

Using SAS to track recruitment in a multi-center study of diabetes

Stephan A. Villavicencio, The George Washington University, Biostatistics Center, Rockville, MD

Therese B. Gibson, The Lockheed Martin Company, Rockville, MD

Patricia A. Cleary, The George Washington University, Biostatistics Center, Rockville, MD

ABSTRACT

This paper describes how SAS® (Version 8.1) was used to track recruitment of potential participants in the Genetics of Kidneys in Diabetes (GoKinD) Study, a multi-center study of the genetic susceptibility to kidney disease among type 1 diabetics. Data came to the data coordinating center (COC) from a recruitment call center (RRC), 21 clinics, and a central laboratory. To create a central database, staff at the COC merged data from various sources, checked data for how they matched between different sources, and used SAS to determine eligibility status from participants' characteristics. Calculations on age at diabetes diagnosis, length of diabetes duration, and a measure of kidney function were done at the COC. In addition, the array statement was used to calculate multiple kidney-function measurements for each participant. Finally, this paper outlines the reasons why over half of the potential participants were not in the final collection.

Keywords: Cross-sectional studies, case-control studies, family studies, SAS, eligibility, exclusionary criteria, subject attrition, type 1 diabetes, diabetic nephropathy.

INTRODUCTION

Under the leadership of The Biostatistics Center of George Washington University (GWU), the Genetics of Kidneys in Diabetes (GoKinD) Study recruited 1403 participants at 21 clinics (GWU Sites), including 277 case singletons, 346 control singletons, 91 case trio families, and 169 control trio families. GoKinD is a Juvenile Diabetes Research Foundation International (JDRF) initiative aimed at the identification of genes involved in diabetic nephropathy (kidney disease caused by diabetes). A large number of individuals with type 1 diabetes (T1D) 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 were also enrolled to form family trios. As of April 2006, GoKinD included 3065 participants, comprising 664 case singletons, 622 control singletons, 271 case trios, and 322 control trios (see Table 1). Extra participants were recruited through the Joslin Diabetes Center in Boston. GoKinD Sites included Joslin Diabetes Center, while GWU Sites consisted of the other 21 clinics.

DEFINITIONS

A *proband* was a type 1 diabetic between the ages of 18 and 59 who fit either case or control criterion. A proband was considered a *singleton* if it was known that at least one parent was deceased, unable, or unwilling to participate. A *trio* is a group consisting of an adult proband with both parents. Trios were sought after in the recruitment effort, because trios allow geneticists to study the inheritance of genes within a family and compare the alleles found within a case family to the alleles found within a control family. This comparison will be used to determine which alleles are protective against diabetic nephropathy (DN) and which alleles are associated with susceptibility toward DN.

TRIO was also the name of the variable used to track the status of probands and their parents (if available). If a proband had parents who were alive, then those individuals were considered *incomplete trios* until the end of the study. Once a proband and his/her parents completed required data and samples, then that group was considered a *complete trio* (as opposed to an *incomplete trio*).

DN is kidney disease caused by the long-term high blood sugar (hyperglycemic) condition of diabetes. It is manifested as either proteinuria or End Stage Renal Disease (ESRD), which includes patients on dialysis or kidney transplant recipients. Participants with ESRD were not required to submit a urine screen for study eligibility; they could be admitted into the study as cases.

Proteinuria is measured by the Albumin Creatinine Ratio (ACR, ugs of albumin per mg creatinine). An ACR greater than 300 ug/mg is proteinuric. However, ACR levels of 20 ug/mg or less are *normoalbuminuric* (within normal ranges of healthy kidney function). Between normoalbuminuria and proteinuria is a pre-clinical condition, known as *microalbuminuria*. People who had microalbuminuria were not included in the final GoKinD collection.

To prevent hyperglycemia in type 1 diabetics, insulin is administered at every meal. Insulin is used as the main treatment for type 1 diabetes. As part of the exclusionary criteria, all participants must have started insulin within a year of diabetes diagnosis.

