec:diag)). Meanwhile, in response to researcher questions including my own, CIRD has identified thousands of NAACCR entries and surgery billing records that got excluded from i2b2 because they are not associated with visits to UT Health clinics. This problem has been corrected. After the next i2b2 refresh I expect an increased number of patients and better agreement of surgery dates between EMR and NAACCR.

For external data I will request non-aggregated limited/deidentified records from the Texas Cancer Registry. I will also look at the NCDB dataset obtained by Urology to see if it has the elements listed in [sec. 2.2](#sec:reqelmnts).

In the remainder of Aim 2 and Aim 3 I will need the following additional variables: (NAACCR only) stage and grade; (EMR only) analgesics, smoking and alcohol, family history of cancer or diabetes, lab results, vital signs, Miperamine (as per Dr. Michalek), frequency of lab and image orders, frequency and duration of visits, and participation in adjuvant trials; (both) birthplace, language, and diabetes; and (census data in i2b2) income and education. Each of these will require a workup similar to that reported in [sec. 2](#sec:dataprep) and refs. I can work independently on many of these but I will need guidance from experts in Urology on interpreting the stage and grade data. If genomic data from the Urology biorepository becomes available for these patients in the course of this study it also will become an important variable for Aim 2.

The use of TCR or NCDB data is *not* a substitute for UT Health and MGH i2b2 data. The registries allow me to test the replicability of high-level findings to State and National populations but they will not have the detailed additional variables I will need to investigate the causes of disparate patient outcomes.

Nor are the R scripts I wrote for this project a substitute for DataFinisher[5](#ref-bokov_denormalize_2016) and i2b2. On the contrary, the reason I was able to make this much progress in one month is that the data linkage and de-identification was done by the CIRD i2b2 ETL, the data selection was simplified by the i2b2 web client, and an enormous amount of post-processing was done by my DataFinisher app which is integrated into our local i2b2. In writing my scripts I have identified a number of additional post-processing steps that generalize to other studies and I will integrate those into DataFinisher so that the data it outputs is even more analysis-ready. This, in turn, will make DataFinisher more valuable for other sites to adopt and simplify logistics of Aim 3.

While I am incorporating the new methods into DataFinisher, I will also reorganize and document the code so I can present it to Dr. Murphy and his informatics team for review and input.

# 6 References

1. Pinheiro, P. S. *et al.* High cancer mortality for US-born Latinos: Evidence from California and Texas. *BMC Cancer***17,** (2017).

2. Murphy, S. *et al.* Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Research***19,** 1675–1681 (2009).

3. Adagarla, B. *et al.* SEINE: Methods for Electronic Data Capture and Integrated Data Repository Synthesis with Patient Registry Use Cases. (2015).

4. Rodriguez, R. *personal communication* (2018).

5. Bokov, A., Manuel, L., Cheng, C., Bos, A. & Tirado-Ramos, A. Denormalize and Delimit: How not to Make Data Extraction for Analysis More Complex than Necessary. *Procedia Computer Science***80,** 1033–1041 (2016).

# Appendix 1 : Example of stage/grade data

Table II: This is proof of feasibility for extracting stage and grade at diagnosis for each NAACCR patient for import into the EMR system (e.g. Epic/Beacon). Clinical and pathology stage descriptors are also available in NAACCR. Here the [Patient Number (anonymized)](#patient_num) are de-identified but with proper authorization they can be mapped to MRNs or internal database index keys.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 2 : Next steps

