Kidney Cancer Data Exploration

KL2 Aim 2

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

October 10, 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. Ready to proceed to chart review of existing data, acquisition of independent NAACCR data, development of additional variables, and working on Aim 1.

###### 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. | * [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) |
| Yellow highlights are items with which I know I need to deal soon. Verbatim names of files, variables/elements, or values are displayed in a special style, like this. Data element names are in addition linked to a glossary at the end of this document, e.g. [Surgical Oncology](#e_surgonc). This is where any relevant cleaning or tranformation steps will be described (in progress). Data elements from NAACCR usually have a NAACCR ID preceding them, e.g. [1780 Quality of Survival](#n_qsurv). I try to use the word ‘data element’ to describe data in its raw state and ‘variable’ to refer to analysis-ready data that I have already processed. Often one variable incorporates information from multiple data elements. Tables, figures, and sections are also linked from text that references them. If you have a Word version of this document, to follow a link, please hold down the ‘control’ key and click on it. The most current version of this document can be found online at <https://rpubs.com/bokov/kidneycancer> and it has a built-in chat session. | |

# 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 returns 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" \l "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 narrow the initial cohort to patients that have NAACCR, defined as having a non-missing [0390 Date of Diagnosis](#n_ddiag) and one or both 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. Actually a total of 673 patients had NAACCR records but 187 of them had kidney cancer documented *only* in the EMR, but neither [Kidney, NOS](#n_kcancer) or [Kidney and Renal Pelvis](#n_seer_kcancer) in NAACCR. Next time I re-run my i2b2 query I will include all site of occurrence information from NAACCR not just kidney. This will allow me to find out what types of cancer these patients do in fact have. 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% with race, Hispanic ancestry, and marital status expected to be especially variable. Nonetheless, if record linkage is correct, when patient counts for NAACCR and EMR are tabulated against each of the above variables, then *most* of the values should agree.

I confirmed that this *is* the case for marital status ([table III](#tbl:xc_marital)), sex ([table IV](#tbl:xc_sex)), race ([table V](#tbl:xc_race)), and Hispanic ancestry ([table VI](#tbl:xc_hisp0)). 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](#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 elements 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 [Appendix 3.2](#sec:vartrn). In addition the following will almost certainly be needed for covariates or mediators: [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, AJCC, or TNM. For crosschecking it will 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 **the elements listed so far should be sufficient to replicate my analysis on 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 unavailable 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. However, 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 (** [**table VI**](#tbl:xc_hisp0) **and** [**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, EMR dates lag considerably behind NAACCR dates ( [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 ([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 pending verification by chart review in cases where 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 to 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 test whether my scripts succeeded in turning the raw data elements into a time-to-event (TTE) variables to which Kaplan-Meier curves can be fit without numeric errors or grossly implausible results. All the plots below are from a small random sample of the data– N=127, 82 Hispanic and 45 non-Hispanic white, 5 unknown excluded. This is further reduced in some cases as described in the figure captions. These sample sizes are not sufficient to detect clinically significant differences and, again, **this is not the goal yet**. The intent is only to insure that my software performs correctly while keeping myself blinded to 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 or stage at diagnosis. There are also refinements planned to 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.

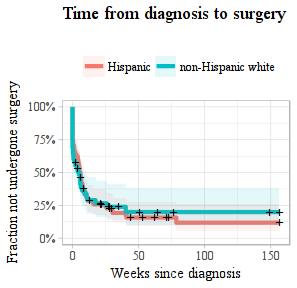


Figure 1: Number of weeks elapsed from [Diagnosis](#a_tdiag) (time 0) to [Surgery](#a_tsurg) for 82 Hispanic and 45 non-Hispanic white patients with a 3-year follow-up period (any surgeries occurring more than 3 years post-diagnosis are treated as censored).

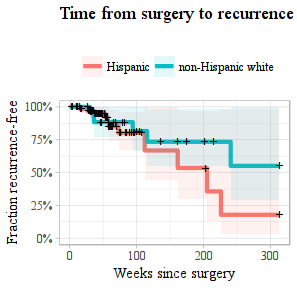


Figure 2: Number of weeks elapsed from [Surgery](#a_tsurg) (time 0) to [Recurrence](#a_trecur) for 67 Hispanic and 34 non-Hispanic white patients. The numbers are lower than for [fig. 1](#fig:surg_survfit) because patients not undergoing surgery are excluded. Here the follow-up period is six years.

###### blank

Typically 2-4 weeks elapse diagnosis from surgery and providers try to not exceed 4 weeks. Nevertheless years may sometimes elapse due to factors such as an indolent tumors or loss of contact with the patient. About 15% of patients never undergo surgery[4](#ref-pcRodriguez2018). [Fig. 1](#fig:surg_survfit) is in agreement with this.

It can also be seen in [fig. 1](#fig:surg_survfit) that 34 surgeries seem to happen on the day of diagnosis. This is plausible if NAACCR diagnosis is based on pathology rather than clinical examination where a positive result is usually coded as a renal mass, not a cancer. In my next data update I intend to also include all ICD9/10 codes for renal mass at which point I will revisit the question of using EMR data to fill in missing diagnosis dates (see [sec. 5](#sec:nextsteps)).

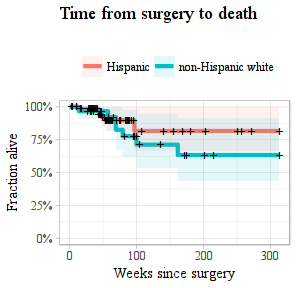


Figure 3: Like [fig. 2](#fig:recur_survfit) except now the outcome is [1760 Vital Status](#n_vtstat) for 67 Hispanic and 34 non-Hispanic white patients. Six-year follow-up.

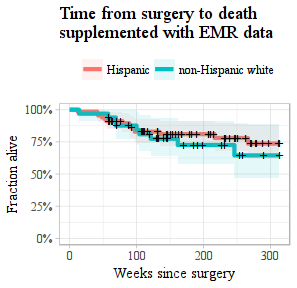


Figure 4: Like [fig. 3](#fig:naaccrdeath_survfit) but now supplemented EMR information to see how much of a difference it makes. For the predictor [Hispanic (broad)](#a_hsp_broad) is used instead of [Hispanic (NAACCR)](#a_hsp_naaccr) and for the outcome [Death](#a_tdeath) is used instead of [1760 Vital Status](#n_vtstat) . There were 68 Hispanic and 33 non-Hispanic white patients. There were 10 fewer censored events than in [fig. 3](#fig:naaccrdeath_survfit) which may improve sensitivity in the actual analysis.

