Dataset Integrity Check for Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) Allele Data

Prepared by NIDDK-CR
February 1, 2022

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 Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) was established to develop and implement studies to test whether imaging techniques can provide accurate and reproducible markers of progression of renal disease in patients with polycystic kidney disease. Autosomal-dominant polycystic kidney disease (ADPKD) is characterized by gradual renal enlargement and cyst growth prior to loss of renal function; however, standard radiographic imaging has not provided the resolution and accuracy necessary to detect small changes in renal volume or to reliably measure renal cyst volumes. The CRISP cohort study longitudinally observed ADPKD individuals using high-resolution magnetic resonance imaging to determine if change in renal and cyst volumes can be detected over a short period of time, and if they correlate with decline in renal function early in disease.

Identification of the ADPKD genes has additionally allowed for molecular diagnostics, although this is complicated by a high level of unclassified variants (UCV). The CRISP cohort also was analyzed by molecular analysis to systematically classify PKD1 and PKD2 UCVs.

3 Archived Datasets

All data files, as provided by the Data Coordinating Center (DCC) and ancillary researchers, are located in the CRISP folder in the data package. For this replication, variables were taken from the "repository090408.sas7bdat" and "crisp_ids_genetic data.xlsx" datasets.

4 Statistical Methods

Analyses were performed to replicate results for the data published by Rossetti et al. [1] for Comprehensive Molecular Diagnostics in Autosomal Dominant Polycystic Kidney Disease. To verify the integrity of the dataset, only descriptive statistics were computed for the provided variables.

Calculations to replicate genetic results, including identifying recurrent versus novel mutations, were not computed.

5 Results

For Table 1 in the publication [1], Summary of mutations in the CRISP families, 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 1. The results of the replication are within expected variation to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the CRISP Allele data files to be distributed are a true copy of the study data as the results of the replication are within expected variation to the published results.

7 References

[1] Rossetti S, Consugar MB, Chapman AB, Torres VE, Guay-Woodford LM, Grantham JJ, Bennett WM, Meyers CM, Walker DL, Bae K, Zhang QJ, Thompson PA, Miller JP, Harris PC. Comprehensive Molecular Diagnostics in Autosomal Dominant Polycystic Kidney Disease. Journal of the American Society of Nephrology, 18(7), 2143-2160, July 2007.
DOI: https://doi.org/10.1681/ASN.2006121387

Table A: Variables used to replicate Table 1 – Summary of mutations in the CRISP families

Table Variable	**dataset.variable**
Definite (MG = A)	crisp_ids_genetic data. Final Mutation Mutation Strength Group
Highly likely (MG = B)	crisp_ids_genetic data. Final Mutation Mutation Strength Group
Likely (MG = C)	crisp_ids_genetic data. Final Mutation Mutation Strength Group
FS deletion/insertion	crisp_ids_genetic data. Final Mutation Type
Nonsense	crisp_ids_genetic data. Final Mutation Type
Splicing	crisp_ids_genetic data. Final Mutation Type
IF deletion/insertion	crisp_ids_genetic data. Final Mutation Type
Missense	crisp_ids_genetic data. Final Mutation Type
Truncating	crisp_ids_genetic data. Final Mutation Fuctional Effect
Small IF	crisp_ids_genetic data. Final Mutation Fuctional Effect
Total different mutations	crisp_ids_genetic data. Final Mutation (nt)
Total mutations	crisp_ids_genetic data. Final Mutation (nt)
No mutation defined	crisp_ids_genetic data. Final Gene

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

Mutation Type	Manuscript PKD1	DSIC PKD1	Diff.	Manuscript PKD2	DSIC PKD2	Diff.	Manuscript Total	DSIC Total	Diff.
Definite (MG = A)	106 (69.3%)	100 (65.8%)	6 (3.5%)	21 (77.8%)	20 (76.9%)	1 (0.9%)	127 (70.6%) (62.9%)[b]	120 (67.4%) (59.7%)	7 (3.2%) (3.2%)
Highly likely (MG = B)	34 (22.2%)	34 (22.4%)	0 (0.2%)	3 (11.1%)	4 (15.4%)	1 (4.3%)	37 (20.6%) (18.3%)[b]	38 (21.4%) (18.9%)	1 (0.8%) (0.6%)
Likely (MG = C)	13 (8.5%)	18 (11.8%)	5 (3.3%)	3 (11.1%)	2 (7.7%)	1 (3.4%)	16 (8.9%) (7.9%)[b]	20 (11.2%) (10.0%)	4 (2.3%) (2.1%)
FS deletion/insertion	49 (32.0%)	46 (30.3%)	3 (1.7%)	7 (25.9%)	7 (26.9%)	0 (1.0%)	56 (31.1%)	53 (29.8%)	3 (1.3%)
Nonsense	38 (24.8%)	39 (25.7%)	1 (0.9%)	9 (33.3%)	9 (34.6%)	0 (1.3%)	47 (26.1%)	48 (27.0%)	1 (0.9%)
Splicing	16 (10.5%)	17 (11.2%)	1 (0.7%)	6 (22.2%)	6 (23.1%)	0 (0.9%)	22 (12.2%)	23 (12.9%)	1 (0.7%)
IF deletion/insertion	9 (5.9%)	9 (5.9%)	0 (0%)	2 (7.4%)	2 (7.7%)	0 (0.3%)	11 (6.1%)	11 (6.2%)	0 (0.1%)
Missense	41 (26.8%)	41 (27.0%)	0 (0.2%)	3 (11.1%)	2 (7.7%)	1 (3.4%)	44 (24.4%)	43 (24.2%)	1 (0.2%)
Truncating[a]	107 (69.9%)	100 (65.8%)	7 (4.1%)	22 (81.5%)	20 (76.9%)	2 (4.6%)	129 (71.7%)	120 (67.4%)	9 (4.3%)
Small IF	46 (30.1%)	52 (34.2%)	6 (4.1%)	5 (18.5%)	6 (23.1%)	1 (4.6%)	51 (28.3%)	58 (32.6%)	7 (4.3%)
Total different mutations	136	134	2	24	24	0	160	158	2
Total mutations	153 (85.0%)	152 (85.3%)	1 (0.3%)	27 (15.0%)	26 (14.6%)	1 (0.4%)	180 (89.1%)[b]	178 (88.6%)	2 (0.5%)
No mutation defined	--	--	--	--	--	--	22 (10.9%)[b]	23 (11.4%)	1 (0.5%)

