Dataset Integrity Check for the PALF Data Files

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
September 11, 2019

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

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 “he.sas7bdat”, “supp_tests.sas7bdat”, and “si_diag.sas7bdat” datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Narkewicz, et al in Pediatr Gastroenterol Nutr. on Feb 2017 [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], Demographics and liver biopsy status by autoAB status and diagnostic group, 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 an exact match to the published results.

7 References

[1] Michael R. Narkewicz, MD, Simon Horslen, MB ChB, Steven H. Belle, PhD MScHyg, David A. Rudnick, MD, PhD, Vicky L. Ng, MD, Philip Rosenthal, MD, Rene Romero, MD, Kathleen M. Loomes, MD, Song Zhang, MS, Regina M Hardison, MS, and Robert H. Squires, MD for the Pediatric Acute Liver Failure Study Group. Pediatr Gastroenterol Nutr 2017 February ; 64(2): 210–217.

Table A: Variables used to replicate Table 1: Demographics and liver biopsy status by autoAB status and diagnostic group.

Table Variable	**dataset.variable**
Antinuclear Ab (ANA)	supp_tests.supana
Anti-Smooth Muscle Ab (SMA) (Pos/Neg)	supp_tests.supasma
Liver/Kidney Microsome Type 1 Ab (LKM)	supp_tests.supalkm
Antinuclear Ab (ANA)	he.ana
Anti-Smooth Muscle Ab (SMA) (Pos/Neg)	he.asma
Liver/Kidney Microsome Type 1 Ab (LKM)	he.alkm
Race American Indian or Alaska Native	he.racaa
Race asian	he.racea
Race black	he.raceb
Race other	he.raceo
Race Native Hawaiian or Pacific Islander	he.racnp
Race White	he.racew
Liver Biopsy	he.lbx
Liver Transplant Date	he.trpdate
Liber Biopsy Date	he.lbxdate
Sex	he.sex
Hispanic	he.hsp
Final Diag	si_diag.fdiag
Enrollment Date	he.erldate
DOB	he.dobdate

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

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	Characteristic	**Median**			**Q1**		
AIH AutoAB+	Age	9.9	9.9	0	3.3	3.3	0
Indeterminate AutoAB+	Age	5.6	5.6	0	2.0	2.0	0
Indeterminate AutoAB-	Age	4.5	4.5	0	1.5	1.5	0
Other AutoAB+	Age	9.6	9.6	0	4.6	4.6	0
Other AutoAB-	Age	7.8	7.8	0	1.2	1.2	0
	Characteristic	**Q3**			**Min**		
AIH AutoAB+	Age	13.9	13.9	0	.6	.6	0
Indeterminate AutoAB+	Age	10.3	10.3	0	.1	.1	0
Indeterminate AutoAB-	Age	11.1	11.1	0	0	0	0
Other AutoAB+	Age	14.6	14.6	0	0	0	0
Other AutoAB-	Age	14.8	14.8	0	0	0	0
	Characteristic	**Max**					
AIH AutoAB+	Age	17.5	17.5	0			

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Indeterminate AutoAB+	Age	17.4	17.4	0			
Indeterminate AutoAB-	Age	17.9	17.9	0			
Other AutoAB+	Age	17.9	17.9	0			
Other AutoAB-	Age	17.9	18.0	.1			
		N			Percent		
AIH AutoAB+	**Sex**						
	Male	30	30	0	47.6	47.6	0
	Female	33	33	0	52.4	52.4	0
	Race						
	Unknown	1	1	0			
	White	45	45	0	72.6	72.6	0
	Other	17	17	0	27.4	27.4	0
	Hispanic						
	No	45	45	0	71.4	71.4	0
	Yes	18	18	0	28.6	28.6	0
	Liver Biopsy						
	Unknown	2	2	0			
	Not Done	19	19	0	31.1	31.1	0
	Prior to LTx	39	39	0	63.9	63.9	0

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	Explant	3	3	0	4.9	4.9	0
		N			Percent		
Indeterminate AutoAB+	**Sex**						
	Male	34	34	0	45.3	45.3	0
	Female	41	41	0	54.7	54.7	0
	Race						
	Unknown	2	2	0			
	White	53	53	0	72.6	72.6	0
	Other	20	20	0	27.4	27.4	0
	Hispanic						
	No	55	55	0	73.3	73.3	0
	Yes	20	20	0	26.7	26.7	0
	Liver Biopsy						
	Unknown	3	3	0			
	Not Done	36	36	0	50	50	0
	Prior to LTx	31	31	0	43.1	43.1	0
	Explant	5	5	0	6.9	6.9	0