###### blank

# 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" \l "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 principal 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 four priority 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 [Appendix 3.1](#sec:xchecks), [Appendix 3.2.1](#sec:diag), [Appendix 3.2.2](#sec:surg) 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 improvements in the CIRD ETL process. As per Dr. Rodriguez I also plan on adding all ICD codes for ‘renal mass’ to my i2b2 query ([Appendix 3.2.1](#sec:diag)). Meanwhile, in response to researcher questions including my own, CIRD staff have 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. After the next i2b2 refresh we expect an increased number of patients and possible improved agreement of event 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 [Appendix 3](#sec:supp). 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) development planned for Aim 1. 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 that is integrated into our local i2b2. During the work I present here I found several 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 will, in turn, will simplify the 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\_num](#patient_num) are de-identified but with proper authorization they can be mapped to MRNs or internal database index keys.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [patient\_num](#patient_num) | [3430 Derived AJCC-7 Stage Grp](#v022_drvd_ajcc_stg) | [3422 Derived AJCC-7 M Descript](#v023_drvd_dscrpt) | [3420 Derived AJCC-7 M](#v024_drvd_ajcc_m) | [3412 Derived AJCC-7 N Descript](#v025_drvd_dscrpt) | [3410 Derived AJCC-7 N](#v026_drvd_ajcc_n) | [3402 Derived AJCC-7 T Descript](#v027_drvd_dscrpt) | [3400 Derived AJCC-7 T](#v028_drvd_ajcc_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: I wrote a function called e\_table() that simultaneously tabulates missingness, frequency of discrepancy, and magnitude of discrepancy. I need to replace many of the *ad-hoc* tables in [Appendix 3.2.1](#sec:diag)-[Appendix 3.2.3](#sec:recur) with e\_table() as I have already done for [table XX](#tbl:etabledeath) in [Appendix 3.2.4](#sec:death).
* TODO: Standardize the naming scheme for temporary tables, chunk labels, anchors, and captions before things get out of hand.
* TODO: Go through this document and add any in-line TODOs to this section perhaps linking them bidirectionally to the text
* TODO: Increase existing data
  + Meet with NAACCR regisrar
  + Update i2b2 query with renal mass
  + Finish credentialing process
* TODO: Get outside data
  + Submit request to TCR
* TODO: Develop covariates/mediators
* TODO: Finish DataFinisher (Aim 1)
* 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: 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~~ [~~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~~](#n_rx1260)
    - [~~1270 Date of 1st Crs RX--CoC~~](#n_rx1270)
    - [~~3170 RX Date--Most Defin Surg~~](#n_rx3170)
  + DONE: [~~1880 Recurrence Type--1st~~](#n_rectype)
* 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: Systematically convert to e\_table() in [Appendix 3.2.1](#sec:diag)-[Appendix 3.2.3](#sec:recur)
* TODO: Variable glossary
  + Migrate detailed variable descriptions there.
  + Create list object for adding on commentary to glossary entries within exploration.R, outside the data dictionary proper.
  + Fill in more of the variable descriptions in [Appendix 4](#sec:vars)
  + Figure out how to make fs() fail gracefully perhaps with a generic ‘this variable is not yet documented’ link target.
  + ~~Automatically translate and link variable names in table column or row names~~
* TODO: Remove the crossed-off stuff in [Appendix 3](#sec:supp) but note someplace what was removed and why (move it to a long-term location on RPubs.com)
* TODO: Organize the inclusion/exclusion criteria into a single named list and analyze their correlation.
* 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: 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: Resume effort to link Mays Center historic trial records from IDEAS to get information about enrollment in adjuvant trials
* 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: ~~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~~](#sec:surg)~~,~~ [~~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.~~
      * [~~Surgery~~](#sec:surg)~~: turn the outline at the beginning into a more concise paragraph.~~
  + [~~Hispanic variable recoding~~](#whether-or-not-the-patient-is-hispanic)~~: turn into paragraphs,~~
  + ~~[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).
* DONE: ~~Create combined (if applicable) variables for each of the following:~~
  + ~~Initial diagnosis~~ [Diagnosis](#a_tdiag), [a\_cdiag](#a_cdiag)
  + ~~Surgery~~ [Surgery](#a_tsurg), [a\_csurg](#a_csurg)
  + ~~Re-ocurrence~~ [Recurrence](#a_trecur), [a\_crecur](#a_crecur)
  + *~~Last follow-up ?~~*
  + ~~Death~~ [Death](#a_tdeath), [a\_cdeath](#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" \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 3 Supplementary results

## Appendix 3.1 Consistency checks

In this section are patient counts for all 2327 patients in the overall set, broken down by various NAACCR variables (rows) and equivalent EMR variables (columns). The **bold** values are counts of patients for whom NAACCR and EMR are in agreement. Patients in the NA are the ones with only EMR and no NAACCR records, so they count as missing rather than discrepant.

Table III: Marital status has good agreement between NAACCR and EMR.

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

Table IV: Sex has good agreement between NAACCR and EMR.

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

Table V: Race has good agreement between NAACCR and EMR.

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

Table VI: Hispanic designation has good agreement between NAACCR and EMR. Here the [0190 Spanish/Hispanic Origin](#n_hisp) variable was simplified by binning into Hispanic and non-Hispanic.

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

Table VII: As [table VI](#tbl:xc_hisp0) but with all the different levels of [0190 Spanish/Hispanic Origin](#n_hisp) shown.

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

Table VIII: Below is a summary of [birth\_date](#birth_date) - [0240 Date of Birth](#n_dob) (in years) for the patients with non-matching dates of birth mentioned in [sec. 2.1](#sec:linkagever). Though there are only 15 of them those few deviate by multiple years from the EMR records.

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

**Divergent birth dates may become an exclusion criterion for analyses where age is a covariate**. However there is no evidence for increased disagreement between NAACCR and EMR for other variables among the patients from [table VIII](#tbl:xc_dob) according to  
tables [IX](#tbl:xc_dob_marital), [X](#tbl:xc_dob_sex), [XI](#tbl:xc_dob_race), [XII](#tbl:xc_dob_hisp) which are like tables [III](#tbl:xc_marital), [IV](#tbl:xc_sex), [V](#tbl:xc_race), [VI](#tbl:xc_hisp0) but tallied up for only those 15 patients.

Table IX: Marital status, for DOB-mismatched patients.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | divorced | legally sepa | married | other | significant | single | unknown | widowed |
| **Divorced** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **Separated** | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 |
| **Married** | 0 | 0 | 0 | **5** | 0 | 0 | 1 | 0 | 1 |
| **Domestic Partner** | 0 | 0 | 0 | 0 | 0 | **0** | 0 | 0 | 0 |
| **Single** | 0 | 0 | 0 | 0 | 0 | 0 | **2** | 0 | 0 |
| **Unknown** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | **0** | 0 |
| **Widowed** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** |

Table X: Sex, for DOB-mismatched patients

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | f | u | Sum |
| **m** | **6** | **1** | 0 | 7 |
| **f** | **0** | **8** | 0 | 8 |
| **Sum** | 6 | 9 | 0 | 15 |

Table XI: Race, for DOB-mismatched patients

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | White | Black | Asian | Pac Islander | Other | Unknown | Sum |
| **White** | **11** | **0** | **0** | **0** | 0 | 2 | 13 |
| **Black** | **0** | **0** | **0** | **0** | 0 | 0 | 0 |
| **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** | 13 | 0 | 0 | 0 | 0 | 2 | 15 |

Table XII: Hispanic ethnicity, for DOB-mismatched patients

|  |  |  |  |
| --- | --- | --- | --- |
|  | Non\_Hispanic | Hispanic | Sum |
| **Non\_Hispanic** | **8** | **1** | 9 |
| **Hispanic** | **0** | **6** | 6 |
| **Sum** | 8 | 7 | 15 |

Table XIII: There is also no increase in EMR nephrectomy dates preceding diagnosis among the DOB-mismatched patients (see [table XVI](#tbl:neph_b4_diag)). This suggests that other dates associated with these patients are not systematically wrong.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | before | same-day | after | NA |
| **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?

For each of the main event variables [Diagnosis](#a_tdiag), [Surgery](#a_tsurg), [Recurrence](#a_trecur), and [Death](#a_tdeath) / [1760 Vital Status](#n_vtstat) there were multiple candidate data elements in the raw data. If such a family of elements is in good agreement overall then individual missing dates can be filled in with the earliest non-missing dates from other data elements in that family (except for mortality where the *latest* non-missing date would make more sense). But to do this I needed not only to establish qualitative agreement as I did for demographic variables in [sec. 2.1](#sec:linkagever) and [Appendix 3.1](#sec:xchecks) but also determine how often these dates lag or lead each other and by how much. The plots in this section use the y-axis to represent time for patient records arranged along the x-axis. They are arranged in an order that varies from one plot to another, chosen for visual interpretability. Each vertical slice of a plot represents one patient’s history, with different colors representing events as documented by different data elements. The goal is to see the frequency, magnitude, and direction of divergence for several variables at the same time.

