Dataset Integrity Check for the Treatment Options for Type 2 Diabetes in Adolescents and Youth Genetics Study (TODAY Genetics)

Prepared by NIDDK-CR
September 17, 2021

Contents

1 Standard Disclaimer

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 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. It is thus not our policy to resolve every discrepancy that is observed in an integrity check. Specifically, 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 in footnotes to the integrity check those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

2 Study Background

The TODAY Genetics study was designed to utilize blood and phenotypic information to explore relationships between participant genes and type 2 diabetes (T2D) in addition to obesity, insulin resistance, and cardiovascular issues associated with insulin resistance. The TODAY Genetics study was a satellite protocol conducted under the auspices of the TODAY study group. Individuals who were diagnosed with T2D and had BMI ≥ 85th percentile while under the age of 18 were recruited from TODAY clinical centers and collaborators. Once enrolled, participants provided a blood sample for analysis of diabetes type and DNA extraction. Information on participant and family medical history was also collected.

3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the TODAY Genetics folder in the data package. For this replication, variables were taken from the geninfo_20190206 dataset.

4 Statistical Methods

Analyses were performed to replicate results for a subset of the data published by Srinivasan et al. [1] for The First Genome-Wide Association Study for Type 2 Diabetes in Youth: The Progress in Diabetes Genetics in Youth (ProDiGY) Consortium. To verify the integrity of the dataset, descriptive statistics were computed. Secondary materials were provided by the DCC to assist with the replication in this DSIC. These include a breakdown of the ProDiGY participants from TODAY Genetics, a table with ancestry breakdowns of the TODAY Genetics participants, as well as information allowing for the stratification of participants by ancestry. Given the limited information published or available from the DCC, only a few variables were able to be replicated and included for this DSIC.

5 Results

For Table 1 in the publication [1], Baseline demographics of ProDiGY participants, Table A lists the variables that were used in the replication, and Table B compares the results calculated from the archived data files to the results published in Table 1.

6 Conclusions

Due to the limited nature of this DSIC, the NIDDK-CR has insufficient evidence that the data to be distributed are a true copy of the study data. For the variables that were able to be replicated, only minor differences were found.

Given that the TODAY Genetics study was an extension of the main TODAY study, any requestors interested in TODAY Genetics should also review the TODAY study (https://repository.niddk.nih.gov/studies/today/).

7 References

[1] Srinivasan S, Chen L, Todd J, Divers J, Gidding S, Chernausek S, Gubitosi-Klug RA, Kelsey MM, Shah R, Black MH, Wagenknecht LE, Manning A, Flannick J, Imperatore G, Mercader JM, Dabelea D, Florez JC. The First Genome-Wide Association Study for Type 2 Diabetes in Youth: The Progress in Diabetes Genetics in Youth (ProDiGY) Consortium. Diabetes, 70(4), 996-1005, April 2021. doi: https://doi.org/10.2337/db20-0443

Table A: Variables used to replicate Table 1 – Baseline demographics of ProDiGY participants

Table Variable	**dataset.variable**
Age	geninfo_20190206.age
Female	geninfo_20190206.sex
Fasting glucose	geninfo_20190206.glu

Table B: Comparison of values computed in integrity check to reference article Table 1 values

Variable	European Ancestry (EUR)	DSIC EUR	Diff.	African American Ancestry (AFR)	DSIC AFR	Diff.	Hispanic Ancestry (AMR)	DSIC AMR	Diff.
TODAYGEN_case (%)	442 (21.2)	442 (21.2)	0 (0)	713 (34.2)	713 (34.2)	0 (0)	929 (44.6)	929 (44.6)	0 (0)
Age, mean (SD)[1]	14.8 (2.8)	14.5 (1.3)	0.3 (1.5)	15.5 (3.2)	14.6 (1.3)	0.9 (1.9)	15.1 (2.8)	14.6 (1.3)	0.5 (1.5)
Female (%)	61.5	61.5	0	66.9	66.9	0	61.0	61.0	0
Fasting glucose, mean (SD)	155.9 (79.0)	155.9 (79.0)	0 (0)	171.3 (109.4)	171.3 (109.4)	0 (0)	160.7 (93.0)	160.7 (93.0)	0 (0)

[1] Age information provided in the data are 1 (≤ 13 years), 14, 15, and 3 (> 15 years). For the purposes of replication with DCC results, “1” was recoded to 13, and “3” was recoded to 16.

