Dataset Integrity Check for the Chronic Renal Insufficiency Cohort (CRIC)

Prepared by Sabrina Chen
3901 Calverton Blvd, Suite 200 Calverton MD 20705
October 12, 2020

Contents

1 Standard Disclaimer

The intent of this DSIC is to provide confidence that the data distributed by the NIDDK repository is a true copy of the study data. Our intent is not to assess the integrity of the statistical analyses reported by study investigators. As with all statistical analyses of complex datasets, complete replication of a set of statistical results should not be expected in secondary analysis. This occurs for a number of reasons including differences in the handling of missing data, restrictions on cases included in samples for a particular analysis, software coding used to define complex variables, etc. Experience suggests that most discrepancies can ordinarily be resolved by consultation with the study data coordinating center (DCC), however this process is labor-intensive for both DCC and Repository staff. It is thus not our policy to resolve every discrepancy that is observed in an integrity check. Specifically, we do not attempt to resolve minor or inconsequential discrepancies with published results or discrepancies that involve complex analyses, unless NIDDK Repository staff suspect that the observed discrepancy suggests that the dataset may have been corrupted in storage, transmission, or processing by repository staff. We do, however, document in footnotes to the integrity check those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

2 Study Background

The Chronic Renal Insufficiency Cohort (CRIC) Study is an observational study that examined risk factors for progression of chronic renal insufficiency (CRI) and cardiovascular disease (CVD) among CRI patients. The study enrolled adults aged 21 to 74 years with a broad spectrum of renal disease severity, half of whom were diagnosed with diabetes mellitus. Subjects underwent extensive clinical evaluation at baseline and at annual clinic visits and via telephone at 6 month intervals. Data on quality of life, dietary assessment, physical activity, health behaviors, depression, cognitive function, health care resource utilization, as well as blood and urine specimens were collected. The primary renal outcome measure was reduction in estimated GFR. Renal events were defined as the need for renal replacement therapy (ESRD), an estimated halving of GFR, and/or a 25 ml/min per 1.73 m2 decline in GFR from baseline.

3 Archived Datasets

All data files, as provided by the Data Coordinating Center (DCC), are located in the CRIC study data package. For this replication, variables were taken from the derived datasets: “personlevel.sas7bdat”, “visitlevel.sas7bdat”, and “anc_vitd.sas7bdat”.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Reese et al. [1] in the American Journal of Nephrology, October 2013. To verify the integrity of the datasets, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Table 1- Participant characteristics, overall and according to eGFR category, Table A lists the variables that were used in the replication and Table B compares the results calculated from the archived data file to the results published in Table 1.

There was a discrepancy found in the education and income variables listed in Table B below for the overall group.

6 Conclusions

There were some discrepancies compared to published results. The DCC confirmed the formats for INCOME and EDUCATION were listed in reverse order and that an erratum was sent to the journal.

7 References

[1] Peter P. Reese Anne R. Cappola Justine Shults Raymond R. Townsend Crystal A. Gadegbeku Cheryl Anderson Joshua F. Baker Dean Carlow Michael J. Sulik Joan C. Lo Alan S. Go Bonnie Ky Laura Mariani Harold I. Feldman Mary B. Leonard CRIC Study Investigators. Physical Performance and Frailty in Chronic Kidney Disease. Am J Nephrol 2013;38:307–315.