### Appendix 3.2.1 Initial diagnosis

At this time only [0390 Date of Diagnosis](#n_ddiag) is usable for calculating [Diagnosis](#a_tdiag). Initially [0580 Date of 1st Contact](#n_fc) was considered as an additional NAACCR source along with the earliest EMR records of [189.0 Malignant neoplasm of kidney, except pelvis](#e_kc_i9) and [C64 Malignant neoplasm of kidney, except renal pelvis](#e_kc_i10). [0443 Date Conclusive DX](http://datadictionary.naaccr.org/default.aspx?c=10" \l "443) is never used by our NAACCR. All other NAACCR data elements containing the word ‘date’ seem to be retired or related to events after initial diagnosis. [0580 Date of 1st Contact](#n_fc) was disqualified because it never precedes [0390 Date of Diagnosis](#n_ddiag) but often trails behind [1200 RX Date--Surgery](#n_dsurg), see [fig. 11](#fig:diag2lc_eventplot). I will need to consult with a NAACCR registrar about what [0580 Date of 1st Contact](#n_fc) actually means but it does not appear to be a first visit nor first diagnosis. As can be seen in [fig. 5](#fig:diag_plot) and [table XIV](#tbl:diag_lag), the first ICD9 or ICD10 code most often occurs after initial diagnosis, sometimes before the date of diagnosis, and coinciding with the date of diagnosis rarest of all. Several of the ICD9/10 first observed dates lead or trail the [0390 Date of Diagnosis](#n_ddiag) by multiple years.

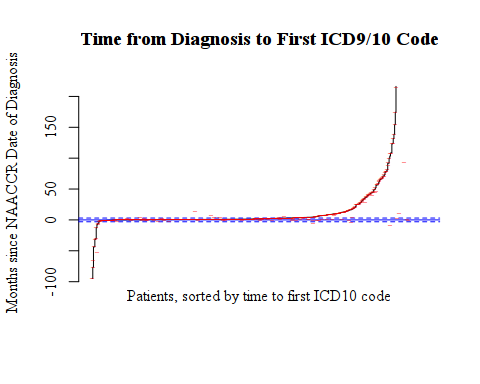


Figure 5: Here is a plot centered on [0390 Date of Diagnosis](#n_ddiag) (blue horizontal line at 0) with black lines indicating ICD10 codes for primary kidney cancer from the EMR and dashed red lines indicating ICD9 codes. The dashed horizontal blue lines indicate +- 3 months from [0390 Date of Diagnosis](#n_ddiag).

###### blank

Table XIV: For patients with NAACCR records, how often do ICD9 or ICD10 codes for kidney cancer in the EMR lead or trail [0390 Date of Diagnosis](#n_ddiag) and by how much?

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

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 *after* first [0390 Date of Diagnosis](#n_ddiag). Only 13 patients have ICD9/10 diagnoses that precede their [0390 Date of Diagnosis](#n_ddiag) by more than 3 months. An additional 54 patients have first EMR diagnoses that precede [0390 Date of Diagnosis](#n_ddiag) 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 disagreement between the the two sources, and the EMR records are currently biased toward later dates.

I will need to meet with the MCC NAACCR registrar to see how she obtains her dates of initial diagnosis and I will need to do a chart review of a sample of NAACCR patients to understand what information visible in Epic sets them apart from kidney cancer patients without NAACCR records. I will also need to do a chart review of the patients with ICD9/10 codes for kidney cancer that seemingly pre-date their [0390 Date of Diagnosis](#n_ddiag). There are 75 patients with multiple NAACCR records. I will need to learn how NAACCR distinguishes their first occurrences and see if **restricting the NAACCR data to just first occurrences will diminish the number of EMR diagnoses preceding those in NAACCR.** It will also be helpful to learn whether there is anything in the EMR distinguishes first kidney cancer occurrences besides lack of previous diagnosis.

### Appendix 3.2.2 Surgery

To construct the [Surgery](#a_tsurg) analytic variable I considered [1200 RX Date--Surgery](#n_dsurg), [1260 Date of Initial RX--SEER](#n_rx1260), [1270 Date of 1st Crs RX--CoC](#n_rx1270), and [3170 RX Date--Most Defin Surg](#n_rx3170) from NAACCR as well as earliest occurrences of [V45.73 Acquired absence of kidney](#e_i9neph), [Z90.5 Acquired absence of kidney](#e_i10neph), or [HX NEPHRECTOMY](#e_hstneph) from the EMR. In the plots and tables below I show why I decided to use [3170 RX Date--Most Defin Surg](#n_rx3170) as the surgery date and when that is unavailable, to fall back on [1200 RX Date--Surgery](#n_dsurg). The other data elements are not used **except to flag potentially incorrect records if they occur earlier than the date of diagnosis**.

#’ ###### blank

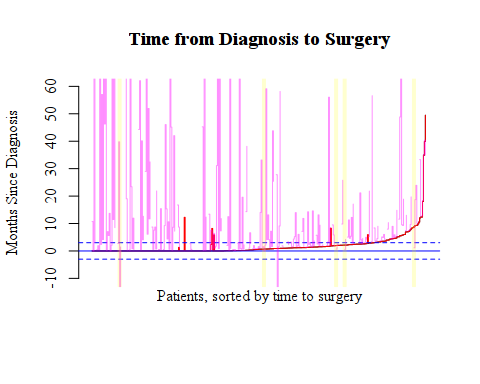


Figure 6: Above is a plot of all patients sorted by [1200 RX Date--Surgery](#n_dsurg) (black line). On the same axis is [3170 RX Date--Most Defin Surg](#n_rx3170) (red line) which is almost identical to [1200 RX Date--Surgery](#n_dsurg) except for a small number of cases where it occurs later than [1200 RX Date--Surgery](#n_dsurg) . It never occurs earlier. The violet 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). The blue horizontal line is [0390 Date of Diagnosis](#n_ddiag) with the dashed lines representing a 3-month window in both directions..

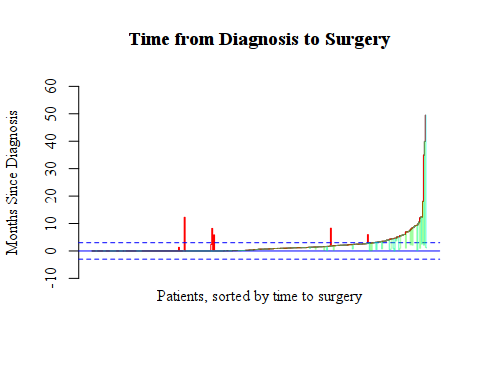


Figure 7: In the above plot the [1270 Date of 1st Crs RX--CoC](#n_rx1270) (green) and [1260 Date of Initial RX--SEER](#n_rx1260) (cyan) events are superimposed on time till [1200 RX Date--Surgery](#n_dsurg) like in [fig. 6](#fig:surg0_plot0) (but violet lines for nephrectomy EMR codes are omitted for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

###### blank

In [fig. 6](#fig:surg0_plot0) the 5 patients for which the earliest EMR nephrectomy code occurs before the earliest NAACCR possible record of surgery are highlighted in yellow. Among the remaining 181 patients who have an EMR code for nephrectomy, there are  
129 for whom it happens more than 3 months after [1200 RX Date--Surgery](#n_dsurg) and those lags have a median of 14.3 months. This level of discrepancy disqualifies [V45.73 Acquired absence of kidney](#e_i9neph), [Z90.5 Acquired absence of kidney](#e_i10neph), and [HX NEPHRECTOMY](#e_hstneph) from being used to fill in missing NAACCR dates. This may change after the next i2b2 update in which the fix to the “visit-less patient” problem will be implemented ([sec. 5](#sec:nextsteps))

###### blank

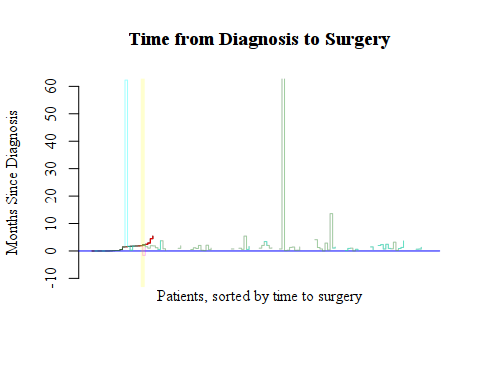


Figure 8: Above is a plot equivalent to [fig. 7](#fig:surg0_plot1) but for patients who do **not** have a [1340 Reason for No Surgery](#n_surgreason) code equal to Surgery Performed. There are many [1270 Date of 1st Crs RX--CoC](#n_rx1270) and [1260 Date of Initial RX--SEER](#n_rx1260) events but only 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 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 and cannot be used to fill in missing values.

