Persons using assistive technology may not be able to fully access information in this file. For assistance, e-mail niddk-cr@imsweb.com. Include the Web site and filename in your message.

Norma Pugh
October 2010

Dataset Integrity Check for the Family Investigation of Nephropathy and Diabetes (FIND) Consortium CIDR I Data Files

As a partial check of the integrity of the FIND CIDR I datasets archived in the NIDDK data repository, a set of tabulations was performed to verify that published results can be reproduced using the archived dataset. The Center for Inherited Disease Research (CIDR) I datasets include genetic linkage data from microsatellite genotyping of 1,400 individuals. Analyses were performed to duplicate results for the data published by Iyengar et al [1] in *Diabetes* in June 2007. The results of this integrity check are described below. The full text of the *Diabetes* article can be found in Attachment 1, and the SAS code for our tabulations is included in Attachment 2. A future DSIC will report the integrity of the FIND CIDR II datasets archived in the NIDDK data repository. The CIDR II datasets include the same 1,400 individuals included in CIDR I, re-genotyped along with several thousand additional individuals, using an SNP linkage panel. The phenotypic data are unchanged, aside from data cleaning.

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 on a first (or second) exercise 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.

Background. The Family Investigation of Nephropathy and Diabetes (FIND) is a multicenter study designed to identify genetic determinants of diabetic nephropathy. Patients are recruited from eight U.S. clinical centers, across four ethnic groups (European Americans, African Americans, Mexican Americans and American Indians) [2].

Two strategies are used to localize susceptibility genes: a family-based linkage study and a case–control study using mapping by admixture linkage disequilibrium (MALD).

In the family-based study, probands with diabetic nephropathy are recruited with their parents and selected siblings. Linkage analyses are conducted to identify chromosomal regions containing genes that influence the development of diabetic nephropathy or related quantitative traits such as serum creatinine concentration, urinary albumin excretion, and plasma glucose concentrations. Regions showing evidence of linkage are examined further with both genetic linkage and association studies to identify genes that influence diabetic nephropathy or related traits.

Two types of MALD studies are being done. One is a case–control study of unrelated individuals of Mexican American heritage in which both cases and controls have diabetes, but only the cases have nephropathy. The other is a case–control study of African American patients with nephropathy (cases) and their spouses (controls) unaffected by diabetes and nephropathy; offspring are genotyped when available to provide haplotype data.

The Iyengar paper reports findings from the first-phase genetic analyses in participants from pedigrees of all four ethnic groups [1].

Participant Characteristics. Table 1 in the publication [1] reports on demographic and clinical characteristics of the genotyped individuals stratified by proband and diabetic nephropathy status. Table A lists the variables we used in our replication. All variables were taken from the ASCII file, FIND_CIDR_I_phenotypes.txt, provided by the DCC.

Table A: Variables Used to Replicate Table 1

Table 1 Variable	Variables Used in Replication
Treatment group	Share, Proband, DM, DN
Gender	Sex
Age	Age
BMI	BMI
Ethnicity	Ethnic_Group
ESRD	N/A
Diabetes: Age diagnosed	DM_age
Diabetes: Duration	DM_duration
A1C	HbA1c
Serum creatinine	Serum_Creatinine
Blood urea nitrogen	BUN
Urine protein-to-creatinine ratio	PCR
Urine ACR	MAC
GFR	GFR

Table B1: Comparison of Values Computed in Dataset Integrity Check to Reference Article Table 1 Values Stratum = Diabetic nephropathy probands

Characteristic	Iyengar	Integrity Check
Sample size	349	313
Male	169 (46.0)	142 (45.4)
Age (years)	57 ± 10.7	58.1 ± 10.8
BMI (kg/m^2)	30 ± 7.2	30.5 ± 7.1
Ethnicity		
European American	50 (14.3)	59 (18.8)
African American	90 (25.8)	85 (27.2)
Mexican American	179 (51.3)	169 (54.0)
American Indian	30 (8.6)	N/A*
ESRD	282 (80.8)	N/A*
Diabetes		
Age diagnosed (years)	34 ± 12.3	34.7 ± 12.3
Duration (years)	23 ± 8.4	23.4 ± 8.5
Biochemistry		
A1C (%)	7.8 ± 1.8	6.5 ± 2.7
Serum creatinine (mg/dl)	8.58 ± 3.02	2.73 ± 2.60
Blood urea nitrogen (mg/dl)	72.9 ± 7.2	44.5 ± 22.8
Urine protein-to-creatinine ratio (g/g)	3.29 ± 0.69	4.46 ± 4.94
Urine ACR (g/g)		
Mean	2.75 ± 0.7	6.34 ± 33.7
Median	3 (3-3)	1.3 (0.3-3.4)
GFR (ml/min per 1.73 m^2)		
Mean	10.6 ± 15.0	40.4 ± 28.4
Median	5 (5-5)	32.9 (21.9-50.4)

