Dataset Integrity Check for the TrialNet (10) Data Files

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
June 12, 2020

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 objective of this study was to describe a pilot trial of using an omega-3 fatty acid (docosahexaenoic acid [DHA]) to prevent islet cell autoimmunity in infants with an increased risk for developing type 1 diabetes (T1D). Infants from pregnant mothers who either have T1D (or the father or a previous child has T1D) and who entered the study in the third trimester or infants younger than age 5 months having a first-degree family member with T1D were eligible for the study. Infants from either group also had to have an increased genetic (HLA) risk for T1D (or multiple first-degree relatives with T1D) to be eligible. The study is a multicenter, 2-arm, randomized, double-masked clinical trial that will last 4 years (1 year of recruitment and 3 years of treatment). Treatment with DHA (or control) began in the last trimester of pregnancy or in the first 5 months after birth. Inflammatory mediators, including cytokines, chemokines, eicosanoids, and C-reactive protein, are being measured along with fatty acids in maternal and infant blood. Ninety-eight infants were enrolled (41 during pregnancy and 57 in the 5 months after birth). HLA results of the 97 eligible infants (1 infant had a protective 0602 allele and was thus ineligible) showed that 90 have DR3 and/or DR4. Seven infants were enrolled without DR3/4 but who instead had multiple first-degree relatives with T1D. Compliance has been excellent, and no families have discontinued participation. Intervention trials in this high-risk group are feasible but require significant effort to identify potential participants.

3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the data folder in the data package. For this replication, variables were taken from the “mastable.sas7bdat”, “autoab.sas7bdat”, “relation.sas7bdat”, “hba1c.sas7bdat”, and “cbc.sas7bdat” datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Herold, et al in The New England Journal of Medicine on June 2019 [1]. To verify the integrity of the datasets, descriptive statistics of baseline characteristics were computed, by entry group (Table B).

5 Results

For Table 1 in the publication [1], Baseline Characteristics of the Participants, Table A lists the variables that were used in the replication and Table B compares the results calculated from the archived data file to the results published in Table 1. The results of the replication are very similar to the published results.

6 Conclusions

The results of the replication are almost an exact match to the published results.

7 References

[1] Kevan C. Herold, M.D., Brian N. Bundy, Ph.D., S. Alice Long, Ph.D., Jeffrey A. Bluestone, Ph.D., Linda A. DiMeglio, M.D., Matthew J. Dufort, Ph.D., Stephen E. Gitelman, M.D., Peter A. Gottlieb, M.D., Jeffrey P. Krischer, Ph.D., Peter S. Linsley, Ph.D., Jennifer B. Marks, M.D., Wayne Moore, M.D., Ph.D., Antoinette Moran, M.D., Henry Rodriguez, M.D., William E. Russell, M.D., Desmond Schatz, M.D., Jay S. Skyler, M.D., Eva Tsalikian, M.D., Diane K. Wherrett, M.D., Anette‑Gabriele Ziegler, M.D., and Carla J. Greenbaum, M.D., for the Type 1 Diabetes TrialNet Study Group. An Anti-CD3 Antibody, Teplizumab, in Relatives at Risk for Type 1 Diabetes. New England Journal of Medicine June 2019. N Engl J Med 2019; 381:603-613.

Table A: Variables used to replicate Table 1: Baseline Characteristics of the Participants.

Table Variable	**dataset.variable**
Treatment description	mastable.rxdesc
sex	mastable.sex
age	mastable.age
Anti-GAD65, harmonized	autoab.gad65h_pos
Micro insulin	autoab.miaa_pos
Anti–IA-2, harmonized	autoab.ia2h_pos
ICA	autoab.ica_pos
Anti-ZnT8	autoab.znt8_pos
relationship	relation.relation
glycated hemoglobin level	Hba1c.hba1c

Table B-1 Baseline interquartile range, minimum and maximum values**:** Comparison of values computed in integrity check to reference article Table 1 values.