Case participants (ELIG=1) were selected for having biochemical evidence of DN after having type 1 diabetes (T1D) for at least ten years, while *control* participants (ELIG=0) were selected for having evidence of kidney function within normal limits after having T1D for a minimum of fifteen years. In addition, control participants did not take ACE-inhibitors or other anti-hypertensives, as these medications help protect kidneys from the damage of hyperglycemia (high blood sugar). Those participants who underwent urine screens but were not categorized into case or control groups were placed into a *microalbuminuric* group (ELIG=2), which was not

included into the final collection. In addition, there were categories for *ineligible* (ELIG=3) and *inactive* (ELIG=6) participants. Moreover, during the study, participants were categorized into temporary groups, based on whether they had started the series of urine screens. Those who did not do one urine screen were considered *potential* (ELIG=4) while those who did start but did not yet finish were considered *pending* participants (ELIG=5).

STUDY RECRUITMENT

To reach out to the type 1 diabetic population of the USA and Canada, the GoKinD recruitment call center (RRC) operated a toll-free call operation to receive callers who learned about the GoKinD Study through community outreach. The call center screened for some basic inclusion criteria: age, having type 1 diabetes, duration of diabetes, use of certain medications, the ability to give consent, having kidney disease not caused by T1D, status for HIV or active tuberculosis, and not being adopted, pregnant, or nursing. Potential participants meeting basic inclusion criteria were then referred to a nearby or remote-collection GoKinD Site to give informed consent and begin data and sample collection. In addition, participants were asked to refer their parents to the call center so that their parents could also join the study.

Failure to participate was a problem for recruitment from the start. Initially, some callers decided not to participate, or the call-center determined that the caller did not meet basic inclusion criteria. Then, at the GoKinD Site, the participant was interviewed by a study coordinator, who also assisted the participant in study activities for data and sample collection. At the GoKinD Site, more participants were found to be ineligible, typically resulting from use of specific anti-hypertensive medications. Further more, since the study required at least two urine samples to further ascertain proteinuria status for some participants, some participants did not return to give their last urine sample. These participants were lost to follow-up (LTFU). In addition, some participants decided in the middle of the study that they were no longer interested in participating. Study coordinators then completed "Notification of Refusal" forms and these forms documented their reason for refusal. The result of "Reasons for non-participation" for the study can be found on Table 2.

DATA COLLECTION

To keep track of the trio status, eligibility status, and the number of urine screens done for each participant, a central data management system was established at the George Washington University (GWU) Biostatistics Center. This system used SAS® Basic Statistical Packages, Version 8.1. SAS not only has the ability to analyze data, but it also can store data and allow data to be viewed in many different interfaces. This paper describes how SAS (Version 8.1) was used to track data on study participants in a multi-center study of the kidney complications of type 1 diabetes. Specifically, SAS was used to determine eligibility status based on a few variables.

Data came to GWU from the RRC, 22 clinics (GoKinD Sites), and the Central Biochemistry Laboratory (CBL: GoKinD's designated laboratory for biochemical analysis, located at the University of Minnesota). Data were sent to the GoKinD Coordinating Center (COC) and SAS programming was done by staff at the COC to organize the data management system.

Data was organized into libraries, based on where the data came from (RRC, CBL, or GWU) and the specific use of the datasets. Datasets were used: to identify participants, to answer to medical history questionnaires, and to determine ELIG (eligibility) and TRIO (family trio) status. Data collected for the GoKinD Study included participant identification numbers (PINs), which facilitated the de-identification of participants for the purpose of privacy protection. In addition, data included: participant demographic information (date of birth, DOB; age at baseline exam; original and current GoKinD Site of participation), history of diabetes (age at diabetes onset, length of time with diabetes), history of kidney disease (proteinuria, dialysis, or transplant), medication records (use of antihypertensive or ACE-Inhibitor), urine tests (ACR), information on exclusionary criteria (being adopted, nursing, or pregnant; having psychiatric disease, HIV, or tuberculosis), eligibility status (ineligible, potential, pending, inactive, microalbuminuric, case, or control), and trio status (singleton, trio, or incomplete trio).