		N			Percent		
Indeterminate AutoAB-	**Sex**						
	Male	171	171	0	58.4	58.2	0.2
	Female	123	123	0	41.8	41.8	0
	Race						
	Unknown	3	3	0			
	White	212	212	0	72.9	72.9	0
	Other	79	79	0	17.1	17.1	0
	Hispanic						
	No	241	241	0	82	82	0
	Yes	53	53	0	18	18	0
	Liver Biopsy						
	Unknown	18	18	0			
	Not Done	147	147	0	53.3	53.3	0
	Prior to LTx	109	109	0	39.5	39.5	0
	Explant	20	20	0	7.3	7.3	0
		N			Percent		
Other AutoAB+	**Sex**						
	Male	27	27	0	42.2	42.2	0
	Female	37	37	0	57.8	57.8	0
	Race						

	Unknown	2	2	0			
	White	41	41	0	66.1	66.1	0
	Other	21	21	0	33.9	33.9	0
	Hispanic						
	No	53	53	0	82.8	82.8	0
	Yes	11	11	0	17.2	17.2	0
	Liver Biopsy						
	Unknown	3	3	0			
	Not Done	44	44	0	72.1	72.1	0
	Prior to LTx	14	14	0	23	23	0
	Explant	3	3	0	5	5	0
		N			Percent		
Other AutoAB-	**Sex**						
	Male	104	104	0	47.1	47.1	0
	Female	117	117	0	52.9	52.9	0
	Race						
	Unknown	4	4	0			
	White	165	165	0	76	76	0
	Other	52	52	0	24	24	0
	Hispanic						
	No	190	190	0	86	86	0
	Yes	31	31	0	14	14	0
	Liver Biopsy						

	Unknown	9	9	0			
	Not Done	158	158	0	74.5	74.5	0
	Prior to LTx	42	42	0	19.8	19.8	0
	Explant	12	12	0	5.7	5.7	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase nofmterr ls=250;

title '/prj/niddk/ims_analysis/PALF/prog_initial_analysis/palf_dsic.sas';
run;


************;
* INPUT     ;
************;
libname palf1 '/prj/niddk/ims_analysis/PALF/private_orig_data/Pediatric Acute Liver Failure (PALF) NIDDK Data Repository File_20180727/NIDDK Repository
Archive/PALF Phase 1 and 2/SAS Datasets/Supplemental Tests/Diagnostic Labs/';
libname palf2 '/prj/niddk/ims_analysis/PALF/private_orig_data/Pediatric Acute Liver Failure (PALF) NIDDK Data Repository File_20180727/NIDDK Repository
Archive/PALF Phase 1 and 2/SAS Datasets/Registry/';
libname palf3 '/prj/niddk/ims_analysis/PALF/private_orig_data/Pediatric Acute Liver Failure (PALF) NIDDK Data Repository File_20180727/NIDDK Repository
Archive/PALF Phase 1 and 2/SAS Datasets/Supplemental Tests/Supplemental Information and final diagnoses/';


************;
* FORMATS   ;
************;
proc format;
  value sexf
  1 = "1 Male"
  2 = "2 Female"
  ;

  value racef
  1 = "1 White"
  2 = "2 Other"
  3 = "3 Unknown"
  ;

  value hispf
  0 = "No"
  1 = "Yes"
  ;

  value liverf
  . = " Unknown"
  1 = "1 Not Done"
  2 = "2 Explant"
  3 = "3 Prior to LTx"
  ;
run;


************;
* MACROS    ;
************;
%macro readin(lib, ds);
  data &ds;
```

```
    set &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(palf1, supp_tests);
```

```
%readin(palf2, he);

proc freq data=he;
  table diag
age
HSP
RACEW
sex
LBX
  /missing;
  run;

%readin(palf2, el);


%readin(palf3, si_diag);


proc freq data=el noprint;
  tables id/out=ids_el;
run;

proc freq data=ids_el;
  table count/missing;
run;

proc freq data=supp_tests noprint;
  tables id/out=ids_supp;
run;

proc freq data=ids_supp;
  table count/missing;
run;

proc freq data=he noprint;
  tables id/out=ids_he;
run;

proc freq data=ids_he;
  table count/missing;
run;

proc freq data=si_diag noprint;
  tables id/out=ids_si;
run;

proc freq data=ids_si;
  table count/missing;
run;



proc sort data=supp_tests (rename=(ana  = supana
                                   asma = supasma
                                   alkm = supalkm));
```