* TODO: Update and clean up the plots and tables, including labels.
  + ~~[Consistency-Checks]~~
    - ~~Marital status, sex, race, hispanic(2):\*\* shorten text and move to captions.~~
    - ~~Write motivation and summary.~~
  + [~~Testing/Interpreting Variables~~](#which-emr-and-naaccr-variables-are-reliable-event-indicators)
    - ~~Write motivation, intro, summary. Incorporate edits.~~
    - [Initial diagnosis](#sec:diag), Surgery, [Re-occurrence](#sec:recur), [Death](#sec:death)
      * Move plots to the top of each
      * Shorten text and move to captions.
      * For each plot state what the conclusions are.
      * concise paragraph.
  + [Hispanic variable recoding](#whether-or-not-the-patient-is-hispanic): turn into paragraphs, think about moving to variable glossary.
  + [Descriptive Plots (Preliminary)]
    - ~~Move them to right after the overview.~~
    - ~~Write intro mentioning that these are the relationships of interest among the four main variables.~~
    - ~~Expand why there are two versions of the survival plot~~
* DONE: ~~Update and streamline the narrative.~~
  + ~~Intro~~
  + ~~Motivation~~
  + ~~Summary of results~~
  + ~~Summary of next steps~~
  + ~~Move questions to after the [Descriptive Plots (Preliminary)] but before the [Consistency-Checks], and place the answered questions at the bottom.~~ [~~Domain expert questions~~](#questions-for-mentors-and-other-domain-experts) still go ahead of [empirical questions](#questions-to-answer-empirically).
* TODO: Remove the crossed-off stuff in [Appendix III: Supplementary tables] ’ but note someplace what was removed and why.
* TODO: Organize the inclusion/exclusion criteria into a single named list
* TODO: Overhaul the existing TableOne in [Cohort Characterization](#sec:cohorchar) – use data dictionary for renaming instead of *ad-hoc* .
* TODO: Migrate everything that uses ~~dat2~~ and dat3 to using dat2a.
* TODO: Create a TableOne for [Hispanic (NAACCR)](#a_hsp_naaccr) (that specific one because then the conclusions can be directly applied to TCR data) to find possible confounding variables. Age, perhaps? Income?
* TODO: Fill in more of the variable descriptions in [Appendix IV: Variable descriptions]
* 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#580) ~~as the diagnosis date if earlier than~~ [~~0390 Date of Diagnosis~~](#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#1260)
    - [~~1270 Date of 1st Crs RX--CoC~~](http://datadictionary.naaccr.org/default.aspx?c=10#1270)
    - [~~3170 RX Date--Most Defin Surg~~](http://datadictionary.naaccr.org/default.aspx?c=10#3170)
  + DONE: ~~Recurrence:~~ [~~1880 Recurrence Type--1st~~](http://datadictionary.naaccr.org/default.aspx?c=10#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)
* TODO: Separate script-level calls to instrequire() to reduce the number of libraries that get loaded unnecessarily.
* TODO: Create a light version of data.R.rdata that has only the minimal necessary stuff for, e.g. exploration.R
* DONE: ~~Create combined (if applicable) variables for each of the following:~~
  + ~~Initial diagnosis~~[Diagnosis](#a_tdiag), [``][a\_cdiag]
  + ~~Surgery~~[Surgery](#a_tsurg), [``][a\_csurg]
  + ~~Re-ocurrence~~[Recurrence](#a_trecur), [``][a\_crecur]
  + *~~Last follow-up ?~~*
  + ~~Death~~[Death](#a_tdeath), [``][a\_cdeath]
  + ~~Strict Hispanic designator~~[Hispanic (strict)](#a_hsp_strict)
  + ~~Lenient Hispanic designator~~[Hispanic (broad)](#a_hsp_broad)
  + ~~NAACCR-only Hispanic designator~~[Hispanic (NAACCR)](#a_hsp_naaccr)
* DONE: ~~Verify that the~~ [~~ETL~~](http://www.hostedredmine.com/issues/719444#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#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 3 Supplementary results

### Appendix 3.1.1 Marital Status

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

Table III: Marital status from NAACCR (columns) and EMR (rows).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

### Appendix 3.1.2 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 |

### Appendix 3.1.3 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 |

### Appendix 3.1.4 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 |

### Appendix 3.1.5 Birth Date

If we look at all patients with NAACCR records rather than just the ones meeting inclusion criteria, the number of birth date mismatches is still low: 24. There are also a few [0240 Date of Birth](#n_dob) birthdates for patients who do *not* have an [0390 Date of Diagnosis](#n_ddiag).

A total of 24 patients have a mismatch between their NAACCR and EMR birthdates and 15 are part of the

Below is a summary of [birth\_date](#birth_date) minus [0240 Date of Birth](#n_dob) for the 15 patients with non-matching dates of birth.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 below in [sec. Appendix 3.1.5.1](#sec:xcdobdem) the 24 total patients with DOB mismatches are not particularly enriched for other mismatches tested so far which is consistent with isolated errors in those respective variables rather than some subset of patients continuing to be incorrectly linked.

#### 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\_inactive** | 0 | 0 | 0 | 15 |
| **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\_i9neph** | 0 | 0 | 7 | 8 |
| **e\_hstneph** | 0 | 0 | 4 | 11 |
| **e\_surgonc** | 0 | 0 | 0 | 15 |
| **e\_i10neph** | 0 | 0 | 7 | 8 |

## Appendix 3.2 Which EMR and NAACCR variables are reliable event indicators?

When multiple elements in the raw data refer to the same observation, a decisions and they may need to be merged into a single variable for analysis… assuming that they Whey the observation is an event like

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.

