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.

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.

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			

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.

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			

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.

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			

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

Full Text of Article

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.

Special Feature

Genetics of Kidneys in Diabetes (GoKinD) Study: A Genetics Collection Available for Identifying Genetic Susceptibility Factors for Diabetic Nephropathy in Type 1 Diabetes

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†

*Centers for Disease Control and Prevention, Diabetes and Molecular Risk Assessment Laboratory, Atlanta, Georgia; †Research Division, Joslin Diabetes Center, Boston, Massachusetts; ‡George Washington University Biostatistics Center, Washington, DC; §University of Minnesota, Minneapolis, Minnesota; ‖Aspen Systems, Inc., Rockville, Maryland; and ¶Juvenile Diabetes Research Foundation, New York, New York

The Genetics of Kidneys in Diabetes (GoKinD) study is an initiative that aims to identify genes that are involved in diabetic nephropathy. A large number of individuals with type 1 diabetes were screened to identify two subsets, one with clear-cut kidney disease and another with normal renal status despite long-term diabetes. Those who met additional entry criteria and consented to participate were enrolled. When possible, both parents also were enrolled to form family trios. As of November 2005, GoKinD included 3075 participants who comprise 671 case singletons, 623 control singletons, 272 case trios, and 323 control trios. Interested investigators may request the DNA collection and corresponding clinical data for GoKinD participants using the instructions and application form that are available at http://www.gokind.org/access. Participating scientists will have access to three data sets, each with distinct advantages. The set of 1294 singletons has adequate power to detect a wide range of genetic effects, even those of modest size. The set of case trios, which has adequate power to detect effects of moderate size, is not susceptible to false-positive results because of population substructure. The set of control trios is critical for excluding certain false-positive results that can occur in case trios and may be particularly useful for testing gene–environment interactions. Integration of the evidence from these three components into a single, unified analysis presents a challenge. This overview of the GoKinD study examines in detail the power of each study component and discusses analytic challenges that investigators will face in using this resource.

J Am Soc Nephrol 17: 1782–1790, 2006. doi: 10.1681/ASN.2005080822

Diabetes is the leading cause of treated ESRD, accounting for almost half of the new cases each year (1–3). Among European Americans with type 1 diabetes, approximately one in three develops severe nephropathy that leads to ESRD (4–6). Evidence that genetic susceptibility plays an important role in diabetic nephropathy in type 1 diabetes first was presented more than a decade ago by Seaquist *et al.* (7) and Borch-Johnsen (8), and subsequent studies by researchers at the Joslin Diabetes Center (9) and The Diabetes Control and Complications Trial Research Group (10) further characterized the nature of the genetic effect.

Despite the strong evidence for genetic susceptibility factors, success in identifying the responsible genetic variants has been limited by the modest data collections that individual research groups have been able to assemble. The Genetics of Kidneys in Diabetes (GoKinD) study, an initiative supported by the Juvenile Diabetes Research Foundation (JDRF) and by the National Institute of Diabetes and Digestive and Kidney Diseases and the Centers for Disease Control and Prevention, was conceived to address this bottleneck by assembling a large DNA collection that is suitable for genetic association studies of nephropathy in type 1 diabetes.

The resulting collection includes nearly 1900 individuals with long-term (10+ yr) type 1 diabetes, half with nephropathy (943 case patients) and half without (946 control subjects). The set of case patients includes two subgroups: 328 patients with persistent proteinuria and 615 with ESRD. The set of control subjects consists only of individuals with normoalbuminuria despite 15 yr of type 1 diabetes. Both sets can be partitioned into two subsets: Those with neither parent enrolled (singletons) and those with both parents enrolled (trios). The totals as of November 2005 included 671 case singletons, 272 case trios, 623 control singletons, and 323 control trios.

Published online ahead of print. Publication date available at www.jasn.org.

P.W.M. and J.J.R. contributed equally to this work.

Address correspondence to: Dr. James H. Warram, Section on Genetics and Epidemiology, Joslin Diabetes Center, One Joslin Place, Boston, MA 02215. Phone: 617-732-2668; Fax: 617-732-2667; E-mail: james.warram@joslin.harvard.edu

ISSN: 1046-6673/1707-1782

The concept of using family trios to detect genetic association was developed more than a decade ago by various researchers who were wary of implicating a genetic variant simply because it happens to occur with greater frequency in a subset of the study participants who also have a relatively high occurrence of disease. To illustrate, consider a study of osteoporosis in individuals of European descent. If, in general, osteoporosis is more common in those of northern European descent compared with southern European descent, then any genetic variant that is more common in the former will tend to exhibit association with case-control analysis. The gold standard that has emerged for addressing such population stratification is the transmission/disequilibrium test (TDT) (11). The TDT procedure evaluates case trios in such a way that only relevant genetic variants are identified. An excellent review of the TDT has been written by two of the pioneers of the field, Ewens and Spielman (12). Recently, Scott and Rogus (13) examined the utility of control trios and found that they are useful in special situations, such as when a disease is highly prevalent or when certain types of gene–environment interaction exist.