```
  by id;
run;

proc sort data=he;
  by id;
run;

proc sort data=si_diag;
  by id;
run;

data analy;
  merge supp_tests (in=in1 keep=id supana supasma supalkm)
        he         (in=in2)
       si_diag    (in=in3 keep=id fdiag);
  by id;
  if in1 or in2;

  if in1 then in_supp=1;
  if in2 then in_he=1;
  if in3 then in_si=1;

  if (ana => 0) or (ana in(.f, .g))   then ana_valid = 1;
  if (asma => 0) or (asma in(.f, .g)) then asma_valid = 1;
  if (alkm => 0) or (alkm in(.f, .g)) then alkm_valid = 1;

  if (supana  => 0) or (supana in(.f, .g))  then supana_valid = 1;
  if (supasma => 0) or (supasma in(.f, .g)) then supasma_valid = 1;
  if (supalkm => 0) or (supalkm in(.f, .g)) then supalkm_valid = 1;


  * flag analytic subset;
  if max(ana_valid, asma_valid, alkm_valid, supana_valid, supasma_valid, supalkm_valid) = 1 then subset722=1;


  * ANA, ASMA, ALKM from HE ds are already binary. Create binary pos/neg for Supplemental results;
    *ANA:  Negative if value was .F (below the level of detection) or less than or equal to 2.9.  Positive if value was .G (above the level of detection) or
greater than 2.9.;
    if (0 <= supana <= 2.9) or (supana in(.f))  then ana_supp = 0;
    else if (2.9 < supana) or (supana in(.g))  then ana_supp = 1;

    *ALKM: Negative if value was .F or less than or equal to 20.  Positive if value was .G or greater than 20.;
    if (0 <= supalkm <= 20) or (supalkm in(.f)) then alkm_supp = 0;
    else if (20 < supalkm) or (supalkm in(.g)) then alkm_supp = 1;


  * Then create the auto antibody variable such that if the ana, asma, alkm from the he dataset or supp_tests dataset is positive the autoantibody ;
  * variable is 1 else it is 0.  You should end up with 202 positive and 520 negative. ;
  if max(ana, asma, alkm) = 1 or max(ana_supp, supasma, alkm_supp) = 1 then autoab_posneg = 1;
  else autoab_posneg = 0;

  * create race group: uk, white, other;
  if max(RACAA,   RACEA,   RACEB,   RACEO,   RACNP)=1 then racegrp = 2;
  else if RACEW=1 then racegrp = 1;
  else racegrp = 3;
```

```
  * create AIH, Indeterminate, other group;
  if fdiag=20 then dxgroup = 1;
  else if fdiag = 24 then dxgroup = 2;
  else if fdiag ne . then dxgroup = 3;

  * create Liver biopsy group;
  if lbx = 1 then do;
    if . < trpdate <= lbxdate then liverbiopgp = 2;
    else liverbiopgp = 3;
  end;
  else if lbx = 0 then liverbiopgp = 1;

  * create age;
  age = (erldate-dobdate)/365.25;


  * create analytic subsets;
  if subset722=1 then do;
    if dxgroup = 1 and autoab_posneg = 1 then aih_abpos = 1;
    if dxgroup = 2 then do;
      if autoab_posneg = 1 then indeter_abpos = 1;
      else if autoab_posneg = 0 then indeter_abneg = 1;
    end;
    if dxgroup = 3 then do;
      if autoab_posneg = 1 then other_abpos = 1;
      else if autoab_posneg = 0 then other_abneg = 1;
    end;
  end;

run;

proc freq data=analy;
  tables in_supp*in_he*in_si*subset722/list missing;
  tables ana_valid*  ana
        asma_valid* asma
        alkm_valid* alkm/list missing;
  tables supana_valid*  supana
         supasma_valid* supasma
         supalkm_valid* supalkm/list missing;
 tables ana_supp * supana
        alkm_supp* supalkm/list missing;
  tables subset722*ana_valid*asma_valid*alkm_valid*supana_valid*supasma_valid*supalkm_valid/list missing;
  tables autoab_posneg*ana* asma*alkm* ana_supp* supasma*  alkm_supp/list missing;
  tables racegrp*racew*racaa*   racea*   raceb*   raceo*    racnp/list missing nopercent;
  tables dxgroup*fdiag/missing list;
  tables subset722*dxgroup*autoab_posneg*aih_abpos*indeter_abpos*indeter_abneg*other_abpos*other_abneg/list missing;
  tables liverbiopgp*lbx*lbxdate*trpdate/list missing;
  format lbxdate trpdate monyy.;
  title3 "check analytic subset (n=722)";
run;

proc sort data=analy;
  by id;
run;

proc sort data=el out=el_enrol nodupkey;
```

```
  where id ne "!B!";
  by id;
run;

data analy;
  merge analy (in=in1) el_enrol (in=in2 keep=id);
  by id;
  if in1 or in2;
  if in2 then enrolled=1;
run;

proc freq data=analy;
  tables enrolled * subset722/list missing;
title3 "cross check w/ subjects enrolled according to EL file";
run;