### Appendix 3.2.1 Initial diagnosis

The c\_kcdiag group of columns in dct0.

* NAACCR: [0390 Date of Diagnosis](#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#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** [**0390 Date of Diagnosis**](#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 [0390 Date of Diagnosis](#n_ddiag). Unfortunately, as can be seen from the below table, for the 486 patients who have non-missing [0390 Date of Diagnosis](#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 [0390 Date of Diagnosis](#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 [0390 Date of Diagnosis](#n_ddiag).
* May need to exclude all but the first occurence for patients with multiple NAACCR entries.
* Question: 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?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |

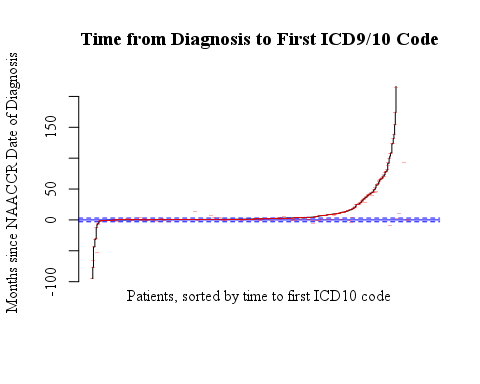


Figure 5: Here is a plot centered on [0390 Date of Diagnosis](#n_ddiag) (blue horizontal line at 0) with black lines indicating ICD10 codes for primary kidney cancer from the EMR and dashed red lines indicating ICD9 codes. The dashed horizontal blue lines indicate +- 3 months from [0390 Date of Diagnosis](#n_ddiag).

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.

### Appendix 3.2.2 Surgery

* NAACCR:
  + In addition to [1200 RX Date--Surgery](#n_dsurg)) the following possibly relevant fields are available in our local NAACCR:
    - [1260 Date of Initial RX--SEER](#n_rx1260)
    - [1270 Date of 1st Crs RX--CoC](#n_rx1270)
    - [3170 RX Date--Most Defin Surg](#n_rx3170)
  + Here are the questions raised:
    - Do they agree with [1200 RX Date--Surgery](#n_dsurg) sufficiently that missing [1200 RX Date--Surgery](#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 [1340 Reason for No Surgery](#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?
* Question: Where in the chart would one positively establish the date of the patient’s first nephrectomy…
  + …in Epic?
  + …in Sunrise?
* 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 do not undergo surgery?
  + Answer (RR): Around 15%

As can be seen in the table below, the variables [e\_i9neph][e\_i9neph], [e\_hstneph][e\_hstneph], [e\_surgonc][e\_surgonc], and [e\_i10neph][e\_i10neph] *sometimes* precede [0390 Date of Diagnosis](#n_ddiag) by many weeks. However, they *usually* follow [0390 Date of Diagnosis](#n_ddiag) by more weeks than the two NAACCR variables [3180 RX Date--Surgical Disch](#n_dsdisc) and [1200 RX Date--Surgery](#n_dsurg). Those two NAACCR variables never occur before [0390 Date of Diagnosis](#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 [0390 Date of Diagnosis](#n_ddiag) than the non-inactive codes.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA’s |
| **v080\_acqrd\_absnc\_inactive** | 2.571 | 83.29 | 205.3 | 233.5 | 375.4 | 708.3 | 477 |
| **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\_i9neph** | -361.1 | 8.143 | 31.43 | 69.5 | 82.71 | 957.4 | 261 |
| **e\_hstneph** | -91.86 | 10.11 | 37.07 | 77.85 | 93.96 | 758.1 | 318 |
| **e\_surgonc** | -194.9 | 0.2143 | 4.714 | 23.58 | 46 | 236.6 | 455 |
| **e\_i10neph** | -20.14 | 9.607 | 37.86 | 85.12 | 111.2 | 957.4 | 226 |
| **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 [0390 Date of Diagnosis](#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\_inactive** | 0 | 0 | 9 | 477 |
| **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\_i9neph** | 3 | 0 | 222 | 261 |
| **e\_hstneph** | 3 | 2 | 163 | 318 |
| **e\_surgonc** | 7 | 1 | 23 | 455 |
| **e\_i10neph** | 1 | 0 | 259 | 226 |

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.