Data was collected and updated in real-time throughout the duration of the study. Each participant underwent several phases of recruitment and study activities. These phases included: outreach, recruitment, screening, collecting data and biological samples. Each participant underwent these phases at his/her own pace. Thus, participants were at different phases and moving at different rates through these phases relative to each other throughout the duration of the study. Keeping track of the different rates of progress of each participant necessitated an effective data management system. To meet this need, SAS was used. With the help of SAS, GWU was able to combine all sources of data to determine if a participant had completed collection of blood, urine samples, and data.

INELIGIBILITY AND REFUSAL

Reasons for *ineligibility* were collected through the SAS program that determined eligibility. In addition, reasons for *refusal-to-participate* were captured on questionnaires with multiple choice and fill-in answers. Reasons for ineligibility and refusal were combined to create "*reasons for non-participation*" (see Table 2). This was done by taking two variables on reasons for refusal and making a third variable with a new format that combined the formats from the first two variables. The collection of this information was intended to guide multi-center recruitment strategies in subsequent studies on type 1 diabetics or studies with family trios.

DATA CHECKS

As it was essential to keep up-to-date records on eligibility status on probands to know which parents were still eligible to form trios, data on eligibility status had to be continually matched between the RRC and GWU. If eligibility data did not match between these entities, it was essential to find out for which participants there was a mismatch. But, before the variables could be compared, the variables needed to be compatible for comparison.

At first, RRC's and GWU's variables were not compatible. RRC's variable, R_RES50C, was a 50-character, fill-in variable, while GWU's variable, ELIG, was a single-digit numeric-value variable. While the ELIG variable can easily be converted into a character variable, the answers to R_RES50C must be summarized into a limited number of values. However, the values of R_RES50C could be summarized by: (1) adopting a new variable and (2) programming this new variable to have a limited number of values. To brainstorm how to summarize the answers for R_RES50C, it was essential to browse these answers within the dataset. One can browse the dataset in tabular form within SAS/ASSIST® software. However, running PROC EXPORT was a better option, as it was able to create a Microsoft® Office Excel Spreadsheet of the dataset of interest (see below). This spreadsheet could be edited and sorted in alphabetical order, and then be imported back into SAS by running PROC IMPORT (not shown).

```
PROC EXPORT DATA=OTHELIG
```

```
OUTFILE="C:\sav\sesug2007\xls\OTHELIG.xls"
```

```
DBMS=EXCEL2000 REPLACE; RUN;
```

R_RES50C	A_STATRR	LABEL OF A_STATRR	X_ID_PAT (PIN)	ELIG	LABEL OF ELIG	A_STATGW	LABEL OF A_STATGW
SCREEN FAILED	F	SCREEN FAILED	5184082703	0	CONTROL	A	ACTIVE
DECLINED	D	DECLINED	5162800403	1	CASE	A	ACTIVE
ENROLLED	A	ACTIVE	5172800903	3	INELIGIBLE	I	INELIGIBLE
ENROLLED	A	ACTIVE	5114109403	6	INACTIVE	J	INACTIVE
ENROLLED	A	ACTIVE	5134086503	0	CONTROL	A	ACTIVE
ENROLLED	A	ACTIVE	5192751903	1	CASE	A	ACTIVE
SCREEN FAILED	F	SCREEN FAILED	5124075403	3	INELIGIBLE	I	INELIGIBLE
SCREEN FAILED	F	SCREEN FAILED	5144104003	6	INACTIVE	J	INACTIVE
DECLINED	D	DECLINED	5152550503	6	INACTIVE	J	INACTIVE

As it turned out, the best way to summarize R_RES50C was to take the first eight characters of R_RES50C and to use the SUBSTR function. Using this function, the first eight characters of R_RES50C were read and then summarized to a few values under one variable, called A_STATRR. The values for A_STATRR included: 'A' for active, 'D' for declined, 'F' for screen failed, and 'U' for unknown (see below).

```
IF SUBSTR(R_RES50C,1,8)="ENROLLED" THEN A_STATRR = "A";
ELSE IF SUBSTR(R_RES50C,1,8)="DECLINED" THEN A_STATRR = "D";
ELSE IF SUBSTR(R_RES50C,1,8)="SCREEN F" THEN A_STATRR = "F";
ELSE
A_STATRR = "U";
```

