Norma Pugh
April 12, 2007

Integrity Check for the Genetics of Kidneys in Diabetes Study (GoKinD) Files

As a partial check of the integrity of the GoKinD datasets archived in the NIDDK data repository, a set of tabulations was performed to verify that published results from the GoKinD study can be reproduced using the archived datasets. Analyses were performed to duplicate published results for the data reported by Mueller et al [1] in the *Journal of the American Society of Nephrology* in July 2006. The results of this integrity check are described below. The full text of the *Journal of the American Society of Nephrology* article can be found in Attachment 1, and the SAS code for our tabulations is included in Attachment 2.

Background. The purpose of the GoKinD Study was to establish a repository of DNA and clinical information from adults with long-term Type 1 diabetes, with or without kidney disease, including information from their parents. This repository is meant to facilitate investigator-driven research into the genetic basis of diabetic kidney disease as well as other issues concerning Type 1 diabetes. Recruitment of new families for the study was closed as of November 2004 [2].

In summary, the eligibility criteria included: (1) people aged 18-59, who have had Type 1 diabetes for at least 15 years and do not have diabetic kidney disease; (2) people aged 18-54, who have had Type 1 diabetes for at least 10 years and who also have diabetic kidney disease; and (3) both parents of participants are asked to join the study as well, whether or not they have diabetic kidney disease. All probands for the data collection must have Type 1 diabetes and either presence or absence of diabetic nephropathy [3].

Preliminary Tabulations. Initial tabulations of the archived datasets showed 10 fewer probands than the number reported in published results. The Data Coordinating Center (DCC) was notified of the discrepancy and reported that 16 GoKinD participants had their eligibility or trio status changed between the baseline paper data freeze (December 21, 2005) and the final Phase 1 data freeze (April 18, 2006).

This was due to a decision made by the GoKinD Executive Committee (February 2, 2006) to exclude any case/control participants whose DNA was not available for at least 1 round of distribution to external researchers - largely due to low DNA concentration or failure to transform (FTT). In addition, there were 2 instances of change in eligibility status for reasons other than DNA issues. These included an issue with a urine screen and a missed exclusion criterion.

Of these 16 participants whose status was changed, 12 are study probands as defined for purposes of this replication analysis. Of these 12 probands, 10 are missing from the datasets archived at the NIDDK data repository; hence, 10 fewer probands than reported in published results. The DCC has confirmed that the 2 remaining probands are a legitimate part of the datasets archived at the repository.

Due to these different sample sizes, none of the published results will match with the replication analysis. However, comparisons of various variables show that the numbers are close. The NIDDK repository therefore has high confidence in the integrity of the GoKinD datasets.

Baseline Nephropathy Data. Table 1 of the 2006 *Journal of the American Society of Nephrology* article reports on nephropathy status at enrollment for all probands. Variables summarized in this baseline table (Table 1. Nephropathy status at enrollment according to study group) can be found in a single analysis dataset (E1A_CORE) created for the GoKinD study. Table A lists the variables used in our replication of the Table 1 variables.

Norma Pugh
April 12, 2007

Table A: Variables Used to Replicate Table 1 Variables

Table 1 Variable	Variables Used in Replication
Sample size	renalst (if value = 1, 2, 3, 4)
Kidney transplant	renalst (if value = 4)
ESRD duration (yr)	diabdur, timeevnt (calculated as diabdur-timeevnt)
ACR (µg albumin / mg creatinine)	acr1, acr2, acr3 (calculated as : mean(acr1, acr2, acr3))
MDRD GFR (ml/min per 1.73 m^2)	gfr1
Estimated GFR <60 ml/min per 1.73 m^2	gfr1 (if value < 60)

In Table B, we compare the results for sample size and nephropathy status at enrollment calculated from the archived dataset to the results published in the 2006 *Journal of the American Society of Nephrology* article. As Table B shows, the results obtained from the archived data are similar to those in the published tabulations. We conclude that the discrepancies are due to the 10 probands missing from the repository datasets.