###### blank

Table XV: As can be seen in the table below, the variables [V45.73 Acquired absence of kidney](#e_i9neph), [HX NEPHRECTOMY](#e_hstneph), [Surgical Oncology](#e_surgonc), and [Z90.5 Acquired absence of kidney](#e_i10neph) *sometimes* precede [0390 Date of Diagnosis](#n_ddiag) by many weeks but they *usually* follow [0390 Date of Diagnosis](#n_ddiag) by more weeks than do [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. This is another way of summarizing how much the EMR variables lag behind NAACCR variables.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA’s |
| [**3170 RX Date--Most Defin Surg**](#n_rx3170) | 0 | 0 | 3 | 8.461 | 9.643 | 215.1 | 119 |
| [**1270 Date of 1st Crs RX--CoC**](#n_rx1270) | 0 | 0 | 2.929 | 6.431 | 6.964 | 318.3 | 28 |
| [**1260 Date of Initial RX--SEER**](#n_rx1260) | 0 | 0 | 3.857 | 8.213 | 8.571 | 270.9 | 198 |
| [**1200 RX Date--Surgery**](#n_dsurg) | 0 | 0 | 2.857 | 7.83 | 9 | 215.1 | 109 |
| [**V45.73 Acquired absence of kidney**](#e_i9neph) | -361.1 | 8.143 | 31.43 | 69.5 | 82.71 | 957.4 | 261 |
| [**HX NEPHRECTOMY**](#e_hstneph) | -91.86 | 10.11 | 37.07 | 77.85 | 93.96 | 758.1 | 318 |
| [**Surgical Oncology**](#e_surgonc) | -194.9 | 0.2143 | 4.714 | 23.58 | 46 | 236.6 | 455 |
| [**Z90.5 Acquired absence of kidney**](#e_i10neph) | -20.14 | 9.607 | 37.86 | 85.12 | 111.2 | 957.4 | 226 |
| [**1860 Recurrence Date--1st**](#n_drecur) | 0 | 40.04 | 73.71 | 137.2 | 205.3 | 935.9 | 402 |

It makes sense that the Epic EMR lags behind NAACCR. As an outpatient system, it’s probably recording visits after the original surgery, and perhaps we are not yet importing the right elements from Sunrise EMR. In [sec. 5](#sec:nextsteps) I outline possible remedies to that. For now, [V45.73 Acquired absence of kidney](#e_i9neph), [HX NEPHRECTOMY](#e_hstneph), [Surgical Oncology](#e_surgonc), and [Z90.5 Acquired absence of kidney](#e_i10neph) can still be used to exclude cases as not first-time occurrences if it precedes diagnosis. Would I lose a lot of cases to such a criterion?

Table XVI: How often ICD9/10 or surgical history codes for nephrectomy precede diagnosis and by how much

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | before | same-day | after | NA |
| [**3170 RX Date--Most Defin Surg**](#n_rx3170) | 0 | 138 | 229 | 119 |
| [**1270 Date of 1st Crs RX--CoC**](#n_rx1270) | 0 | 149 | 309 | 28 |
| [**1260 Date of Initial RX--SEER**](#n_rx1260) | 0 | 83 | 205 | 198 |
| [**1200 RX Date--Surgery**](#n_dsurg) | 0 | 146 | 231 | 109 |
| [**V45.73 Acquired absence of kidney**](#e_i9neph) | 3 | 0 | 222 | 261 |
| [**HX NEPHRECTOMY**](#e_hstneph) | 3 | 2 | 163 | 318 |
| [**Surgical Oncology**](#e_surgonc) | 7 | 1 | 23 | 455 |
| [**Z90.5 Acquired absence of kidney**](#e_i10neph) | 1 | 0 | 259 | 226 |

Only a small number of cases would be disqualified. Another important question is the level of agreement between [1340 Reason for No Surgery](#n_surgreason) and the NAACCR data elements that are candidates for comprising the surgery variable.

Table XVII: Every NAACCR candidate data element (columns) tabulated against [1340 Reason for No Surgery](#n_surgreason) (rows). The bold cells are ones consistent with their respective data elements indicating the primary surgery. The second row is italicized because surgery may still occur as a non-primary course of treatment. Nevertheless the counts in the FALSE columns should be greater than the counts in the TRUE columns for every row except the first. [3170 RX Date--Most Defin Surg](#n_rx3170) and [1200 RX Date--Surgery](#n_dsurg) are in close agreement with each other and have the fewest deviations from expected behavior of a primary surgery data element

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

In summary, based on [fig. 6](#fig:surg0_plot0) and [table XIV](#tbl:diag_lag) [V45.73 Acquired absence of kidney](#e_i9neph), [HX NEPHRECTOMY](#e_hstneph), [Surgical Oncology](#e_surgonc), and [Z90.5 Acquired absence of kidney](#e_i10neph) can only be used to disqualify patients for having erroneous records or previous history of kidney cancer but cannot fill in missing diagnosis dates. Based on figs. [7](#fig:surg0_plot1), [8](#fig:surg1_plot), and [table XVIII](#tbl:rectype_cstatus) [1270 Date of 1st Crs RX--CoC](#n_rx1270) and [1260 Date of Initial RX--SEER](#n_rx1260) are not necessarily always surgery events. This leaves [3170 RX Date--Most Defin Surg](#n_rx3170) with [0390 Date of Diagnosis](#n_ddiag) as a fallback. When I meet with the NAACCR regisrar I will seek their feedback about this approach and I will ask them about the most reliable way to identify the first kidney cancer occurrence for a patient if they have several (overlapping?) NAACCR entries. I also need to ask a chart abstraction expert about the best way to find in Epic and in Sunrise the date of a patient’s first nephrectomy

### Appendix 3.2.3 Re-occurrence

Candidate data elements for constructing the [Recurrence](#a_trecur) variable were [1770 Cancer Status](#n_cstatus), [1880 Recurrence Type--1st](#n_rectype), and [1860 Recurrence Date--1st](#n_drecur) from NAACCR. Our site is on NAACCR v16, not v18, so 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, [1750 Date of Last Contact](#n_lc) should be used instead. From the EMR the candidates were 14 ICD9/10 codes for secondary tumors. In [table XVIII](#tbl:rectype_cstatus) I reconcile [1770 Cancer Status](#n_cstatus) and [1880 Recurrence Type--1st](#n_rectype).