GoKinD uses both case trios and control trios as well as a set of unrelated case and control singletons. The advantage of including singletons is that, in addition to being much easier to identify and ascertain, they offer exceptionally high power to detect genetic association. The tradeoff, of course, is that they are prone to false-positive results if population stratification exists.

The GoKinD Collection of DNA and clinical documentation of case patients and control subjects are available to the research community through an application process that is accessible on the GoKinD web site (https://www.gokind.org/access). Nonrenewable samples also will become available at a later date. Broad distribution of the collection is intended to spark creativity with regard to both the genetic variants studied and the analytic approaches used. These approaches are not limited to those that require the whole collection. The large collection also may be used as a sampling frame for selecting narrowly defined groups for testing very specific hypotheses. Here, we summarize the clinical characteristics of the study groups and provide detailed power calculations for each of the collection's design components. Finally, we discuss some analytic challenges that await potential users of the collection.

Materials and Methods

Organization of GoKinD

The collaborative effort to build the GoKinD collection was organized through a coordinating center, housed jointly at the Joslin Diabetes Center (JDC) and the George Washington University Biostatistics Center (GWU), a Central Biochemistry Laboratory at the University of Minnesota (CBL), and a genetics laboratory and specimen repository at the Centers for Disease Control and Prevention (CDC).

Recruitment and Study Groups

Patients for this study were recruited through two centers. The Section of Genetics and Epidemiology at the JDC recruited and examined patients of the Joslin Clinic who were already enrolled in the Joslin Kidney Study on Genetics of Diabetic Nephropathy. All these patients resided in New England. In total, 320 case singletons, 180 case trios, 346 control singletons, and 154 control trios were recruited through the JDC. The George Washington Biostatistics Center worked with Matthews Media Group to identify for this study in the United States and Canada volunteers who subsequently were directed to one of 27 clinical centers around the United States for examination. In total, 351 case singletons, 92 case trios, 277 control singletons, and 169 control trios were recruited through GWU. The principal investigators and recruitment staff who contributed to the collection are listed in the Acknowledgments. All data management was centralized at GWU.

To be eligible as a case patient, a patient had to have type 1 diabetes (minimum 10 yr duration) and severe diabetic nephropathy (ESRD or persistent proteinuria). To be eligible as a control subject, a patient had to have type 1 diabetes for at least 15 yr and have normoalbuminuria despite never having been treated with angiotensin-converting enzyme inhibitors or angiotensin receptor blockers and not receiving current treatment with antihypertensive medication. Persistent proteinuria and normoalbuminuria were defined by the urinary albumin to creatinine ratio (ACR) (≥300 and <20 μg/mg, respectively). Further details of the eligibility criteria are summarized in Figure 1. When both parents of a participant were alive and willing to participate, both were examined to

Type 1 diabetes	Diabetes diagnosed before age 31, insulin treatment begun within one year of diagnosis and continued uninterrupted since diagnosis. Tests for GAD antibodies were not performed.
Severe diabetic nephropathy	Persistent proteinuria or ESRD not attributable to a condition other than diabetes and arising after at least 10 years of diabetes duration.
ESRD	Chronic dialysis or kidney transplant. The onset of ESRD is defined as the date of the first dialysis or kidney transplant, whichever occurred first.
Persistent proteinuria	At least two of the last three urine samples positive for albuminuria in specimens taken at least one month apart. One test could be a historical result from the medical record documenting a urinary albumin/creatinine ratio (ACR) exceeding 300 μg albumin/mg creatinine or a 1+ dipstick (e.g. Multistix). All others had to be confirmed by the CBL as a urinary ACR exceeding 300 μg albumin/mg creatinine
Normoalbuminuria:	At least two of the last three ACR measurements in random urine specimens taken at least one month apart being less than 20 μg albumin/mg creatinine. If 3 ACR measurements are needed, the highest must be less than 40 μg albumin/mg creatinine. One could be a historical result from the medical record. All others had to be confirmed by the CBL as a urinary ACR less than 20 μg albumin/mg creatinine

Figure 1. Definitions of Genetics of Kidneys in Diabetes (GoKinD) study eligibility criteria.

form complete trios (proband and both parents) for TDT analysis. Additional eligibility requirements were age 18 through 59 yr at the time of enrollment. Patients were recruited regardless of gender, race, or ethnic origin. However, patients were excluded when they could not communicate with staff or reported HIV infection or active tuberculosis. Pregnant women were excluded, but they became eligible for screening 3 mo postpartum.

All participants signed informed consent forms that explained the purpose of the collection and the intention to share their DNA and other biologic samples with investigators who were approved by a scientific review process that was established by JDRF. The project and consent procedures were approved by local Institutional Review Boards of all recruitment centers, the coordinating centers, CBL, and the CDC.