[a] Including IF changes of five amino acids or greater and atypical splicing

[b] Percentage of total families in study

Attachment A: SAS Code

```
* CRISP Allele Data DSIC;

* Library for CRISP data;
libname crisp "Z:\NIDDK\niddk-dr_studies1\CRISP\private_orig_data\StandardRelease\Data";

proc contents data=crisp.repository090408 VARNUM; run;

data work.crisp_full;
        set crisp.repository090408 (keep=pkdid famid fmps genetype);
run;

proc contents data=work.crisp_full VARNUM; run; *964 obs;

proc sort data=crisp_full out=crisp_fam nodupkey;
        by pkdid;
run;

proc contents data=work.crisp_fam VARNUM; run; *241 obs;

* Import CRISP alleles genetic data;
proc import DATAFILE = "Z:\NIDDK\niddk-
dr_studies1\CRISP\private_orig_data\delivery_12_18_2020\crisp_ids_genetic data.xlsx"
 DBMS = xlsx
 OUT = crisp_alleles;
run;

proc contents data= crisp_alleles VARNUM; run;

proc sort data= crisp_alleles; by pkdid; run;

*Combine genetic data with family ids;
data combined; *241 obs;
        merge crisp_fam crisp_alleles;
        by pkdid;
run;

*Sort by famid, comfirm mutations are the same within famid;
proc sort data=combined;
        by famid;
run;

*Restrict to one obs per family;
proc sort data=combined out=combined_fam nodupkey; *204 obs;
        by famid;
run;
```

```
proc contents data=combined_fam; run;

data combined_fam2;
        set combined_fam;
        if 'Final Gene'n = "        " then delete;
run;

proc freq data=combined_fam2 nlevels;
        tables famid;
run;

*Exclude mutation negative cases, as per Table V;
data analysis;
        set combined_fam2;
        if famid in      (100002, 100006, 100009, 103227, 118641,
                                       157925, 160928, 188086, 213454, 235752,
                                       244111, 252086, 259940, 299663, 313893,
                                       364664, 369941, 406737, 407132, 464923,
                                       476972, 493328)
                then delete;
        if 'Final Gene'n = "        " then delete;
run;

* Identify mutation type as definite, highly likely, or likely;
proc freq data=analysis;
        tables 'Final Mutation Mutation Strength'n;
        tables 'Final Mutation Mutation Strength'n*'Final Gene'n;
run;

proc freq data=analysis nlevels;
        tables famid;
run;

* Identify mutation type;
data analysis2;
        set analysis;
        if 'Final Mutation Type'n = "Frameshift" OR 'Final Mutation Type'n = "Large Deletion" then type
= "FS deletion/insertion";
        if 'Final Mutation Type'n = "Nonsense" then type = "nonsense";
        if 'Final Mutation Type'n = "Splice" then type = "splicing";
        if 'Final Mutation Type'n = "InFrame D/I" OR 'Final Mutation Type'n = "Inframe D/I" then type =
"IF deletion/insertion";
        if 'Final Mutation Type'n = "Missense" then type = "missense";
run;

proc freq data=analysis2;
        tables type;
        tables type*'Final Gene'n;
run;

* Identify mutation functional effect as truncating;
proc freq data=analysis2;
        tables 'Final Mutation Fuctional Effect'n;
        tables 'Final Mutation Fuctional Effect'n*'Final Gene'n;
run;

* Total different mutations;
proc sort data=analysis2 out=mutation_nt nodupkey;
        by 'Final Mutation (nt)'n;
run;

proc freq data=mutation_nt;
        tables 'Final Gene'n;
run;

* Total mutations;
proc freq data=analysis2;
        tables 'Final Gene'n;
run;
```