###### blank

Table XVIII: [1770 Cancer Status](#n_cstatus) is in good agreement with [1880 Recurrence Type--1st](#n_rectype). Almost all [1770 Cancer Status](#n_cstatus) Tumor\_Free patients also have Disease-free in their [1880 Recurrence Type--1st](#n_rectype) column, the Tumor ones have a variety of values, and the Unknown ones are 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 |

###### blank

[1880 Recurrence Type--1st](#n_rectype) can be simplified by leaving values of Disease-free (0), Never disease-free (70), and Unknown if recurred or was ever gone (99) as they are; if there were multiple values for the same case and one of those values was 70 then defaulting to Never disease-free; and recoding all other values as simply Recurred. I named this analytic variable [Recurrence Status](#a_n_recur).

###### blank

Table XIX: 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) 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 with a [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.

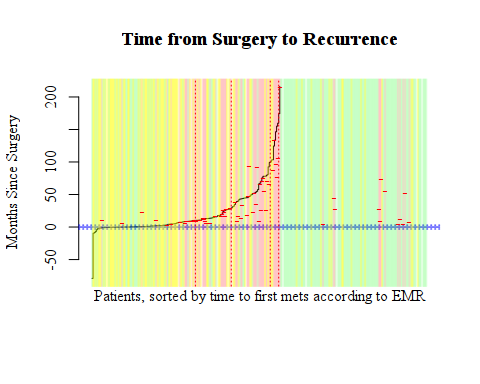


Figure 9: In the above plot, the black line represents months elapsed between surgery and the first occurence of an EMR code for secondary tumors, if any. The horizontal red line segments indicate individual [1860 Recurrence Date--1st](#n_drecur) . The dotted vertical red lines denote Recurred patients who are missing a [1860 Recurrence Date--1st](#n_drecur) . The blue horizontal line is the date of surgery and the dotted horizontal lines above and below it are +- 3 months. Patients whose [1880 Recurrence Type--1st](#n_rectype) is Disease-free are highlighted in green, Never disease-free in yellow, and Recurred in red. There are 75 patients with multiple NAACCR records, and all records for these patients have been excluded from this plot.

###### blank

The green highlights in [fig. 9](#fig:recur_plot) 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? These are questions to ask the NAACCR registrar. The EMR codes are in better agreement with [1860 Recurrence Date--1st](#n_drecur) than the data elements in [Appendix 3.2.1](#sec:diag) and [Appendix 3.2.2](#sec:surg) so it might make sense to back-fill the few [1860 Recurrence Date--1st](#n_drecur) that are missing but first I want to make sure I understand how to reliably distinguish on the EMR side genuine recurrences from secondary tumors that existed at presentation. The small number of cases affected either way lowers the priority of this isuse. For now I will rely only on [1860 Recurrence Date--1st](#n_drecur) in constructing the analytical variable [Recurrence](#a_trecur).

### Appendix 3.2.4 Death

Unlike diagnosis ([Appendix 3.2.1](#sec:diag)), surgery ([Appendix 3.2.2](#sec:surg)), and recurrence ([Appendix 3.2.3](#sec:recur)) death dates exhibit good agreement between various sources and can be used to supplement the data available from NAACCR.

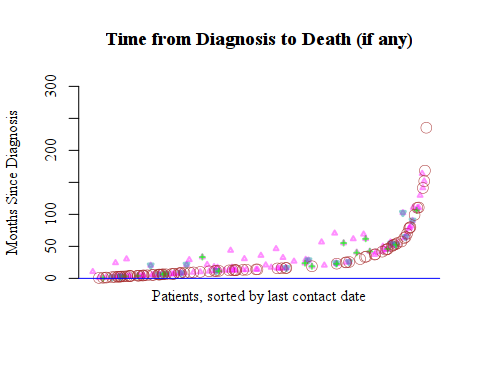


Figure 10: Above are plotted times of death (if any) relative to [0390 Date of Diagnosis](#n_ddiag) (horizontal blue line). The four data sources are [Death, i2b2](#e_death) (), [Deceased per SSA](#s_death) (), [Expired](#e_dscdeath) (), and [1760 Vital Status](#n_vtstat) ().

###### blank

Table XX: Date associated with [1760 Vital Status](#n_vtstat) compared to death dates from each source (rows). The first five columns represent the number of patients falling into each of the time-bins (in days) relative to [1760 Vital Status](#n_vtstat). The last four columns indicate the number of patients for each possible combination of missing values (Left means the variable indicated in the row name is missing and Right means [1760 Vital Status](#n_vtstat) is missing). The parenthesized values below the counts are percentages (of the total number of patients with both variables non-missing for the first five columns and of the total number of patients for the last four columns). Where available, the median difference in days is shown below the count and percentage. This table has only the 486 patients having a kidney cancer diagnosis in NAACCR. The last two rows represent the earliest and latest documentation of death, respectively, from [Deceased per SSA](#s_death), [Expired](#e_dscdeath), [Death, i2b2](#e_death), [Earliest Death](#Earliest Death), and [Latest Death](#Latest Death)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Below -30 | -30 to 0 | same | 0 to 30 | Above 30 | Neither missing | Left missing | Right missing | Both missing |
| [**Deceased per SSA**](#s_death) | 1 (10.0%) -31.0 | 0 ( 0.0%) | 9 (90.0%) 0.0 | 0 ( 0.0%) | 0 ( 0.0%) | 10 ( 2.1%) 0.0 | 83 (17.1%) | 8 ( 1.6%) | 385 (79.2%) |
| [**Expired**](#e_dscdeath) | 1 (11.1%) -34.0 | 7 (77.8%) -5.0 | 1 (11.1%) 0.0 | 0 ( 0.0%) | 0 ( 0.0%) | 9 ( 1.9%) -5.0 | 84 (17.3%) | 8 ( 1.6%) | 385 (79.2%) |
| [**Death, i2b2**](#e_death) | 1 ( 1.3%) -31.0 | 0 ( 0.0%) | 73 (96.1%) 0.0 | 2 ( 2.6%) 5.5 | 0 ( 0.0%) | 76 (15.6%) 0.0 | 17 ( 3.5%) | 46 ( 9.5%) | 347 (71.4%) |
| [**Earliest Death**](#Earliest Death) | 1 ( 1.1%) -34.0 | 7 ( 7.5%) -5.0 | 85 (91.4%) 0.0 | 0 ( 0.0%) | 0 ( 0.0%) | 93 (19.1%) 0.0 | 0 ( 0.0%) | 47 ( 9.7%) | 346 (71.2%) |
| [**Latest Death**](#Latest Death) | 0 ( 0.0%) | 0 ( 0.0%) | 91 (97.8%) 0.0 | 2 ( 2.2%) 5.5 | 0 ( 0.0%) | 93 (19.1%) 0.0 | 0 ( 0.0%) | 47 ( 9.7%) | 346 (71.2%) |

In [table XX](#tbl:etabledeath) the sum of the Neither missing and Left missing is always 93 which is the number of deceased patients according to NAACCR records alone. The Right missing column is the number of patients whose deceased status is recorded in the external source but not in NAACCR. For the last two rows Right missing means the total number of deceased patients not recorded in NAACCR but which can be filled in from one or more of the other sources. There are 47 such patients. Finally the last column, Both missing, is the number of patients presumed to be alive because none of the sources have any evidence for being deceased. The Left missing column indicates how many patients are reported deceased in NAACCR but *not* the other source. Though there are some missing for each individual data source, NAACCR is never the only source reporting them deceased– the values in the bottom two rows are both 0.

The left-side columns of [table XX](#tbl:etabledeath) show the prevalence and magnitude of discrepancies in death dates of the 93 patients that NAACCR and at least one other source agree are deceased. There are at most 10 such patients and for 9 of them the discrepancy is less than one month, with a median difference ranging from -5 to 5.5 days. **The small number of discrepancies and the small magnitude of the ones that do occur justify filling in missing NAACCR death dates from the other sources.**

### Appendix 3.2.5 Whether or not the patient is Hispanic

Despite the overall agreement between [0190 Spanish/Hispanic Origin](#n_hisp) and [Hispanic or Latino](#e_hisp) there needs to be some way to adjudicate the minority of cases where the sources disagree. The following additional data elements can provide relevant information to form a final consensus variable for analysis: [language\_cd](#language_cd), [Language](#e_lng), [Ethnicity](#e_eth), [race\_cd](#race_cd), and [Race (NAACCR 0160-0164)](#a_n_race) First, each of these variables is re-coded to Hispanic, non-Hispanic, and Unknown.