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Arm	**Characteristic**	**Median**			**Q1**			**Q3**			**Min**			**Max**		
TEPLIZUMAB	Age — yr	14	14	0	12	12	0	22	22	0	8.5	8.5	0	49.5	49.5	0
	Median glycated hemoglobin level (IQR) — %	5.2	5.2	0	4.9	4.9	0	5.4	5.4	0						
PLACEBO	Age — yr	13	13	0	11	11	0	16	16	0	8.6	8.6	0	45	45	0
	Median glycated hemoglobin level (IQR) — %	5.3	5.3	0	5.1	5.1	0	5.4	5.4	0						

Table B-2 Baseline counts and percentages**:** Comparison of values computed in integrity check to reference article Table 1 values.

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Arm	**Characteristic**	**N**			**Percent**		
TEPLIZUMAB	Age <18 yr — no. (%)	29	29	0	66	66	0
	Male sex — %				57	57	0
	Relationship to person with type 1 diabetes — no. (%)						
	Sibling†	28	28	0	64	64	0
	Offspring	6	6	0	14	14	0
	Parent	6	6	0	14	14	0
	Sibling and another first-degree relative	2	2	0	5	5	0
	Second-degree relative	2	2	0	5	5	0
	Third-degree relative or further removed	0	0	0	0	0	0
	Autoantibodies — no. of participants positive (%)‡						
	Anti-GAD65, harmonized	40	40	0	91	91	0
	Micro insulin	20	20	0	45	45	0
TEPLIZUMAB	Anti–IA-2, harmonized	27	27	0	61	61	0
	ICA	29	28	1	66	64	2
	Anti-ZnT8	32	29	3	73	66	7
PLACEBO	Age <18 yr — no. (%)	26	26	0	81	81	0
	Male sex — %				53	53	0
	Relationship to person with type 1 diabetes —						
	no. (%)						
	Sibling†	16	16	0	50	50	0
	Offspring	6	6	0	19	19	0
	Parent	3	3	0	9	9	0
	Sibling and another first-degree relative	3	3	0	9	9	0
	Second-degree relative	3	3	0	9	9	0
	Third-degree relative or further removed	1	1	0	3	3	0
	Autoantibodies — no. of participants positive						
	(%)‡						
	Anti-GAD65, harmonized	28	28	0	88	88	0
	Micro insulin	11	11	0	34	34	0
	Anti–IA-2, harmonized	24	24	0	75	75	0
	ICA	28	28	0	88	88	0
	Anti-ZnT8	24	24	0	75	75	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TrialNet_10/prog_initial_analysis/trialnet10.dsic.20200527.sas';
run;

libname sasfile1
'/prj/niddk/ims_analysis/TrialNet_10/private_orig_data/Data.Extraction.10.ProventionBio.Version_A
nalysis191120PW/10/sasv9';

proc format;
 value val
 .     = "no value"
 other = "   value"
 ;

 value oneplus
 . = "no value"
 0 = "0"
 1-high = "1+"
 ;

 value zerohi
 . = "no value"
 0-high = "0-high"
 ;

 value sexf
 1= "Female"
 2= "Male"
 ;

 value relation
 1 = '1 Sibling'
 2 = '2 Offspring'
 3 = '3 Parent'
 4 = '4 Sibling and another first-degree relative'
 5 = '5 Second-degree relative'
 6 = '6 Third-degree relative or further'
 ;

 value posneg
 0 = "neg"
 1 = "pos"
 ;

run;

%macro readin(lib, ds);
  data &ds;
    set sasfile&lib..&ds;
  run;

  proc contents data=&ds;
  title3 "&ds";
  run;
%mend;

* produce n and %;
%macro npercent(rownum, var, varf, subset, subsetname);
  proc freq data=analy noprint;
    where &subset = 1;
    tables &var/list missing out=tbl1&subsetname;
    format &var &varf..;
  run;
  data tbl1&subsetname;
    length covar covarf $100;
    set tbl1&subsetname;
    covar = "&var";
    covarf = put(&var,&varf..);
    rownum = &rownum;
  run;

  data prnt&subsetname;
    set prnt&subsetname tbl1&subsetname;
  run;

%mend;

%macro univ(rownum, var, subset, subsetname);

  proc univariate data=analy outtable= univ&subsetname noprint;
    where &subset=1 and &var not in(.,0);
    var &var
       ;
  run;

  data univ&subsetname;
    length covarf $100;
    set univ&subsetname;
        covarf = "&subset";
        rownum = &rownum;
  run;

  data prntuniv&subsetname;
    set prntuniv&subsetname univ&subsetname;
  run;

