# Dataset Integrity Check for TrialNet Pathway to Prevention (TN01)

**Prepared by NIDDK-CR**
**February 17, 2022**

# 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

TrialNet is a network of 19 clinical centers worldwide that conduct research on the study, prevention, and early treatment of type 1 diabetes (T1D). The TrialNet Pathway To Prevention study (TN01) is a screening and monitoring study that was established to provide a source of subjects for enrollment into prevention trials. In addition, the prospectively followed cohort is used to gain new insights into the natural history of pre-type 1 diabetes in patients with increased risk for the disease. The primary aim of the study is to identify subjects for TrialNet prevention and investigation trials, but can also be used for assessing the predictive value of existing and novel risk markers of T1D, and for examining the demographic, immunologic, and metabolic characteristics of individuals at risk for developing T1D.

# 3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the TN01 folder in the data package. For this replication, variables were taken from the tn01_nh04_baselineriskassess.sas7bdat, tn01_researchlabsbaa_cy2004_2012.sas7bdat, tn01_researchlabsbaa_cy2013_2018.sas7bdat, and tn01_nh01_screening.sas7bdat datasets.

# 4 Statistical Methods

Analyses were performed to replicate results for the data published by Jacobsen et al. [1] for The Risk of Progression to Type 1 Diabetes is Highly Variable in Individuals with Multiple Autoantibodies Following Screening. To verify the integrity of the data, descriptive statistics were computed.

# 5 Results

For Table 1 in the publication [1], Baseline characteristics of mAB-positive 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. A list of the 1815 participants included in this publication were provided to the NIDDK Central Repository by the DCC, however, the exact breakdown of mAB-positive participants was not known. The results of the replication vary from the publication due to the nature of the data and the limited guidelines outlined in the publication.

# 6 Conclusions

The NIDDK Central Repository worked with the DCC to ensure that the TN01 data files to be distributed are a true copy of the study data.

# 7 References

[1] Jacobsen LM, Bocchino L, Evans-Molina C, DiMeglio L, Goland R, Wilson DM, Atkinson MA, Aye T, Russell WE, Wentworth JM, Boulware D, Geyer S, Sosenko JM. The Risk of Progression to Type 1 Diabetes is Highly Variable in Individuals with Multiple Autoantibodies Following Screening. Diabetologia, 63(3), 588-596, March 2020. doi: https://doi.org/10.1007/s00125-019-05047-w

**Table A**: Variables used to replicate Table 1 – Baseline characteristics of mAB-positive participants

| **Table Variable** | **dataset.variable** |
|---|---|
| Age, years | tn01_nh01_screening.age<br>tn01_researchlabsbaa_cy2004_2012.visit<br>tn01_researchlabsbaa_cy2013_2018.visit |
| BMI, $kg/m^2$ | tn01_nh04_baselineriskassess.heightcm<br>tn01_nh04_baselineriskassess.weightkg<br>tn01_researchlabsbaa_cy2004_2012.visit<br>tn01_researchlabsbaa_cy2013_2018.visit |
| Sex, n (%) | tn01_nh01_screening.sex<br>tn01_researchlabsbaa_cy2004_2012.visit<br>tn01_researchlabsbaa_cy2013_2018.visit |
| Antibodies, n (%) | tn01_researchlabsbaa_cy2004_2012.result<br>tn01_researchlabsbaa_cy2004_2012.test_name<br>tn01_researchlabsbaa_cy2004_2012.visit<br>tn01_researchlabsbaa_cy2013_2018.result<br>tn01_researchlabsbaa_cy2013_2018.test_name<br>tn01_researchlabsbaa_cy2013_2018.visit |

