Norma Pugh
March 2011

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

As a partial check of the integrity of the FIND CIDR II data files archived in the NIDDK data repository, a set of tabulations was performed to verify that manuscript results can be reproduced using the archived data files. Analyses were performed to duplicate results for the manuscript authored by Igo et al [1] (2010). The results of this integrity check are described below. The full text of the manuscript is not included because it is unpublished. The SAS code for our tabulations is included in Attachment 1.

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 case has 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 Igo manuscript reports findings from the genetic analyses in participants from pedigrees of all four ethnic groups [1]. The first genetic data files that were generated (CIDR I) include genetic linkage data from microsatellite genotyping of 1400 individuals. The CIDR II data files include the same 1400

individuals re-genotyped along with several thousand additional individuals, using a SNP linkage panel. The phenotypic data are unchanged, other than corrections/additions that have been made to the data.

Participant Characteristics. Table 1 in the manuscript [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. The data file, CIDR_II_Pedigrees.xls, contains demographic variables and proband status, FIND_CIDR2_ForRepository_30Sep2010.xls contains additional demographic variables and medical history, laboratory results, and key variables that were included in a review of medical records.

Table A: Variables Used to Replicate Table 1

Table Variable	Variables Used in Replication
Treatment group	CIDR_II_Pedigrees: proband, share, sex FIND_CIDR2_ForRepository_30Sep2010: DM, macratio_2, UPR, serumcreat, fail, dm_duration
Ethnicity	CIDR_II_Pedigrees: group
Gender	CIDR_II_Pedigrees: sex
Age	FIND_CIDR2_ForRepository_30Sep2010: age
ESRD	FIND_CIDR2_ForRepository_30Sep2010: fail
Diabetes: diagnosis age	FIND_CIDR2_ForRepository_30Sep2010: diagage
Diabetes: duration	FIND_CIDR2_ForRepository_30Sep2010: dm_duration
BMI	FIND_CIDR2_ForRepository_30Sep2010: ht, wt
HbA1c	FIND_CIDR2_ForRepository_30Sep2010: hba1c
Serum creatinine	Results are not available
BUN	Results are not available
Urine ACR	FIND_CIDR2_ForRepository_30Sep2010: macratio_2
Urine PCR	FIND_CIDR2_ForRepository_30Sep2010: urineprotein, urinecreat
eGFR	CIDR_II_Pedigrees: sex, group FIND_CIDR2_ForRepository_30Sep2010: serumcreat, age

In Table B, we compare the results calculated from the archived data files to the results presented in Table 1 of the manuscript, Description of the overall FIND sample.

Table B: Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values, Stratum = Probands

Characteristic	Igo	Integrity Check
Sample size	1277	901
Ethnicity		
African American	348 (27.3)	324 (36.0)
American Indian	254 (19.9)	NA
European American	196 (15.3)	191 (21.2)
Mexican American	479 (37.5)	386 (42.8)
Female	692 (54.2)	487 (54.1)
Age, y	58 ± 11	59 ± 11
ESRD	509 (40.0)	743 (82.5)
DM diagnosis age, y	35 ± 12	36 ± 12
DM duration, y	22.7 ± 9.4	22.9 ± 8.8
BMI, kg/m^2	30.2 ± 7.2	30.3 ± 7.3
HbA1c, %	7.2 ± 1.7	7.2 ± 1.7
Serum creatinine, mg/dl	NA	NA
BUN, mg/dl	NA	NA
Urine ACR, g/g	2.9 ± 1.0	0.6 ± 8.9
Urine PCR, g/g	3.5 ±1.4	4.4 ± 4.7
eGFR, ml/min/1.73 m^2	11 ± 17	63 ± 21

Note: Results for Serum creatinine and BUN were not available at time of replication.

Table B: Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values, continued
Stratum = DN+ Relatives

Characteristic	Igo	Integrity Check
Sample size	731	358
Ethnicity		
African American	183 (25.0)	165 (46.1)
American Indian	200 (27.4)	NA
European American	58 (7.9)	45 (12.6)
Mexican American	290 (39.7)	148 (41.3)
Female	425 (58.1)	192 (53.6)
Age, y	58 ± 12	62 ± 12
ESRD	53 (7.3)	126 (35.2)
DM diagnosis age, y	39 ± 13	37 ± 12
DM duration, y	17.2 ± 10.8	20.7 ± 10.4
BMI, kg/m^2	31.6 ± 8.3	31.2 ± 8.1
HbA1c, %	8.3 ± 2.3	7.4 ± 1.9
Serum creatinine, mg/dl	NA	NA
BUN, mg/dl	NA	NA
Urine ACR, g/g	2.3 ± 2.0	0.9 ± 1.7
Urine PCR, g/g	3.2 ± 3.0	3.4 ± 5.2
eGFR, ml/min/1.73 m^2	49 ± 43	46 ± 11

Note: Results for Serum creatinine and BUN were not available at time of replication.