Sample Processing

Detailed descriptions of the methods that were used at the CBL and CDC are available in Supplementary Appendix A (available online). In brief, biologic samples were shipped from each recruitment facility to the CBL for analysis of albumin and creatinine in urine, hemoglobin A_{1c} (HbA_{1c}) in blood, and total cholesterol, HDL cholesterol, cystatin C, and creatinine in serum. The CBL also prepared whole-blood lysates for DNA extraction and transformed peripheral blood lymphocytes to establish cell lines for additional DNA supplies. Cryopreserved cell lines; whole-blood lysates; and saved urine, serum, and plasma samples were shipped to the CDC, which is the repository for all GoKinD biologic samples. The CDC extracted DNA from both whole-blood cell lysates and transformed lymphocyte lysates and genotyped the HLA DQA1, DQB1, and DRB1 loci; the −23 insulin gene single-nucleotide polymorphism (14); and additional microsatellite markers to test for sample mix-ups and verify family relationships.

Tracking of specimens from recruitment facilities to the repository at the CDC was the responsibility of GWU. Distribution of the collection to approved investigators will be handled jointly by the CDC (DNA) and GWU (clinical data).

Quality Control

Replicate samples were collected from 5% of the participants to permit quality control analysis of study procedures from sample collection through DNA genotyping. For the seven clinical measurements at the CBL, the coefficients of reliability ranged from 95 to 99% except for urine albumin (93%) and urine ACR (91%) (15). The lymphocyte cell transformation success rate at the CBL was 99.8% (2354 of 2360 samples). For testing sample mix-ups and nonpaternity, three microsatellites and a gender-specific locus were genotyped at the CDC (in addition to the HLA and insulin loci). All problematic samples subsequently were genotyped for nine additional microsatellites to resolve the issue. Noteworthy is that these microsatellite tests are sensitive enough to detect even slight sample contamination. The CDC genotyped 3302 potentially eligible individuals for the collection. Not one instance of contamination of a blood sample with a second individual's blood was found. We detected 17 instances of sample labeling errors and, allowing for undetected errors, estimate that labeling errors occurred between five and seven times per 1000 individuals. The 17 detected errors were resolved or removed from the collection. After these corrections, we estimate that the final collection of 3076 individuals may include three to six undetected sample mix-ups in the 1291 singletons and none in the trios (error rate between one and two per 1000 for the whole collection).

Statistical Analyses

Case patients and control subjects were compared using Wilcoxon rank-sums tests for quantitative variables and χ^2 or Fisher exact test for categorical variables.

Power Calculations

For purposes of calculating power, diabetic nephropathy was considered as a dichotomy. Patients with ESRD were not distinguished from patients with proteinuria. For each component of the collection (singletons, case trios, and control trios), power was estimated for a range of scenarios with regard to the underlying genetic model. To do so, we used the first approximation suggested by Knapp (16) for case trios, the extensions of Scott and Rogus (13) for control trios, and Rogus *et al.* (17) for singletons. The required input parameters include the frequency of the risk allele (P) and the relative risks (RR) for those who are homozygotes (ψ_2) or heterozygotes (ψ_1) with respect to the risk allele. For control trios and singletons, prevalence also must be specified. Our calculations assumed 35% prevalence of renal disease in type 1 diabetes and risk allele frequencies (P) of 0.1, 0.3, and 0.5. Sensitivity analysis also was performed assuming prevalence of either 30 or 40%. RR were set according to four modes of inheritance: Multiplicative (ψ_2

Table 1. Nephropathy status at enrollment according to study group[a]

Characteristic	Case Probands		Control Probands
	ESRD (n = 615)	Proteinuria[b] (n = 328)	Normoalbuminuria (n = 946)
Kidney transplant (%)	91	NA	NA
ESRD duration (yr)	8.5 ± 5.3	NA	NA
ACR (μg albumin/mg creatinine)			
median	NA	1061	5.8
interquartile range	NA	606 to 1966	4.0 to 8.5
MDRD GFR (ml/min per 1.73 m^2) (18)	NA	52 ± 26	88 ± 17
Estimated GFR <60 ml/min per 1.73 m^2 (%)	NA	65	5

[a]Data are mean ± SD or % except for ACR, for which median and interquartile range are given. ACR, urinary albumin/creatinine ratio; MDRD, Modification of Diet in Renal Disease; NA, not applicable.

[b]An additional 139 screened probands (99 singletons and 40 trios) were found to have microalbuminuria. Their DNA and biologic samples, although excluded from the Genetics of Kidneys in Diabetes (GoKinD) collection, were retained in the repository for potential future use.

$= \gamma$, $\psi_1 = \gamma^{1/2}$), additive ($\psi_2 = \gamma$, $\psi_1 = [\gamma + 1]/2$), recessive ($\psi_2 = \gamma$, $\psi_1 = 1$), and dominant ($\psi_2 = \gamma$, $\psi_1 = \gamma$). A feature of this parameterization is the consistency of homozygote RR (γ) across all modes of inheritance. We report power estimates for homozygote RR values of $\gamma = 3.0$, 2.5, and 2.0, assuming a one-sided test with $\alpha = 5\%$. We also examine models with $\gamma = 1.5$ in the context of power to detect genes with modest effects.

Results

Characteristics of the Study Groups

The characteristics of probands whose parents were unavailable for completing trios (singletons) were, in general, similar to the characteristics of those with parents (trio probands). Therefore, singleton and trio probands were combined in Tables 1 through 3 to focus attention on the differences between case patients and control subjects. All the characteristics shown here, as well as many additional characteristics that were omitted for brevity, are available according to study group and separately for singletons and trios in Supplementary Appendix B (available online).

