Data Set Integrity Check for the Trial Net Natural History (TN-01) Study

Prepared by

RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194
December 18, 2011

Revision History

Version	Author/Title	Date	Comments
0.1	SM Rogers	December, 2011	Original
0.2	Formatted draft – C. Hollingsworth	February 2012	
1.0			First version
			New data received March 2012, age missing
1.1	Version 1.0 revised	August 2012	Revised with new data received June 2012
1.2	Version 1.1 revised	August 2012	Revised based on new data received August 2012

Table of Contents

1 Introduction

The Trial Net Natural History Study is a prospective cohort study of relatives of persons with type 1 diabetes (T1D). Eligible persons for the TN-01 study do not have T1D but are at increased risk because they have a family member with the disease. TN-01 data archived in the NIDDK repository include Phase 1 screening, Phase 2 baseline assessment, and Phase 3 follow-up. As a partial check of the TN-01 data archived in the NIDDK data repository, a dataset integrity check (DSIC) was performed. This DSIC replicates a small number of analyses performed in the publication by Mahon et al. (2009) in Pediatric Diabetes [1].

2 Background/ Purpose

The intent of this DSIC is to provide confidence that the data distributed by the NIDDK repository is a true copy of the study data. Our intent is not to assess the integrity of the statistical analyses reported by study investigators. As with all statistical analyses of complex datasets, complete replication of a set of statistical results should not be expected on a first (or second) exercise in secondary analysis. This occurs for a number of reasons including differences in the handling of missing data, restrictions on cases included in samples for a particular analysis, software coding used to define complex variables, etc. Experience suggests that most discrepancies can ordinarily be resolved by consultation with the study data coordinating center (DCC), however this process is labor-intensive for both DCC and Repository staff. We do not attempt to resolve minor or inconsequential discrepancies with published results or discrepancies that involve complex analyses, unless NIDDK Repository staff suspect that the observed discrepancy suggests that the dataset may have been corrupted in storage, transmission, or processing by repository staff. We do, however, document those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

3 Archived Datasets

The DCC submitted 13 SAS (.sas7bdat) datasets which correspond to 10 study data collection forms (Form NH20, Participant Contact Change and Site Transfer, was not associated with a data file), a patient registration file, a laboratory file, and an adverse event file. An analysis file for the manuscript by Mahon et al. was not provided and reconstruction of the analysis dataset and derived variables was conducted by repository analysts. Datasets used to create this analysis file include: nh01_screening form, nh01f_familyhistoryform, nh04_baselineriskassessment, and research_labs (version 120712).

4 DSIC Analysis

The publication by Mahon et al. reviews the baseline characteristics and antibody test results of Phase 1 and Phase 2 subjects. We present our DSIC results to published results in Table 1 (Phase 1 and Phase 2 subjects' characteristics) and Table 2 (Antibody test results). For this DSIC, SAS datasets were converted to Stata using Stat/Transfer (Circle Systems Inc). Stata v12.0 was used for all analyses (Appendix 1).

The Mahon analysis was restricted to subjects who entered Phase 1 between March 2004 and January 2006 and subjects who entered Phase 2 by December 2006. Data provided to the NIDDK repository include Phase 1, Phase 2, and Phase 3 subjects who entered the study between March 2004 and July 2011. In Phase 1 (Screening), pancreatic autoantibodies (glutamic acid decarboxylase, insulin, ICA-512 and ICA) were measured. According to the original study eligibility criteria, subjects with at least two positive tests for any one of the four antibodies in Phase 1 were eligible for Phase 2 (Baseline risk assessment). Subjects with discordant results in Phase 1 (ex, ICA-512 positive on the first test and ICA-512 negative on the second) received a third test. The subject was confirmed positive if two of three tests for that antibody were positive. Phase 3 involved follow-up risk assessments of subjects completing Phase 2.