%mend;

%readin(1, autoab);
%readin(1, relation);
%readin(1, mastable);
%readin(1, cbc);
%readin(1, hba1c);

proc freq data=hba1c;
  tables HBA1C/missing;
  run;

proc freq data=mastable;
  tables rxdesc/missing;
  run;

** pull together table 1 vars;
proc sort data=mastable;
  by maskid;
  run;

data dups1;
  set mastable;
  by MASKID;
  if not (first.MASKID and last.MASKID);
run;

proc sort data=autoab;
  by maskid DATEOFDRAW;
  run;

data dups2;
  set autoab;
  by MASKID;
  if not (first.MASKID and last.MASKID);
run;
```

```
proc print data=autoab (obs=20);
  by maskid;
  id maskid;
  run;


* "To get the correct timepoint you would need to cross reference the treatment date in the
mastable against the closest previous draw date in the autoab table.";
proc sort data=mastable;
  by maskid TREATMENTDATE;
run;

data autoab (keep=dateofdraw  datesampleevaluation  gad65h      gad65h_pos  ia2h        ia2h_pos
ica         ica_pos     maskid      miaa        miaa_pos    znt8        znt8_pos
                 treatmentdate deltadts);
  merge autoab (in=in1) mastable (in=in2 keep=maskid TREATMENTDATE);
  by maskid;
  if in1;

  * delta days where draw date is before treatment date;
  if dateofdraw < treatmentdate then deltadts = dateofdraw - treatmentdate;

run;

proc sort data=autoab;
  where deltadts ne .;
  by maskid deltadts;
run;

data autoab_sel;
  set autoab;
  by maskid;
  if last.maskid;
run;

data autoab (keep=dateofdraw  datesampleevaluation  gad65h      gad65h_pos  ia2h        ia2h_pos
ica         ica_pos     maskid      miaa        miaa_pos    znt8        znt8_pos
                 treatmentdate deltadts randomtp);
  merge autoab (in=in1) autoab_sel (in=in2 keep=maskid dateofdraw);
  by maskid dateofdraw;
  if in1;
  if in2 then randomtp=1;
run;

/*
proc print data=autoab;
  by maskid;
  id maskid;
  var treatmentdate dateofdraw deltadts randomtp;
  format treatmentdate dateofdraw mmddyy10.;
run;
*/

proc sort data=autoab out=autoab_sel;
  where randomtp=1;
  by maskid;
run;

proc sort data=relation;
  by maskid ;
  run;

data dups3;
  set relation;
  by MASKID;
  if not (first.MASKID and last.MASKID);
run;
```

```
proc freq data=relation;
  table relation/missing;
  run;

data relationper;
  set relation;
  by maskid;

  retain sibling offspring parent second third;

  if first.maskid then do;
    sibling = 0;
    offspring = 0;
    parent = 0;
    second = 0;
    third = 0;
  end;

  if substr(relation,1,2) in('FS','IT','NT') then sibling = sum(sibling,1);
  if substr(relation,1,2) in('CH') then Offspring = sum(offspring,1);
  if substr(relation,1,2) in('P=') then Parent = sum(parent,1);
  if substr(relation,1,2) in('AU','HS','GP') then second = sum(second,1);
  if substr(relation,1,2) in('C=') then third = sum(third,1);

  if last.maskid then do;
     *Sibling and another first-degree relative  ;
     if sibling > 0 and (max(offspring, parent) > 0) then relationshipt1d = 4;
     else if sibling > 0 then relationshipt1d = 1;
     else if offspring > 0 then relationshipt1d = 2;
     else if parent > 0 then  relationshipt1d = 3;
     else if second > 0 then  relationshipt1d = 5;
     else if third > 0 then relationshipt1d = 6;
     output;
  end;
run;

