10. GOKIND DATA AND INFORMATION TRANSFER

10.1 OVERVIEW OF GOKIND DATA COLLECTION

The GoKinD data are centrally processed and stored as SAS datasets on a secure server at the GoKinD Coordinating Center (COC) at George Washington University Biostatistics Center. Electronic data feeds of administrative, recruitment, phenotypic and genotypic data occur in a variety of text and Microsoft formats, which are then converted to SAS at the COC. See Figure 10.1 for a depiction of the GoKinD electronic data collection. In the following section we will describe the data source and attributes for the GoKinD data collection management system.

GW clinic participant medical and demographic information are keyed from copies of multipart paper forms at the COC in a Visual FoxPro application. Integration of the data from MMG, Joslin, GW clinics, CBL, and CDC is performed in SAS. Data quality evaluation, reconciliation and cleaning are ongoing processes. Feedback reports of biochemical results for participants and eligibility status reports for the clinics are generated at the COC in SAS on a weekly basis. The GoKinD quality control analysis for biochemical and genetic results was performed on replicate samples taken from 5% of the participants.

Copies of the multipart mailing forms which accompany blood and urine samples shipped from clinics to CBL are also keyed at the COC. The mailing forms are the first part of the specimen tracking system for GoKinD which permits real-time monitoring of movement of individual specimens by unique barcode identifiers as they are received by CBL, then boxed and shipped at CBL and received by CDC and CASPIR. Specimen tracking is done in a web-based application at the COC in SAS and SQL for CBL and CDC and then for monitoring shipment of both renewable (DNA) and nonrewable resources (saved plasma, serum, and urine samples) to approved external researchers.

Figure 10.2 is a spreadsheet that presents the overview of the data collection for GoKinD. For each source the figure shows the type of data that originates from the source, the format, how the data is transferred, the frequency of the transfer, and information regarding edits, backup, and security.

10.2 MATTHEWS MEDIA GROUP

The GoKinD study call center staff at Matthews Media Group has created a database consisting of the eligibility screening questions that comprise the JDRF patient screening questionnaire. The information is managed at MMG in SQL Server 2000 and is transferred by secure FTP to the COC as Microsoft Access databases on a weekly basis during the recruitment phase of GoKinD (both an incremental file and a full refresh database). Reconciliation of recruitment status of family members occurs monthly by way of reports from MMG back to the GW Clinic Study Coordinators. Periodic reconciliation of key information collected by phone in the JDRF questionnaire at MMG and the medical and demographic data collected on multipart forms for GW clinic participants and keyed at the COC occurs as screening and recruitment process progresses for a family. MMG plays a key role in working with the GW Clinic Study Coordinators in completing the GK207 form (Notification of Transfer, Remote Site Collection, or Refusal to Participate).

10.3 CLINICS

The GW clinics (excluding Joslin) use NCR multipart data forms to collect medical and demographic and mail one of the copies of the forms weekly to the COC, where the data are entered in a Visual FoxPro application and converted to SAS datasets. The GK203 form (Notification of Death) is used by all clinics, including Joslin. All of the clinics use NCR multipart forms to record the administrative data necessary for monitoring biological sample collection and mailing to the CBL located at the Fairview University Medical Center. Biolocal sample collection and shipment includes DNA, lymphocytes, whole blood, plasma, serum, and urine, as well as replicate quality control samples collected using a masked patient id and separate set of accession numbers from the primary samples.

10.4 CBL

The CBL will analyze the specimens for basic biochemistries and renal evaluations (blood hemoglobin A1c, serum lipid total cholesterol, serum lipid HDL, serum creatinine, serum cystatin, and urinary albumin and creatinine). The biochemical results will be transmitted weekly to GWU by way of a fixed text ASCII file using a secure file transfer protocol. The CBL text file consists of one to several records per lab result as an incremental file of all records changed or added since the previous file transfer to the COC. The results field on the text record is an 80-character text field, which normally contains the numeric results of the assay, but can contain free text comments regarding sample issues. A new field is created in SAS at the COC that stores a numeric lab result and converts the text values to SAS special missing values reflecting the content of the original 80 character text field (e.g., SAS special missing ".A" = "abnormal hemoglobin variant observed", ".B" = "below detection limit", ".E"=excess triglycerides / lipemic", and so on.