[language\_cd](#language_cd) and [Language](#e_lng) are interpreted as being evidence in favor of Hispanic ethnicity if the language includes Spanish. English, ASL, and unknown values are all treated as Unknown ethnicity. However, a language *other* than the above (e.g. German) is interpreted as evidence for being non-Hispanic.

[0190 Spanish/Hispanic Origin](#n_hisp) already have explicit designations of non-Hispanic and Unknown and all other values are interpreted as Hispanic. [Hispanic or Latino](#e_hisp) is interpreted as Hispanic if TRUE and Unknown if FALSE (in contrast with most of the other elements, there is no way to distinguish a genuinely FALSE value of [Hispanic or Latino](#e_hisp) from a missing one).

[Ethnicity](#e_eth) is the whole ethnicity variable from i2b2 OBSERVATION\_FACT and suprprisingly it sometimes disagrees with [Hispanic or Latino](#e_hisp). A value of hispanic is interpreted directly. The values other,unknown, unknown/othe,i choose not, and @ are all interpeted as Unknown and any other value (at our site, arab-amer and non-hispanic) is interpreted as non-Hispanic. Rules are then applied to create unified variables from all these data elements. I have three such variables– [Hispanic (NAACCR)](#a_hsp_naaccr), [Hispanic (broad)](#a_hsp_broad), and [Hispanic (strict)](#a_hsp_strict)

[Hispanic (NAACCR)](#a_hsp_naaccr) only uses information from NAACCR.

[Hispanic (broad)](#a_hsp_broad) errs on the side of assigning Hispanic ethnicity if there is any evidence for it at all, then non-Hispanic, and Unknown only if there is truly no information from any source about the patient’s ethnicity. In particular, Hispanic is assigned 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 a value of 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 ; and non-Hispanic otherwise.

Finally, [Hispanic (strict)](#a_hsp_strict) only assigns 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 is assigned if *all* non-missing values of [0190 Spanish/Hispanic Origin](#n_hisp) and [Ethnicity](#e_eth) are unanimous for non-Hispanic (the [Hispanic or Latino](#e_hisp) element is not used for the reasons explained above) *and* neither [Language](#e_lng) nor [language\_cd](#language_cd) vote for Hispanic. If neither of these conditions are met, Unknown is assigned.

There is an additional step for patients coded as non-Hispanic where they are further classified into non-Hispanic white and Other. For [Hispanic (NAACCR)](#a_hsp_naaccr) this is determined by whether or [Race (NAACCR 0160-0164)](#a_n_race) is White. For [Hispanic (broad)](#a_hsp_broad) the criterion is whether *at least one* of [Race (NAACCR 0160-0164)](#a_n_race) or [race\_cd](#race_cd) is White. For [Hispanic (strict)](#a_hsp_strict) it’s whether *both* [Race (NAACCR 0160-0164)](#a_n_race) and [race\_cd](#race_cd) are White.

In the end, [Hispanic (NAACCR)](#a_hsp_naaccr), [Hispanic (broad)](#a_hsp_broad), and [Hispanic (strict)](#a_hsp_strict) all have the same levels, but differ in the proportion of patients assigned to each.

Table XXI: The agreement and disagreement between [Hispanic (NAACCR)](#a_hsp_naaccr), [Hispanic (broad)](#a_hsp_broad), and [Hispanic (strict)](#a_hsp_strict) The bottom 7 rows represent the kidney cancer patients currently without NAACCR records, so for them [Hispanic (NAACCR)](#a_hsp_naaccr) does not exist.

|  |  |  |  |
| --- | --- | --- | --- |
| [Hispanic (NAACCR)](#a_hsp_naaccr) | [Hispanic (broad)](#a_hsp_broad) | [Hispanic (strict)](#a_hsp_strict) | N Patients |
| Hispanic | Hispanic | Hispanic | 213 |
| Hispanic | Hispanic | Unknown | 141 |
| non-Hispanic white | non-Hispanic white | non-Hispanic white | 190 |
| non-Hispanic white | non-Hispanic white | Unknown | 59 |
| non-Hispanic white | Hispanic | Unknown | 11 |
| non-Hispanic white | non-Hispanic white | Other | 3 |
| Other | Other | Other | 23 |
| Other | Other | Unknown | 13 |
| Other | Hispanic | Unknown | 2 |
| Other | non-Hispanic white | Other | 1 |
| Unknown | Unknown | Unknown | 9 |
| Unknown | Hispanic | Unknown | 4 |
| Unknown | non-Hispanic white | Unknown | 3 |
| Unknown | Other | Unknown | 1 |
| - | Hispanic | Hispanic | 512 |
| - | non-Hispanic white | - | 440 |
| - | Unknown | Unknown | 363 |
| - | Hispanic | Unknown | 254 |
| - | - | Other | 76 |
| - | - | Unknown | 6 |
| - | non-Hispanic white | Unknown | 3 |

Of the 673 with NAACCR records (all, not just the 486 meeting the current criteria, see [sec. 1](#sec:overview)) only 22 have differences between [Hispanic (NAACCR)](#a_hsp_naaccr) and [Hispanic (broad)](#a_hsp_broad) but 229 have differences between [Hispanic (NAACCR)](#a_hsp_naaccr) and [Hispanic (strict)](#a_hsp_strict).

According to [Hispanic (NAACCR)](#a_hsp_naaccr), [Hispanic (broad)](#a_hsp_broad), and [Hispanic (strict)](#a_hsp_strict) respectively, 52.6%, 55.1%, and 31.6% of the NAACCR patients are Hispanic. At 55.1% [Hispanic (broad)](#a_hsp_broad) comes the closest to the [2016 Census estimates for San Antonio](https://www.census.gov/quickfacts/fact/table/sanantoniocitytexas/HSD410216). Also, anecdotal evidence suggests that Hispanic ethnicity is under-reported. This argues for using [Hispanic (broad)](#a_hsp_broad) when possible, but I will keep [Hispanic (strict)](#a_hsp_strict) available for sensitivity analysis.

## Appendix 3.3 What is going on with the first contact variable?

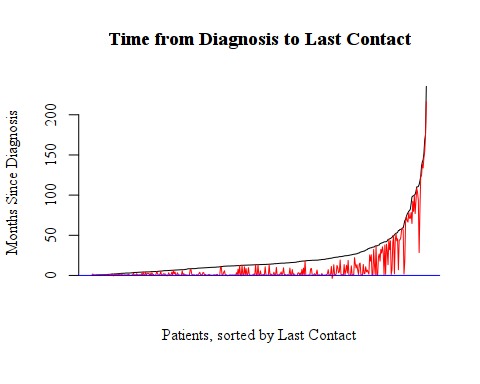


Figure 11: Wierd observation– [0580 Date of 1st Contact](#n_fc) (red) is almost always between [1750 Date of Last Contact](#n_lc) (black) and [0390 Date of Diagnosis](#n_ddiag) (blue) though diagnosis is usually on a biopsy sample and that’s why it’s dated as during or after surgery we thought. If first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

###### blank

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

*This section is no longer relevant but is still available for reference in the [kidneycancer\_181009 snapshot of this document](https://rstudio-pubs-static.s3.amazonaws.com/427637_7a87dcd0aeab42daa309e7d158ef59a2.html" \l "what-is-the-coverage-of-valid-records-in-each-data-source)*

## Appendix 3.5 Which variables are near-synonymous?

*This section is no longer relevant but is still available for reference in the [kidneycancer\_181009 snapshot of this document](https://rstudio-pubs-static.s3.amazonaws.com/427637_7a87dcd0aeab42daa309e7d158ef59a2.html" \l "which-variables-are-near-synonymous)*

# Appendix 4 Variable descriptions

Here are descriptions of the variables referenced in this document.