5 Results

Phase 1 and Phase 2 Subject Characteristics. Table 1 presents the characteristics of Phase 1 and Phase 2 subjects in the published manuscript and our DSIC calculations. The manuscript reports results for 12636 TrialNet subjects in Phase 1 and 322 subjects in Phase 2; our DSIC reports on 12641 subjects in Phase 1 and 318 in Phase 2. The variables in each of the datasets used in deriving our DSIC estimates can be found in Appendix 1. Our DSIC calculations of the distribution of subject characteristics by age, gender, race, ethnicity, and family history of Type 1 diabetes are similar to the published results.

Table 1: Phase 1 and Phase 2 Subject Characteristics.

	Mahon et al (2009)		DSIC	
	Phase 1 (n=12 636)	**Phase 2 (n=322)**	**Phase 1 (n=12,641)**	**Phase 2 (n=318)**
Mean age (SD) (yr)	21.8 (14.8)	19.4 (14.3)	21.7 (14.8)	19.8 (14.2)
Age range (yr) %				
1-5	14	15	14	12
6-10	19	26	19	29
11-15	15	16	15	17
16-20	7	6	6	6
21-25	2	2	2	3
26-45	43	35	43	33
Female %	59	58	59	59
Race %				
White	88	90	89	92
African-American	3	2	3	3
Asian	1	2	2	3
Others	9	6	6[b]	2[b]
Ethnicity %				
Hispanic	13	9	13	10
Others	87	91	87	90
Family history of T1D % [a]				
First degree relative	91	94	92	93
Second degree relative	17	18	8[c]	7
Third degree relative	9	8		
One affected relative	77	73	78 [d]	72 [d]
Two or more affected relatives	20	27	22	28
HLA DQBT*0602 %	Not done	5	Not done	

[a] Subjects could have two or /more relatives, of different degrees, with T1D.
[b] Excludes subjects who refused to report race or reported race unknown.
[c] Data in nh01_screeningform combine 2nd and 3rd degree relatives in the variable, DegreeOfRelative_2ndor3rd.
[d] Excludes subjects with missing data.

Note: Published results from Mahon et al. (2009), Pediatric Diabetes 10: 97-104.

Antibody Test Results. Table 2 of the manuscript provides results of antibody screening for subjects in the Phase 1 sample and for a subset of subjects screened in the Diabetes Prevention Trial-Type 1 (see manuscript for details of DPT-1 sample). Our DSIC tabulations vary from the published results in several ways. First, our DSIC analyses do not include tabulations for DPT-1 subjects. DPT-1 data are provided in a separate database on the NIDDK repository website. Second, in addition to tabulations of the number of Phase 1 subjects with positive antibody tests at screening, Mahon et al. also report confirmation rates by antibody type and by number of positive antibodies. Initially, the Natural History study protocol specified that Phase 1 subjects with at least two positive tests for any of four antibodies (MIAA, GADA, ICA-512 and ICA) were eligible for Phase 2. Subjects with discordant results for a specific antibody on the first two samples were asked to provide a third 'confirmatory' sample. Therefore, if two out of three tests were positive for GADA, the subject was defined a 'confirmed GADA positive'. We note that 'confirmed' positive or negative results are indicated in data (research_labs) provided to the NIDDK repository; the individual sample test results (first and second samples, see *Figure 1* in manuscript) to derive the confirmed test result do not appear to be included in the lab data. Hence, we are unable to provide comparable tabulations of the confirmation rates as in Mahon et al. As an alternate form of comparison, we tabulated the number and type of confirmed positive antibody tests at screening (Visit="Screening") and baseline (Visit="Baseline (Phase 2)". Our independent calculations are in agreement with the published manuscript. These tabulations suggest that subjects with two or three confirmed positive antibody tests at screening were more likely to test antibody positive at baseline than subjects with only one positive antibody screening test. Furthermore, subjects who tested GADA positive at screening were most likely to test Ab positive at baseline – also in accordance with findings by Mahon et al. These supplementary results are provided by Stata code in Appendix 1. Finally, we note for users of the TN-01 data that the antibody criteria for determining Phase 2 eligibility were revised in February 2007. The original criteria required that at least one specific antibody be positive on two separate tests, the new definition required positive results for any two antibody tests. The Mahon et al. manuscript reports on the original antibody criteria.