Nephropathy Status

Case patients included two subgroups: Those with ESRD (65%) and those with proteinuria (35%). For highlighting the differences between these two subgroups, renal characteristic of probands are summarized in Table 1 according to three categories: Case patients with ESRD, case patients with protein-

Table 2. Characteristics of probands according to study group[a]

	Case Probands ($n = 943$)	Control Probands ($n = 946$)	*P*
Demographic characteristics			
white race (%)	90	97	<0.0001
male gender (%)	50	41	<0.0001
age at entry (yr)	42.6 ± 7.2	38.1 ± 8.6	<0.0001
body mass index (kg/m^2)	25.7 ± 5.3	26.2 ± 4.4	<0.0001
ever smoked cigarettes	48%	33%	<0.0001
Diabetes history			
age at diabetes diagnosis (yr)	11.9 ± 6.7	12.9 ± 7.3	0.0095
diabetes duration (yr)	30.7 ± 7.9	25.3 ± 7.7	<0.0001
PTX (%)	33	0	<0.0001
HbA_{1c} (%) with PTX	5.8 ± 1.5[b]	NA	
HbA_{1c} (%) without PTX	8.3 ± 1.6	7.5 ± 1.2	<0.0001
insulin pump (%)	23	40	<0.0001

[a]PTX, pancreas transplant; HbA_{1c}, glycosylated hemoglobin.
[b]$P < 0.0001$, for case patients without PTX *versus* those who had transplants.

Table 3. Other characteristics related to diabetes

Characteristic	Case Probands ($n = 943$)	Control Probands ($n = 946$)	*P*
Hypertension	85%	6%	<0.0001
Antihypertensive treatment	83%	NA	
Systolic BP (mmHg)	131 ± 19	118 ± 12	<0.0001
Diastolic BP (mmHg)	74 ± 11	71 ± 8	<0.0001
Total cholesterol (mg/dl)	189 ± 46	185 ± 32	0.1575
HDL cholesterol (mg/dl)	54 ± 18	58 ± 16	<0.0001
Use of lipid-lowering drugs	45%	15%	<0.0001
No. of parents living			<0.0001[a]
0	26%	13%	
1	23%	20%	
2	48%	64%	
unknown	2%	3%	
Laser therapy for retinopathy[b]	85%	16%	<0.0001
Cardiovascular disease[b]	89%	11%	<0.0001
Neuropathy[b]	68%	11%	<0.0001

[a]χ^2 for the difference in the distribution of the number of living parents.
[b]The complications are self-reported.

uria, and control subjects with normoalbuminuria. At enrollment, case patients with ESRD had survived 8.5 ± 5.3 yr after the onset of ESRD, and 91% of them had received a kidney transplant; the remainder were on dialysis. Urinary albumin excretion of case patients with proteinuria generally was well above the lower limit for proteinuria (ACR ≥ 300 μg/mg). Median ACR was 1061 μg/mg (interquartile range 602 to 1941). For control subjects, albumin excretion generally was well below the upper limit of normoalbuminuria (ACR < 20 μg/mg). Median ACR was 5.8 μg/mg (interquartile range 4.0 to 8.4). Renal function, as estimated by the Modification of Diet in Renal Disease equation from serum creatinine, was significantly reduced in case patients with proteinuria as compared with control subjects, with 65% having an estimated GFR <60 ml/min as compared with only 5% of control subjects. Alternative estimates of renal function as based on serum creatinine and cystatin C are available in Supplementary Appendix B.

Demographic Characteristics

The GoKinD collection is primarily a white collection: 90% of case patients and 97% of control subjects (Table 2). Most of the study groups are approximately 40% male, with the exception of the case singletons, which is 53% male. On average, case patients were 4 yr older than control subjects, and this difference was due largely to the age of case patients with ESRD (43.9 ± 6.5), which is 3 yr older than the age of case patients with proteinuria (40.3 ± 7.8). Regardless of renal status, the age of trio probands was younger than singletons, presumably because the availability of both parents was age related. A positive smoking history was reported by almost half of the case patients as compared with one third of the control subjects ($P < 10^{-4}$).

Diabetes History

The age at diagnosis of type 1 diabetes was similar in control subjects and case patients, but the duration of diabetes at enrollment was 5 yr longer, on average, for case patients ($P < 10^{-4}$). This difference was due partly to the longer diabetes duration of case patients with ESRD (32.0 ± 7.3) as compared with case patients with proteinuria (28.3 ± 8.4). However, the diabetes duration of case patients with ESRD at the onset of ESRD was 23.9 ± 6.7, which was significantly ($P < 10^{-4}$) less than the diabetes duration at enrollment for case patients with proteinuria and similar to that for control subjects ($P = 0.0022$). The level of glycemic control at enrollment was significantly affected ($P < 10^{-4}$) by whether the proband had a pancreas transplant, a procedure reported only by case patients. The HbA_{1c} of the 33% of case patients with a pancreas transplant was 5.8 ± 1.5%, whereas it was 8.3 ± 1.6% for case patients without a pancreas transplant as compared with 7.5 ± 1.2% for control subjects ($P = 0.0001$). Noteworthy, at enrollment, insulin pumps were being used by 40% of control subjects but only 23% of case patients ($P < 10^{-4}$).