At this point, there was another problem that needed a solution: the ELIG variable was a numeric variable, not a character variable like A_STATRR. This was resolved by creating a new variable A_STATGW and setting its character values based on the numeric values of ELIG. The values for A_STATGW included: 'A' for active, 'I' for ineligible, 'J' for inactive, 'P' for pending, and 'Q' for potential (see below).

```
IF ELIG NE . THEN DO;
IF ELIG = 0 THEN A_STATGW = "A";
ELSE IF ELIG = 1 THEN A_STATGW = "A";
```

```

ELSE IF ELIG = 2 THEN A_STATGW = "A";
ELSE IF ELIG = 3 THEN A_STATGW = "I";
ELSE IF ELIG = 6 THEN A_STATGW = "J";
ELSE IF ELIG = 4 THEN A_STATGW = "P";
ELSE IF ELIG = 5 THEN A_STATGW = "Q";  END;

```

Now that there were two variables with 8 different values between the variables, these variables could be compared. Another variable was created, AGR_STAT, to make this comparison. (This name stands for the agreement of active status between GWU and RRC). To create this variable, it was determined that both A_STATGW and A_STATRR were not empty:

```

IF A_STATGW EQ " " OR A_STATRR EQ " " THEN AGR_STAT=0;
ELSE IF A_STATGW NE " " AND A_STATRR NE " " THEN DO;

```

Next, the agreement between A_STATGW and A_STATRR was tested through the new variable, AGR_STAT. This variable was programmed to have two values: 0=no agreement between RRC and GWU, and 1=agreement (see below).

```

IF A_STATGW = "A" THEN DO;
    IF A_STATRR EQ "A" THEN AGR_STAT=1;
    ELSE AGR_STAT=0;  END;
ELSE IF A_STATGW IN ("I", "J") THEN DO;
    IF A_STATRR EQ "D" THEN AGR_STAT=1;
    ELSE IF A_STATRR EQ "F" THEN AGR_STAT=1;
    ELSE AGR_STAT=0;  END;
ELSE IF A_STATGW IN ("P", "Q") THEN DO;
    IF A_STATRR EQ "U" THEN AGR_STAT=1;
    ELSE AGR_STAT=0;  END; END;

```

A spreadsheet, created by PROC EXPORT, showing the new variable (AGR_STAT) between the variables it had compared (A_STATRR and A_STATGU) is shown below:

R_RES50C	A_STATRR	LABEL OF A_STATRR	AGR_STAT	A_STATGW	LABEL OF A_STATGW	ELIG	LABEL OF ELIG	X_ID_PAT (PIN)
SCREEN FAILED	F	SCREEN FAILED	0	A	ACTIVE	0	CONTROL	5184082703
DECLINED	D	DECLINED	0	A	ACTIVE	1	CASE	5162800403
ENROLLED	A	ACTIVE	0	I	INELIGIBLE	3	INELIGIBLE	5172800903
ENROLLED	A	ACTIVE	0	J	INACTIVE	6	INACTIVE	5114109403
ENROLLED	A	ACTIVE	1	A	ACTIVE	0	CONTROL	5134086503
ENROLLED	A	ACTIVE	1	A	ACTIVE	1	CASE	5192751903
SCREEN FAILED	F	SCREEN FAILED	1	I	INELIGIBLE	3	INELIGIBLE	5124075403
SCREEN FAILED	F	SCREEN FAILED	1	J	INACTIVE	6	INACTIVE	5144104003
DECLINED	D	DECLINED	1	J	INACTIVE	6	INACTIVE	5152550503

CALCULATING ACR WITH ARRAY STATEMENTS

One of the most important tasks done by SAS was to process data to determine participant eligibility. Eligibility determination involved several parameters, which included: kidney function, type of diabetes, diabetes duration, status and time of taking insulin. To determine kidney function, those undergoing weekly dialysis or who had a transplant of kidney or pancreas were defined as patients with End Stage Renal Disease (ESRD) and were eligible to be cases without needing to undergo a urine screen. However, those who did not have ESRD conducted up to 3 urine screens at the GoKinD Site to verify their ACR level. Thus, there were 3 different variables for ACR, taken on three different occasions. ACR was not directly derived from the urine tests. Each urine test resulted in two values: (1) urinary albumin and (2) urinary creatinine. These values were used to calculate ACR: ACR equaled Albumin (ALBUMN#) multiplied by 100 and divided by Creatinine (URNCRT#). However, it was humanly impossible to do the same ACR calculation on the same person three times, and to do it for a large fraction of all potential participants. This is where SAS was most helpful. An array statement was created in SAS to run the ACR calculation up to three times for all participants without ESRD. An array statement is a calculation that is done to many variables and many values. This can be done to any group of variables, as long as the variables are either all numeric or all character (Delwiche, L. & Slaughter, S., 2003).