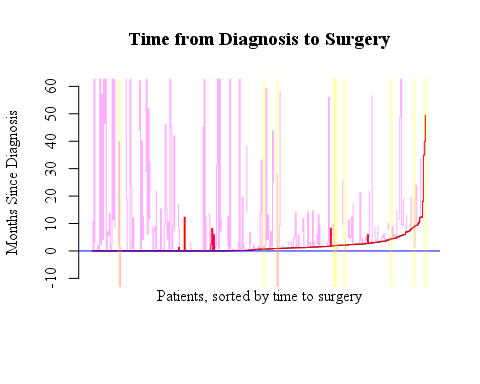


Figure 6: Above is a plot of all patients sorted by [1200 RX Date--Surgery](#n_dsurg) (black line). On the same axis is [3170 RX Date--Most Defin Surg](#n_rx3170) (red line) which is almost identical to [1200 RX Date--Surgery](#n_dsurg) except for a small number of cases where it occurs later than [1200 RX Date--Surgery](#n_dsurg) . It never occurs 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 [fig. 6](#fig:surg0_plot0) the 9 patients for which one or more EMR codes are recorded prior to [1200 RX Date--Surgery](#n_dsurg) are highlighted in yellow.

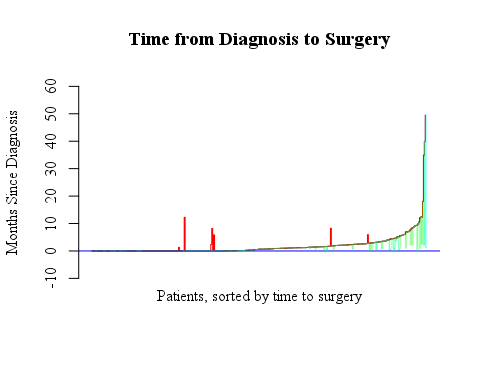


Figure 7: In the above plot the [1270 Date of 1st Crs RX--CoC](#n_rx1270) (green) and [1260 Date of Initial RX--SEER](#n_rx1260) (cyan) events are superimposed on time till [1200 RX Date--Surgery](#n_dsurg) like in [fig. 6](#fig:surg0_plot0) (but purple lines for nephrectomy EMR codes are omitted for readability). The [1270 Date of 1st Crs RX--CoC](#n_rx1270) and [1260 Date of Initial RX--SEER](#n_rx1260) variables trend earlier than [1200 RX Date--Surgery](#n_dsurg).

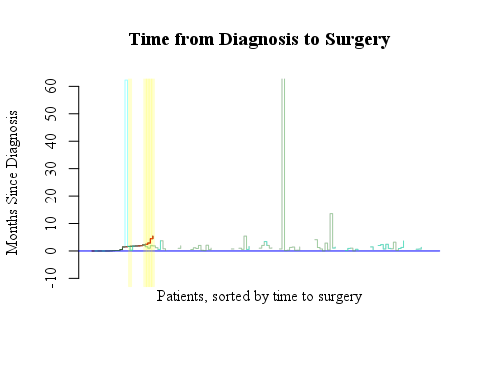


Figure 8: Above is a plot equivalent to [fig. 7](#fig:surg0_plot1) but for patients who do *not* have a [1340 Reason for No Surgery](#n_surgreason) code equal to Surgery Performed. There are many [1270 Date of 1st Crs RX--CoC](#n_rx1270) and [1260 Date of Initial RX--SEER](#n_rx1260) events but only a small number of [1200 RX Date--Surgery](#n_dsurg) (black) and [3170 RX Date--Most Defin Surg](#n_rx3170) (red). The [1200 RX Date--Surgery](#n_dsurg) and [3170 RX Date--Most Defin Surg](#n_rx3170) that do occur track each other perfectly. Together with NAACCR data dictionary’s description this suggests that [3170 RX Date--Most Defin Surg](#n_rx3170) is the correct principal surgery date in close agreement with [1200 RX Date--Surgery](#n_dsurg) , so perhaps missing [3170 RX Date--Most Defin Surg](#n_rx3170) values can be filled in from [1200 RX Date--Surgery](#n_dsurg) . However [1270 Date of 1st Crs RX--CoC](#n_rx1270) and [1260 Date of Initial RX--SEER](#n_rx1260) seem like non-primary surgeries or other events.