Other Conditions Related to Diabetes

Hypertension was present in 85% of case patients, and almost all were treated with antihypertensive medication. A history of antihypertensive medication was an exclusion criterion for control subjects; therefore, hypertension was infrequent (6%). The few control subjects who were recruited with hypertension had not yet begun treatment because the hypertension was diagnosed in conjunction with the enrollment examination. Despite treatment, measured systolic and diastolic BP were higher in case patients than in control subjects ($P < 10^{-4}$ for both). Total cholesterol and HDL cholesterol were similar in case patients and control subjects despite more frequent use of lipid-lowering drugs by case patients than control subjects (45 and 15%, respectively; $P < 10^{-4}$). Parental mortality was higher among case patients than control subjects ($P < 10^{-4}$) and was the chief reason that probands (among control subjects as well as case patients) were not available for forming trios. Both parents were living for only 48% of case patients and 64% of control subjects.

Other Complications of Diabetes

A history of laser treatment for retinopathy and diagnosed cardiovascular disease were reported by most case patients but only a few control subjects ($P < 10^{-4}$; Table 3). Self-reported neuropathy was less prevalent but reported mainly by case patients. The prevalence of all three was somewhat higher in case patients with ESRD than in case patients with proteinuria.

Power Calculations

The goal of GoKinD is to identify genetic variants that play a role in diabetic nephropathy. A variant may exert an independent effect on nephropathy or an interacting effect that involves other genes or nongenetic factors. In this article, the simplest situation of a single genetic locus is presented in depth. More complex situations are addressed in the Discussion section.

Power calculations were performed separately for each of the three study components assuming a lifetime cumulative risk of 35% for diabetic nephropathy in patients with type 1 diabetes (4). Parameters of a single locus model were varied to include all combinations of four alternative modes of inheritance (dominant, recessive, additive, and multiplicative), three choices for the frequency of the risk allele (0.1, 0.3, and 0.5), and three values for the disease risk for individuals who carry two risk alleles (homozygotes) relative to individuals who carry none. Results are summarized in Table 4. Results also were obtained for models with a RR of 1.5 for the homozygotes. These are not shown in Table 4 but are described in the text.

In almost all circumstances, the set of 1294 singletons has excellent power (>99%). The lone exception, the recessive model with 10% allele frequency, has power of only 30 to 70%. This situation, however, represents an unlikely scenario that assigns approximately 34% risk for nephropathy to 99% of the population and almost 100% risk to the remaining 1%. Excluding this unlikely case, good power (>80%) is maintained even for models with RR of 1.5. Therefore, the set of singletons is sufficient to detect genetic effects of even modest size.

The set of case trios has ample power to detect most effects of moderate size (RR for homozygotes for the risk allele ≥2). For example, power ranges from 77 to 99% for the dominant mod-

Table 4. Power for each of the GoKinD study design components to detect genetic association[a]

	Multiplicative Model			Additive Model			Recessive Model			Dominant Model		
P[b]	Case Trios (n = 272)	Control Trios (n = 323)	Singletons (n = 1294)	Case Trios (n = 272)	Control Trios (n = 323)	Singletons (n = 1294)	Case Trios (n = 272)	Control Trios (n = 323)	Singletons (n = 1294)	Case Trios (n = 272)	Control Trios (n = 323)	Singletons (n = 1294)
γ[b] = 3.0[c]												
0.10	0.91	0.69	0.99	0.98	0.85	0.99	0.23	0.13	0.70	0.99	0.99	0.99
0.30	0.99	0.87	0.99	0.99	0.89	0.99	0.97	0.76	0.99	0.99	0.94	0.99
0.50	0.99	0.83	0.99	0.99	0.78	0.99	0.99	0.95	0.99	0.96	0.61	0.99
γ = 2.5												
0.10	0.79	0.52	0.99	0.89	0.66	0.99	0.18	0.11	0.54	0.99	0.97	0.99
0.30	0.97	0.73	0.99	0.98	0.76	0.99	0.88	0.56	0.99	0.99	0.86	0.99
0.50	0.98	0.71	0.99	0.97	0.67	0.99	0.99	0.85	0.99	0.91	0.52	0.99
γ = 2.0												
0.10	0.57	0.33	0.99	0.66	0.40	0.99	0.13	0.08	0.32	0.96	0.77	0.99
0.30	0.85	0.51	0.99	0.88	0.54	0.99	0.64	0.34	0.99	0.96	0.68	0.99
0.50	0.88	0.51	0.99	0.87	0.49	0.99	0.95	0.62	0.99	0.77	0.40	0.99

[a]One-sided test, α = 0.05.
[b]P is the allele frequency of the risk allele in the population γ is the ratio of the disease risk for carriers of two risk alleles to the disease risk for carriers of zero risk alleles.
[c]For recessive model with P = 0.1, γ was set to 2.9 to maintain legal penetrance values.

els, 66 to 99% for the additive models, and 57 to 99% for the multiplicative models. For the recessive models, excluding those with a 10% risk allele, power ranges from 64 to 99%. For RR of 1.5, the maximum power for the models considered is only 66%.