Norma Pugh
April 12, 2007

Table B: Comparison of Nephropathy Status Table Values Computed in Integrity Check to Reference Article Values

Variable	Case Probands: ESRD		
	Mueller et al (2006)	**Integrity Check**	**Difference**
Sample size	615	609	6
Kidney transplant (%)	91	91	0
ESRD duration (yr)	8.5 ± 5.3	7.1 ± 5.4	1.4 ± 0.1
ACR (µg albumin / mg creatinine)			
Median	NA	NA	NA
Interquartile range	NA	NA	NA
MDRD GFR (ml/min per 1.73 m^2)	NA	NA	NA
Estimated GFR <60 ml/min per 1.73 m^2 (%)	NA	NA	NA
Legend: ESRD, end-stage renal disease; ACR, urinary albumin/creatinine ratio; MDRD, Modification of Diet in Renal Disease; NA, not applicable			

Variable	Case Probands: Proteinuria		
	Mueller et al (2006)	**Integrity Check**	**Difference**
Sample size	328	326	2
Kidney transplant (%)	NA	NA	NA
ESRD duration (yr)	NA	NA	NA
ACR (µg albumin / mg creatinine)			
Median	1061	1062	1
Interquartile range	606 – 1966	599 - 1921	7 - 45
MDRD GFR (ml/min per 1.73 m^2)	52 ± 26	52 ± 26	0
Estimated GFR <60 ml/min per 1.73 m^2 (%)	65	65	0
Legend: ESRD, end-stage renal disease; ACR, urinary albumin/creatinine ratio; MDRD, Modification of Diet in Renal Disease; NA, not applicable			

Variable	Control Probands		
	Mueller et al (2006)	**Integrity Check**	**Difference**
Sample size	946	944	2
Kidney transplant (%)	NA	NA	NA
ESRD duration (yr)	NA	NA	NA
ACR (µg albumin / mg creatinine)			
Median	5.8	5.8	0
Interquartile range	4.0 – 8.5	4.0 – 8.5	0
MDRD GFR (ml/min per 1.73 m^2)	88 ± 17	88 ± 17	0
Estimated GFR <60 ml/min per 1.73 m^2 (%)	5	5	0
Legend: ESRD, end-stage renal disease; ACR, urinary albumin/creatinine ratio; MDRD, Modification of Diet in Renal Disease; NA, not applicable			

Norma Pugh
April 12, 2007

Baseline Characteristics Data. Table 2 of the 2006 *Journal of the American Society of Nephrology* article reports on baseline characteristics for all probands. Variables summarized in this baseline table (Table 2. Characteristics of probands according to study group) can be found in a single analysis dataset (E1A_CORE) created for the GoKinD study. Table C lists the variables used in our replication of the Table 2 variables.

Table C: Variables Used to Replicate Table 2 Variables

Table 2 Variable	Variables Used in Replication
Sample size	renalst (if value = 1, 2, 3, 4)
White race	race (if value = 1)
Male gender	sex (if value = 1)
Age at entry (yr)	age
Body mass index (kg/m^2)	bmi
Ever smoked cigarettes	eversmk (if value = 2)
Age at diabetes diagnosis (yr)	age, diabdur (calculated as: age-diabdur)
Diabetes duration (yr)	diabdur
PTX	pancr (if value = 2)
HbA$_{1c}$ with PTX	hba1c, pancr (hba1c where pancr = 2)
HbA$_{1c}$ without PTX	hba1c, pancr (hba1c where pancr = 1)
Insulin pump	insregmn (if value = 3)

In Table D, we compare the results for sample size and baseline characteristics calculated from the archived dataset to the results published in the 2006 *Journal of the American Society of Nephrology* article. As Table D shows, the results obtained from the archived data are similar to those in the published tabulations. We conclude that the discrepancies are due to the 10 probands missing from the repository datasets.

Norma Pugh
April 12, 2007

Table D: Comparison of Baseline Characteristics Table Values Computed in Integrity Check to Reference Article Values

Variable	Case Probands		
	Mueller et al (2006)	Integrity Check	Difference
Sample size	943	935	8
Demographic Characteristics			
White race (%)	90	89	1
Male gender (%)	50	50	0
Age at entry (yr)	42.6 ± 7.2	42.6 ± 7.2	0
Body mass index (kg/m^2)	25.7 ± 5.3	25.7 ± 5.3	0
Ever smoked cigarettes (%)	48	48	0
Diabetes history			
Age at diabetes diagnosis (yr)	11.9 ± 6.7	11.9 ± 6.7	0
Diabetes duration (yr)	30.7 ± 7.9	30.8 ± 7.9	0.1 ± 0
PTX (%)	33	33	0
HbA$_{1c}$ (%) with PTX	5.8 ± 1.5	5.8 ± 1.5	0
HbA$_{1c}$ (%) without PTX	8.3 ± 1.6	8.3 ± 1.6	0
Insulin pump (%)	23	18	5
Legend: PTX, pancreas transplant; HbA$_{1c}$, glycosylated hemoglobin; NA, not applicable			