Table B2: Comparison of Values Computed in Dataset Integrity Check to Reference Article Table 1 Values
Stratum = Diabetic nephropathy relatives

Characteristic	Iyengar	Integrity Check
Sample size	390	391
Male	180 (46.2)	181 (46.3)
Age (years)	56 ± 11.4	58.4 ± 12.2
BMI (kg/m^2)	32 ± 8.6	31.4 ± 7.7
Ethnicity		
European American	42 (10.8)	52 (13.3)
African American	95 (24.4)	108 (27.6)
Mexican American	210 (53.8)	231 (59.1)
American Indian	43 (11.0)	N/A*
ESRD	53 (13.6)	N/A*
Diabetes		
Age diagnosed (years)	41 ± 13.1	43.4 ± 14.4
Duration (years)	16 ± 10.1	15.1 ± 10.6
Biochemistry		
A1C (%)	8.6 ± 2.4	7.5 ± 3.1
Serum creatinine (mg/dl)	2.6 ± 3.2	1.3 ± 1.5
Blood urea nitrogen (mg/dl)	29.9 ± 25.5	22.9 ± 17.4
Urine protein-to-creatinine ratio (g/g)	1.38 ± 1.4	1.66 ± 3.2
Urine ACR (g/g)		
Mean	0.93 ± 1.2	0.68 ± 1.2
Median	0.1 (0.05-0.8)	0.2 (0.05-0.74)
GFR (ml/min per 1.73 m^2)		
Mean	68.5 ± 42.0	77.6 ± 36.9
Median	76.2 (47.4-98.0)	77.0 (49.6-102.7)

Table B3: Comparison of Values Computed in Dataset Integrity Check to Reference Article Table 1 Values
Stratum = Diabetes without nephropathy relatives

Characteristic	Iyengar	Integrity Check
Sample size	147	153
Male	43 (29.3)	48 (31.4)
Age (years)	59 ± 10.1	59.6 ± 11.1
BMI (kg/m^2)	32 ± 7.3	31.5 ± 7.1
Ethnicity		
European American	27 (18.4)	36 (23.5)
African American	33 (22.4)	34 (22.2)
Mexican American	80 (54.4)	83 (54.2)
American Indian	7 (4.8)	N/A*
ESRD	0 (0.0)	N/A*
Diabetes		
Age diagnosed (years)	42 ± 11.5	43.4 ± 12.0
Duration (years)	17 ± 7.0	16.6 ± 7.0
Biochemistry		
A1C (%)	7.8 ± 2.0	7.0 ± 2.9
Serum creatinine (mg/dl)	0.99 ± 0.16	0.80 ± 0.33
Blood urea nitrogen (mg/dl)	15.7 ± 4.9	15.6 ± 5.0
Urine protein-to-creatinine ratio (g/g)	0.17 ± 0.26	0.27 ± 1.31
Urine ACR (g/g)		
Mean	0.01 ± 0.01	0.01 ± 0.01
Median	0.009 (0.006-0.01)	0.01 (0.008-0.02)
GFR (ml/min per 1.73 m^2)		
Mean	85.6 ± 24.5	85.6 ± 24.6
Median	89.8 (73.9-105.1)	81.0 (67.7-101.8)

In Tables B1 through B3, we compare the results calculated from the archived data file to the results published in Table 1, Clinical characteristics of the genotyped individuals stratified by proband and diabetic nephropathy status. As the tables show, the results are of the replication are not similar to published results.

The American Indian population did not give permission to share their data with outside investigators. Additionally, 39 subjects did not agree to data sharing. Therefore, these populations are excluded from both the archived data file and the NIDDK replication, resulting in smaller sample sizes.

Also of note, the ESRD variable was not replicated because "Transplant" and "Dialysis" variables were not included in the ASCII file provided by the FIND DCC.