Table IV: Table of every NAACCR surgery event variable versus [1340 Reason for No Surgery](#n_surgreason)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | n\_dsurg = FALSE | n\_dsurg = TRUE | e\_i9neph = FALSE | e\_i9neph = TRUE | e\_hstneph = FALSE | e\_hstneph = TRUE | e\_surgonc = FALSE | e\_surgonc = TRUE |
| **Surgery Performed** | 14 | 458 | 226 | 246 | 281 | 191 | 433 | 39 |
| **Surgery Not First Course** | 122 | 24 | 130 | 16 | 129 | 17 | 133 | 13 |
| **No Surgery, Contra Indicated** | 16 | 2 | 15 | 3 | 15 | 3 | 16 | 2 |
| **No Surgery, Deceased** | 4 | 0 | 4 | 0 | 4 | 0 | 4 | 0 |
| **No Surgery, No Reason Given** | 5 | 0 | 3 | 2 | 2 | 3 | 4 | 1 |
| **No Surgery, Refused** | 4 | 4 | 8 | 0 | 8 | 0 | 6 | 2 |
| **Unknown Whether Surgery Done** | 15 | 2 | 16 | 1 | 14 | 3 | 17 | 0 |
| **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 [3170 RX Date--Most Defin Surg](#n_rx3170) supplemented by [1200 RX Date--Surgery](#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 ([3170 RX Date--Most Defin Surg](#n_rx3170) and [1200 RX Date--Surgery](#n_dsurg)) variable as above with the additional criterion that only cases where the [1340 Reason for No Surgery](#n_surgreason) is Surgery Performed be included.

### Appendix 3.2.3 Re-occurrence

The current available variables are: [1770 Cancer Status](#n_cstatus), [1880 Recurrence Type--1st](#n_rectype) and [1860 Recurrence Date--1st](#n_drecur). 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#1772). According to the v16 standard, instead [1750 Date of Last Contact](#n_lc) should be used.

Now we can reconcile [1770 Cancer Status](#n_cstatus) and [1880 Recurrence Type--1st](#n_rectype). We can see below that almost all [1770 Cancer Status](#n_cstatus)Tumor\_Free patients also have a Disease-free in their [1880 Recurrence Type--1st](#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 | 246 | 0 |
| **Recurred but no other info** | 0 | 2 | 0 |
| **Unknown if recurred or was ever gone** | 0 | 2 | 31 |

This suggest the following rules for binning them:

* [1880 Recurrence Type--1st](#n_rectype) is Disease-free (disease free)
* [1880 Recurrence Type--1st](#n_rectype) is Never disease-free (never disease free)
* [1880 Recurrence Type--1st](#n_rectype) raw code includes 70 then assume never diease free
* [1880 Recurrence Type--1st](#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 [1860 Recurrence Date--1st](#n_drecur) (recurrence date) are the ones which also have a Recurred status for [Recurrence Status](#a_n_recur) (with 19 missing an [1860 Recurrence Date--1st](#n_drecur)). The only exception is 1 Never diease-free patient that had an [1860 Recurrence Date--1st](#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 [1860 Recurrence Date--1st](#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.

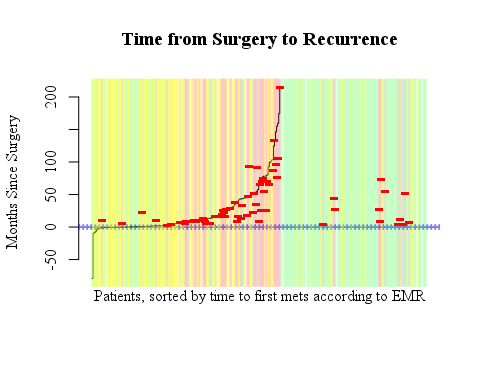


Figure 9: In the above plot, the black line represents months elapsed between surgery and the first occurence of an EMR code for secondary tumors, if any. The horizontal red line segments indicate individual [1860 Recurrence Date--1st](#n_drecur) . The blue horizontal line is the date of surgery. Patients whose status [1880 Recurrence Type--1st](#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 [1200 RX Date--Surgery](#n_dsurg) and of those 5 occur more than 3 months prior to [1200 RX Date--Surgery](#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.

### Appendix 3.2.4 Death

When more than one source has a death date, they are in agreement. To be fair, the agreement between [Death, i2b2](#e_death), [Expired[7,579 facts; 7,544 patients]](#e_dscdeath), and [Deceased per SSA](#s_death) is probably due to our i2b2 ETL already merging [Expired[7,579 facts; 7,544 patients]](#e_dscdeath) and [Deceased per SSA](#s_death) into [Death, i2b2](#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.