Table A: Variables used to replicate Table 1- Baseline Characteristics of the Derivation and Validation CRIC Populations

Characteristic	**dataset.variable**
Black	personlevel.black
Sex	personlevel.sex
Education	personlevel.edu_cat_3
Income	personlevel.income_cat_1
Vitamin D	anc_vitd.vitd_25_ms
Age	visitlevel.age_integer
Stroke	visitlevel.stroke
Anemia	visitlevel.anemia
BMI	visitlevel.bmi
KDQOL: SF-12 Mental Composite	visitlevel.sf12_mcs
KDQOL: SF-12 Physical Composite	visitlevel.sf12_pcs
Diabetes	visitlevel.diabetes
Congestive Heart Failure	visitlevel.chf
MI/Prior revascularization	visitlevel.mirevasc

Table B-1: Comparison of values computed in integrity check to reference article Table 1 values (n, %)

COVAR	COVARF	COUNT			PERCENT		
		DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
AGE_INTEGERgp	Yes	535	559	-24	48	50	-2
sex	Female	526	526	0	47	47	0
BLACK	Black	396	396	0	36	36	0
DIABETES	Yes	474	473	1	43	43	0
STROKE	Yes	99	99	0	9	9	0
Cardiovascular disease	Yes	286	286	0	26	26	0
ANEMIA	Yes	413	413	0	37	37	0
BMI Category	<25	204	204	0	18	19	-1
	>=25 and <30	354	353	1	32	32	0
	>=30 and <35	282	282	0	25	26	-1
	35+	265	263	2	24	24	0
Education	Less than high school	81	550	-469	7	50	-43
	High school graduate	158	321	-163	14	29	-15
	Some college	321	158	163	29	14	15
	College graduate or higher	550	81	469	50	7	43
Income	$20,000 or under	145	214	-69	13	19	-6
	$20,001 - $50,000	281	309	-28	25	28	-3
	$50,001 - $100,000	309	281	28	28	25	3
	More than $100,000	214	145	69	19	13	6
	Don't wish to answer	162	162	0	15	15	0
Vit D deficiency	Yes	233	232	1	21	21	0

Table B-2: Comparison of values computed in integrity check to reference article Table 1 values (median, IQR)

VAR	_Q1_			_MEDIAN_			_Q3_		
	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
Age	57	57	0	64	65	-1	70	71	-1
BMI, kg/m2	26	26	0	30	30	0	35	35	0
KDQOL score Physical component	37	37	0	48	48	0	54	54	0
KDQOL score Mental component	46	46	0	55	55	0	59	59	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase nofmterr ls=200 pageno=1 /*nolabel*/ ;

title '/prj/niddk/ims_analysis/CRIC/prog_initial_analysis/CRIC.dsic.sas';
run;

/*                                                                                                                                    */
/* DSIC check: Table 1 for P. Reese 2013 Nephrology manuscript.
*/
/*                                                                                                                                    */


************;
* INPUT     ;
************;

* The Aug delivery has an updated VISITS file;
libname sasaug  '/prj/niddk/ims_analysis/CRIC/private_orig_data/Final_Data_for_Update_2020_AUG/Final_Data_for_Update_2020_AUG/';
libname sasmay  '/prj/niddk/ims_analysis/CRIC/private_orig_data/Final_Data_for_Transfer_2020_MAY/Final_Data_for_Transfer_2020_MAY/Study_Data/Derived_Data';

* the 1,111 analysis cohort;
proc import datafile="/prj/niddk/ims_analysis/CRIC/private_orig_data/Final_Data_for_Update_2020_AUG/MS127_cohort_09212020.xlsx"
  dbms=xlsx replace
  out=work.cohort;
  getnames=yes;
run;

libname fmts '/prj/niddk/ims_analysis/CRIC/private_orig_data/Final_Data_for_Transfer_2020_MAY/Final_Data_for_Transfer_2020_MAY/Study_Data/Derived_Data';

PROC FORMAT CNTLIN = fmts.formats64;
run;


************;
* MACROS    ;
************;
%macro readin(sasfile, ds, dsname);
  data &dsname;
    set &sasfile..&ds;
  run;

  proc contents data=&dsname;
  title3 "&ds";
  run;
%mend;


* produce n and %;
%macro tbl1all(rownum, var, varf, subset, subsetname);
  proc freq data=table1_s noprint;
    where analy in(&subset);
    tables &var/list missing out=tbl1&subsetname;
  run;
```