Table B: Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values, continued
Stratum = DN- Relatives

Characteristic	Igo	Integrity Check
Sample size	608	298
Ethnicity		
African American	146 (24.0)	108 (36.2)
American Indian	95 (15.6)	NA
European American	116 (19.1)	61 (20.5)
Mexican American	251 (41.3)	129 (43.3)
Female	431 (70.9)	188 (63.1)
Age, y	59 ± 11	60 ± 12
ESRD	0 (0.0)	3 (1.0)
DM diagnosis age, y	41 ± 12	50 ± 16
DM duration, y	17.7 ± 7.6	17.0 ± 7.1
BMI, kg/m^2	33.0 ± 7.9	32.4 ± 7.8
HbA1c, %	7.8 ± 1.8	7.7 ± 1.8
Serum creatinine, mg/dl	NA	NA
BUN, mg/dl	NA	NA
Urine ACR, g/g	0.03 ± 0.40	0.01 ± 0.01
Urine PCR, g/g	0.13 ± 0.64	0.10 ± 0.10
eGFR, ml/min/1.73 m^2	88 ± 24	86 ± 25

Note: Results for Serum creatinine and BUN were not available at time of replication.

Replication Limitations. In Table B, we compare the results calculated from the archived data file to results presented in the manuscript, Table 1, Description of the overall FIND sample. As Table B shows, the results of the replication are not similar to those presented in the manuscript. This is expected for many reasons.

The American Indian population did not give permission to share their data with outside investigators. Additionally, a number of subjects across the other ethnic groups denied permission for data sharing. Therefore, these populations are excluded from both the archived data files and the NIDDK replication.

Also of note, proband status was determined by use of an additional file provided by the FIND DCC. Proband status could not be determined from the archived data because the repository does not include the list of concomitant drugs considered to be for diabetes, nor the date of initiation of replacement therapy. We are aware that dates are commonly stripped from archived data files, for purposes of de-identification.

References

1. Robert P. Igo, Jr., PhD., et al; **Genomewide linkage scan for diabetic renal failure loci and albuminuria in four populations: The FIND Study**.