Attachment A: SAS Code

```
libname dsic "X:\NIDDK\niddk-dr_studies6\TODAY_Genetics\private_orig_data\TODAY GENETICS Files
for Repository\sas_data";

****************************;
*Pulling in the list of IDs*;
****************************;

DATA WORK.Prodigy_TODAY_Genetics_IDs_Updat;
   LENGTH
      FID            $ 10
      IID            $ 7 ;
   FORMAT
      FID            $CHAR10.
      IID            $CHAR7. ;
   INFORMAT
      FID            $CHAR10.
      IID            $CHAR7. ;
   INFILE 'C:\Users\616045\AppData\Roaming\SAS\EnterpriseGuide\EGTEMP\SEG-12356-
efdfbb42\contents\Prodigy_TODAY_Genetics_IDs_Updated 080421-
0ca2c813793e459d9601be3bcf233d4d.txt'
      LRECL=18
      ENCODING="WLATIN1"
      TERMSTR=CRLF
      DLM='7F'x
      MISSOVER
      DSD ;
   INPUT
      FID            : $CHAR10.
      IID            : $CHAR7. ;
RUN;

/*************************************/
/*Pulling in the Ancestry Information */
/*************************************/
DATA WORK.EUR;
   LENGTH
      ID             $ 7 ;
   FORMAT
      ID             $CHAR7. ;
   INFORMAT
      ID             $CHAR7. ;
   INFILE 'C:\Users\616045\AppData\Roaming\SAS\EnterpriseGuide\EGTEMP\SEG-12356-
efdfbb42\contents\EUR-65f3b2b5ecd240a987b08103b57492e3.txt'
      LRECL=7
      ENCODING="WLATIN1"
```

```
      TERMSTR=CRLF
      DLM='7F'x
      MISSOVER
      DSD ;
   INPUT
      ID            : $CHAR7. ;
RUN;

DATA WORK.AMR;
   LENGTH
      ID            $ 7 ;
   FORMAT
      ID            $CHAR7. ;
   INFORMAT
      ID            $CHAR7. ;
   INFILE 'C:\Users\616045\AppData\Roaming\SAS\EnterpriseGuide\EGTEMP\SEG-12356-
efdfbb42\contents\AMR-8359d9843c9d4c0abd4fec3363b8e97e.txt'
      LRECL=7
      ENCODING="WLATIN1"
      TERMSTR=CRLF
      DLM='7F'x
      MISSOVER
      DSD ;
   INPUT
      ID            : $CHAR7. ;
RUN;

DATA WORK.AFR_V_Youth_Control;
   LENGTH
      ID            $ 7 ;
   FORMAT
      ID            $CHAR7. ;
   INFORMAT
      ID            $CHAR7. ;
   INFILE 'C:\Users\616045\AppData\Roaming\SAS\EnterpriseGuide\EGTEMP\SEG-12356-
efdfbb42\contents\AFR_V_Youth_Control-28b08d8041cc4932a0ab1b3602093e4c.txt'
      LRECL=7
      ENCODING="WLATIN1"
      TERMSTR=CRLF
      DLM='7F'x
      MISSOVER
      DSD ;
   INPUT
      ID            : $CHAR7. ;
RUN;

DATA WORK.AFR_V_T2D_Control;
   LENGTH
```

```
      ID            $ 7 ;
   FORMAT
      ID            $CHAR7. ;
   INFORMAT
      ID            $CHAR7. ;
   INFILE 'C:\Users\616045\AppData\Roaming\SAS\EnterpriseGuide\EGTEMP\SEG-12356-
efdfbb42\contents\AFR_V_T2D_Control-5dfbeefc96674cc9974332a701473254.txt'
      LRECL=7
      ENCODING="WLATIN1"
      TERMSTR=CRLF
      DLM='7F'x
      MISSOVER
      DSD ;
   INPUT
      ID            : $CHAR7. ;
RUN;

proc contents data=dsic.geninfo_20190206;
run;

*checking the ID var;
proc print data=dsic.geninfo_20190206;
var ID;
run;

*the FID var is the corresponding ID variable;
proc contents data=work.prodigy_today_genetics_ids_updat;
run;

*renaming the FID variable for the purposes of merging;
data one; set work.prodigy_today_genetics_ids_updat;
ID = FID;
run;

*checking the new ID variable;
proc print data=one;
var ID;
run;

*creating a temp dataset for merging;
data two; set dsic.geninfo_20190206;
run;

*merging data sets in order to subset the the participants to those included in the publication;
proc sort data=one;
by ID;
run;
```

```
proc sort data=two;
by ID;
run;

data three;
merge one (keep= ID iid in=a)
            two (in=b);
by ID ;
if a=1;
run;

*****************************************;
**     Replication of Subset Tables     **;
*****************************************;
proc contents data=three;
run;

proc freq data=three;
tables age;
run;

*recodeing age variable;
data three_1; set three;
if age = 1 then age1 = 13;
if age = 3 then age1 = 16;
if age = 14 then age1 = 14;
if age = 15 then age1 = 15;
run;

/***********************************/
/* Combining the Ancestry Datasets  */
/***********************************/
data eur_1; set eur;
Ancestry = "EUR";
iid = id;
run;

data amr_1; set amr;
Ancestry = "AMR";
iid = id;
run;

data AFR_V_T2D_Control_1; set AFR_V_T2D_Control;
Ancestry = "AFR_V_T2D";
iid = id;
run;

data ancestry; set eur_1 amr_1 AFR_V_T2D_Control_1;
```

```
run;

/***********************************************/
/* Merging the Ancestry Data with the full data */
/***********************************************/
proc sort data=ancestry;
by iid;
run;

proc sort data=three_1;
by iid;
run;

data three_2;
merge three_1
        ancestry;
by iid;
run;

/****************************/
/*   Replicating the tables  */
/****************************/

*Breakdown of the Ancestry groups;
proc freq data=three_2;
tables ancestry;
run;

*Age breakdown by Ancestry groups;
proc means data=three_2 mean std;
var age1;
class ancestry;
run;

*Sex breakdown by Ancestry groups;
proc freq data=three_2;
tables sex*ancestry/norow nopercent;
run;

*Fasting Glucose breakdwon by Ancestry groups;
proc means data=three_2;
var glu;
class Ancestry;
run;
```