Table 2: Antibody Test Results.

	Mahon et al. (2009)		DSIC
	NHS (n=12 636)	**DPT-1* (n=17 207)**	**NHS (n=7 069)±**
No. of subjects with a positive Ab test on first sample, n (%)			
One Ab	347 (2.8)	1009 (5.9)	195 (2.8)
Two Ab	125 (1)	193 (1.1)	75 (1.1)
Three Abs	91 (0.7)	147 (0.8)	47 (0.7)
Four Abs	42 (0.3)	54 (0.3)	28 (0.4)
No. of subjects positive by specific Abs on 1st sample, n (%)			
GADA	452 (3.6)	688 (4)	259 (3.7)
ICA512A	205 (1.6)	276 (1.6)	107 (1.5)
MIAA	203 (1.6)	437 (2.5)	121 (1.7)
ICA	178*	651 (3.8)	111
No. of subjects positive for 1 or more biochemical Abs on 1st sample, n (%)	605 (4.8)	1076 (6.2)	343(4.8)
Confirmation rates by Ab type			
GADA	318/362 (88)		Na
ICA512A	124/145 (86)		Na
MIAA	117/157 (75)		Na
ICA	102/139 (74)		Na
Confirmation rates by no. of positive Abs on 1st sample†, n (%)			
One Ab pos	211/284 (74)		Na
Two Abs pos	91/94 (97)		Na
Three Abs pos	67/67 (100)		Na
Four Abs pos	33/33 (100)		Na
Two or more Abs pos	191/194 (96)		Na

Note: Ab, antibody; DPT-1 Diabetes Prevention Trial Type-1; GADA glutamic acid decarbyoxylase antibodies; ICA islet cell antibodies; ICA512A antibodies to ICA-512; mIAA insulin autoantibodies.
*ICA is only tested in TrialNet subjects with at least one positive biochemical antibody but was tested in all DPT-1 subjects.
† Based on up to three samples being tested for antibodies in Phase 1 and Phase 2 more than 1 yr, where a confirmed positive antibody required that a specific antibody be positive on at least two samples. Subjects could be confirmed positive for more than one antibody.
± 7069 valid results for phase 1 subjects.
Note: Published results from Mahon et al. (2009), Pediatric Diabetes 10: 97-104.

[na]: Not available; data supplied to the repository include the final confirmed antibody test results for each Ab type only; the results of multiple antibody testing to achieve this final outcome ('positive' or 'negative') were not provided. See notes in text above.

6 Conclusion.

Our DSIC results of subject characteristics are similar to those reported by Mahon et al. We note that minor protocol modifications and the posting of final antibody test results [1] to the repository database limit our performing fully comparable tabulations of antibody test results. Otherwise, the results of this DSIC suggest that the data provided to the NIDDK repository include the range of study variables and data collection instruments from the TN01 study and show no obvious evidence of corruption in storage, transmission, or processing by repository staff.

7 Reference.

[1] Mahon JF, Sosenko JM, Rafkin-Mervis L, et al for the TrialNet Natural History Committee and Type 1 Diabetes TrialNet Study Group. 2009. The TrialNet Natural History Study of the Development of Type 1 Diabetes: objectives, design, and initial results. Pediatric Diabetes 10:97-104

