# Dataset Integrity Check for Chronic Renal Insufficiency Cohort Study (CRIC) Data

# 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 was 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 evaluations at baseline and annual clinic visits, as well as via telephone at 6-month intervals. Data on quality of life, dietary assessment, physical activity, health behaviors, depression, cognitive function, health care resource utilization, and blood and urine specimens were collected. Outcomes such as measures of kidney function and occurrence of new and worsening CVD were assessed.

# 3 Archived Datasets

A full listing of archived datasets included in the package can be found in the Roadmap document. All data files, as provided by the Data Coordinating Center (DCC) and ancillary researchers, are located in the CRIC folder in the data package. For this replication, variables were taken from the "personlevel.sas7bdat" and "visitlevel.sas7bdat"datasets.

# 4 Statistical Methods

Analyses were performed to replicate results for the data published by Schrauben et al. [1]. To verify the integrity of the data, only descriptive statistics were computed.

# 5 Results

For Table 2 in the publication [1], 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 files to the results published in Table 2. The results of the replication are an exact match to the published results.

# 6 Conclusions

The NIDDK Central Repository is confident that the CRIC data files to be distributed are a true copy of the study data as the results of the replication are an exact match to the published results.

# 7 References

[1] Schrauben SJ, Hsu JY, Nunes JW, Fischer MJ, Srivastava A, Chen J, Charleston J, Steigerwalt S, Tan TC, Fink JC, Ricardo AC, Lash JP, Wolf M, Feldman HI, Anderson AH. Health Behaviors in Younger and Older Adults With CKD: Results From the CRIC Study. Kidney International Reports, 4(1), 80-93, September 2018. doi: https://doi.org/10.1016/j.ekir.2018.09.003

**Table A:** Variables used to replicate 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 Variable | dataset.variable |
|---|---|
| Age | visitlevel.age_integer |
| Gender | personlevel.sex |
| Race | personlevel.race_ethnicity_cat2 |
| Education | personlevel.edu_cat_3 |
| Insurance | personlevel.hins_cat1 |
| Income | personlevel.income_cat_1 |

**Table B:** Comparison of values computed in integrity check to reference article Table 2 values (For "Overall" only)

| Characteristic | Pub: Age < 65 Overall (n = 3552) | DSIC: Age < 65 Overall (n = 3552) | Diff. (n = 0) | Pub: Age ≥ 65 Overall (n = 1947) | DSIC: Age ≥ 65 Overall (n = 1947) | Diff. (n = 0) |
|---|---|---|---|---|---|---|
| Age, mean (SD) | 54 (9) | 54 (9) | 0 (0) | 70 (3) | 70 (3) | 0 (0) |
| Gender, male (%) | 55 | 55 | 0 | 58* | 58 | 0 |
| Race, (%) | | | | | | |
|    Non-Hispanic white | 39 | 39 | 0 | 45.9 | 45.9 | 0 |
|    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 |
|    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 | 35 | 0 |
|    > $50,000 | 37 | 37 | 0 | 32 | 32 | 0 |

*Note: The DCC informed the NIDDK Central Repository that there was an error in the publication, the correct value was provided by the DCC.

# Attachment A: SAS Code

libname dsic "X:\NIDDK\niddk-dr_studies1\CRIC\private_created_data\CRIC Final 2023 Data Redactions";

```
/***************************/
/* CRIC DSIC Schrauben et al. */
/***************************/

*defining the cohort;
data temp;
set dsic.personlevel ;
/*(keep=pid enrol_phs hcric);*/
if ^(enrol_phs=3 AND hcric=1);
run;

data visit; set dsic.visitlevel;
if vnum = 3;
if age_integer < 65 then age_65 = 1; if age_integer >= 65 then age_65 = 2;
run;

*age;
proc sort data=temp;
by pid;
run;

proc sort data=visit;
by pid;
run;

data one; merge
temp (in=a)
visit (in=b);
by pid;
if a=b;
run;

proc means data=one n mean std;
var age_integer;
class age_65;
run;

*Gender;
proc freq data=one;
tables sex*age_65/norow nopercent;
run;
```

```
*race;
proc freq data=one;
tables race_ethnicity_cat2*age_65/norow nopercent missing;
run;

*Education;
proc freq data=one;
tables edu_cat_3*age_65/norow nopercent;
run;

*health insurance;
data two; set one;
ins = .;
if hins_cat1 = 2 then ins = 1;
if hins_cat1 = 3 OR hins_cat1 = 4 then ins = 2;
if hins_cat1 = 5 then ins = 3;
run;

proc freq data=two;
tables ins*age_65/norow nopercent;
run;

*income;
data three; set one;
income = .;
if income_cat_1 = 1 then income = 1;
if income_cat_1 = 2 then income = 2;
if income_cat_1 = 3 OR income_cat_1 = 4 then income = 3;
run;

proc freq data=three;
tables income*age_65/norow nopercent;
run;
```