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

| Variable | Manuscript 2Ab+ (n=804) | DSIC 2Ab+ (n=728) | Diff. (n=76) | Manuscript >2Ab+ (n=1011) | DSIC >2Ab+ (n=1071) | Diff. (n=60) |
|---|---|---|---|---|---|---|
| Age, years | 13.29 ± 10.44 | 11.65 ± 10.31 | 1.64 ± 0.13 | 11.61 ± 8.39 | 11.36 ± 8.93 | 0.25 ± 0.54 |
| BMI, kg/m$^2$ | 19.86 ± 5.72 | 19.51 ± 5.33 | 0.35 ± 0.39 | 19.49 ± 5.12 | 19.59 ± 5.34 | 0.10 ± 0.22 |
| Sex, n (%) | | | | | | |
| Female | 381 (47.74) | 387 (48.62) | 6 (0.88) | 455 (45.14) | 449 (44.46) | 6 (0.68) |
| Male | 417 (52.26) | 409 (51.38) | 8 (0.88) | 553 (54.86) | 561 (55.54) | 8 (0.68) |
| Antibodies, n (%) | | | | | | |
| GADA | | | | | | |
| Negative | 122 (15.17) | 316 (43.11) | 194 (27.94) | 44 (4.35) | 47 (4.41) | 3 (0.06) |
| Positive | 682 (84.83) | 417 (56.9) | 265 (27.93) | 967 (95.65) | 1019 (95.59) | 52 (0.06) |
| IA-2A | | | | | | |
| Negative | 586 (72.89) | 539 (74.04) | 47 (1.15) | 164 (16.22) | 294 (27.45) | 130 (11.23) |
| Positive | 218 (27.11) | 189 (25.96) | 29 (1.15) | 847 (83.78) | 777 (72.55) | 70 (11.23) |
| IAA | | | | | | |
| Negative | 414 (51.40) | 522 (71.80) | 108 (20.40) | 320 (31.65) | 381 (35.57) | 61 (3.92) |
| Positive | 390 (48.51) | 205 (28.20) | 185 (20.31) | 691 (68.35) | 690 (64.43) | 1 (3.92) |
| ZnT8A | | | | | | |
| Negative | 486 (60.45) | 308 (42.31) | 178 (18.14) | 107 (10.58) | 100 (9.34) | 7 (1.24) |
| Positive | 318 (39.55) | 420 (57.70) | 102 (18.15) | 904 (89.42) | 971 (90.66) | 67 (1.24) |