In the GoKinD Study, the ALB array read the albumin measurements on three different occasions (ALBUMN1, ALBUMN2, and ALBUMN3). Then the URN array read the Urinary Creatinine Measurements on three different occasions (URNCRT1, URNCRT2, URNCRT3). Finally, the ACRS array read the result of calculations on the ALB and URN, matched for the same date of sample testing (see below).

```
ARRAY ALB(*) ALBUMN1-ALBUMN3;
ARRAY URN(*) URNCRT1-URNCR3;
ARRAY ACRS(*) ACR1-ACR3;
DO I=1 TO 3;
  IF ALB(I)>0 AND URN(I)>0 THEN ACRS(I)=ROUND((ALB(I)*100/URN(I)),.1);
  ELSE ACRS(I)=.; END;
DROP ALBUMN1-ALBUMN3 URNCRT1-URNCR3; RUN;
```

The result was three ACR values per participant. The next step was to program SAS to take the results of these three ACR values and to make a determination of eligibility for each participant.

ELIGIBILITY DETERMINATION

Now that ACR values were calculated, the next step was to determine if the participant has type 1 diabetes (T1DIAB=1). This was done by verifying inclusion criteria: (a) age at diabetes diagnosis was under 31 ($0 < \text{DIAGAGE} < 31$), (b) participant took insulin regularly ($\text{BOINSUL1}=2$), and (c) participant began insulin therapy within a year of diagnosis ($0 < \text{INSLAG} < 1$). However, if participant did not take insulin ($\text{BOINSUL1}=1$), then if he/she was diagnosed with diabetes before age 31 and was either on dialysis ($\text{BODIAL}=2$) or had a transplant of kidney ($\text{BOTRANS}=2$) or pancreas ($\text{BOPANCR}=2$), then the participant was also considered a type 1 diabetic (see below).

```
IF N(DIAGAGE,BOINSUL1,INSLAG)<3 THEN T1DIAB=99; /*99=PENDING INFO*/
ELSE IF 0<=DIAGAGE<31 AND (BOINSUL1=2) OR
(BOINSUL1=1 AND BODIAL=2 OR BOTRANS=2 OR BOPANCR=2))
AND 0<=INSLAG<=1 THEN T1DIAB=1; ELSE T1DIAB=0;
```

Once it was determined that the patient had T1D, the next step was to screen for age eligibility. If the participant was under 18 or over 59, then he/she was ineligible ($\text{ELIG}=3$) to participate as a proband. If a participant was pending a sample collection, then he/she was pending ($\text{ELIG}=5$) (see below).

```
IF 0<AGE<18 OR AGE>59 THEN DO;
  ELIG=3; INELIG=1; END;
ELSE IF AGE=. THEN ELIG=5;
ELSE DO;
```

```
IF T1DIAB=99 THEN ELIG=5;
```

USING ACRs TO DETERMINE STUDY GROUP

Once it was determined that the proband had T1D, the next step was to determine if the diabetes duration made the participant eligible to be a case or control participant. This eligibility was summarized in new variables called CASEDUR (case duration) and CNTLDUR (control duration) (see below).

```
ELSE IF T1DIAB=1 THEN DO;  
  IF DIABDUR=. THEN ELIG=5;  
  ELSE IF DIABDUR<10 THEN DO;  
    ELIG=3; INELIG=3; END;  
  ELSE IF DIABDUR>=10 THEN DO;
```

Once diabetes duration was captured in the DIABDUR variable, it was used to give value to the CASEDUR and CNTLDUR variables (see below). If DIABDUR was not equal or over 10 years, then CNTLDUR=0 (participant was not eligible).

```
  IF DIABDUR>=10 THEN CASEDUR=1;  
  IF DIABDUR>=15 THEN CNTLDUR=1;  
  ELSE CNTLDUR=0;
```

Next, each participant's ACR values was measured against the criteria for case or control groups. If the participant's ACR values did not meet either criterion, then the participant was declared microalbuminuric.