Variable	Control Probands		
	Mueller et al (2006)	Integrity Check	Difference
Sample size	946	944	2
Demographic Characteristics			
White race (%)	97	97	0
Male gender (%)	41	41	0
Age at entry (yr)	38.1 ± 8.6	38.1 ± 8.6	0
Body mass index (kg/m^2)	26.2 ± 4.4	26.2 ± 4.4	0
Ever smoked cigarettes (%)	33	33	0
Diabetes history			
Age at diabetes diagnosis (yr)	12.9 ± 7.3	12.9 ± 7.3	0
Diabetes duration (yr)	25.3 ± 7.7	25.3 ± 7.7	0
PTX (%)	0	0	0
HbA$_{1c}$ (%) with PTX	NA	NA	NA
HbA$_{1c}$ (%) without PTX	7.5 ± 1.2	7.5 ± 1.2	0
Insulin pump (%)	40	39	1
Legend: PTX, pancreas transplant; HbA$_{1c}$, glycosylated hemoglobin; NA, not applicable			

Norma Pugh
April 12, 2007

Other Characteristics. Table 3 of the 2006 *Journal of the American Society of Nephrology* article reports on other characteristics for all probands. Variables summarized in this baseline table (Table 3. Other characteristics related to diabetes) can be found in a single analysis dataset (E1A_CORE) created for the GoKinD study. Table E lists the variables used in our replication of the Table 3 variables.

Table E: Variables Used to Replicate Table 3 Variables

Table 3 Variable	Variables Used in Replication
Sample size	renalst (if value = 1, 2, 3, 4)
Hypertension	hyperten (if value = 2)
Antihypertensive treatment	anth (if value = 2), ace (if value = 2)
Systolic BP (mmHg)	sysbp
Diastolic BP (mmHg)	diabp
Total cholesterol (mg/dl)	cholstr
HDL cholesterol (mg/dl)	hdl
Use of lipid-lowering drugs	lip (if value = 2)
Number of parents living	*(variable not included in distributed dataset)*
Laser therapy for retinopathy	flaser (if value = 2), plaser (if value = 2)
Cardiovascular disease	cardiov (if value = 2)
Neuropathy	neurp (if value = 2)

In Table F, we compare the results for sample size and other characteristics calculated from the archived dataset to the results published in the 2006 *Journal of the American Society of Nephrology* article. As Table F shows, the results obtained from the archived data are similar to those in the published tabulations. We conclude that the discrepancies are due to the 10 probands missing from the repository datasets.

Norma Pugh
April 12, 2007

Table F: Comparison of Other Characteristics Table Values Computed in Integrity Check to Reference Article Values

Variable	Case Probands		
	Mueller et al (2006)	**Integrity Check**	**Difference**
Sample size	943	935	8
Hypertension (%)	85	85	0
Antihypertensive treatment (%)	83	82	1
Systolic BP (mmHg)	131 ± 19	131 ± 19	0
Diastolic BP (mmHg)	74 ± 11	74 ± 11	0
Total cholesterol (mg/dl)	189 ± 46	190 ± 46	1 ± 0
HDL cholesterol (mg/dl)	54 ± 18	54 ± 17	0 ± 1
Use of lipid-lowering drugs (%)	45	45	0
Number of parents living (%)		*not compared*	*not compared*
0	26	.	.
1	23	.	.
2	48	.	.
unknown	2	.	.
Laser therapy for retinopathy (%)	85	83	2
Cardiovascular disease (%)	89	86	3
Neuropathy (%)	68	68	0
Legend: NA, not applicable			

Variable	Control Probands		
	Mueller et al (2006)	**Integrity Check**	**Difference**
Sample size	946	944	2
Hypertension (%)	6	6	0
Antihypertensive treatment (%)	NA	NA	NA
Systolic BP (mmHg)	118 ± 12	118 ± 12	0
Diastolic BP (mmHg)	71 ± 8	71 ± 8	0
Total cholesterol (mg/dl)	185 ± 32	185 ± 32	0
HDL cholesterol (mg/dl)	58 ± 16	58 ± 15	0 ± 1
Use of lipid-lowering drugs (%)	15	15	0
Number of parents living (%)		*not compared*	*not compared*
0	13	.	.
1	20	.	.
2	64	.	.
unknown	3	.	.
Laser therapy for retinopathy (%)	16	16	0
Cardiovascular disease (%)	11	9	2
Neuropathy (%)	11	11	0
Legend: NA, not applicable			