###### patient\_num

patient\_num :

patient\_num

###### n\_rectype

1880 Recurrence Type–1st :

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 :

3170 RX Date–Most Defin Surg; Date 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\_surgreason

1340 Reason for No Surgery :

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\_ddiag

0390 Date of Diagnosis :

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 :

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 :

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\_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](http://datadictionary.naaccr.org/default.aspx?c=10" \l "1760)

###### n\_cstatus

1770 Cancer Status :

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 :

1860 Recurrence Date–1st

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

###### n\_seer\_kcancer

Kidney and Renal Pelvis :

Kidney and Renal Pelvis; SEER site

###### n\_kcancer

Kidney, NOS :

Kidney, NOS; KC, Registry

###### e\_surgonc

Surgical Oncology :

Surgical Oncology; Visit to Surgical Oncology; Visit to Surgical Oncology (UT Health)

###### n\_dsdisc

3180 RX Date–Surgical Disch :

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)

###### v022\_drvd\_ajcc\_stg

3430 Derived AJCC-7 Stage Grp :

3430 Derived AJCC-7 Stage Grp

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3430](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3430)

###### v023\_drvd\_dscrpt

3422 Derived AJCC-7 M Descript :

3422 Derived AJCC-7 M Descript

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3422](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3422)

###### v024\_drvd\_ajcc\_m

3420 Derived AJCC-7 M :

3420 Derived AJCC-7 M

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3420](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3420)

###### v025\_drvd\_dscrpt

3412 Derived AJCC-7 N Descript :

3412 Derived AJCC-7 N Descript

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3412](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3412)

###### v026\_drvd\_ajcc\_n

3410 Derived AJCC-7 N :

3410 Derived AJCC-7 N

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3410](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3410)

###### v027\_drvd\_dscrpt

3402 Derived AJCC-7 T Descript :

3402 Derived AJCC-7 T Descript

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3402](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3402)

###### v028\_drvd\_ajcc\_t

3400 Derived AJCC-7 T :

3400 Derived AJCC-7 T

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3400](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3400)

###### v029\_drvd\_ajcc\_stg

3000 Derived AJCC-6 Stage Grp :

3000 Derived AJCC-6 Stage Grp

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#3000](http://datadictionary.naaccr.org/default.aspx?c=10" \l "3000)

###### v030\_drvd\_dscrpt

2990 Derived AJCC-6 M Descript :

2990 Derived AJCC-6 M Descript

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2990](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2990)

###### v031\_drvd\_ajcc\_m

2980 Derived AJCC-6 M :

2980 Derived AJCC-6 M

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2980](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2980)

###### v032\_drvd\_dscrpt

2970 Derived AJCC-6 N Descript :

2970 Derived AJCC-6 N Descript

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2970](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2970)

###### v033\_drvd\_ajcc\_n

2960 Derived AJCC-6 N :

2960 Derived AJCC-6 N

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2960](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2960)

###### v034\_drvd\_dscrpt

2950 Derived AJCC-6 T Descript :

2950 Derived AJCC-6 T Descript

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2950](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2950)

###### v035\_drvd\_ajcc\_t

2940 Derived AJCC-6 T :

2940 Derived AJCC-6 T

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2940](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2940)

###### v037\_tnm\_pth\_dscrptr

0920 TNM Path Descriptor :

0920 TNM Path Descriptor

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#920](http://datadictionary.naaccr.org/default.aspx?c=10" \l "920)

###### v051\_tnm\_cln\_t

0940 TNM Clin T :

0940 TNM Clin T

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#940](http://datadictionary.naaccr.org/default.aspx?c=10" \l "940)

###### v052\_tnm\_cln\_n

0950 TNM Clin N :

0950 TNM Clin N

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#950](http://datadictionary.naaccr.org/default.aspx?c=10" \l "950)

###### v053\_tnm\_cln\_m

0960 TNM Clin M :

0960 TNM Clin M

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#960](http://datadictionary.naaccr.org/default.aspx?c=10" \l "960)

###### v054\_tnm\_cln\_stg\_grp

0970 TNM Clin Stage Group :

0970 TNM Clin Stage Group

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#970](http://datadictionary.naaccr.org/default.aspx?c=10" \l "970)

###### v055\_tnm\_cln\_dscrptr

0980 TNM Clin Descriptor :

0980 TNM Clin Descriptor

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#980](http://datadictionary.naaccr.org/default.aspx?c=10" \l "980)

###### v062\_tnm\_pth\_stg\_grp

0910 TNM Path Stage Group :

0910 TNM Path Stage Group

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#910](http://datadictionary.naaccr.org/default.aspx?c=10" \l "910)

###### v073\_tnm\_pth\_m

0900 TNM Path M :

0900 TNM Path M

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#900](http://datadictionary.naaccr.org/default.aspx?c=10" \l "900)

###### v074\_tnm\_pth\_n

0890 TNM Path N :

0890 TNM Path N

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#890](http://datadictionary.naaccr.org/default.aspx?c=10" \l "890)

###### v079\_tnm\_pth\_t

0880 TNM Path T :

0880 TNM Path T

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#880](http://datadictionary.naaccr.org/default.aspx?c=10" \l "880)

###### n\_dob

0240 Date of Birth :

0240 Date of Birth

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

###### birth\_date

birth\_date :

birth\_date

###### 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](http://datadictionary.naaccr.org/default.aspx?c=10" \l "150)

###### e\_marital

Marital Status :

Marital Status; Marital Status, i2b2

###### n\_sex

0220 Sex :

0220 Sex; Sex, Registry

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#220](http://datadictionary.naaccr.org/default.aspx?c=10" \l "220)

###### sex\_cd

sex\_cd :

sex\_cd; Sex, i2b2

###### 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, Black Asian, Pac Islander, Other, and Unknown

###### race\_cd

race\_cd :

race\_cd; Race, i2b2

###### n\_hisp

0190 Spanish/Hispanic Origin :

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)

###### e\_hisp

Hispanic or Latino :

Hispanic or Latino; Hispanic Origin, i2b2

###### e\_death

Death, i2b2 :

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

###### s\_death

Deceased per SSA :

Deceased per SSA; Death, SSN

###### e\_dscdeath

Expired :

Expired; Discharge Disposition

###### n\_brthplc

0250 Birthplace :

0250 Birthplace

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#250](http://datadictionary.naaccr.org/default.aspx?c=10" \l "250)

###### n\_mets

2850 CS Mets at DX :

2850 CS Mets at DX

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#2850](http://datadictionary.naaccr.org/default.aspx?c=10" \l "2850)

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

###### n\_mult

0446 Multiplicity Counter :

0446 Multiplicity Counter

Link: [http://datadictionary.naaccr.org/default.aspx?c=10#446](http://datadictionary.naaccr.org/default.aspx?c=10" \l "446)

###### a\_tdeath

Death :

Death; Death

###### 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\_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\_tdiag

Diagnosis :

Diagnosis; Diagnosis

###### a\_trecur

Recurrence :

Recurrence; Recurrence; Analytic master variable for time to recurrence. Based on [n\_drecur](#n_drecur)

###### a\_tsurg

Surgery :

Surgery; Surgery

###### a\_hsp\_naaccr

Hispanic (NAACCR) :

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

###### 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.

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

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

###### NA

NA :

###### NA

NA :

###### NA

NA :

###### NA

NA :