For the set of control trios, power was more model dependent. Excluding the rare recessive case, 30% of the models had power that exceeded 80%, another 36% had power between 60 to 80%, and the remaining 33% had power <60%.

European-Americans constitute 1757 (93%) of the probands, and the remaining 134 probands represent a collection of small numbers from other ethnic/racial groups. When the analysis is restricted to European Americans, power is consistently reduced by approximately 2 percentage points for any given scenario (*i.e.*, power of 86% in the entire data set would decrease to 84% if only white individuals are considered).

As noted above, power calculations were based on an assumed lifetime risk for diabetic nephropathy of 35%. This figure was based on cohort studies of European American children with type 1 diabetes in New England. That risk for patients with type 1 diabetes may vary geographically or between ethnic/racial groups. Therefore, we conducted a sensitivity analysis by varying the assumed lifetime risk. The results of this analysis are unique for each of the three study design components. For case trios, power does not depend at all on lifetime risk. This property, which has been described previously (13), suggests that the power calculations for case trios in Table 4 apply regardless of the actual lifetime risk. For singletons, when the rare recessive case is excluded, the change in lifetime risk to 30 or to 40% is immaterial, because power exceeds 99% in all circumstances. Even at a lifetime risk of 20%, the power of singletons exceeds 98% for all scenarios. Where lifetime risk matters more, as predicted by Scott and Rogus (13), is for control trios. If risk is 30%, then the actual power of the control trios is approximately 20% less than the values in Table 4 (range 73 to 100%). If risk is 40%, then the actual power is approximately 20% greater than the values in Table 4 (range 100 to 136%).

Discussion

Value of Three Study Design Components

The GoKinD collection represents a unique opportunity for scientists to use three complementary study designs to uncover the genetic basis of diabetic nephropathy. Although the probands of the trio families could be combined with the singleton subset, this strategy would sacrifice the independence of the trio components as validation sets. Moreover, the benefit of this strategy would be small because the set of singletons already has excellent power for a wide range of genetic models, even loci with small effects. Because of the vulnerability of singleton analysis to spurious findings as a result of population stratification, confirmation of positive findings must be sought in independent data sets.

The set of case trios, which is immune to population stratification effects, was recruited for just this purpose. However, the usefulness of this remedy is limited to situations in which the hypothesized gene effect is relatively large. In situations in which trio analysis does not have adequate power, an alternative is to test and adjust for population stratification (19). Although straightforward in principle, testing for stratification may be more difficult than anticipated. As demonstrated recently, standard methods for detecting it failed in a study involving European Americans (20). Two alternatives that do not rely on tests for population stratification may be considered. One is to match case patients and control subjects for

country of origin of their grandparents. Unfortunately, this information is not available for GoKinD participants. The second alternative provides a more versatile and comprehensive solution. This would entail collection of a panel of DNA that comprises diverse European populations. Then, when a positive association is found with diabetic nephropathy in the singleton case patients and control subjects in GoKinD, the frequency of the risk allele would be examined in the European panel. If the frequency varies little among European populations, then the association is unlikely to be due to stratification in a European American population. Conversely, if the frequency varies widely across European populations, then the association is more plausibly attributed to stratification. Although the set of control trios has the least power to identify nephropathy genes, it plays important roles in other respects. First, it provides protection from false-positive results that arise in case trios from the phenomenon of segregation distortion, the preferential transmission of an allele irrespective of disease phenotype. For example, any allele related to the phenotype of type 1 diabetes will be preferentially transmitted in GoKinD case trios. Because the same will be true in control trios, the locus can be recognized as a type 1 diabetes locus rather than a nephropathy locus (11). The second important role for control trio analysis is the evaluation of models that involve gene–environment interaction, where control trios can outperform case trios (13). This class of genetic models is of particular relevance to diabetic nephropathy, a phenotype that develops only in the presence of a diabetic milieu, particularly with poorer glycemic control (10).

The availability of three complementary designs is a major strength; however, it also presents challenges in interpreting results that are not consistent across all study designs. When all three components are genotyped (3075 samples), one of eight outcomes will occur (Table 5). Patterns 1 and 8 represent clear-cut scenarios in which all components are in agreement. In pattern 2, only control trios yield statistically nonsignificant results, a scenario that is likely to be common given the generally lower power of this component. This pattern's mirror image, pattern 7, in which significance is found only in control trios, is consistent with certain patterns of gene–gene or gene–environment interaction (13). Pattern 3 also is consistent with this possibility. The remaining patterns have less clear interpretations. Although the observation of pattern 4 may be expected because of the higher statistical power of the singleton component, it also is consistent with population stratification, an interpretation that would be strengthened if this pattern were seen across many loci. Pattern 6 or 7 would be expected if segregation distortion exists, but this interpretation is tested easily when case and control trios are considered together (11,21).