Norma Pugh
April 12, 2007

Notes

1. One of the four analysis datasets (combined data across sites) provided is examined in this replication analysis (E1A_CORE). This dataset contains all baseline measurements on all enrolled patients. The remaining analysis datasets (combined data across sites) include family lookup and genetic data. In addition to the analysis datasets containing combined data across sites, datasets for each individual site (George Washington University clinics and Joslin Diabetes Clinic) are housed at the repository.

References

1. Patricia W. Mueller, John J. Rogus, Patricia A. Cleary, Yuan Zhao, Adam M. Smiles, Michael W. Steffes, Jean Bucksa, Therese B. Gibson, Suzanne K. Cordovado, Andrzej S. Krolewski, Concepcion R. Nierras, and James H. Warram
   Genetics of Kidneys in Diabetes (GoKinD) Study: A Genetics Collection Available for Identifying Genetic Susceptibility Factors for Diabetic Nephropathy in Type 1 Diabetes
   J. Am. Soc. Nephrol., Jul **2006**; **17**: **1782** - 1790.

2. Genetics of Kidneys in Diabetes (GoKinD) Website: Home page. Genetics of Kidneys in Diabetes (GoKinD) Study

3. Genetics of Kidneys in Diabetes (GoKinD) Website: Eligibility criteria page. Genetics of Kidneys in Diabetes (GoKinD) Study: Eligibility

ATTACHMENT 1

" The full text of the article referenced will be provided to approved requestors along with the data archive."

Patricia W. Mueller, John J. Rogus, Patricia A. Cleary, Yuan Zhao, Adam M. Smiles, Michael W. Steffes, Jean Bucksa, Therese B. Gibson, Suzanne K. Cordovado, Andrzej S. Krolewski, Concepcion R. Nierras, and James H. Warram. Genetics of Kidneys in Diabetes (GoKinD) Study: A Genetics Collection Available for Identifying Genetic Susceptibility Factors for Diabetic Nephropathy in Type 1 Diabetes. Journal of the American Society of Nephrology, Jul 2006; 17: 1782 - 1790.