Due to the dissimilar results, the FIND DCC requested a second independent replication from the Center for Clinical Investigation (CCI) at Case Western Reserve University. The CCI DSIC (Appendix A) reports 3 sources of data: "Iyengar", the published results; "NIDDK", results obtained by the NIDDK repository using an earlier data delivery that was later determined to be compromised; and "CCI", the second, independent replication. It is not clear if CCI received the same dataset that was most recently provided to NIDDK and used for this current DSIC.

Comparison of published results and the NIDDK replication shows large differences in lab work for the "Diabetic nephropathy probands" group. Comparisons of both the "Diabetic nephropathy relatives" and "Diabetes without nephropathy relatives" groups do not vary as much, yet differences are present.

Comparison of published results and the CCI replication also shows large differences in lab work for the "Diabetic nephropathy probands" group. Comparisons of both the "Diabetic nephropathy relatives" and "Diabetes without nephropathy relatives" groups do not vary as much, yet differences are present

Finally, comparison of the current NIDDK replication and the CCI replication shows good agreement for most of the lab work. There are a minimal number of differences, particularly in the "Diabetic nephropathy probands" group.

References

1. Sudha K. Iyengar, et al; **Genome-Wide Scans for Diabetic Nephropathy and Albuminuria in Multiethnic Populations**; Diabetes; Volume 56(6); June 2007; pages 1577-1585.

2. NIDDK Website: Family Investigation of Nephropathy and Diabetes (FIND) page. Family Investigation of Nephropathy & Diabetes : NIDDK.

Appendix A
Independent DSIC Produced by CCI
Presented as Received, with No Modifications

Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values, Stratum = Diabetic nephropathy probands

Characteristic	Iyengar	NIDDK	CCI
Sample size	349	297	317
Male	169 (46.0)	146 (49.2)	152 (48.0)
Age (years)	57 ± 10.7	58 ± 10.8	57 ± 10.9
BMI (kg/m^2)	30 ± 7.2	30 ± 7.1	30 ± 7.0
Ethnicity			
European American	50 (14.3)	59 (19.9)	61 (19.2)
African American	90 (25.8)	72 (24.2)	90 (28.4)
Mexican American	179 (51.3)	166 (55.9)	134 (42.3)
American Indian	30 (8.6)	N/A	32 (10.1)
ESRD	282 (80.8)	105 (35.4)	N/A
Diabetes			
Age diagnosed (years)	34 ± 12.3	35 ± 12.4	N/A
Duration (years)	23 ± 8.4	23 ± 8.4	23 ± 8.2
Biochemistry			
A1C (%)	7.8 ± 1.8	7.3 ± 3.4	7.2 ± 1.6
Serum creatinine (mg/dl)	8.58 ± 3.02	2.85 ± 2.84	3.34 ± 3.11
Blood urea nitrogen (mg/dl)	72.9 ± 7.2	0.005 ± 0.002	43.7 ±22.9
Urine protein-to-creatinine ratio (g/g)	3.29 ± 0.69	4.85 ± 5.18	N/A
Urine ACR (g/g)			
Mean	2.75 ± 0.7	6.74 ± 34.8	1.90 ± 2.2
Median	3 (3-3)	2 (0-267)	1 (0-10)
GFR (ml/min per 1.73 m^2)			
Mean	10.6 ± 15.0	38.2 ± 25.6	42.9 ± 29.0
Median	5 (5-5)	33 (8-143)	36 (8-143)

CCI documentation:

Diabetic nephropathy probands grouped as follows:
diabetic nephropathy = 1, proband = 1

Group includes:
N=6 with diabetes = 1, and valid durations ranging from 4-8 years
N=311 with diabetes = 2, and valid durations ranging from 9-47 years
***Note that codebook states that diabetes = 2 coincides with a duration greater than or equal to 10 years**

Group excludes (***per email communication with Rob)**:
N=44 with diabetes = -9, but valid durations ranging from 10-50 years

Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values, continued
Stratum = Diabetic nephropathy relatives