The text files of individual laboratory results are converted to SAS at the COC and the individual results transmitted over time for the baseline visits are collapsed to one record per patient id – accession number combination of final laboratory results transmitted by CBL to COC. Quality control analysis of the assays occurs by way of replicate QC samples sent by the clinics to CBL with masked patient ids and different accession numbers than the primary samples. A second aspect of the quality control process is the weekly reconciliation of information keyed by CBL from the original copy of each of mailing forms that accompanied the samples to CBL and that same information keyed at the COC from a second copy of the multipart mailing forms sent by the clinics to COC on a weekly basis – including patient identifiers, accession numbers, date collected by clinic, date shipped to CBL, and the number and type of sample.

The CBL will cryopreserve and transform lymphocytes. The CBL will prepare a stable cell lysate from whole blood for DNA isolation using the Gentra Purgene protocol. The CBL will ship transformed lymphocytes and stable cell lysates to the CDC as soon as possible, just as the DNA tubes were shipped as soon as possible from the clinic to CBL. Shipment of saved samples of non-renewable resources (saved plasma, saved serum, and saved urine) from CBL to CDC for long-term storage and ultimate distribution from the repository to approved external researchers does not occur until the coordinating center has determined the eligibility status of the proband or sibling to be case or control and all families members of potential case or control trios to be available.

Specimen tracking of samples received at CBL and potentially packed and shipped to CDC occurs by way of the COC's web-based specimen tracking system by CBL scanning the machine-readable barcodes on each specimen label for accession number, specimen number and unique astro number. These 3 identifiers as well as sample type and destination are also printed on the labels in human-readable form. As CBL receives specimens from clinics and then prepares boxes and ships boxes of samples to CDC, the specimen tracking system monitors in real time the location of each individual specimen.

10.5 JOSLIN

Joslin Diabetes Center will transmit Joslin participant medical and demographic data and Joslin recruitment data on a monthly basis to COC in the form of SAS datasets which are complete refreshes of the previous month's data. Periodically Joslin will transmit fatality files as SAS data in addition to completing the GK203 Death Notification form. The clinical and demographic information transmitted to the COC is integrated with the laboratory data from the CBL at the COC to determine GoKinD eligibility status. Feedback reports of biochemical results and eligibility summaries are transmitted from the COC back to Joslin as text files and as SAS datasets.

10.6 CDC AND CASPIR

The Center for Disease Control (CDC) is the next destination for DNA, first processed at CBL, which is then stored at CDC. If the participant is deemed to be an eligible case or control by the COC or is quality control sample for DNA it is genotyped at CDC. CDC uploads basic demographic and eligibility data about each participant by way of the web-based specimen tracking system (see below). Since eligibility and trio status change over time, as a participant goes from being potential to pending to case, control, micro, ineligible, or inactive and a family potentially progresses from singleton to incomplete trio to complete trio and back, or as inconsistencies on patient identifiers are corrected, there is a periodic reconciliation of CDC's stored eligibility and trio status with the most current status at the COC.

Genotypic data for cases and controls and their complete trio parents primary samples and quality control samples are transmitted periodically from CDC to COC as Microsoft Excel files. Instances where the genetic data fail to confirm the sex of the participant or the putative familial relationships of trios are investigated at CDC, CBL, and the COC. Expanded genotyping with forensic profiler is employed with L-DNA and repeat L-DNA samples to verify DNA results, and if necessary and possible, repeat DNA samples are collected from the participants and genotyped. Potential non-paternity, non-maternity, sex discrepancies, and disagreement among DNA, L-DNA, rL-DNA, and rDNA samples for participants are investigated and documented in Excel at both CDC and COC until resolved. Should the genetics data indicate that sample mix-up has occurred or familial relationships are not confirmed, the fate of the samples, the genetic data, the corresponding biochemical data, and the participant's status in the GoKinD study are documented in an Excel file, including making the participant inactive, suppressing some or all biochemical results, and so on. This Excel file of potential and actual problem patient ids based on genetics is part of the GoKinD data documentation.

10.6.1 Definitions for CDC Genetics Data

10.6.1.1 HLA Library Release code

The HLA Library release code identifies the list of known HLA alleles that an individual's DNA sequence was compared against. The library release codes are defined by the ImMunoGeneTics (IMGT)/HLA database, which is part of the international IMGT project. This database includes the official HLA sequences named by the WHO HLA Nomenclature Committee For Factors of the HLA System. For a description of all known alleles included in the defined library, please refer to: http://www.ebi.ac.uk/imgt/hla/docs/version.html.