ATTACHMENT 2

SAS Code for Tabulations from GoKinD Datasets in the NIDDK Repository

```
/**************************************************************************************/
/*
/* Program: R:\05_Users\Norma\GoKinD\table1.sas
/* Author:  Norma Pugh
/* Date:    22 January 07
/* Purpose: Replicate results from table 1.
/**************************************************************************************/

/* Libnames and formats */
libname data 'R:\05_Users\Norma\GoKinD\TransportedData';
%include 'R:\05_Users\Norma\GoKinD\formats.sas';

/* Get Table 1 variables */
data table1;
 set data.e1a_core(where=(renalst in(1,2,3,4)));
 length trt $ 9;
 if renalst=1 then trt='Control';
 if renalst=2 then trt='Case_Prot';
 if renalst in(3,4) then trt='Case_ESRD';

 if renalst=4 then transplant=1; else transplant=0;
 meanacr=mean(of acr1-acr3);
 if gfr1<60 then gfr_lt_60=1; else gfr_lt_60=0;
run;

title'Table 1: Treatment counts'; run;
proc freq data=table1; tables trt; run;

title'Table 1: ESRD only - Categorical counts'; run;
proc freq data=table1(where=(trt='Case_ESRD')); tables transplant; run;

title'Table 1: Quantitative stats'; run;
proc sort data=table1; by trt; run;
proc means data=table1(where=(trt^='Case_ESRD')) n median q1 q3;
 by trt;
 var meanacr;
run;

proc means data=table1(where=(trt^='Case_ESRD')) n mean std;
 by trt;
 var gfr1;
run;

title'Table 1: Categorical counts'; run;
proc freq data=table1(where=(trt^='Case_ESRD'));
 by trt;
 tables gfr_lt_60;
run;
```

```
/******************************************************************************/
/*
/* Program: R:\05_Users\Norma\GoKinD\table2.sas
/* Author:  Norma Pugh
/* Date:    22 January 07
/* Purpose: Replicate results from table 2.
/******************************************************************************/

/* Libnames and formats */
libname data 'R:\05_Users\Norma\GoKinD\TransportedData';
%include 'R:\05_Users\Norma\GoKinD\formats.sas';

/* Get Table 2 variables */
data table2;
 set data.e1a_core(where=(renalst in(1,2,3,4)));
 if renalst=1 then trt='Control';
 if renalst in(2,3,4) then trt='Case';

 if race=1 then white='y'; else white='n';
 age_at_diab=age-diabdur;
 if eversmk=2 then smoke='y'; else smoke='n';
 if trt='Case' & pancr=1 then trt_ptx='Case_noptx';
  else if trt='Control' & pancr=1 then trt_ptx='Cntl_noptx';
 if insregmn=3 then pump='y'; else pump='n';
run;

title'Table 2: Categorical Counts & p-values'; run;
proc freq data=table2; tables trt*(white sex smoke pancr pump) / chisq; run;

title'Table 2: Quantitative Means & Standard Deviations'; run;
proc sort data=table2; by trt pancr; run;
proc means data=table2 n mean std;
 by trt;
 var age bmi age_at_diab diabdur;
run;

proc means data=table2 n mean std;
 by trt;
 class pancr;
 var hba1c;
run;

title'Table 2: Quantitative p-values'; run;
proc npar1way data=table2 wilcoxon;
 class trt;
 var age bmi age_at_diab diabdur;
run;

proc npar1way data=table2(where=(trt_ptx in('Case_noptx','Cntl_noptx'))) wilcoxon;
 class trt_ptx;
 var hba1c;
run;
```

```
/******************************************************************************/
/*
/* Program: R:\05_Users\Norma\GoKinD\table3.sas
/* Author:  Norma Pugh
/* Date:    7 February 07
/* Purpose: Replicate results from table 3.
/******************************************************************************/

/* Libnames and formats */
libname data 'R:\05_Users\Norma\GoKinD\TransportedData';
%include 'R:\05_Users\Norma\GoKinD\formats.sas';

/* Get Table 3 variables */
data table3;
 set data.e1a_core(where=(renalst in(1,2,3,4)));
 if renalst=1 then trt='Control';
 if renalst in(2,3,4) then trt='Case';

 if hyperten=2 then hi_bp='y'; else hi_bp='n';
 if anth=2 or ace=2 then med='y'; else med='n';
 if flaser=2 or plaser=2 then laser='y'; else laser='n';
 if neurp=2 then neurpthy='y'; else neurpthy='n';
run;

title'Table 3: Categorical Counts & p-values'; run;
proc freq data=table3;
 tables trt*(hi_bp med lip    laser cardiov neurpthy) / chisq;
run;

title'Table 3: Quantitative Means & Standard Deviations'; run;
proc sort data=table3; by trt; run;
proc means data=table3 n mean std;
 by trt;
 var sysbp diabp cholstr hdl;
run;

title'Table 3: Quantitative p-values'; run;
proc npar1way data=table3 wilcoxon;
 class trt;
 var sysbp diabp cholstr hdl;
run;
```

```
/**********************************************************************************/
/*
/* Program: R:\05_Users\Norma\GoKinD\table1_update.sas
/* Author:  Norma Pugh
/* Date:    29 March 07
/* Purpose: Update replication for table 1 to include 'ESRD duration(yr)'. Use timeevnt
/*           variable per Paddy Cleary e-mail.
/* Revised: 12 April 07 to use (diabdur-timeevnt), per updated Paddy Cleary e-mail.
/**********************************************************************************/

/* Libnames and formats */
libname data 'R:\05_Users\Norma\GoKinD\TransportedData';
%include 'R:\05_Users\Norma\GoKinD\formats.sas';

/* Get Table 1 variables */
data table1;
 set data.e1a_core(where=(renalst in(1,2,3,4)));
 length trt $ 9;
 if renalst=1 then trt='Control';
 if renalst=2 then trt='Case_Prot';
 if renalst in(3,4) then trt='Case_ESRD';
 ESRD_yrs=diabdur-timeevnt;
 run;

proc sort data=table1; by trt; run;

proc means data=table1(where=(trt='Case_ESRD')) n mean std;
 by trt;
 var ESRD_yrs;
 title'Table 1: Quantitative stats'; run;
run;
```