*Table 1;
proc freq data=analy;
  where subset722=1 and max(autoab_posneg,aih_abpos,indeter_abpos,indeter_abneg,other_abpos,other_abneg)=1;
  tables age autoab_posneg sex racegrp hsp /missing;
  tables dxgroup*autoab_posneg/missing list;
  tables liverbiopgp/missing;
run;



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

%npercent(2,  SEX         , sexf   , aih_abpos , aih_abpos);
%npercent(3,  racegrp     , racef  , aih_abpos , aih_abpos);
%npercent(4,  hsp         , hispf  , aih_abpos , aih_abpos);
%npercent(5,  liverbiopgp , liverf , aih_abpos , aih_abpos);


data prntindeter_abpos;
  set _null_;
run;

%npercent(2,  SEX         , sexf   , indeter_abpos , indeter_abpos);
%npercent(3,  racegrp     , racef  , indeter_abpos , indeter_abpos);
%npercent(4,  hsp         , hispf  , indeter_abpos , indeter_abpos);
%npercent(5,  liverbiopgp , liverf , indeter_abpos , indeter_abpos);


data prntindeter_abneg;
  set _null_;
run;

%npercent(2,  SEX         , sexf   , indeter_abneg , indeter_abneg);
%npercent(3,  racegrp     , racef  , indeter_abneg , indeter_abneg);
%npercent(4,  hsp         , hispf  , indeter_abneg , indeter_abneg);
%npercent(5,  liverbiopgp , liverf , indeter_abneg , indeter_abneg);
```

```
data prntother_abpos;
  set _null_;
run;

%npercent(2,  SEX         , sexf   , other_abpos , other_abpos);
%npercent(3,  racegrp     , racef  , other_abpos , other_abpos);
%npercent(4,  hsp         , hispf  , other_abpos , other_abpos);
%npercent(5,  liverbiopgp , liverf , other_abpos , other_abpos);



data prntother_abneg;
  set _null_;
run;

%npercent(2,  SEX         , sexf   , other_abneg , other_abneg);
%npercent(3,  racegrp     , racef  , other_abneg , other_abneg);
%npercent(4,  hsp         , hispf  , other_abneg , other_abneg);
%npercent(5,  liverbiopgp , liverf , other_abneg , other_abneg);



* Table 1;
data npercent;
  length subgroup $25;
  set prntaih_abpos     (in=in1)
      prntindeter_abpos (in=in2)
      prntindeter_abneg (in=in3)
      prntother_abpos   (in=in4)
      prntother_abneg   (in=in5);

  if in1 then subgroup = "AIH AutoAB+";
  if in2 then subgroup = "Indeterminate AutoAB+";
  if in3 then subgroup = "Indeterminate AutoAB-";
  if in4 then subgroup = "Other AutoAB+";
  if in5 then subgroup = "Other AutoAB-";

  percent = round(percent, .1);
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;



* med, q1, q3;
data prntunivaih_abpos;
  set _null_;
```

```
run;
%univ(1   , age        , aih_abpos , aih_abpos);

data prntunivindeter_abpos;
  set _null_;
run;
%univ(1   , age        , indeter_abpos , indeter_abpos);

data prntunivindeter_abneg;
  set _null_;
run;
%univ(1   , age        , indeter_abneg , indeter_abneg);

data prntunivother_abpos;
  set _null_;
run;
%univ(1   , age        , other_abpos , other_abpos);

data prntunivother_abneg;
  set _null_;
run;
%univ(1   , age        , other_abneg , other_abneg);


data alluniv;
  set prntunivaih_abpos     (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ _min_ _max_)
      prntunivindeter_abneg (in=in2 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ _min_ _max_)
      prntunivindeter_abpos (in=in3 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ _min_ _max_)
      prntunivother_abneg   (in=in4 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_ _min_ _max_)
      prntunivother_abpos   (in=in5 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;
```