```
  data tbl1&subsetname;
    length covar covarf $100;
    set tbl1&subsetname;
    covar = "&var";
    covarf = put(&var,&varf..);
    rownum = &rownum;
  run;

  data prnt&subsetname;
    set prnt&subsetname tbl1&subsetname;
  run;

%mend;


%macro univ(rownum, var, subset, subsetname);

  proc univariate data=table1_s outtable= univ&subsetname noprint;
    where &subset=1;
    var &var
       ;
  run;

  data univ&subsetname;
    length covarf $100 _var_ $25;
    set univ&subsetname;
        covarf = "&subset";
        rownum = &rownum;
  run;

  data prntuniv&subsetname;
    set prntuniv&subsetname univ&subsetname;
  run;

%mend;




*************;
* FORMATS    ;
*************;
proc format;
  value novalue
      . = "No Value"
  other = "   Value"
  ;

  value egfr
  . = "Missing"
 10 = "Stage1:>=90"
 20 = "Stage2: 60 to<90"
 30 = "Stage3a: 45 to<60"
 35 = "Stage3b: 30 to<45"
 40 = "Stage4: 15 to<30"
```

```
 50 = "Stage5:<15"
 ;

 value bmigpf
 1 = '<25'
 2 = '>=25 and <30'
 3 = '>=30 and <35'
 4 = '35+'
 ;

run;


%readin(sasmay, anc_vitd, anc_vitd );
%readin(sasmay, personlevel, personlevel );
%readin(sasaug, visitlevel, visitlevel)

proc print data=cohort (obs=10);
run;


* preliminary freqs;
proc freq data=anc_vitd;
  tables    VDBP  /*         Vitamin D binding protein       */
    VITD_1_25     /* 1,25-dihydroxy vitamin D (pg/mL)        */
    VITD_25_MS    /* Serum 25(OH)-Total Vitamin D LC/MSMS    */
    VITD_2_25_MS  /* Serum 25(OH)2 Vitamin D LC/MSMS         */
    VITD_3_25_MS  /* Serum 25(OH)3 Vitamin D LC/MSMS         */ /missing;
run;

proc freq data=personlevel;
  tables BLACK sex DIABETES_AT_BASELINE EDU_CAT_3 INCOME_CAT_1/missing;

proc freq data=visitlevel;
  tables EGFR_CRIC_STAGES AGE_INTEGER STROKE ANYCVD ANEMIA BMI
         SF12_MCS
         SF12_PCS
          vnum  /missing;
  format EGFR_CRIC_STAGES egfr.;
run;




** subset to the 1,111 cohort;
proc sort data=cohort;
  by pid vnum;
run;

proc sort data=visitlevel;
  by pid vnum;
run;

proc sort data=anc_vitd;
  by pid vnum;
run;
```

```
proc sort data=personlevel;
  by pid;
run;

data table1_s;
  merge cohort      (in=in1 keep=pid vnum)
        personlevel (in=in2 keep=pid  BLACK sex  EDU_CAT_3 INCOME_CAT_1);
  by pid;
  if in1;
run;

data table1_s;
  merge table1_s   (in=in1 keep=pid vnum  black sex edu_cat_3 income_cat_1)
        anc_vitd   (in=in2 keep=pid vnum  vitd_25_ms)
        visitlevel (in=in3 keep=pid vnum  egfr_cric age_integer stroke anycvd anemia bmi sf12_mcs sf12_pcs diabetes chf mirevasc
                           egfr_cric_cat1    egfr_cric_cat5    egfr_cric_stages)
        ;
  by pid vnum;
  if (in2 or in3) and not in1 then put "NOTE: unexpected " pid= vnum=;
  if in1;

  analy=1;

  *     Age & Age > 65 years: AGE_INTEGER, defined by AGE_INTEGER>=65    ;
  if AGE_INTEGER => 65 then AGE_INTEGERgp = 1;
  else if 0 < AGE_INTEGER < 65 then AGE_INTEGERgp=0;

