Dataset Integrity Check for the Chronic Renal Insufficiency Cohort (CRIC) Phase 3 Data Files

Prepared by Sabrina Chen
3901 Calverton Blvd, Suite 200 Calverton MD 20705
May 23, 2019

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”.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Schrauben et al. in the Kidney International Reports, January 2019. To verify the integrity of the datasets, descriptive statistics were computed.

5 Results

For Table 2 in the publication [1], Table 2-Demographic characteristics of latent class-defined health behavior engagement patterns in the CRIC phase I and III cohorts overall, and by age group (<65 years and $65 years), 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 gender/male variable listed in Table B below for the "65+" age group. This information was sent to the Data Coordinating Center and an erratum was sent to the journal of the publication.

6 Conclusions

The NIDDK repository is confident that the CRIC data files to be distributed are a true copy of the study data. While there was a discrepancy noted in Table B, the CRIC Data Coordinating Center confirmed that the issue was a typo in the manuscript and not in the CRIC data files to be distributed.

7 References

[1] Sarah J. Schrauben, Jesse Y. Hsu, Julie Wright Nunes, Michael J. Fischer, Anand Srivastava, Jing Chen, Jeanne Charleston, Susan Steigerwalt, Thida C. Tan, Jeffrey C. Fink, Ana C. Ricardo, James P. Lash, Myles Wolf, Harold I. Feldman, and Amanda H. Anderson and the CRIC Study Investigators. Health Behaviors in Younger and Older Adults With CKD: Results From the CRIC Study. Kidney International Reports. January 2019Volume 4, Issue 1, Pages 80–93.

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

Characteristic	**dataset.variable**
Age	age_integer.visitlevel
Education	edu_cat_1.personlevel
Income	income_cat_1.personlevel
Insurance	hins_cat1.personlevel
Gender	sex.personlevel
Race	race_ethnicity_cat2.personlevel

Table B: Comparison of values computed in integrity check to reference article Table 1 values

Characteristic	Manuscript (N=908)	DSIC (N=908)	Difference (N=0)	Manuscript (n=652)	DSIC (n=652)	Difference (n=0)
	Age < 65 years			Age 65+ years		
Age, mean (SD)	54 (9)	54(9)	0(0)	70 (3)	70(3)	0(0)
Gender, male, %	55	55	0	42	58	16
Race, %						
Non-Hispanic white	39	39	0	45.9	46	0.1
Non-Hispanic black	45	45	0	42	42	0
Hispanic	12	12	0	9	9	0
Other	4	4	0	3	3	0
Education, %						
High school or less	36	36	0	41	41	0
College or more	64	64	0	59	59	0
Insurance, %						
Medicaid	24	24	0	13	13	0

Characteristic	Manuscript (N=908)	DSIC (N=908)	Difference (N=0)	Manuscript (n=652)	DSIC (n=652)	Difference (n=0)
Medicare/VA	42	42	0	80	80	0
Private/commercial	34	34	0	7	7	0
Income, %						
<=$20,000	37	37	0	33	33	0
$20-50,000	26	26	0	35	34	-1
>$50,000	37	37	0	32	32	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase nofmterr;

title '/prj/niddk/ims_analysis/CRIC/prog_initial_analysis/explore.2017.vs.V7.files.20190320.sas';
run;

**********;
* INPUT   ;
**********;
LIBNAME PUBLIC  '/prj/niddk/dataset_files/CRIC_V7/Data/CRIC_Study_Data/Derived_Data';

LIBNAME SAS1    '/prj/niddk/ims_analysis/CRIC/private_orig_data/final_data_for_transfer_2017/final_data_for_transfer_2017/Study_Data/Derived_Data/Phase I';
LIBNAME SAS3    '/prj/niddk/ims_analysis/CRIC/private_orig_data/final_data_for_transfer_2017/final_data_for_transfer_2017/Study_Data/Derived_Data/Phase III';

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

  value negpos
  0 = "Negative"
  1 = "Positive"
  ;

  value noyes
  0 = "No"
  1 = "Yes"
  ;

  value zerof
  . = "0"
  ;

  value sexf
  1 = "Male"
  2 = "Female"
  ;

  value educf
  1 = '6th grade or less ?'
  2 = '7th to 12th grade, no highschool diploma?'
  3 = 'High school graduate or equivalent ?'
  4 = 'Technical or vocational school degree ?'
  5 = 'Some college education, but not complete?'
  6 = 'College graduate ?'
  7 = 'Professional or graduate degree(e.g. Mas?'
  ;

  value educ2gf
```

```
  1 = "<= High school"
  2 = "College or more"
  ;

  value racef
  1 = "Non-Hispanic white"
  2 = "Non-Hispanic black"
  3 = "Hispanic"
  4 = "Other"
  ;

  value incomef
  1 = "<= $20,000"
  2 = "$20-50,000"
  3 = "> $50,000"
  ;

  value agelt65f
  1 = "< 65"
  ;

  value agege65f
  1 = ">= 65"
  ;

  value hinsf
  1 = "Medicaid"
  2 = "Medicare/VA"
  3 = "Private/commercial"
  ;

run;