## Attachment A: SAS Code

```
libname dsic "X:\NIDDK\niddk-dr_studies6\TrialNet_01\private_created_data\Redacted Data";
libname mid "X:\NIDDK\niddk-dr_studies6\TrialNet_01\private_created_data";

/*********************************/
/*   TN01 DSIC Jacobson et. al.   */
/*********************************/

*calling the screening dataset;
data screen; set dsic.tn01_NH01_screening;
run;

*calling the baseline risk assessment dataset;
data base; set dsic.tn01_nh04_baselineriskassess;
run;

*calling the MID list dataset;
data mid; set mid.mid_list;
run;

*Calling the Research Labs datasets;
data reslabs1; set dsic.tn01_researchlabsbaa_cy04_12;
run;

data reslabs2; set dsic.tn01_researchlabsbaa_cy13_20;
run;

data reslabs3; set dsic.tn01_researchlabs_other;
run;

proc print data=dsic.tn01_locallababresults;
run;

*Combining the two reslabs datasets;
data reslabs; set reslabs1 reslabs2;
run;

proc contents data=reslabs;
run;

proc freq data=reslabs;
tables Result_Type visit;
run;

*sorting the MID dataset and reslabs datset by MaskID;
proc sort data=mid;
```

```
by MaskID;
run;

proc sort data=reslabs;
by MaskID;
run;

*merging the MID dataset with the reslabs dataset;
data ABLabs;
merge
mid (in=a)
reslabs (in=b);
by maskid;
if a=1;
run;

proc freq data=ABlabs;
tables visit;
run;

*removing observations with missing results;
proc contents data=ablabs;
run;

proc freq data=ablabs;
tables result;
run;

data ABlabs_1; set ABlabs;
if result = "" OR result = "ins-uninhib" /*OR result = ">25000"*/ then delete;
if result = ">25000" then result = 25000;
run;

*converting result from character to numeric;
data ABlabs_2; set ABlabs_1;
result_n = input(result, 8.);
drop result;
rename result_n=result;
if test_name = "IA-2H" then test_name = "IA2H";
run;

proc freq data=ablabs_2;
tables visit MaskID Test_Name;
run;

*Limiting to only screening visits;
data ABlabs_3; set ABlabs_2;
where visit = "Screening";
```

```
run;

*issues with duplicate tests per MaskID with different results, keeping the latest date;
*ZNT8;
data znt8; set ablabs_3;
where Test_Name = "ZNT8";
run;

proc contents data=znt8;
run;

proc sort data=znt8;
by MaskID  Date_at_Test_Results_Reported;
run;

data znt8_1; set znt8;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*GAD65;
data gad65; set ablabs_3;
where Test_Name = "GAD65";
run;

proc contents data=gad65;
run;

proc sort data=gad65;
by MaskID  Date_at_Test_Results_Reported;
run;

data gad65_1; set gad65;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*gad65H;
data gad65h; set ablabs_3;
where Test_Name = "GAD65H";
run;
```

```
proc contents data=gad65h;
run;

proc sort data=gad65h;
by MaskID  Date_at_Test_Results_Reported;
run;

data gad65h_1; set gad65h;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*ICA512;
data ica512; set ablabs_3;
where Test_Name = "ICA512";
run;

proc contents data=ICA512;
run;

proc sort data=ICA512;
by MaskID  Date_at_Test_Results_Reported;
run;

data ICA512_1; set ICA512;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*IA2H;
data ia2h; set ablabs_3;
where Test_Name = "IA2H";
run;

proc contents data=ia2h;
run;

proc sort data=ia2h;
by MaskID  Date_at_Test_Results_Reported;
run;

data ia2h_1; set ia2h;
```

```
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*ICA;
data ica; set ablabs_3;
where Test_Name = "ICA";
run;

proc contents data=ica;
run;

proc sort data=ica;
by MaskID  Date_at_Test_Results_Reported;
run;

data ica_1; set ica;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*MIAA;
data MIAA; set ablabs_3;
where Test_Name = "MIAA";
run;

proc contents data=miaa;
run;

proc sort data=miaa;
by MaskID  Date_at_Test_Results_Reported;
run;

data miaa_1; set miaa;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*combining back into a single dataset;
```

```
data ABlabs_4; set znt8_1 gad65_1 gad65h_1 ica512_1 ica_1 miaa_1 ia2h_1;
run;

/*
*Combining the new Znt8 back in with ABlabs;
data ablabs_4; set ablabs_3;
if Test_Name = "ZNT8" then delete;
run;

data ABlabs_5; set ablabs_4 znt8_1;
run;
*/

*transposing the dataset from long to wide;
data ABlabs_5; set ABLabs_4;
keep MaskID Result Test_Name;
run;

/*
data ablabs_7; set ablabs_6;
if test_name = "IA-2H" then test_name = "IA2H";
run;
*/

proc sort data=ABlabs_5 nodup;
by MaskID;
run;

proc transpose data=ABlabs_5 out=ABlabs_6 prefix=Test_;
by MaskID;
id Test_Name;
var Result;
run;

*identifying positive AB results;
data ABlabs_7; set ABlabs_6;
if Test_MIAA > 0.009 then miaapos = 1; else miaapos = 0;
if Test_ZNT8 > 0.014 then znpos   = 1; else znpos   = 0;
if Test_GAD65 > 0.031 or Test_GAD65H > 19 then gadpos = 1; else gadpos = 0;
if Test_IA2H > 4 OR Test_ICA512 > 0.048 OR Test_ICA > 9 then iapos = 1; else iapos = 0;
run;

/*
data ABlabs_9; set ABlabs_8;
if Test_MIAA >= 0.010 then miaapos = 1; else miaapos = 0;
if Test_ZNT8 >= 0.015 then znpos   = 1; else znpos   = 0;
if Test_GAD65 >= 0.032 or Test_GAD65H >= 20 then gadpos = 1; else gadpos = 0;
if Test_IA2H >= 5 OR Test_ICA512 >= 0.049 OR Test_ICA >= 10 then iapos = 1; else iapos = 0;
```