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

Attachment 1

```
options errorabend;
/*************************************************************************************/
/*
/* Program: R:\05_Users\Norma\FIND_CIDRII\Data\getdata.sas
/* Author:  Norma Pugh
/* Date:    March 2011
/* Purpose: Read-in new analysis datasets (Excel version).
/* NOTE:    AI POPULATION DID NOT GIVE PERMISSION TO SHARE THEIR DATA.
/*
/*************************************************************************************/
LIBNAME OUT 'R:\05_Users\Norma\FIND_CIDRII\Data';

/*****************/
/* Pedigree Data */
/*****************/
FILENAME FILE1 DDE "EXCEL|[CIDR_II_Pedigrees.xls]CIDR_II_Pedigrees!R2C1:R5340C8";

DATA PEDIGREE(KEEP = GROUP StudyID SEX PROBAND SHARE);
 LENGTH ID $9;
 INFILE FILE1 DLM='09'X NOTAB DSD MISSOVER;
 INPUT GROUP $ FAMILY $ ID $ FATHER $ MOTHER $ SEX PROBAND SHARE $;
 StudyID='0'||ID;
RUN;

/***********************/
/* Questionnaire Data */
/***********************/
FILENAME FILE2 DDE "EXCEL|[FIND_CIDR2_ForRepository_30Sep2010.xls]Med Ques!R2C1:R3184C50";

DATA MEDQUES(DROP=TEMP1-TEMP15);
 LENGTH StudyID $10 Drug1-Drug20 $30;
 INFILE FILE2 DLM='09'X NOTAB DSD MISSOVER;
 INPUT StudyID $ Age DM DM_Duration Fail AgeDialysis OrganTrans Organ $ Cause $ OthCause $ WT HT
       TEMP1 $ TEMP2 $ TEMP3 $ TEMP4 $ TEMP5 $ TEMP6 $ TEMP7 $ TEMP8 $ TEMP9 $ Insulin TEMP10 $
TEMP11 $
       TEMP12 $ TEMP13 $ TEMP14 $ TEMP15 $ Lifestyle Meds Drug1 $ Drug2 $ Drug3 $ Drug4 $ Drug5 $
Drug6 $
       Drug7 $ Drug8 $ Drug9 $ Drug10 $ Drug11 $ Drug12 $ Drug13 $ Drug14 $ Drug15 $ Drug16 $
Drug17 $ Drug18 $
       Drug19 $ Drug20 $ Drug21 $ Drug22 $ Drug23 $ Drug24 $ Drug25 $ Drug26 $ Drug27 $ Drug28 $
Drug29 $ Drug30 $;
RUN;

/*****************************/
/* Medical Record Review Data */
/*****************************/
FILENAME FILE3 DDE "EXCEL|[FIND_CIDR2_ForRepository_30Sep2010.xls]Med Rec Review!R2C1:R1010C26";

DATA MEDREVIEW(DROP=TEMP1-TEMP15);
 LENGTH StudyID $10;
 INFILE FILE3 DLM='09'X NOTAB DSD MISSOVER;
 INPUT StudyID $ TEMP1 $ TEMP2 $ MRR_DIABETES TEMP3 $ DIAGAGE $ TEMP4 $ MRR_KIDBIOP TEMP5 $ TEMP6
$ TEMP7 $ TEMP8 $ TEMP9 $ DIALYSIS $
       TEMP10 $ MRR_RETINOP TEMP11 $ TEMP12 $ TEMP13 $ UPRO UPR TEMP14 $ TEMP15 $ U_ALB $ U_CREAT
U_PROT;
```

```
RUN;

/******************/
/* Laboratory Data */
/******************/
FILENAME FILE4 DDE "EXCEL|[FIND_CIDR2_ForRepository_30Sep2010.xls]Lab Results!R2C1:R6798C16";

DATA LABS(DROP=TEMP1);
 LENGTH StudyID $10;
 INFILE FILE4 DLM='09'X NOTAB DSD MISSOVER;
 INPUT StudyID $ Lab $ Collection Assay TEMP1 $ HBA1C BUN SERUMCREAT UTPSIGN $ URINEPROTEIN
URINECREAT GLUCOSE UMSIGN $ URINEMICRO MACRATIO MACRATIO_2;
RUN;

/***************/
/* Output data */
/***************/
DATA OUT.PEDIGREE; SET PEDIGREE; RUN;
DATA OUT.MEDQUES; SET MEDQUES; RUN;
DATA OUT.MEDREVIEW; SET MEDREVIEW; RUN;
DATA OUT.LABS; SET LABS; RUN;
```

```
options errorabend;
/*****************************************************************************/
/*
/* Program: R:\05_Users\Norma\FIND_CIDRII\Analysis\CIDRII.sas
/* Author:  Norma Pugh
/* Date:    March 2011
/* Purpose: Replicate table 1 results.
/*
/*****************************************************************************/
/* DATA SOURCE */
libname data 'R:\05_Users\Norma\FIND_CIDRII\Data';


/* Sort datasets */
proc sort data=data.pedigree out=pedigree; by studyid; run;
proc sort data=data.medques out=medques; by studyid; run;
proc sort data=data.medreview out=medreview; by studyid; run;


/* Obtain appropriate lab values (note: some lab values needed to define treatment group) */
%macro lab(testname);
 data &testname; set data.labs(where=(lab='MS')); if &testname>.; run;
 proc sort; by studyid collection assay;
 data &testname(keep=studyid &testname); set &testname; by studyid collection assay; if
first.studyid; run;
%mend lab;

%lab(hba1c);
%lab(serumcreat);
%lab(bun)