************;
* MACROS    ;
************;
%macro readin(lib, ds, suf);
  data &ds.&suf;
    set &lib..&ds;
  run;

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


* produce n and %;
%macro npercent(rownum, var, varf, subset, subsetname);
  proc freq data=analy noprint;
    where &subset = 1;
    tables &var/list missprint 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=analy outtable= univ&subsetname noprint;
    where &subset=1;
    var &var
       ;
  run;

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

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

%mend;



%readin(sas3, visitlevel, p3);

proc freq data=visitlevelp3;
   tables diabetes/missing;
   tables egfr_roche proteinuria anycvd systolic diastolic controlled1 controlled2 sf12_pcs modminwk bmi bmi_cat_1 bmi_cat_2  bmi_cat_3       diet_fib
/missing;
run;

%readin(sas3, personlevel, p3);

proc freq data=personlevelp3;
  tables diabetes_at_baseline/missing;
  tables sex race_ethnicity_cat2 ccidsite edu_cat_1 edu_cat_2 edu_cat_3  /missing;
  tables income_cat_1  hins_cat1 hins_cat2/missing;
  run;
```

```
%readin(sas1,sa_allc   , p1);
%readin(sas1,sa_cvd_cnsr, p1);
%readin(sas1,sa_cvd    , p1);

%readin(public, visitlevel , p1);

proc freq data=visitlevelp1;
  tables VNUM/missing;
run;

proc sort data=visitlevelp1 nodupkey;
  where vnum = 3;
  by pid;
run;

data checkvisp1;
  set visitlevelp1;
  by pid;
  if not (first.pid and last.pid);
run;


%readin(public, personlevel, p1);


* check overlap;
data overlap;
  merge personlevelp1 (in=in1 keep=pid) personlevelp3 (in=in2 keep=pid);
  by pid;
  if in1 then in_phase1=1;
  if in2 then in_phase3=1;
run;

proc freq data=overlap;
  tables in_phase1*in_phase3/list missing;
run;



** Phase 3 ;
data analyp3;
  merge visitlevelp3  (in=in1 keep= pid diabetes age_integer egfr_roche proteinuria anycvd systolic diastolic controlled1 controlled2 sf12_pcs modminwk bmi
                                    bmi_cat_1 bmi_cat_2  bmi_cat_3 diet_fib)
        personlevelp3 (in=in2 keep= pid sex race_ethnicity_cat2 ccidsite edu_cat_1 edu_cat_2 edu_cat_3 income_cat_1  hins_cat1 hins_cat2);
  by pid;
  if in1 or in2;

  * create subset flag for each row to use in macro call;
  all = 1;

  if diabetes = 0 then diabetes_no=1;
  else if diabetes = 1 then diabetes_yes=1;

  if . < age_integer < 65 then age_lt65 = 1;
  else if age_integer >= 65 then age_ge65 = 1;
```

```
  * regroup;
  if EDU_CAT_1 in (1, 2, 3) then educ2gp = 1;
  else if EDU_CAT_1 in (4, 5, 6, 7) then educ2gp = 2;

  if income_cat_1 = 1 then income_3gp = 1;
  else if income_cat_1 = 2 then income_3gp = 2;
  else if income_cat_1 in(3,4) then income_3gp = 3;

  if income_cat_1 = 1 then income_3gp_v2 = 1;
  else if income_cat_1 = 2 then income_3gp_v2 = 2;
  else if income_cat_1 in(3,4, 97) then income_3gp_v2 = 3;

  if hins_cat1 = 2 then hins_3gp = 1;
  else if hins_cat1 in (3,4) then hins_3gp = 2;
  else if hins_cat1 = 5 then hins_3gp = 3;

run;


proc freq data=analyp3;
  tables diabetes*diabetes_no*diabetes_yes/list missing;
  tables educ2gp*edu_cat_1/list missing;
  tables income_3gp*income_3gp_v2*income_cat_1/list missing;
  tables hins_3gp*hins_cat1/list missing;
  tables age_lt65*age_ge65*age_integer/list missing;
  tables age_lt65*age_ge65*hins_cat1/list missing;
  tables age_lt65*age_ge65*hins_cat2/missing list;
  title3 "check Phase 3 analysis vars";
run;