```
run;

*total AB+;
data ABlabs_10; set ABlabs_7;
ABpos = 0;
array AB[*] miaapos znpos gadpos iapos;

        do i=1 to 4;
                IF AB[i] = 1 then ABpos = ABpos +1;
        end;
 drop i;
run;

proc freq data=ABlabs_10;
tables ABpos;
run;

*creating a flag variable for 2 vs 3/4 AB+;
data ABlabs_11; set Ablabs_10;
if ABpos <= 2 then ABflag = 0; else ABflag = 1;
run;

proc freq data=ABlabs_11;
tables ABflag;
run;

*Merging the screening dataset witht the finalized ABlabs dataset;
proc sort data=ablabs_11;
by maskid;
run;

proc sort data=screen;
by MaskID;
run;

data demo; merge
ablabs_11 (in=a)
screen   (in=b);
by maskid;
if a=1;
run;

/*************************/
/*       AGE         */
/*************************/

proc sort data=demo;
by ABflag;
```

```
run;

proc means data=demo n mean std;
var age;
by ABflag;
run;

/****************/
/*    Sex     */
/****************/

proc freq data=demo;
tables sex*ABflag/norow nopercent;
run;

/********************/
/*   BMI         */
/********************/

proc sort data=base;
by maskid;
run;

proc sort data=demo;
by MaskID;
run;

data bmi;
merge
base (in=a)
demo (in=b);
by maskid;
if b=1;
run;

proc contents data=bmi;
run;

data bmi_1; set bmi;
height_m = (heightcm/100);
run;

proc means data=bmi_1;
var height_m;
run;

data bmi_2; set bmi_1;
heightsq = (height_m*height_m);
```

```
bmi = (weightkg)/(heightsq);
run;

proc sort data=bmi_2;
by ABflag;
run;

proc means data=bmi_2 mean std;
var bmi;
by ABflag;
run;

/*****************/
/*  Antibodies   */
/*****************/
proc contents data=ablabs_2;
run;

proc freq data=demo;
tables (gadpos iapos miaapos znpos)*ABflag/norow nopercent;
run;

*ZNT8;
data znt8_2; set ablabs_2;
where Test_Name = "ZNT8";
run;

data znt8_3; set znt8_2;
znt8pos = 0;
*if result > 0.014 then znt8pos = 1; else znt8pos = 0;
run;

proc freq data=znt8_3;
tables znt8pos;
run;

proc sort data=znt8_3;
by MaskID znt8pos;
run;

data znt8_4; set znt8_3;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.maskID then output;
run;
```

```
*GAD65;
data gad65_2; set ablabs_3;
where Test_Name = "GAD65";
run;

data gad65_3; set gad65_2;
gadpos1 = 0;
if result > 0.032 then gadpos1 = 1; else gadpos1 = 0;
run;

proc freq data=gad65_3;
tables gadpos1;
run;

proc sort data=gad65_3;
by MaskID gadpos1;
run;

data gad65_4; set gad65_3;
by maskid;
retain N;
if first.maskid then N=1;
else N = N+1;
if last.maskid then output;
run;

*gad65H;
data gad65h_2; set ablabs_3;
where Test_Name = "GAD65H";
run;

data gad65h_3; set gad65_2;
gadpos2 = 0;
if result > 0.032 then gadpos2 = 1; else gadpos2 = 0;
run;

proc freq data=gad65h_3;
tables gadpos2 result;
run;

proc sort data=gad65h_3;
by MaskID  gadpos2;
run;

data gad65h_4; set gad65h_3;
by MaskID;
retain N;
if first.MaskID then N=1;
```

```
else N = N+1;
if last.MaskID then output;
run;

*ICA512;
data ica512_2; set ablabs_3;
where Test_Name = "ICA512";
run;

data ica512_3; set ica512_2;
ia2pos1 = 0;
if result > 0.049 then ia2pos1 = 1; else ia2pos1 = 0;
run;

proc sort data=ICA512_3;
by MaskID ia2pos1;
run;

data ICA512_4; set ICA512_3;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*IA2H;
data ia2h_2; set ablabs_3;
where Test_Name = "IA2H";
run;

data ia2h_3; set ia2h_2;
ia2pos2 = 0;
if result > 5 then ia2pos2 = 1; else ia2pos2 = 0;
run;

proc sort data=ia2h_3;
by MaskID  ia2pos2;
run;

data ia2h_4; set ia2h_3;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;
```

```
*ICA;
data ica; set ablabs_3;
where Test_Name = "ICA";
run;

proc contents data=ica;
run;

proc sort data=ica;
by MaskID  Date_at_Test_Results_Reported;
run;

data ica_1; set ica;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

*MIAA;
data MIAA_2; set ablabs_3;
where Test_Name = "MIAA";
run;

data MIAA_3; set MIAA_2;
miaapos = 0;
if result >0.010 then miaapos = 1; else miaapos = 0;
run;

proc sort data=miaa_3;
by MaskID  miaapos;
run;

data miaa_4; set miaa_3;
by MaskID;
retain N;
if first.MaskID then N=1;
else N = N+1;
if last.MaskID then output;
run;

data znt8_5; set znt8_4;
keep MaskID znt8pos;
run;

data gad65_5; set gad65_4;
keep MaskID gadpos1;
```

```
run;

data ica512_5; set ica512_4;
keep MaskID ia2pos1;
run;

data gad65h_5; set gad65h_4;
keep MaskID gadpos2;
run;

data ia2h_5; set ia2h_4;
keep MaskID ia2pos2;
run;

data miaa_5; set miaa_4;
keep MaskID miaapos;
run;

*merge all datasets;
data abtabs;
merge
gad65_5 (in=a)
ica512_5 (in=b)
gad65h_5 (in=c)
ia2h_5 (in=d)
miaa_5 (in=e)
znt8_5 (in=f);
by maskid;
run;

data abtabs1; set abtabs;
if gadpos1 = 1 OR gadpos2 = 1 then gadpos = 1; else gadpos = 0;
if ia2pos1 = 1 OR ia2pos2 = 1 then ia2pos = 1; else ia2pos = 0;
drop gadpos1 gadpos2 ia2pos1 ia2pos2;
run;

*flag variable from above;
data ablabs_11_1; set ablabs_11;
keep MaskID ABflag;
run;

proc sort data=ablabs_11_1;
by maskid;
run;

data abtabs2;
merge
abtabs1 (in=a)
```

```
ablabs_11_1;
by maskid;
run;

proc freq data=abtabs2;
tables ia2pos*ABflag;
run;

proc freq data=abtabs2;
tables miaapos*ABflag;
run;

proc freq data=abtabs2;
tables znt8pos*ABflag;
run;
```