10.6.1.2 HLA Genotypes that include an "X"

An "X" in an HLA genotype indicates that a previously undescribed polymorphism has been identified, however it has not yet been submitted and/or named by the WHO HLA Nomenclature Committee For Factors of the HLA System. CDC has named these sequences according to the HLA allele that the sequence most closely matches with an X to indicate the type of new polymorphism. For example, if a new polymorphism has a DNA sequence that is an exact match (with the exception of the new polymorphism) to a DQB1*030101 allele, it would be named DQB1*03010X (if the DNA change does not change the amino acid sequence) or DQB1*030X (if the DNA change does change the amino acid sequence).

10.6.2 CASPIR

CASPIR is the source of the 95 label label-sets for each accession number barcode (Gxxxx) that are used by the clinics, CBL, and CDC for specimen labeling to uniquely identify the accession number, specimen number and unique astro numbers for each primary and quality control sample. CASPIR is the next destination for transformed lymphocytes and saved samples of plasma, serum, and urine for the GoKinD repository from the clinics to CBL to CASPIR, for ultimate distribution to qualified investigators. Data feeds from CASPIR to the COC are in the form of SAS datasets for pre-assigned label sets and for shipment of saved samples.

10.7 BRITISH DIABETES ASSOCIATION – DIABETES UK

A sister study to GoKinD is in progress in the United Kingdom by Diabetes UK. It is hoped to be able to combine the phenotypic and genotypic data collected in the UK with that collected in the United States and Canada in GoKinD. Genotypic analysis of UK DNA samples is being performed at CDC. Medical and demographic information will be transmitted from the UK to the COC. Test files of phenotypic data as Microsoft Excel files have been received at the COC and test samples of DNA from Diabetes UK trios have been genotyped at the CDC.

10.8 GWU

10.8.1 Biochemical Data Reports

GWU will receive the laboratory data from the CBL weekly. GWU will process the data and fax or FTP a one-page clinical report of biochemical results for each participant and the

guidelines for interpreting laboratory data (see Figures 10.3 and 10.4). Joslin also receives the cumulative biochemical data in the form of SAS datasets in each weekly production cycle.

10.8.2 Eligibility Reports

Once a proband or sibling has been completely evaluated for eligibility, the COC generates an eligibility report (Figure 10.5) for transmission to the clinic, including information on participants who are deemed not to be eligible as cases or controls with the reason for ineligibility. Subsequent changes in eligibility status will trigger a new eligibility report for the clinic.

10.8.3 Reconciliation of Mailing Forms and General Forms

The COC receives and data enters copies of all mailing forms from Joslin and GW clinics and all general medical forms for GW clinic participants. Cross-form checks on identifiers and similar fields and checks against CBL's entry of the mailing forms as they receive specimens is part of the weekly production process. Discrepancies on identifiers (patient id, accession number, and initials), collection dates, and shipment dates are resolved by investigation at the clinics, CBL and COC and corrections are documented and updated (with a signed paper audit trail) at the COC.

10.8.4 Personal Information – Identifiable Participant Data

For GWU clinic participants, there is an optional GK200 Personal Information Form (see appendix) that is completed by the Study Coordinator at the time of enrollment or clinic visit. This for contains particularly sensitive information in terms of parcipant name, address, phone numbers, email, place of birth, etc. To protect participant confidentiality this form is keyed in a separate Visual FoxPro application that is encrypted and does NOT go through the standard weekly processing of conversion from FoxPro to SAS. The Information Specialist at the COC keys the form in FoxPro and it immediately encrypts on a secure server. The Data Manager at the COC is the only person with the ability to unencrypt the file for the purpose of Phase II GoKinD follow-up with participants or comparison of date of birth and sex fields as entered on the GK200 form and the GK202 Checklist form with a highly limited export from the encrypted file. The paper form is stored in locked cabinets in a locked office.

Personal identifier information on Joslin clinic participants is stored at Joslin and is not transmitted to the COC except participant initials.