/*
proc print data=relationper;
  by maskid;
  id maskid;
  var relation sibling offspring parent siblingfirst second;
run;
*/



proc sort data=hba1c;
  by maskid DATEOFDRAW;
  run;

data dups4;
  set hba1c;
  by MASKID;
  if not (first.MASKID and last.MASKID);
run;

data hba1c_bl;
  set hba1c;
  by maskid;
  if first.maskid;
run;


data analy;
  merge mastable autoab_sel relationper hba1c_bl;
  by maskid;

   * create subset flag for each row to use in macro call;
   all = 1;
   * subsets;
   if rxdesc = "Placebo" then placebo=1;
   if rxdesc = "Teplizu" then teplizu=1;

   if sex = "Female" then sexnum=1;
   else if sex = "Male" then sexnum=2;

   if . < age < 18 then age_lt18 = 1;
run;


proc contents data = analy;
title3 'trialnet 10';
run;

* check for dups;
proc sort data=analy;
  by MASKID;
run;

data dups;
  set analy;
  by MASKID;
  if not (first.MASKID and last.MASKID);
run;

data _null_;
  set dups;
  abort;
run;

proc freq data=analy;
  tables age_lt18*age/list missing;
  tables RXDESC relation  sex
 GAD65H_POS
 IA2H_POS
 ICA_POS
 MIAA_POS
 ZNT8_POS
 HBA1C/missing;
 tables relationshipt1d*sibling* offspring* parent* second *third/list missing;
  title3 'check new groupings';
run;

* med, q1, q3;
data prntunivplacebo;
*  length _VAR_ $100;
  set _null_;
run;

%univ(1   , age        , placebo , placebo);
%univ(1   , HBA1C        , placebo , placebo);

data prntunivteplizu;
*  length _VAR_ $100;
  set _null_;
run;

%univ(1   , age        , teplizu , teplizu);
%univ(1   , HBA1C        , teplizu , teplizu);

data alluniv;
  set prntunivplacebo     (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ _min_
 _max_)
      prntunivteplizu     (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ _min_
 _max_)
  ;
  _median_   = round(_median_, .1 );
  _q1_       = round(_q1_    , .1 );
  _q3_      = round(_q3_    , .1 );
  _min_      = round(_min_, .1);
  _max_     = round(_max_, .1);
run;

proc sort data=alluniv;
  by rownum;
  run;

proc print data= alluniv noobs;
    var rownum _var_  covarf _nobs_ _median_ _q1_ _q3_ _min_ _max_ /*_std_*/;
    title3 "Table 1 - median, q1, q3 for each subset";
run;



* n and percent;
data prntplacebo;
  set _null_;
run;

%npercent(1,  age_lt18        , val   , placebo , placebo);
%npercent(1,  SEXnum          , sexf  , placebo , placebo);
%npercent(2,  relationshipt1d , relation, placebo , placebo);
%npercent(7,  gad65h_pos      , posneg  , placebo , placebo);
%npercent(8,  miaa_pos        , posneg  , placebo , placebo);
%npercent(9,  ia2h_pos        , posneg  , placebo , placebo);
%npercent(10, ica_pos         , posneg  , placebo , placebo);
%npercent(11, znt8_pos        , posneg  , placebo , placebo);


data prntteplizu;
  set _null_;
run;

%npercent(1,  age_lt18        , val   , teplizu , teplizu);
%npercent(1,  SEXnum          , sexf  , teplizu , teplizu);
%npercent(2,  relationshipt1d , relation, teplizu , teplizu);
%npercent(7,  gad65h_pos      , posneg  , teplizu , teplizu);
%npercent(8,  miaa_pos        , posneg  , teplizu , teplizu);
%npercent(9,  ia2h_pos        , posneg  , teplizu , teplizu);
%npercent(10, ica_pos         , posneg  , teplizu , teplizu);
%npercent(11, znt8_pos        , posneg  , teplizu , teplizu);


* Table 1;
data npercent;
  length subgroup $10;
  set prntplacebo (in=in1) prntteplizu (in=in2);
  if in1 then subgroup = "placebo";
  if in2 then subgroup = "teplizu";

  percent = round(percent);
run;

proc sort data=npercent;
  by subgroup rownum covarf;
run;

proc print data=npercent;
  var rownum subgroup covar covarf count percent;
  title3 "Table 1 - n, percent";
run;
```