APPENDIX 1

```
/*TABLE 1. Phase 1 and Phase 2 subjects' characteristics
      by Mahon et al, Pediatric Diabetes, 2009*/

**tabulate characteristics of Phase 1 subjects from Screening and Family Hx forms
cd "C:\Documents and Settings\smr\My Documents\TrialNet\Data Extraction_1_June2012\Stata\"
                  use "nh01_screeningform.dta"
  *select Phase 1 subjects screened between March 2004 and January 2006
gen before = cond(Date_at_Screening < td(01Jan2006), 1, 0)
  *N subjects=12674
  *select subjects screened after March 2004
gen after = cond(Date_at_Screening > td(01Mar2004), 1, 0)
  *N subjects==12668
list Date_at_Screening before after in 1/75
keep if before==1 & after==1
tab1 Age Sex Race_White Race_BlackorAfricanAmerican Race_Asian Race_AmericanIndian Race_NativeHawaiian
/// Race_Refused Race_Unknown OtherSpecify Ethnicity DegreeOfRelative_1st DegreeOfRelative_2ndor3rd,
missing
sort MaskID
keep if Age >=1 & Age <=45
keep if DegreeOfRelative_1st==1 | DegreeOfRelative_2ndor3rd==1
gen ph1subject=1
tab ph1subject
  *N=12642
summ Age
recode Age 1/5=1 6/10=2 11/15=3 16/20=4 21/25=5 26/45=6, gen(recage) label(recage)
label define recage 1"1-5" 2"6-10" 3"11-15" 4"16-20" 5"21-25" 6"26-45"
tab Age recage
gen raceother=.
replace raceother=1 if Race_AmericanIndian==1 | Race_NativeHawaiian==1 | OtherSpecify!=""

duplicates report
duplicates list
duplicates drop
          * 1 duplicate observation dropped, N=12641
tab1 Sex Race_White Race_BlackorAfricanAmerican Race_Asian raceother Ethnicity DegreeOfRelative_1st
      DegreeOfRelative_2ndor3rd, missing
gen white=0
replace white=1 if Race_White==1
gen AA=0
replace AA=1 if Race_BlackorAfricanAmerican==1
gen Asian=0
replace Asian=1 if Race_Asian==1
gen other=0
replace other=1 if raceother==1
tab1 white AA Asian other
gen Race=.
replace Race=1 if white==1
replace Race=2 if AA==1
replace Race=3 if Asian==1
replace Race=4 if other==1
tab  Race, missing
```

```
gen degree1=0
replace degree1=1 if DegreeOfRelative_1st==1
gen degree23=0
replace degree23=1 if DegreeOfRelative_2ndor3rd==1
summ Age
tab1 recage Sex Race Ethnicity degree1 degree23
        save "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History Study\REVISED
            DSIC\DSIC_phase1", replace

        /* Tabulation of number of affected relatives from family history form */
use "nh01f_familyhistoryform.dta"
sort MaskID
keep MaskID NumOfRelativesWithT1D RelativeWithT1D1
duplicates report
duplicates list
duplicates drop
   *1 duplicate observation dropped
merge m:1 MaskID using "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History Study\REVISED
DSIC\DSIC_phase1", gen(merge3)
        *matched N=10711 _merge==3
tab1 NumOfRelativesWithT1D RelativeWithT1D1
tab NumOfRelativesWithT1D if ph1subject==1
sort MaskID
        save "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History Study\REVISED
           DSIC\DSIC_phase1rel", replace
  *n=10,515 observations

        /* tabulate characteristics of Phase 2 subjects from Baseline Assessment form */
use "nh04_baselineriskassessment.dta", clear
tab1 ParticipantHasOneRelativeWith ParticipateInDPT1 Visit
        *Visit denotes phase 2 baseline risk assessment subjects
keep MaskID Visit Date_of_Visit ParticipantHasOneRelativeWith ParticipateInDPT1 AntibodySampleCollected ///
  HLASamplesCollected OGTTSampleCollected ParticipantDiagnosedWithT1D
sort MaskID
  **N=1177
        /* per manuscript: keep if subjects entered Phase 2 by 31 Dec 2006      */
gen ph2date = cond(Date_of_Visit < td(31Dec2006), 1, 0)
tab ph2date
gen phase2=0
replace phase2=1 if Visit=="Baseline (Phase 2)"