10.8.5 Specimen Tracking System

Copies of the multipart mailing forms which accompany DNA, blood and urine samples shipped from clinics to CBL are also keyed at the COC. The mailing forms are the first part of the specimen tracking system for GoKinD which permits real-time monitoring of movement of individual specimens by unique barcode identifiers as they are received by CBL, then boxed and shipped at CBL and received by CDC and CASPIR. Specimen tracking is done in a web-

based application at the COC in SAS and SQL for CBL and CDC and then for monitoring shipment of both renewable (DNA) and nonrewable resources (saved plasma, serum, and urine samples) to approved external researchers.

The specimen tracking system also makes current demographic information and eligibility and trio status information available to CBL and to CDC for real-time look-up or fixed text export. Boxes which contain specimens which are not eligible to be shipped to CDC cannot be shipped, until the ineligible specimens are removed. Specimens of a given type based on the unique astro numbers can only go in the appropriate type of box for shipment. Real-time monitoring of specimen location by CBL, CDC, and the COC is possible. Comments on individual samples or entire boxes are possible by both CBL and CDC.

10.8.6 GWU COC Data Receipt Summary

In summary, the data received for GoKinD at the COC at George Washington University Biostatistics Center include:

Source	Data Content	Frequency	Format
MMG	GW clinic recruitment screens	weekly	Microsoft Access
Joslin	Joslin recruitment data	monthly	SAS
CBL	biochemical & administrative	weekly	fixed text ASCII & Microsoft Excel
CDC	genetics & administrative	periodically	Microsoft Excel & SAS datasets
Joslin	medical & demographic data	monthly	SAS
GW Clinic	medical & demographic data	weekly	paper copy from multipart forms
All Clinics	mailing forms for samples	weekly	paper copy from multipart forms
CBL	specimen tracking	real-time	SAS and SQL – web based
CDC	specimen tracking	real-time	SAS and SQL – web based
COC-GW clinic invoicing		monthly	SAS & Oracle