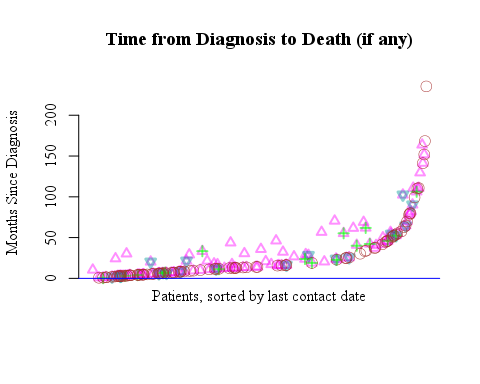


Figure 10: Above are plotted times of death (for patients that have them) relative to [0390 Date of Diagnosis](#n_ddiag) (horizontal blue line). The four data sources are [Death, i2b2](#e_death) (), [Deceased per SSA](#s_death) (), [Expired[7,579 facts; 7,544 patients]](#e_dscdeath) (), and [1760 Vital Status](#n_vtstat) ().

Here are some crosschecks on the NAACCR-only death indicator [1760 Vital Status](#n_vtstat). Overall there are 136 patients that according to [1760 Vital Status](#n_vtstat) are deceased. For all 136 of these patients, *and only for them*, the condition also holds that [1760 Vital Status](#n_vtstat) is equal to [1750 Date of Last Contact](#n_lc) but [1750 Date of Last Contact](#n_lc) happens before or on [age\_at\_visit\_days](#age_at_visit_days). If something is coded as happening *after*[age\_at\_visit\_days](#age_at_visit_days) then because of how the data is summarized in the dat2 section of the data.R script it means that the event never happened. If [1750 Date of Last Contact](#n_lc) never happened it means that patient has evidence of kidney cancer in the EMR data but no accompanying NAACCR record. In short, we are filtering for existence of NAACCR records. That, in turn, means that [1760 Vital Status](#n_vtstat)<=[1750 Date of Last Contact](#n_lc) is a valid censoring criteria (censored if false) provided that the input data is filtered to include only patients with NAACCR records (for other patients, both [1760 Vital Status](#n_vtstat) and [age\_at\_visit\_days](#age_at_visit_days) should be interpreted as missing).

### Appendix 3.2.5 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
* [Language](#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
* [0190 Spanish/Hispanic Origin](#n_hisp) is the [0190 Spanish/Hispanic Origin](http://datadictionary.naaccr.org/default.aspx?c=10#190) variable from NAACCR. Slightly processed by data.R and levels\_map.csv
  + non-Hispanic: Non\_Hispanic
  + Unknown: Unknown
  + Hispanic: any other value
* [Hispanic or Latino](#e_hisp) is the indicator variable for Hispanic ethnicity from i2b2 OBSERVATION\_FACT.
  + Hispanic: TRUE
  + Unknown: FALSE
* [Ethnicity](#e_eth) is the whole ethnicity variable from i2b2 OBSERVATION\_FACT and suprprisingly it is not in full agreement with [Hispanic or Latino](#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 [0190 Spanish/Hispanic Origin](#n_hisp), [Hispanic or Latino](#e_hisp), and [Ethnicity](#e_eth) are unanimous for Hispanic
* non-Hispanic if ALL non-missing values of [0190 Spanish/Hispanic Origin](#n_hisp) and [Ethnicity](#e_eth) are unanimous for non-Hispanic (note that [Hispanic or Latino](#e_hisp) not included here) and neither [Language](#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), [Language](#e_lng), [0190 Spanish/Hispanic Origin](#n_hisp), [Hispanic or Latino](#e_hisp), and [Ethnicity](#e_eth) have value Hispanic
* Unknown if ALL non-missing values of [language\_cd](#language_cd), [Language](#e_lng), [0190 Spanish/Hispanic Origin](#n_hisp), [Hispanic or Latino](#e_hisp), and [Ethnicity](#e_eth) are unanimous for Unknown
* non-Hispanic if any other result

What is the risk of relapse for patients after nephrectomy?

## Appendix 3.3 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 | 1729 | 2218 |
| TRUE | TRUE | TRUE | 67 | 489 |
| FALSE | FALSE | FALSE | 3 | 422 |
| TRUE | FALSE | FALSE | 39 | 419 |
| TRUE | TRUE | FALSE | 380 | 380 |

*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 [e\_i9neph][e\_i9neph], [e\_hstneph][e\_hstneph], [e\_surgonc][e\_surgonc], and [e\_i10neph][e\_i10neph] if they occur prior to* [*0390 Date of Diagnosis*](#n_ddiag) *and those will exclude far fewer records than suggested by this table* .