  *     Cardiovascular disease: defined by CHF=1 OR MIREVASC=1        ;
  if CHF=1 OR MIREVASC=1 then cardiovas = 1;
  else cardiovas = 0;

  *vitamin d??  Vitamin D deficiency categorized as <20 ng/ml. ;
  * VITD_25_MS<20;
  if . < VITD_25_MS<20 then VITD_25_MSgp = 1;
  else if VITD_25_MS>=20 then VITD_25_MSgp = 0;

  * BMI cats;
  if . < bmi < 25       then bmigp = 1;
  else if 25 <= bmi <30 then bmigp = 2;
  else if 30 <= bmi <35 then bmigp = 3;
  else if 35 <= bmi     then bmigp = 4;

run;

proc freq data=table1_s;
  tables vitd_25_ms
         black sex diabetes edu_cat_3 income_cat_1
         age_integer stroke anycvd anemia bmi
         sf12_mcs
         sf12_pcs
         egfr_cric
         egfr_cric_cat1
         egfr_cric_cat5
         egfr_cric_stages/missing;
  tables age_integergp*age_integer/list missing;
```

```
  tables cardiovas *chf* mirevasc/list missing;
  tables vitd_25_msgp*vitd_25_ms/list missing;
  tables bmigp*bmi/list missing;
  format   sf12_mcs  sf12_pcs novalue.;
  title3 "checking";
run;



** Table 1 Overall;
data prntall;
  set _null_;
run;

%tbl1all(1,  AGE_INTEGERgp ,   yesno         , 1 , all);
%tbl1all(2,  sex           ,   sex           , 1 , all);
%tbl1all(3,  BLACK         ,   BLACK         , 1 , all);
%tbl1all(4,  DIABETES      ,   yesno         , 1 , all);
%tbl1all(5,  STROKE        ,   YESNOTYES     , 1 , all);
%tbl1all(6,  cardiovas     ,   yesno         , 1 , all);
%tbl1all(7,  ANEMIA        ,   yesno         , 1 , all);
%tbl1all(8,  bmigp         ,   bmigpf        , 1 , all);
%tbl1all(9, EDU_CAT_3     ,   EDU_CAT_3A    , 1 , all);
%tbl1all(10, INCOME_CAT_1  ,   INCOME_CAT_1A , 1 , all);
%tbl1all(11, VITD_25_MSgp  ,   yesno         , 1 , all);
* skip dialysis;


data prntall;
  set prntall;
   percent     = round(percent    );
run;

proc print data=prntall;
  where covarf not in("No", "Male", "Not Black", "Not Yes", "Missing");
  var rownum covar covarf count percent;
  title3 'Table 1 Overall - n(%)';
run;



data prntunivall;
  set _null_;
run;

%univ(12,   AGE_INTEGER , 1 , all);
%univ(13,   BMI         , 1 , all);
%univ(14,  SF12_PCS    , 1 , all);
%univ(15,  SF12_MCS    , 1 , all);


data prntunivall;
  set prntunivall;
   _median_  = round(_median_ );
   _q1_      = round(_q1_     );
   _q3_      = round(_q3_     );
```

```
   _mean_    = round(_mean_   );
   _std_     = round(_std_    );
run;

proc print data=prntunivall;
 var rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ /* _min_ _max_ _mean_ _std_ */;
 title3 "Table 1 Overall - median IQR";
run;


proc export  outfile='/prj/niddk/ims_analysis/CRIC/private_created_data/CRIC.dsic.n.20201007.xlsx'
  dbms=xlsx
  replace
  data=prntall (keep=rownum covar covarf count percent);


proc export  outfile='/prj/niddk/ims_analysis/CRIC/private_created_data/CRIC.dsic.med.20201007.xlsx'
  dbms=xlsx
  replace
  data=prntunivall (keep=rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ );
```