Characteristic	Iyengar	NIDDK	CCI
Sample size	390	376	425
Male	180 (46.2)	173 (46.0)	189 (44.5)
Age (years)	56 ± 11.4	59 ± 12.1	58 ± 12.1
BMI (kg/m^2)	32 ± 8.6	31 ± 7.3	32 ± 8.5
Ethnicity			
European American	42 (10.8)	52 (13.8)	49 (11.5)
African American	95 (24.4)	95 (25.3)	103 (24.2)
Mexican American	210 (53.8)	229 (60.9)	227 (53.4)
American Indian	43 (11.0)	N/A	46 (10.8)
ESRD	53 (13.6)	17 (4.5)	N/A
Diabetes			
Age diagnosed (years)	41 ± 13.1	44 ± 14.6	N/A
Duration (years)	16 ± 10.1	15 ± 10.7	16 ± 10.4
Biochemistry			
A1C (%)	8.6 ± 2.4	7.3 ± 3.4	8.4 ± 2.3
Serum creatinine (mg/dl)	2.6 ± 3.2	1.2 ± 1.4	1.4 ± 1.5
Blood urea nitrogen (mg/dl)	29.9 ± 25.5	0.002 ± 0.002	22.7 ± 17.6
Urine protein-to-creatinine ratio (g/g)	1.38 ± 1.4	1.63 ± 3.1	N/A
Urine ACR (g/g)			
Mean	0.93 ± 1.2	0.66 ± 1.2	0.72 ± 1.3
Median	0.1 (0.05-0.8)	0.2 (0-7.0)	0.2 (0-7.9)
GFR (ml/min per 1.73 m^2)			
Mean	68.5 ± 42.0	78.0 ± 36.4	78.6 ± 37.2
Median	76.2 (47.4-98.0)	77.8 (4.7-198.9)	78.7 (4.4-198.9)

CCI documentation:

Diabetic nephropathy relatives grouped as follows:
diabetic nephropathy = 1, proband = 0

Group includes:
N=3 with diabetes = 1, and duration = -1 years
***Duration changed to 0 (per email communication with Rob)**
N=28 with diabetes = 1, and duration = 0 years
N=76 with diabetes = 1, and durations ranging from 1 to 8 years
N=318 with diabetes = 2, and durations ranging from 9 to 54 years
***Note that codebook states that diabetes = 2 coincides with a duration greater than or equal to 10 years**

Group excludes (***per email communication with Rob)**:
N=1 with diabetes = -9, but valid duration of 10 years
N=26 with diabetes = 0, and duration = 0 years

Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values, continued
Stratum = Diabetes without nephropathy relatives

Characteristic	Iyengar	NIDDK	CCI
Sample size	147	108	168
Male	43 (29.3)	34 (31.5)	50 (29.8)
Age (years)	59 ± 10.1	55 ± 14.0	59 ± 10.8
BMI (kg/m^2)	32 ± 7.3	33 ± 7.6	32 ± 7.4
Ethnicity			
European American	27 (18.4)	18 (16.7)	36 (21.4)
African American	33 (22.4)	22 (20.4)	34 (20.2)
Mexican American	80 (54.4)	68 (63.0)	88 (52.4)
American Indian	7 (4.8)	N/A	10 (6.0)
ESRD	0 (0.0)	0 (0.0)	N/A
Diabetes			
Age diagnosed (years)	42 ± 11.5	53 ± 14.6	N/A
Duration (years)	17 ± 7.0	3 ± 2.5	17 ± 7.0
Biochemistry			
A1C (%)	7.8 ± 2.0	6.8 ± 2.9	7.7 ± 2.0
Serum creatinine (mg/dl)	0.99 ± 0.16	0.64 ± 0.35	0.88 ± 0.24
Blood urea nitrogen (mg/dl)	15.7 ± 4.9	0.001 ± 0.0004	15.8 ± 5.1
Urine protein-to-creatinine ratio (g/g)	0.17 ± 0.26	0.10 ± 0.11	N/A
Urine ACR (g/g)			
Mean	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01
Median	0.009 (0.006-0.01)	0.010 (0.002-0.036)	0.013 (0.002-0.10)
GFR (ml/min per 1.73 m^2)			
Mean	85.6 ± 24.5	94.8 ± 23.1	85.7 ± 24.5
Median	89.8 (73.9-105.1)	93.8 (47.1-145.2)	81.3 (42.9-179.5)

CCI documentation:

Diabetic nephropathy relatives grouped as follows:
diabetes = 2, diabetic nephropathy = 1, proband = 0

Group includes:
N=168 with diabetes = 2, and durations ranging from 9 to 37 years
***Note that codebook states that diabetes = 2 coincides with a duration greater than or equal to 10 years**

Norma Pugh
October 2010

Note on Appendix A:

The ESRD result presented in the first version of the NIDDK FIND CIDR I DSIC, is based on a crude attempt to construct the variable. Since the necessary variables were not included in the DCC delivery, blood and urine data were used to develop frequency estimates. It was subsequently determined, that this method was not appropriate.