Table 5. Possible outcomes of genetic association analysis using the three designs in GoKinD[a]

Pattern	Case/Control Singletons	Case Trios	Control Trios
1	+	+	+
2	+	+	–
3	+	–	+
4	+	–	–
5	–	+	+
6	–	+	–
7	–	–	+
8	–	–	–

[a] +, results are statistically significant; –, results are not statistically significant.

Duration of Diabetes in Case Patients and Control Subjects

Recent theoretical work has demonstrated the importance of considering duration of diabetes when carrying out either singleton or trio analysis (17). In this context, "duration" refers to duration of diabetes at onset of nephropathy for case patients and duration of diabetes at time of enrollment for control subjects. Because the onset of proteinuria often is undocumented, various approximations or surrogates for this information should be tried. Ignoring duration can result in substantial power loss or even findings that paradoxically implicate nonrisk alleles as causative (17). On the basis of simulation studies, the effect of a risk allele most clearly is demonstrable in a comparison of case patients with short diabetes duration with control subjects with long duration. The simplest analytic strategy for addressing the duration issue is subgroup analysis of reasonably defined duration strata. Another option is to use conditional logistic regression with duration as an independent variable (22). In any event, it will be incumbent on investigators to formulate an appropriate analytic model that is based on the hypothesized duration effect (*e.g.*, threshold, linear, quadratic).

High Mortality among Case Patients and Their Parents

Genetic studies of diabetic complications may be vulnerable to survivor effects as a consequence of the very high mortality rates for patients with diabetic nephropathy, especially ESRD. To participate in GoKinD, case patients with ESRD had survived an average of 8 yr of ESRD, and case patients with proteinuria had survived more intense mortality than control subjects (23). Therefore, any genetic factor that is associated with survival will be enriched to some degree in case patients. Moreover, the known clustering of early mortality in parents of patients with type 1 diabetes and nephropathy (24) resulted in only 48% of case patients having two parents available to form a trio as compared with 64% of control subjects having two (Table 3). As a result, the enrichment of a survival factor may be particularly strong in case trios. The likelihood that such mortality effects would result in a spurious association that requires further investigation. However, an investigator should consider this alternative among the interpretations of pattern 6 in Table 5, an association that is significant only in case trios.

Limitations of the GoKinD Collection

GoKinD represents a major collaborative effort that promises to speed the discovery of the genetic basis of diabetic nephropathy. Nevertheless, several important issues remain to be addressed. One is the development of novel analytic approaches

that will bring together all three design components in a systematic manner. The procedure for doing so will depend profoundly on whether population stratification exists in the collection; therefore, a reasonably large effort is warranted to examine the GoKinD samples for this phenomenon. Moreover, our power calculations considered only single locus models. However, an appealing feature of GoKinD is that sample sizes are likely to be adequate for testing many hypotheses related to gene–gene interaction. Gauderman (25) recently outlined sample size requirements for various types of gene–gene interaction models. Four study designs were considered, including case trio and case only analysis, both of which are possible in GoKinD. QUANTO (http://hydra.usc.edu/gxe), the software package that implements these power calculations, subsequently has been extended for unmatched case-control studies that are relevant to the singletons in GoKinD. Similar power calculations for gene–environment interaction models also are possible using QUANTO (26).

Acknowledgments

The GoKinD Coordinating Center was funded by the JDRF, and the CDC was funded by PL 105-33, 106-554, and 107-360 administered by the National Institutes of Health.

The GoKinD collaborators acknowledge the contributions to recruitment of the JDC, the Clinical Centers associated with GWU, and Matthews Media Group without which the study would not have been possible. GoKinD investigators from these centers include Stephen A. Brietzke, Debbie Eichelberger, and Christine Hogue, University of Missouri; David Brillon and Juan Cordero, New York Presbyterian Hospital, Cornell University; George A. Burghen and Pam LeNoue, University of Tennessee; George W. Burke and Eva P. Herrada, University of Miami; Debra Counts and Sherry Johnsonbaugh, University of Maryland Medical System; James Desemone and Manjula Salgam, Albany Medical Center; Steven V. Edelman and Gayle Lorenzi, University of California San Diego; Carla Greenbaum and Daxa Sabhaya, Virginia Mason Research Center; Richard A. Guthrie and Ann Brenner, Mid-America Diabetes Associates, P.A.; Irene Hramiak and Judith Harth, St. Joseph's Health Care, University of Western Ontario; Mark Johnson and Paula McIver, University of North Carolina at Chapel Hill; Lois Jovanovic and Allison Wollitzer, Sansum Medical Research Center; John I. Malone and Jennifer Steinbrueck, University of South Florida; Michael Mauer, Nick Rabe, and Cathy Bagne, University of Minnesota; Michael E. May and Janie Lipps, Vanderbilt University Medical Center; Larry Melton and Jonnie Feller, Baylor University Medical Center; Mark E. Molitch and Daphne Adelman, Northwestern University; Robert E. Ratner and Evelyn Robinson, Med-Star Clinical Research Center; John Rogus, Adam Smiles, James H. Warram, Andrzej S. Krolewski, Amanda Johnson, Andrea Segal, Josh Rubin, Julie Bonner, Katie Georgitis, Kimberly Prudhomme Fader, Kristen Silva, Matt Niemi, Melissa Sugar, Nicole Wilkinson, Sarah Comeaney, Scott Tucker, Susan Orsillo, Tom Reynolds, and Kellie Anderson, Joslin Diabetes Center; William L. Sivitz and Meg Bayless, University of Iowa; John A. Colwell, Denise Wood, Maria Szpiech, and Kathy Bradbury, Medical University of South Carolina; Neil H. White and Lucy Levandoski, Washington University School of Medicine; Bernard Zinman and Annette Barnie, Mount Sinai Hospital, University of Toronto; and Therese B. Gibson, Aspen Systems, Inc.