%lab(macratio_2);
%lab(urineprotein);
%lab(urinecreat);

/* Treatment group: Proband */
data proband(keep=studyid trt); set pedigree; if proband=1 & share='Yes'; trt=1; run;


/* Treatment group: Affected Relative */
data AffRel(keep=studyid trt);
 merge pedigree(in=x1)
            medques
            macratio_2
            medreview
            serumcreat;
 by studyid; if proband=0 & share='Yes';
 if dm=1 &
 (
 (0<macratio_2>0.3 & upr=3)
 or ((0<serumcreat>=1.4 & sex=2) or (0<serumcreat>=1.6 & sex=1))
 or (OrganTrans=1 or dialysis=1)
 );
 trt=2;
 /* Exclusion criterion */
 if (0.03<=macratio_2<0.3) & (upr=2) then delete;
run;
```

```
/* Treatment group: Unaffected Relative */
data UnAffRel(keep=studyid trt);
 merge pedigree(in=x1)
            medques
            macratio_2;
 by studyid; if proband=0 & share='Yes';
 if dm=1 & (0<macratio_2<0.03) & dm_duration>=10;
 trt=3;
run;


/* Final treatment groups */
data treat;
 merge proband AffRel UnAffRel;
 by studyid;
run;


/*************************/
/* DEFINE ANALYSIS DATASET */
/*************************/
data table1;
 merge treat(in=x1)
       pedigree(in=x2 keep=studyid group sex share)
       medques(in=x3 keep=studyid age OrganTrans dm_duration wt ht)
       medreview(in=x4 keep=studyid diagage dialysis)
       hba1c
       serumcreat
       bun
       macratio_2
       urineprotein
       urinecreat;
 by studyid;
 if x1 & share='Yes';

 /* ESRD */
 if OrganTrans=1 or Dialysis=1 then ESRD=1; else ESRD=0;

 /* BMI */
 bmi=(wt/ht**2)*703;

 /* Diagnosis Age */
 numdiagage=diagage+0;

 /* uPCR */
 uPCR=(urineprotein/urinecreat);

 /* ESTIMATED GFR */
 /* Step A */  /* serumcreat = measurement at baseline */
 if sex=2 & serumcreat>1.5 then eGFRSCreat=.;
  else if sex=1 & serumcreat>2 then eGFRSCreat=.;
  else eGFRSCreat=serumcreat;

 /* Step B */
 if sex>. then do;
  if sex=1 then GenderFact=1;
  if sex=2 then GenderFact=0.742;
```

```
 end;

 /* Step C */
 if group>'' then do;
  if group in('EA','MA') then EthFact=1;
  if group='AA' then EthFact=1.212;
 end;

 /* Step D (using MDRD equation) */
 if age>=18 & eGFRSCreat>0 then
  eGFR = 186 * (eGFRSCreat**(-1.154)) * (Age**(-0.203)) * GenderFact * EthFact;
 else eGFR=.;

 /* character vars to numeric */
 if group='AA' then numgroup=1;
 else if group='EA' then numgroup=2;
 else if group='MA' then numgroup=3;
run;


/* REPLICATE ANALYSIS RESULTS */
proc freq data=table1; tables trt / list nopct nocum; title'Treatment Group Counts: Overall';
run;

%macro frq(var);
proc freq data=table1(where=(&var>.)) noprint; tables trt / out=denom(keep=trt count
rename=(count=denom)); run;
proc freq data=table1(where=(&var>.)) noprint; tables trt*&var / out=frqstats(drop=percent); run;
data frqstats; merge frqstats denom; by trt; pct=(count/denom)*100; run;
proc print data=frqstats; title"Frequency Counts: &var"; run;
proc freq data=table1(where=(&var>. & trt in(1,2))) noprint; tables trt*&var / chisq; output
out=pstats chisq; run;
proc print data=pstats; var p_pchi; title"P-value (Probands vs AffRel): &var"; run;
proc freq data=table1(where=(&var>. & trt in(2,3))) noprint; tables trt*&var / chisq; output
out=pstats chisq; run;
proc print data=pstats; var p_pchi; title"P-value (AffRel vs UnAffRel): &var"; run;
%mend frq;

proc sort data=table1; by trt; run;

%macro mean_(var);
proc means data=table1 n mean std; by trt; var &var; run;
proc ttest data=table1(where=(trt in(1,2))); class trt; var &var; title"P-value (Probands vs
AffRel): &var"; run;
proc ttest data=table1(where=(trt in(2,3))); class trt; var &var; title"P-value (AffRel vs
UnAffRel): &var"; run;
%mend mean_;

%frq(numgroup);
%frq(sex);
%mean_(age);
%frq(ESRD);
%mean_(numDiagAge);
%mean_(DM_Duration);
%mean_(bmi);
%mean_(hba1c);
%mean_(serumcreat);    /* Missing in paper */
%mean_(bun);           /* Missing in paper */
```

```
%mean_(macratio_2);
%mean_(uPCR);
%mean_(eGFR);
```