tab phase2
sort phase2
keep if phase2==1 & ph2date==1
tab Visit
  **N=484
rename Date_of_Visit Date_of_Visitp2
label var Date_of_Visitp2 "Date of Visit Phase 2"

 merge 1:1 MaskID using "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History
Study\REVISED DSIC\DSIC_phase1rel", gen(merge4)

   **calculate age at Phase 2
tab1 DOB_Month DOB_Year
```

```
gen birthdate = mdy(DOB_Month, 1, DOB_Year)
    /* note day of birth not provided, randomly set to 1 */
format birthdate %td
gen agep2= Date_of_Visitp2-birthdate
gen agep2yr=agep2/365
format agep2yr %2.0f
summ agep2yr
recode agep2yr 1/5.99=1 6/10.99=2 11/15.99=3 16/20.99=4 21/25.99=5 26/45.99=6 46/47=7, gen(recph2age)
label(recage2)
label define recage2 1"1-5" 2"6-10" 3"11-15" 4"16-20" 5"21-25" 6"26-45" 7">45
tab recph2age, missing
tab recph2age
   /* N=318 subjects entered Phase 2
          manuscript notes 322 subjects */
list MaskID Age birthdate Date_of_Visit Date_of_Visitp2 agep2yr recph2age in 1/30
summ Age recph2age
tab Sex phase2, col
tab Race phase2, col
tab Ethnicity phase2, col
tab DegreeOfRelative_1st phase2, col
tab DegreeOfRelative_2ndor3rd phase2, col
gen degreeofrel=.
replace degreeofrel=1 if DegreeOfRelative_1st==1
replace degreeofrel=2 if DegreeOfRelative_2ndor3rd==1
tab degreeofrel phase2, col
tab NumOfRelatives phase2, col
          save "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History Study\REVISED DSIC\DSIC
phase2", replace

**************Tabulations for Table 2 Antibody test results*************
use "research_labs", clear
tab1 TEST_NAME Visit event_title
          /* Event_title indicates phase 1 screening, phase 2 baseline, phase 3 baseline
          and followup at each 6 months thereafter.  Sample Ns for visit and
          event_title match */
keep if TEST_NAME=="MIAA" | TEST_NAME=="GAD65" | TEST_NAME=="ICA512" | TEST_NAME=="ICA"
*restrict to autoantibodies and Initial Screening  PRN indicates annual rescreening
tab event_title if Visit=="Screening"
keep if Visit=="Screening"
tab1 SPEC_NAME TEST_NAME
sort TEST_NAME
by TEST_NAME: tab OUTCOME Visit
gen Routcome=.
replace Routcome=1 if OUTCOME=="Pos"
replace Routcome=0 if OUTCOME=="Neg"
by TEST_NAME: tab Routcome Visit

/* Baseline (phase 1) antibody test results */
gen test2=.
replace test2=1 if TEST_NAME=="MIAA"
replace test2=2 if TEST_NAME=="GAD65"
replace test2=3 if TEST_NAME=="ICA512"
replace test2=4 if TEST_NAME=="ICA"
```

```
label define test2 1"MIAA" 2"GAD65" 3"ICA512" 4"ICA"
label values test2 test2
tab test2, missing

  ***reshape data long to wide
sort MaskID
tab SPEC_NAME
drop if SPEC_NAME=="Serum - autoantibodies (stored)"
          keep MaskID Routcome test2 Visit
*drop 12 duplicate observations for this analysis
duplicates list
 duplicates drop
reshape wide Routcome, i(MaskID) j(test2)
label var Routcome1 "MIAA test result_baseline"
label var Routcome2 "GAD65 test result_baseline"
label var Routcome3 "ICA512 result_baseline"
label var Routcome4 "ICA result_baseline"
tab1 Routcome*, missing

merge 1:1 MaskID using "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History Study\REVISED
            DSIC\DSIC_phase1", gen(merge5)
 tab1 Routcome* ph1subject, missing
 tab1 Routcome* if ph1subject==1
          *Select Phase 1 subjects who entered study between March 2004 and January 2006
keep if ph1subject==1       /* 12641 phase 1 subjects, 5572 with missing OUTCOME values */
tab1 Routcome*, missing
egen totab = rowtotal (Routcome*)
tab totab, missing
    save "C:\Documents and Settings\smr\My Documents\TrialNet\Natural History Study\REVISED DSIC\Table
          2_Aug2012", replace
```