References

1. Jones CA, Krolewski AS, Rogus JJ, Xue JL, Collins A, Warram JH: Epidemic of end-stage renal disease in people with diabetes in the United States population: Do we know the cause? *Kidney Int* 67: 1684–1691, 2005
2. American Diabetes Association: National diabetes fact sheet. Available: http://www.diabetes.org/diabetes-statistics.jsp. Accessed May 13, 2005
3. Centers for Disease Control and Prevention: National diabetes fact sheet. Available: http://www.cdc.gov/diabetes/pubs/estimates.htm. Accessed May 13, 2005
4. Krolewski AS, Warram JH, Cristlieb AR, Busick EJ, Kahn CR: The changing natural history of nephropathy in type 1 diabetes. *Am J Med* 78: 785–793, 1985
5. Krolewski M, Eggers PW, Warram JH: Magnitude of end-stage renal disease in IDDM: A 35 year follow-up study. *Kidney Int* 50: 2041–2046, 1996
6. National Institute of Diabetes and Digestive and Kidney Diseases: Kidney disease of diabetes. Available: http://kidney.niddk.nih.gov/kudiseases/pubs/kdd/index.htm. Accessed May 13, 2005
7. Seaquist ER, Goetz FC, Rich S, Barbosa J: Familial clustering of diabetic kidney disease. Evidence for genetic susceptibility to diabetic nephropathy. *N Engl J Med* 320: 1161–1165, 1989
8. Borch-Johnsen K, Norgaard K, Hommel E, Mathiesen ER, Jensen JS, Deckert T, Parving HH: Is diabetic nephropathy an inherited complication? *Kidney Int* 41: 719–722, 1992
9. Quinn M, Angelico MC, Warram JH, Krolewski AS: Familial factors determine the development of diabetic nephropathy in patients with IDDM. *Diabetologia* 39: 940–945, 1996
10. The Diabetes Control and Complications Trial Research Group: Clustering of long-term complications in families with diabetes in the diabetes control and complications trial. *Diabetes* 46: 1829–1839, 1997
11. Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52: 506–516, 1993
12. Ewens WJ, Spielman RS: The transmission/disequilibrium test: History, subdivision, and admixture. *Am J Hum Genet* 57: 455–464, 1995
13. Scott L, Rogus JJ: Using unaffected child trios to test for transmission distortion. *Genet Epidemiol* 19: 381–394, 2000
14. Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, Nerup J, Bouzekri N, Cambon-Thomsen A, Ronningen KS, Barnett AH, Bain SC, Todd JA: Susceptibility to the human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat Genet* 9: 284–292, 1995
15. Haggard E: *Intraclass Correlation and the Analysis of Variance*, New York, Dryden Press, 1958
16. Knapp M: A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet* 64: 1177–1185, 1999
17. Rogus JJ, Warram JH, Krolewski AS: Genetic studies of late diabetic complications: The overlooked importance of diabetes duration before complication onset. *Diabetes* 51: 1655–1662, 2002
18. Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes

MW, Hogg RJ, Perrone RD, Lau J, Eknoyan G: National Kidney Foundation practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. *Ann Intern Med* 139: 137–147, 2003
19. Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65: 220–228, 1999
20. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN: Demonstrating stratification in a European American population. *Nat Genet* 37: 868–872, 2005
21. Deng HW, Chen WM: The power of the transmission disequilibrium test (TDT) with both case-parent and control-parent trios. *Genet Res* 78: 289–302, 2001
22. Mokliatchouk O, Blacker D, Rabinowitz D: Association tests for traits with variable age at onset. *Hum Hered* 51: 46–53, 2001
23. Warram JH, Laffel LM, Ganda OP, Christlieb AR: Coronary artery disease is the major determinant of excess mortality in patients with insulin-dependent diabetes mellitus and persistent proteinuria. *J Am Soc Nephrol* 3[Suppl]: S104–S110, 1992
24. Tarnow L, Cambien F, Rossing P, Nielsen FS, Hansen BV, Lecerf L, Poirier O, Danilov S, Boelskifte S, Borch-Johnsen K: Insertion/deletion polymorphism in the angiotensin-1-converting enzyme gene is associated with coronary artery disease in IDDM patients with nephropathy. *Diabetologia* 38: 798–803, 1995
25. Gauderman WJ: Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 155: 478–484, 2002
26. Gauderman WJ: Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med* 21: 35–50, 2002

Supplemental information for this article is available online at www.jasn.org.

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;
```