** Phase 1;
data analyp1;
  merge visitlevelp1  (in=in1 keep= pid diabetes age_integer egfr_roche proteinuria anycvd systolic diastolic controlled1 controlled2 sf12_pcs modminwk bmi
                                    bmi_cat_1 bmi_cat_2  bmi_cat_3 diet_fib)
        personlevelp1 (in=in2 keep= pid sex race_ethnicity_cat2 ccidsite edu_cat_1 edu_cat_2 edu_cat_3 income_cat_1  hins_cat1 hins_cat2);
  by pid;
  if in1 or in2;

  * create subset flag for each row to use in macro call;
  all = 1;

  if diabetes = 0 then diabetes_no=1;
  else if diabetes = 1 then diabetes_yes=1;

  if . < age_integer < 65 then age_lt65 = 1;
  else if age_integer >= 65 then age_ge65 = 1;

  * regroup;
  if EDU_CAT_1 in (1, 2, 3) then educ2gp = 1;
  else if EDU_CAT_1 in (4, 5, 6, 7) then educ2gp = 2;

  if income_cat_1 = 1 then income_3gp = 1;
  else if income_cat_1 = 2 then income_3gp = 2;
  else if income_cat_1 in(3,4) then income_3gp = 3;
```

```
  if income_cat_1 = 1 then income_3gp_v2 = 1;
  else if income_cat_1 = 2 then income_3gp_v2 = 2;
  else if income_cat_1 in(3,4, 97) then income_3gp_v2 = 3;

  if hins_cat1 = 2 then hins_3gp = 1;
  else if hins_cat1 in (3,4) then hins_3gp = 2;
  else if hins_cat1 = 5 then hins_3gp = 3;

run;

proc freq data=analyp1;
  tables diabetes*diabetes_no*diabetes_yes/list missing;
  tables educ2gp*edu_cat_1/list missing;
  tables income_3gp*income_3gp_v2*income_cat_1/list missing;
  tables hins_3gp*hins_cat1/list missing;
  tables age_lt65*age_ge65*age_integer/list missing;
  tables age_lt65*age_ge65*hins_cat1/list missing;
  tables age_lt65*age_ge65*hins_cat2/missing list;
  title3 "check Phase 1 analysis vars";
run;


** Overall - need to combine phase 1 and 3 ;
data analy;
  set analyp1 (keep=pid age_integer age_lt65 age_ge65  sex  race_ethnicity_cat2  educ2gp income_3gp income_3gp_v2 hins_3gp hins_cat1  hins_cat2)
      analyp3 (keep=pid age_integer age_lt65 age_ge65  sex  race_ethnicity_cat2  educ2gp income_3gp income_3gp_v2 hins_3gp hins_cat1  hins_cat2);
run;


data prntunivagelt65f;
  set _null_;
run;

%univ(1  , age_integer, age_lt65     , agelt65f);


data prntunivagelt65f;
  set prntunivagelt65f;
  _mean_ = round(_mean_);
  _std_  = round(_std_);
run;

proc print data= prntunivagelt65f noobs;
    var rownum _var_  covarf _nobs_ _median_ _min_ _max_ _mean_ _std_;
     title3 "Overall < 65 years";
run;

* n and percent;
data prntagelt65f;
  set _null_;
run;

%npercent(2, sex                   , sexf    , age_lt65, agelt65f);
%npercent(3, race_ethnicity_cat2   , racef      , age_lt65, agelt65f);
%npercent(4, educ2gp               , educ2gf , age_lt65, agelt65f);
%npercent(5, hins_3gp            , hinsf     , age_lt65, agelt65f);
```

```
%npercent(6, income_3gp                , incomef    , age_lt65, agelt65f);

data prntagelt65f;
  set prntagelt65f;
  percent = round(percent);
run;

proc print data=prntagelt65f;
  var rownum covar covarf count percent;
  title3 "Overall < 65 years";
run;


data prntunivagege65f;
  set _null_;
run;

%univ(1  , age_integer, age_ge65     , agege65f);

data prntunivagege65f;
  set prntunivagege65f;
  _mean_ = round(_mean_);
  _std_  = round(_std_);
run;

proc print data= prntunivagege65f noobs;
    var rownum _var_  covarf _nobs_ _median_ _min_ _max_ _mean_ _std_;
  title3 "Overall >= 65 years";
run;

data prntagege65f;
  set _null_;
run;

%npercent(2, sex                   , sexf    , age_ge65, agege65f);
%npercent(3, race_ethnicity_cat2   , racef      , age_ge65, agege65f);
%npercent(4, educ2gp               , educ2gf , age_ge65, agege65f);
%npercent(5, hins_3gp            , hinsf     , age_ge65, agege65f);
%npercent(6, income_3gp          , incomef   , age_ge65, agege65f);

data prntagege65f;
  set prntagege65f;
  percent = round(percent);
run;

proc print data=prntagege65f;
  var rownum covar covarf count percent;
  title3 "Overall >= 65 years";
run;
```