FIGURE 10.1

GOKIND Electronic Data Collection

January 13, 2003 Draft

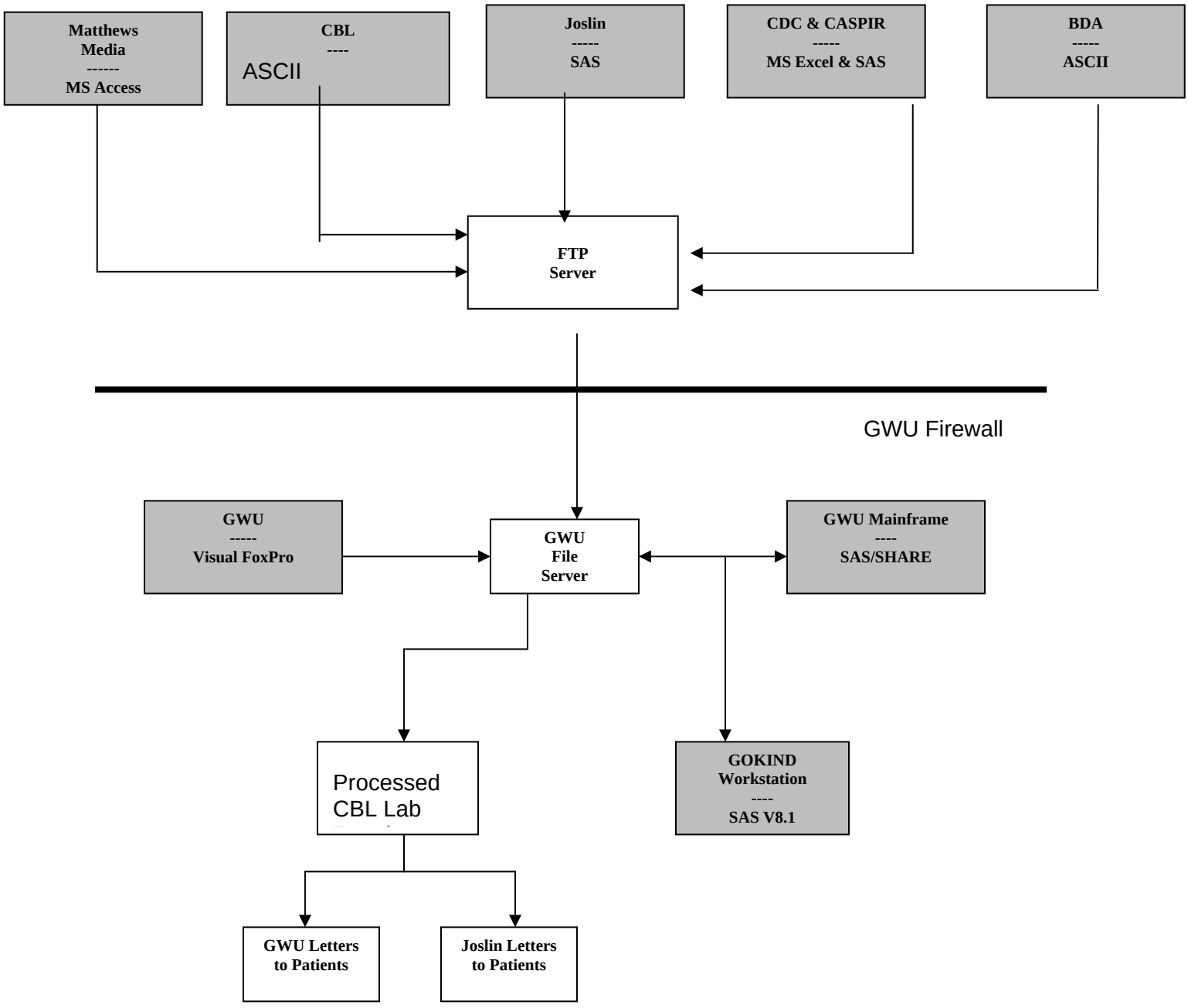

FIGURE 10.2
GOKIND Data Collection

	DATA	RECORDS	FORMAT	HOW	FREQUENCY	EDITS	BACKUP	SECURITY
Clinics	Data Forms	New	Paper	Mailers	Weekly	N/A	On file	Locked cabinet
	Corrected Forms	Changed	Paper	Mailers	Weekly	N/A	On file	Locked cabinet
	Corrected Edits	New	Paper	Mailers	Weekly	N/A	On file	Locked cabinet
	Tubes # 1-3	New	Specimens	FedEx	Immediate	N/A	None	Coded ID
	Hemoglobin A1c	New	Specimens	FedEx	Within 5 Days	N/A	None	Coded ID
	Plasma, serum, etc	New	Specimens	FedEx	Weekly	N/A	None	Coded ID
	10% QC Specimens	New	Specimens	FedEx	Same Schedule	N/A	None	Coded ID
	5% QC Genotypes	New	Genotypes	FedEx	Same Schedule	N/A	None	Coded ID
	Specimen Tracking	New	Paper	Mailers	Weekly	N/A	On file	Locked cabinet
Matthews Media	Recruitment	New	Access	FTPS	Weekly	No	Daily	Password & Firewalls
	Follow-up	New/Changed	Access	FTPS	Weekly	No	Daily	Password & Firewalls
CBL	Measurements	New	ASCII	FTPS	Weekly	No	Electronic	Password
	Lymph/Lysates	New	Specimens	FedEx	Weekly	N/A	Stored	?
Joslin	Recruitment	New	SAS	FTPS	Monthly	No	Daily	Password
	Patient Data	New	SAS	FTPS	Monthly	Yes	Weekly	Password
	Specimens	New	Specimens	FedEx	Immediate	N/A	No	Coded ID
	Specimen Tracking	New	Paper	Mailers	Weekly	N/A	On file	Locked cabinet
GWU	Patient Data	New/Changed	FoxPro	Network	Weekly	Yes	Daily	Password
	Specimen Tracking	New/Changed	FoxPro	Network	Weekly	Yes	Daily	Password
	GK202 Personal Data	New/Changed	Paper	Mail	Weekly	N/A	N/A	Encryption
	Weekly Database	Cumulative	SAS	Network	Weekly	Yes	Electronic	Archived
	Lab Results	New/Changed	Electronic	E-mail/Fax	Weekly	Yes	Electronic	Password
CDC	DNA	New	Specimens	Store	N/A	N/A	N/A	Coded ID
	Initial Analyses	New	Excel/SAS	FTPS	Periodic	N/A	Electronic	Password
	Specimen Tracking	New	Specimens	Web	Immediate	Yes	Electronic	Password
BDA	Patient Data Lab Measurements Analyses							

FIGURE 10.3

Date:
Initials:
Barcode:
Specimen Date:
Screen Number:

Study ID#:

Patient Recruitment Center:
Coordinating Center: George Washington University

Urinary Albumin1 = mg/L
Urinary Albumin2 = mg/L
Urinary Albumin 3 = mg/L

Urinary Creatinine1 = mg/dL
Urinary Creatinine2 = mg/dL
Urinary Creatinine3 = mg/dL

Albumin/Creatinine Ratio1 = ug/mg
Albumin/Creatinine Ratio2 = ug/mg
Albumin/Creatinine Ratio3 = ug/mg

Total Cholesterol = mg/dL

HDL = mg/dL

Glycohemoglobin A1C = %

Serum Creatinine = mg/dL

Serum Cystatin = mg/L

FIGURE 10.4

GoKinD Guidelines for Interpreting Laboratory Data

Albumin/Creatinine Ratio (ACR)
If ACR <20 ug/mg, this result is in the normal range.

If ACR elevated (20-300 ug/mg), this is considered out of the normal range. This result must be re-checked and confirmed by your doctor. Please contact your physician to discuss this result.

If ACR >300 ug/mg, this result is proteinuric. This result must be re-checked and confirmed by your doctor. Please contact your physician to discuss this result.

Hemoglobin A1c (HbA1c)
Normal range is 5.0 ± 1.0 ($\bar{X} \pm 2$ St. Dev.)[1]

A value under 6.0% is considered normal.

If a parent is not diabetic with an abnormal A1c level (>6.0%), it is important for you to follow-up this elevated result with your physician.

Serum Creatinine (mg/dL)
If serum creatinine <1.5 mg/dL, the level is within normal range for men.
If serum creatinine <1.3 mg/dL, the level is within the normal range for women.

If serum creatinine >1.5 mg/dL, the result is out of the normal range for men. Please follow-up this result with your doctor.

If serum creatinine >1.3 mg/dL, the result is out of the normal range for women. Please follow-up this result with your doctor.

Serum Cystatin
Serum cystatin is considered experimental data, and we do not yet have a way to interpret the value. It is not necessary to release this value to the study participants or their physicians.

Total Cholesterol and HDL Cholesterol (mg/dL)
If total cholesterol <200 and HDL cholesterol >40, total cholesterol and HDL cholesterol levels are within normal ranges.

If total cholesterol >200, your cholesterol level is considered out of the normal range. Please follow-up your cholesterol level with your doctor.

If HDL cholesterol <40, your HDL level is below the normal range. Please follow-up your cholesterol with your doctor.

[1]The DCCT Research Group. Feasibility of Centralized Measurements of Glycated Hemoglobin in the Diabetes Control and Complications Trial: A Multicenter Study. Clinical Chemistry 1987; 33:2267-71.

FIGURE 10.5

Family ID #: _____ Date of Today: __/__ /____
Date of Birth: __ /__ /____ Initials: ___
Age at Screening: __ Year
(must be between 18-54)

Diagnosed Before 31:
Insulin Taken Within 1 Year of Diagnosis:
PATIENT HAS TYPE 1 DIABETES:
Duration of Type 1 Diabetes: __ YEAR

Presence of Diabetic Nephropathy:
Duration of T1D >=10 Years:
ESRD:
Kidney Transplant:
Kidney/Pancreas Transplant:
Dialysis:
Persistent Proteinuria:
Historic ACR Positive without an ACR value:

	ACR RATIO VALUE
USCR1:	___ _____ ug/mg
USCR2:	___ _____ ug/mg
USCR3:	_____ ug/mg

** 2 out of 3 values must be positive or greater than 300 ug albumin/mg urine creatinine

PATIENT IS ELIGIBLE AS A CASE:

Absence of Diabetic Nephropathy:
Duration of T1D >= 15 years:
Treated with ACE inhibitor:
Using antihypertensives:
Persistent Normoalbuminuria:
Historic ACR Positive without an ACR value:

	ACR RATIO VALUE
USCR1:	_____ ug/mg
USCR2:	_____ ug/mg
USCR3:	_____ ug/mg

** 2 out of 3 values must be less than 20 ug albumin/mg urine creatinine and if a third is needed the highest value must also be less than 40 ug albumin/mg urine creatinine.

PATIENT IS ELIGIBLE AS A CONTROL:

PATIENT IS ELIGIBLE AS A MICROALBIMINURIC:
