Dataset Integrity Check for the CRISP3 Data Files

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
July 9, 2020

Contents

1 Standard Disclaimer 2

2 Study Background 2

3 Archived Datasets 2

4 Statistical Methods 2

5 Results 2

6 Conclusions 3

7 References 3

Table A: Variables used to replicate Table 1: Patient characteristics at baseline 4

Table B-1 Counts and Percentages: Comparison of values computed in integrity check to reference article Table 1 values 5

Table B-2 Mean and SD: Comparison of values computed in integrity check to reference article Table 1 values 5

Table B-3 Median and IQR: Comparison of values computed in integrity check to reference article Table 1 values 5

SAS Code 6

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 Consortium for Radiological Imaging Studies of Polycystic Kidney Disease (CRISP) studied the progression of autosomal dominant polycystic kidney disease (ADPKD), and compared radiological techniques for measuring increases in renal volume during the progression of ADPKD. The CRISP study tested whether magnetic resonance (MR) can detect changes in renal volume, cyst volume, or changes in % cystic involvement in ADPKD individuals over a short period of time (1 to 2 years). CRISP participants had ADPKD with relatively normal renal function and creatinine clearances. The recruitment goal was for two thirds of the participants to have a high risk of progression to ESRD and one third to have an absence of risk factors for ESRD.

3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the CRISP3 data package. For this replication, variables were taken from the “baseline_tkv_tmp_cjasn” and “baseline_tkv_pkd_msg_7_6_2017” datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Yu et al [1] in Kidney International March 2018. To verify the integrity of the dataset, descriptive statistics were computed (table 1 from the paper).

5 Results

Table 1 in the publication [1], Patient characteristics at baseline. Our Table A lists the variables we used in our replication and Table B compare the results calculated from the archived data file to the results published in Table 1. The results of the replication are almost an exact match to the published results. The CRISP DCC confirmed that the count of 37 for ‘PKD2 and NMD’ category is correct (Table B-1).

6 Conclusions

The NIDDK repository is confident that the CRISP3 data files to be distributed are within expected results.

7 References

Alan S.L. Yu, Chengli Shen, Douglas P. Landsittel, Peter C. Harris, Vicente E. Torres, Michal Mrug, Kyongtae T. Bae, Jared J. Grantham, Frederic F. Rahbari-Oskoui, Michael F. Flessner, William M. Bennett and Arlene B. Chapman; for the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP). Baseline total kidney volume and the rate of kidney growth are associated with chronic kidney disease progression in Autosomal Dominant Polycystic Kidney Disease. Kidney International (2018) 93, 691–699.

Table A: Variables used to replicate Table 1: Patient characteristics at baseline

Table Variable	Variables Used in Replication from the "Table 1" Dataset
Visit	baseline_tkv_tmp_cjasn.vis
F10Q4. Corrected Iothalamate Clearance (ml/min/1.73m2)	baseline_tkv_tmp_cjasn.cic_c
Chronic Kidney Disease Stage 3	baseline_tkv_tmp_cjasn.ckd3_1
Gender	baseline_tkv_tmp_cjasn.gender
Race	baseline_tkv_tmp_cjasn.race4
Hypertension Yes/No	baseline_tkv_tmp_cjasn.hypertension_yn
PKD Gene Mutation	baseline_tkv_tmp_cjasn.genetype
Mutation Strength Group	baseline_tkv_pkd_msg_7_6_2017.msg
Age	baseline_tkv_tmp_cjasn.age
F9Q1. Height in cm	baseline_tkv_tmp_cjasn.height_c
Body Mass Index	baseline_tkv_tmp_cjasn.bmi_c
Serum creatinine, mg/dl	baseline_tkv_tmp_cjasn.serumcreat
Measured GFR, ml/min per 1.73 m2	baseline_tkv_tmp_cjasn.cic_c
eGFR, ml/min per 1.73 m2	baseline_tkv_tmp_cjasn.ckd_epi
htTKV, ml/m	baseline_tkv_tmp_cjasn.httkv

Table B-1 Counts and Percentages: Comparison of values computed in integrity check to reference article Table 1 values

	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	N			**%**		
Total no. of patients	184	184	0			
Male, no. (%)	76	76	0	41.3	41.3	0
Race, no. (%)						
White	162	162	0	88	88	0
African American	19	19	0	10.3	10.3	0
Other	3	3	0	1.7	1.7	0
Hypertension, no. (%)	75	75	0	41	41	0
PKD genotype,d no. (%)						
PKD1, truncating (MSG1) and nontruncating MSG2	129	129	0	75	70.9	4.1
PKD1, nontruncating MSG3	16	16	0	9.3	8.8	0.5
PKD2 and NMD	27	37	-10	15.7	20.3	-4.6

Table B-2 Mean and SD: Comparison of values computed in integrity check to reference article Table 1 values

	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	Mean			**SD**		
Age, yr	32.1	31.6	0.5	8.9	8.9	0
Height, cm	172.5	172.5	0	11	11	0
BMI, kg/m2	25.8	25.8	0	5.3	5.3	0
Serum creatinine, mg/dl	0.9	0.95	-0.05	0.2	0.2	0
Measured GFR,b ml/min per 1.73 m2	97.6	97.6	0	23.5	23.5	0
eGFR,c ml/min per 1.73 m2	93.6	93.6	0	22.6	22.6	0

Table B-3 Median and IQR: Comparison of values computed in integrity check to reference article Table 1 values

	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	Median			**IQR**		
htTKV, ml/m	501.4	501.4	0	442.9	442.9	0

SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/CRISP3/prog_initial_analysis/crisp3.dsic.sas';
run;

libname sasin  "/prj/niddk/ims_analysis/CRISP3/private_created_data/5_1_2020/3 - Analysis Data set for primary study publication(s) (1)/";

data base;
  set sasin.baseline_tkv_tmp_cjasn;
run;

proc contents data=base;
title3 "baseline_tkv_tmp_cjasn";
run;

data msg;
  set sasin.baseline_tkv_pkd_msg_7_6_2017;
run;

proc contents data=msg;
title3 "baseline_tkv_pkd_msg_7_6_2017";
run;

proc freq data=msg;
  tables gene msg/missing;
  run;

proc freq data=base;
  where vis=0 and cic_c>60;
  tables ckd3_1/list missing;
  title3 "Table 1 - id baseline subset";
run;


proc sort data=base out=tbl1;
  where vis=0 and cic_c>60 and ckd3_1 ne .;
  by PKDID;
run;

proc sort data=msg;
  by PKDID;
run;

data tbl1;
  merge tbl1 (in=in1 keep=PKDID gender race4 hypertension_yn age height_c bmi_c serumcreat cic_c ckd_epi httkv genetype)
        msg  (in=in2 keep=PKDID gene msg);
  by PKDID;
```

```
  if in1;
  if in2 then in_msg=1;

  * create PKD genotype (based on program: BASELINE_TKV_KI paper9_13_2017_REPOSITORY.do);
  if genetype="PKD2" or genetype="NMD" then genecom=0;
  else if genetype="PKD1" & msg=3 then genecom=1;
  else if genetype="PKD1" & (msg=1 or msg=2) then genecom=2;
run;

proc freq data=tbl1;
  table in_msg/missing;
  tables genecom*genetype*msg/list missing;
  title3 "check genecom";
run;

proc freq data=tbl1;
  tables gender
        race4
        hypertension_yn
        genecom /list missprint;
title3 "Table 1 (n, %)";
run;

proc univariate data=tbl1;
  var age
      height_c
      bmi_c
      serumcreat
      cic_c
      ckd_epi
      httkv ;
title3 "Table 1 (mean, SD, median, IQR)";
run;
```