## Appendix 3.4 What is going on with the first contact variable?

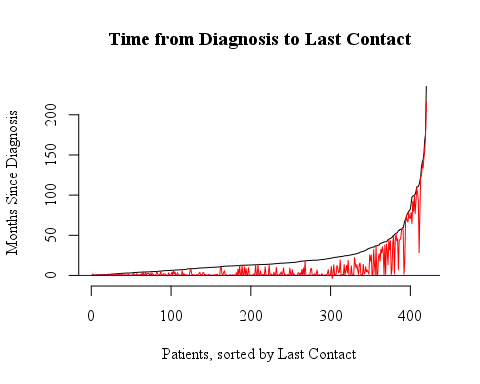


Figure 11: Wierd observation– [0580 Date of 1st Contact](#n_fc) (red) is almost always between [1750 Date of Last Contact](#n_lc) (black) and [0390 Date of Diagnosis](#n_ddiag) (blue) though diagnosis is usually on a biopsy sample and that’s why it’s dated as during or after surgery we thought. If first contact is some kind of event after first diagnosis, what is it?.

Surgery [1200 RX Date--Surgery](#n_dsurg) seems to happen in significant amounts both before and after first contact [0580 Date of 1st Contact](#n_fc).

## Appendix 3.5 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 4 Variable descriptions

Here are descriptions of the variables referenced in this document.

###### start\_date

start\_date :

start\_date

###### birth\_date

birth\_date :

birth\_date

###### sex\_cd

sex\_cd :

sex\_cd; Sex, i2b2

###### language\_cd

language\_cd :

language\_cd; Language, i2b2

###### race\_cd

race\_cd :

race\_cd; Race, i2b2

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

age\_at\_visit\_days :

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

###### n\_rectype

1880 Recurrence Type–1st :

1880 Recurrence Type–1st

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1880>

###### n\_rx3170

3170 RX Date–Most Defin Surg :

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

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#3170>

###### n\_rx1270

1270 Date of 1st Crs RX–CoC :

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>

###### n\_rx1260

1260 Date of Initial RX–SEER :

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>

###### n\_fc

0580 Date of 1st Contact :

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>

###### n\_dsdisc

3180 RX Date–Surgical Disch :

3180 RX Date–Surgical Disch

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#3180>

###### n\_surgreason

1340 Reason for No Surgery :

1340 Reason for No Surgery

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1340>

###### n\_mets

2850 CS Mets at DX :

2850 CS Mets at DX

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#2850>

###### n\_marital

0150 Marital Status at DX :

0150 Marital Status at DX; Marital Status, Registry

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#150>

###### n\_hisp

0190 Spanish/Hispanic Origin :

0190 Spanish/Hispanic Origin; Hispanic Origin, Registry

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#190>

###### n\_sex

0220 Sex :

0220 Sex; Sex, Registry

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#220>

###### n\_dob

0240 Date of Birth :

0240 Date of Birth

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#240>

###### n\_brthplc

0250 Birthplace :

0250 Birthplace

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#250>

###### n\_ddiag

0390 Date of Diagnosis :

0390 Date of Diagnosis

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#390>

###### n\_dsurg

1200 RX Date–Surgery :

1200 RX Date–Surgery

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1200>

###### n\_lc

1750 Date of Last Contact :

1750 Date of Last Contact; Last Contact

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1750>

###### n\_vtstat

1760 Vital Status :

1760 Vital Status; Vital Status, Registry; This gets individually converted to a TTE variable by data.R

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1760>

###### n\_cstatus

1770 Cancer Status :

1770 Cancer Status; Cancer Status, Registry

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1770>

###### n\_drecur

1860 Recurrence Date–1st :

1860 Recurrence Date–1st

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#1860>

###### n\_mult

0446 Multiplicity Counter :

0446 Multiplicity Counter

Link: <http://datadictionary.naaccr.org/default.aspx?c=10#446>

###### s\_death

Deceased per SSA :

Deceased per SSA; Death, SSN

###### e\_hisp

Hispanic or Latino :

Hispanic or Latino; Hispanic Origin, i2b2

###### e\_dscdeath

Expired[7,579 facts; 7,544 patients] :

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

###### e\_eth

Ethnicity :

Ethnicity; EMR demographics

###### n\_seer\_kcancer

Kidney and Renal Pelvis :

Kidney and Renal Pelvis; SEER site

###### e\_marital

Marital Status :