Historic ACR values (CASEHACR for cases and CNTLHACR for controls) could also be counted toward eligibility determination. However, if one ACR value was pending, then the participant's status could not be immediately determined. Thus, the participant was considered to be pending. Those participants who were pending ACR measurements were categorized as pending (99) using the six lines below:

```
IF N(ACR1,ACR2,ACR3)=0 OR (ACR1 NE . AND N(ACR2,ACR3)=0)  
  OR (HISTACR NE 1 AND N(OF ACR1-ACR3)<2)  
  THEN DO;  
  CASEACR=99;  
  CNTLACR=99;  
  MICROACR=99; END;
```

However, those who had a historic ACR that placed them in the control group had another category from other pending participants. This was CNTLACR=98.

```
  IF (CNTLHACR=1 AND N(ACR1,ACR2,ACR3)=0)  
  OR (CNTLHACR NE 1 AND N(ACR1,ACR2,ACR3)=1 AND  
    MAX(ACR1,ACR2,ACR3)<40)  
    THEN CNTLACR=98;      /***CONTROL PENDING***/
```

The formula below was used to determine if a participant was eligible to participate as a case. To be a case, two out of three ACR values had to be over 300 (CASEACR=1).

```
ELSE DO;  
  IF (ACR1>300 AND ACR2>300) OR (ACR1>300 AND ACR3>300) OR  
    (ACR2>300 AND ACR3>300) THEN CASEACR=1;
```

```

ELSE IF CASEHACR=1 AND ACR1=. AND (ACR2>300 OR ACR3>300)
  THEN CASEACR=1;
ELSE IF CASEHACR=1 AND ACR1=. AND 0<ACR2<=300 AND ACR3=.
  THEN CASEACR=99; /****PENDING****/
ELSE IF (CASEHACR NE 1)
  AND N(ACR1,ACR2,ACR3)=2 AND
  MAX(ACR1,ACR2,ACR3)>300 AND MIN(ACR1,ACR2,ACR3)<=300
  THEN CASEACR=99; /****PENDING****/

```

The formula below was used to determine if the participant was eligible to participate as a control. Control participants had to have two ACR values under 20 with a third ACR value that was under 40. Once it was determined (after three ACRs) that the participant did not qualify for either the case or control group, the participant was considered microalbuminuric.

```

IF (0<=ACR1<20 AND 0<=ACR2<20 AND ACR3=.) OR
  (0<=ACR1<20 AND 0<=ACR3<20 AND ACR2=.) OR
  (0<=ACR2<20 AND 0<=ACR3<20 AND ACR1=.) OR
  (0<=ACR1<20 AND 0<=ACR2<20 AND ACR3=.)
  THEN CNTLACR=1;
ELSE IF (0<=ACR1<20 AND 0<=ACR3<20 AND 0<=ACR2<40) THEN CNTLACR=1;
ELSE IF (0<=ACR2<20 AND 0<=ACR3<20 AND 0<=ACR1<40) THEN CNTLACR=1;
ELSE IF (0<=ACR1<20 AND 0<=ACR2<20 AND 0<=ACR3<40) THEN CNTLACR=1;
ELSE IF CNTLHACR=1 AND 0<=ACR2<20 THEN CNTLACR=1;
ELSE IF CNTLHACR=1 AND 0<=ACR2<40 AND 0<=ACR3<20 THEN CNTLACR=1;
ELSE IF (CNTLHACR IN (.,0) AND N(ACR1,ACR2,ACR3)=2 AND
  MAX(ACR1,ACR2,ACR3)<40 AND MIN(ACR1,ACR2,ACR3)<20)
  OR (CNTLACR=1 AND 0<ACR2<40 AND ACR3=.)
  THEN CNTLACR=98; /***** CONTROL PENDING *****/
IF CASEACR NOT IN (1, 99) AND CNTLACR NOT IN (1,98,99) THEN
  MICROACR=1; END;