###### start\_date

start\_date :

start\_date

###### e\_kc\_i9

189.0 Malignant neoplasm of kidney, except pelvis :

189.0 Malignant neoplasm of kidney, except pelvis; KC ICD9, i2b2; 189.0 Malignant neoplasm of kidney, except pelvis

###### e\_kc\_i10

C64 Malignant neoplasm of kidney, except renal pelvis :

C64 Malignant neoplasm of kidney, except renal pelvis; KC ICD10, i2b2; C64 Malignant neoplasm of kidney, except renal pelvis

###### e\_i9neph

V45.73 Acquired absence of kidney :

V45.73 Acquired absence of kidney; V45.73 Acquired absence of kidney

###### e\_i10neph

Z90.5 Acquired absence of kidney :

Z90.5 Acquired absence of kidney

###### e\_hstneph

HX NEPHRECTOMY :

HX NEPHRECTOMY; Surgical history

###### v008\_scndr\_nrndcrn\_inactive

C7B-C7B Secondary neuroendocrine tumors (C7B) :

C7B-C7B Secondary neuroendocrine tumors (C7B); C7B-C7B Secondary neuroendocrine tumors (C7B)

###### v009\_mlgnt\_unspcfd

C79 Secondary malignant neoplasm of other and unspecified sites :

C79 Secondary malignant neoplasm of other and unspecified sites; C79 Secondary malignant neoplasm of other and unspecified sites

###### v009\_mlgnt\_unspcfd\_inactive

C79 Secondary malignant neoplasm of other and unspecified sites :

C79 Secondary malignant neoplasm of other and unspecified sites; C79 Secondary malignant neoplasm of other and unspecified sites

###### v010\_rsprtr\_dgstv

C78 Secondary malignant neoplasm of respiratory and digestive organs :

C78 Secondary malignant neoplasm of respiratory and digestive organs; C78 Secondary malignant neoplasm of respiratory and digestive organs

###### v010\_rsprtr\_dgstv\_inactive

C78 Secondary malignant neoplasm of respiratory and digestive organs :

C78 Secondary malignant neoplasm of respiratory and digestive organs; C78 Secondary malignant neoplasm of respiratory and digestive organs

###### v011\_unspcfd\_mlgnt

C77 Secondary and unspecified malignant neoplasm of lymph nodes :

C77 Secondary and unspecified malignant neoplasm of lymph nodes; C77 Secondary and unspecified malignant neoplasm of lymph nodes

###### v011\_unspcfd\_mlgnt\_inactive

C77 Secondary and unspecified malignant neoplasm of lymph nodes :

C77 Secondary and unspecified malignant neoplasm of lymph nodes; C77 Secondary and unspecified malignant neoplasm of lymph nodes

###### v012\_unspcfd\_mlgnt

196 Secondary and unspecified malignant neoplasm of lymph nodes :

196 Secondary and unspecified malignant neoplasm of lymph nodes; 196 Secondary and unspecified malignant neoplasm of lymph nodes

###### v012\_unspcfd\_mlgnt\_inactive

196 Secondary and unspecified malignant neoplasm of lymph nodes :

196 Secondary and unspecified malignant neoplasm of lymph nodes; 196 Secondary and unspecified malignant neoplasm of lymph nodes

###### v013\_rsprtr\_dgstv

197 Secondary malignant neoplasm of respiratory and digestive systems :

197 Secondary malignant neoplasm of respiratory and digestive systems; 197 Secondary malignant neoplasm of respiratory and digestive systems

###### v013\_rsprtr\_dgstv\_inactive

197 Secondary malignant neoplasm of respiratory and digestive systems :

197 Secondary malignant neoplasm of respiratory and digestive systems; 197 Secondary malignant neoplasm of respiratory and digestive systems

###### v014\_mlgnt\_spcfd

198 Secondary malignant neoplasm of other specified sites :

198 Secondary malignant neoplasm of other specified sites; 198 Secondary malignant neoplasm of other specified sites

###### v014\_mlgnt\_spcfd\_inactive

198 Secondary malignant neoplasm of other specified sites :

198 Secondary malignant neoplasm of other specified sites; 198 Secondary malignant neoplasm of other specified sites

###### NA

NA :

###### NA

NA :

###### language\_cd

language\_cd :

language\_cd; Language, i2b2

###### e\_lng

Language :

Language

###### e\_eth

Ethnicity :

Ethnicity; EMR demographics

##### v055\_tnm\_cln\_dscrptr

Test section

# Appendix 5 Audit trail

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sequence | time | type | name | hash |
| 0001 | 2018-10-10 15:19:49 | info | sessionInfo | - |
| 0002 | 2018-10-10 15:19:49 | this\_script | exploration.spin.Rmd | TEST\_OUTPUT\_DO\_NOT\_USE |
| 0003 | 2018-10-10 15:20:02 | rdata | .depdata[ii] = “dictionary.R.rdata” | 6cfeb3856eb31a14bfc158517ebda554 |
| 0004 | 2018-10-10 15:20:10 | rdata | .depdata[ii] = “data.R.rdata” | c12830910b4a5c1d236305e6c90f238e |
| 0003.0001 | 2018-10-10 10:59:50 | info | sessionInfo | - |
| 0003.0002 | 2018-10-10 10:59:50 | this\_script | dictionary.R | 8faecd3 |
| 0003.0003 | 2018-10-10 11:00:02 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0003.0004 | 2018-10-10 11:00:14 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0003.0005 | 2018-10-10 11:00:14 | file | rawdct = “local/in/meta\_HSC20170563N\_kc\_v200.int.csv” | 77226290495672d030798e64327fe10a |
| 0003.0006 | 2018-10-10 11:00:14 | file | tpldct = “datadictionary\_static.csv” | e1e120a2efd284d032682d8ff89235c6 |
| 0003.0007 | 2018-10-10 11:00:17 | info | sessionInfo | - |
| 0003.0008 | 2018-10-10 11:00:17 | save | save | - |
| 0004.0001 | 2018-10-10 11:00:46 | info | sessionInfo | - |
| 0004.0002 | 2018-10-10 11:00:46 | this\_script | data.R | 8faecd3 |
| 0004.0003 | 2018-10-10 11:00:58 | rdata | .depdata = “dictionary.R.rdata” | 6cfeb3856eb31a14bfc158517ebda554 |
| 0004.0004 | 2018-10-10 11:00:58 | file | levels\_map\_file = “levels\_map.csv” | 4c66bc0cd1fd35eb9e64c3c49296a05f |
| 0004.0005 | 2018-10-10 11:01:34 | seed | project\_seed | - |
| 0004.0006 | 2018-10-10 11:03:25 | info | sessionInfo | - |
| 0004.0007 | 2018-10-10 11:03:26 | save | save | - |
| 0004.0003.0001 | 2018-10-10 10:59:50 | info | sessionInfo | - |
| 0004.0003.0002 | 2018-10-10 10:59:50 | this\_script | dictionary.R | 8faecd3 |
| 0004.0003.0003 | 2018-10-10 11:00:02 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0004.0003.0004 | 2018-10-10 11:00:14 | file | inputdata = “local/in/HSC20170563N\_kc\_v200.int.csv” | caa0a30bd87cd77659b118986cab73a4 |
| 0004.0003.0005 | 2018-10-10 11:00:14 | file | rawdct = “local/in/meta\_HSC20170563N\_kc\_v200.int.csv” | 77226290495672d030798e64327fe10a |
| 0004.0003.0006 | 2018-10-10 11:00:14 | file | tpldct = “datadictionary\_static.csv” | e1e120a2efd284d032682d8ff89235c6 |
| 0004.0003.0007 | 2018-10-10 11:00:17 | info | sessionInfo | - |
| 0004.0003.0008 | 2018-10-10 11:00:17 | save | save | - |

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