Marital Status; Marital Status, i2b2

###### n\_kcancer

Kidney, NOS :

Kidney, NOS; KC, Registry

###### e\_lng

Language :

Language

###### patient\_num

Patient Number (anonymized) :

Patient Number (anonymized); Patient

###### e\_death

Death, i2b2 :

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

###### a\_n\_recur

Recurrence Status :

Recurrence Status; 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 (broad) :

Hispanic (broad); Hispanic (broad); 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) :

Hispanic (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) :

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

###### a\_tdiag

Diagnosis :

Diagnosis; Diagnosis

###### a\_trecur

Recurrence :

Recurrence; Recurrence

###### a\_tsurg

Surgery :

Surgery; Surgery

###### a\_tdeath

Death :

Death; Death

###### a\_n\_race

Race (NAACCR 0160-0164) :

Race (NAACCR 0160-0164); Race, registry; To obtain a combined NAACCR race code for analysis, it is necessary to combine NAACCR variables 0160 Race - 0164 Race into one and then recode it to the closest match among White, BlackAsian, Pac Islander, Other, and Unknown

# Appendix 5 Audit trail

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sequence | time | type | name | hash |
| 0001 | 2018-10-01 14:43:16 | info | sessionInfo | - |
| 0002 | 2018-10-01 14:43:16 | this\_script | exploration.spin.Rmd | TEST\_OUTPUT\_DO\_NOT\_USE |
| 0003 | 2018-10-01 14:43:29 | rdata | .depdata[ii] = “dictionary.R.rdata” | e5c97a66fae2d9fafddbe6c0acd71ef4 |
| 0004 | 2018-10-01 14:43:36 | rdata | .depdata[ii] = “data.R.rdata” | 0f480d5124b18ea7d00bea26ee37be43 |
| 0003.0001 | 2018-10-01 11:26:35 | info | sessionInfo | - |
| 0003.0002 | 2018-10-01 11:26:35 | this\_script | dictionary.R | b35ff87 |
| 0003.0003 | 2018-10-01 11:27:04 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0003.0004 | 2018-10-01 11:27:19 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0003.0005 | 2018-10-01 11:27:19 | file | rawdct = “local/in/meta\_HSC20170563N\_kc\_v200.int.csv” | 77226290495672d030798e64327fe10a |
| 0003.0006 | 2018-10-01 11:27:19 | file | tpldct = “datadictionary\_static.csv” | b682c19f8a0530ffe7f2d689769fdb53 |
| 0003.0007 | 2018-10-01 11:27:22 | info | sessionInfo | - |
| 0003.0008 | 2018-10-01 11:27:22 | save | save | - |
| 0004.0001 | 2018-10-01 11:27:50 | info | sessionInfo | - |
| 0004.0002 | 2018-10-01 11:27:50 | this\_script | data.R | b35ff87 |
| 0004.0003 | 2018-10-01 11:28:01 | rdata | .depdata = “dictionary.R.rdata” | e5c97a66fae2d9fafddbe6c0acd71ef4 |
| 0004.0004 | 2018-10-01 11:28:01 | file | levels\_map\_file = “levels\_map.csv” | 4c66bc0cd1fd35eb9e64c3c49296a05f |
| 0004.0005 | 2018-10-01 11:28:59 | seed | project\_seed | - |
| 0004.0006 | 2018-10-01 11:31:07 | info | sessionInfo | - |
| 0004.0007 | 2018-10-01 11:31:08 | save | save | - |
| 0004.0003.0001 | 2018-10-01 11:26:35 | info | sessionInfo | - |
| 0004.0003.0002 | 2018-10-01 11:26:35 | this\_script | dictionary.R | b35ff87 |
| 0004.0003.0003 | 2018-10-01 11:27:04 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0004.0003.0004 | 2018-10-01 11:27:19 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0004.0003.0005 | 2018-10-01 11:27:19 | file | rawdct = “local/in/meta\_HSC20170563N\_kc\_v200.int.csv” | 77226290495672d030798e64327fe10a |
| 0004.0003.0006 | 2018-10-01 11:27:19 | file | tpldct = “datadictionary\_static.csv” | b682c19f8a0530ffe7f2d689769fdb53 |
| 0004.0003.0007 | 2018-10-01 11:27:22 | info | sessionInfo | - |
| 0004.0003.0008 | 2018-10-01 11:27:22 | save | save | - |

1. UT Health San Antonio [↑](#footnote-ref-2)