```

To determine control eligibility, participant had to have 10 years of diabetes (CASEDUR=1) and have ESRD or have proteinuria (CASEACR=1).

```

IF CASEDUR=1 AND (ESRD=1 OR CASEACR=1) THEN ELIG=1;

```

CATEGORIZATION OF MICROALBUMINURICS AND INELIGIBLES

If a participant was not considered a case or control, he/she would have to be categorized as a microalbuminuric or an ineligible participant. If the participant had 15 years of diabetes (CNTLDUR=1) and not take an ACE inhibitor or an anti-hypertensive (CNTLDRG=0), had an ACR value less than 20 (CNTLACR=1), and did not have ESRD (ESRD not in (1,99)), then he/she was considered a microalbuminuric participant (ELIG=2).

```

ELSE IF CNTLDUR=1 AND CNTLDRG=0 AND CNTLACR=1 AND ESRD NOT IN (1,99)
  THEN ELIG=2;

```

Otherwise, the participant could be ineligible (ELIG=3). This status was determined if the participant had 15 years of diabetes (CNTLDUR=1) and not take an ACE inhibitor or an anti-hypertensive (CNTLDRG=0), and did not have ESRD. In addition, a participant could be considered ineligible if he/she had 10 years of diabetes (CASEDUR=1), did not have ESRD (ESRD NOT IN (1,99)), had an ACR in the microalbuminuric range (MICROACR=1), and had unknown ACE-inhibitor or antihypertensive status (CNTLDRG NE 99).

```
ELSE IF CNTLDUR=1 AND CNTLDRG=0 AND ESRD NOT IN (1,99)
AND MICROACR=1 THEN ELIG=3;
ELSE IF CASEDUR=1 AND ESRD NOT IN (1,99) AND MICROACR=1
AND CNTLDRG NE 99 THEN ELIG=3;
```

Moreover, if someone had an ACR above the normoalbuminuric (CNTLACR in (1,98)) and did take ACE inhibitor or anti-hypertensive (INELIG=4), then he/she was categorized as an ineligible participant (ELIG=3) (see below).

```
ELSE IF CNTLACR IN (1,98) AND CNTLDRG=1 AND ESRD NOT IN (1,99)
THEN DO;
    ELIG=3; INELIG=4; END;
```

If it was not known if a participant had ESRD (ESRD=99) or it was not known if a participant was in the case (CASEACR=99), control groups (CNTLACR IN (98,99)), or it was not known if the participant took an anti-hypertensive (CNTLDRG IN (.,99)), then the participant was considered a pending participant (ELIG=5). This means that the participant had given at least one urine sample but so far there was no indication to which group the participant might end up in. Otherwise, if a participant did not give his/her first urine sample, then he/she was a potential participant (ELIG=4) (see below).

```
ELSE IF ESRD=99 OR CASEACR=99 OR CNTLACR IN (98,99) OR
CNTLDRG IN (.,99)
THEN ELIG=5;
ELSE DO; ELIG=4;
    IF T1DIAB=1 AND CNTLACR=1 AND CNTLDRG=0 AND DIABDUR<15
    THEN INELIG=5; ELSE INELIG=8;
```

Above, the reason for ineligibility was that the diabetes duration for a control was under 15 years (INELIG=5). If there was another reason for ineligibility, then INELIG=8.

The values for INELIG will help determine the reasons why participants were not eligible for the GoKinD Study.

REASONS FOR NON-PARTICIPATION

Finally, there were 1498 (51.4%) potential participants who did not make it into the final collection (Table 2). The reasons for non-participation included: 253 (8.9%) callers who were not interested from the start, 460 (15.8%) participants who met exclusionary criteria, and 585 (20.1%) participants who either lost interest during the middle of the study or were lost to follow-up (LTFU). In addition, there were 386 (42.1%) potential parents who withdrew from the study. Among this group of parents, 68 (17.6%) could not participate because of the absence of another family member.

CONCLUSION

SAS® Statistical Packages (Version 8.1) provided effective statistical software to organize data for a multi-center study of diabetes. Data came to the GoKinD Coordinating Center (COC) from the recruitment call center (RRC), 22 GoKinD Sites, and the Central Biochemistry Laboratory (CBL). An organized central database was created from data from these different sources. Staff at the COC used SAS to check data for matching between RRC and GWU and to determine eligibility for participants, based on participant characteristics and exclusionary criteria. Array statements were used with mathematical calculations in determining which study group a participant belonged to.

There were 1498 (51.4%) potential participants who did not make it into the final collection. The reasons for non-participation included: 253 (8.9%) callers who were initially not interested, 460 (15.8%) participants who were ineligible, and 585 (20.1%) participants who either lost interest during the study or were lost to follow-up.

REFERENCES

1. Delwiche, L. & Slaughter, S. (2003). *The Little SAS Book* (3rd ed.). SAS Institute Inc.: Cary, NC.
2. Mueller, P., Rogus, J., Cleary, P., Zhao, Y., Smiles, A., Steffes, M., Bucksa, J., Gibson, T., Cordovado, S., Krolewski, A., Nierras, C., Warram, J. (2006). *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 Nephrology, 17(7):1782-90.
3. SAS Institute Inc. (1990). *SAS Language, Version 6* (1st ed.). SAS Institute Inc.: Cary, NC.
4. SAS Institute Inc. (2007). *SAS OnlineDoc® 9.1.3*. SAS Institute Inc.: Cary, NC. Retrieved July 5, 2007 from <http://support.sas.com/techsup/faq/>.

ACKNOWLEDGMENTS

We thank the GoKinD Study Group, especially Karen Anderson, for her work on GoKinD data. We would also like to thank Venita DePuy Bowden and Joe Kelley for their guidance in preparing the paper for the SESUG 2007 Conference. Finally, we thank our family and friends for their support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors through:

Stephan Villavicencio
6110 Executive Blvd., Ste. 750
Rockville, MD 20852
301 881-9260
stephanv@biostat.bsc.gwu.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Table 1. Final Result of Participant Recruitment in the GoKinD Study: Trios and Singletons in Case/Control Categories

		Case			Control		
Characteristic		Trio Proband	Singleton	Total Case	Trio Proband	Singleton	Total Control
N=		271	664	935	322	624	946
Joslin Diabetes Center		180 66.42%	318 47.89%	498 53.26%	153 47.52%	345 55.29%	498 52.64%
GWU Sites		91 33.58%	346 52.11%	437 46.74%	169 52.48%	279 44.71%	448 47.36%

Table 2: Reasons for non-participation at GWU Sites

Reasons For Probands And Parents Not Participating In Study	Number Of Potential Participants Not Participating	Percent Not Participating*
Person was not interested in participating from the start	253	8.69%
Ineligible for study	460	15.79%
'Ineligible' Because Taking Anti-Hypertensive	169	5.80%
Microalbuminuria	71	2.44%
Reasons For 'Ineligibility' Other Than 'Taking Anti-Hypertensive'	220	7.55%
Diabetes Duration Too Short	27	0.93%
Outside Of Age Criteria	17	0.58%
Diagnosis Late Or Not Type 1 Diabetic	14	0.48%
Has Psychiatric Disorder, HIV+, Or TB Positive Test	7	0.24%
Nursing Or Pregnant	6	0.21%
Other Kidney Disease	3	0.10%
Caller Is Adopted	2	0.07%
Ineligible For Other Reasons	144	4.94%
Lost interest in the middle of the study or were lost to follow-up (LTFU)	673	23.10%
Proband is not interested	215	7.38%
Relative is not interested	90	3.09%
Parent is willing to participate, but other family member is not interested	68	2.33%
Lost to follow-up	300	10.30%
Miscellaneous reasons	112	3.84%
Problems Completing Urine Screen And Blood Collection	44	1.51%
Participant Is Too Elderly Or Ill For Study	16	0.55%
Transferred To Joslin	20	0.69%
Administrative Problem	14	0.48%
Participant Is Concerned With Confidentiality And The Potential For Misuse Of DNA And Exam Results	14	0.48%
No reason reported	4	0.14%
Total	1498	51.42%

* "Percent Not Participating" - as a proportion of all 2913 potential participants.