Dataset Integrity Check for The Diabetes Prevention Program (DPP) Ancillary Data Release

Prepared by NIDDK-CR
July 19, 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 Diabetes Prevention Program (DPP) was a randomized clinical trial with the focus on determining which interventions could help in the prevention or delay of type 2 diabetes. The ancillary data release includes measurements conducted on biospecimens collected during the course of the main DPP study. The focus of this ancillary study, a secondary analysis of the DPP multicenter randomized clinical trial, was to identify and evaluate the relationships between steroid sex hormones, sex hormone binding globulin (SHGB), SHGB single-nucleotide polymorphisms (SNPs), and diabetes risk factors to determine how these relationships impact the progression to diabetes within DPP.

3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the DPP study data package. For this replication, variables were taken from the following dataset: “hormone.sas7bdat”.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Dr. Mather et al. [1] for Steroid Sex Hormones, Sex Hormone-Binding Globulin, and Diabetes Incidence in the Diabetes Prevention Program. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Table 1 – Baseline Characteristics and Sex Hormone Levels of Subjects Entering DPP With Measurable Sex Hormone Concentrations, Table A below lists the variables that were

used in the replication, and Table B (B1-B2) compares the results calculated from the archived data files to the results published in Table 1.

6 Conclusions

The results of the replication are within expected variation to the published results.

7 References

[1] Mather KJ, Kim C, Christophi CA, Aroda VR, Knowler WC, Edelstein SE, Florez JC, Labrie F, Kahn SE, Goldberg RB, Barrett-Connor E. Steroid Sex Hormones, Sex Hormone-Binding Globulin, and Diabetes Incidence in the Diabetes Prevention Program. The Journal of Clinical Endocrinology and Metabolism, 100(10), 3778-3786, October 2015. PMCID: PMC4596040 doi: 10.1210/jc.2015-2328

Table A: Variables used to replicate Table 1 – Baseline Characteristics and Sex Hormone Levels of Subjects Entering DPP With Measurable Sex Hormone Concentrations

Characteristic	**dataset.variable**
Dehydroepiandrosterone (DHEA), pg/mL	hormone.DHEA
DHEA-S, pg/mL	hormone.DHEA_S
Estrone, pg/mL	hormone.E1
Estrone-s, pg/mL	hormone.E1_s
Estradiol, pg/mL	hormone.ESTRADIOL
Testosterone (T), pg/mL	hormone.TESTOSTERONE
Dihydrotestosterone (DHT), pg/mL	hormone.DHT
Sex hormone-binding globulin (SHBG), nmol/L	hormone.SHBG

Table B1: Comparison of values computed in integrity check to reference article Table 1 values for Men and Premenopausal Women

Variable	Manuscript Men (n=969)	DSIC Men (n=969)	Diff. (n=0)	Manuscript Premenopausal (n=948)	DSIC Premenopausal (n=891)	Diff. (n=57)
DHEA, pg/mL	3068 [2461, 4103]	3053.4 [2454.8, 4088.4]	14.6 [6.2, 14.6]	2080 [1395, 3052]	2090.7 [1406.7, 3073.6]	10.7 [11.7, 21.6]
DHEA-S, pg/mL	999 [603, 1551]	995.3 [600.4, 1544.7]	3.7 [2.6, 6.3]	866 [556, 1234]	876 [562.6, 1248.9]	10 [6.6, 14.9]
Estrone, pg/mL	38.2 [31.5, 47.4]	38.2 [31.5, 47.4]	0 [0, 0]	86.6 [54.4, 129.7]	86.8 [54.8, 127.6]	0.2 [0.4, 2.1]
Estrone-S, pg/mL	630 [427, 918]	630.6 [427.2, 917.5]	0.6 [0.2, 0.5]	1179 [593, 2182]	1185.8 [604, 2165.3]	6.8 [11, 16.7]
Estradiol, pg/mL	22.3 [18.2, 27.3]	22.4 [18.2, 27.3]	0.1 [0, 0]	53.5 [23.1, 103.5[1]]	53.5 [23.6, 102.1]	0 [0.5, 1.4]
T, pg/mL	10953 [8579, 13658]	10937.8 [8579.2, 13658]	15.2 [0.2, 0]	611 [462, 864]	611.1 [461.9, 859.3]	0.1 [0.1, 4.7]
DHT, pg/mL	234.32 [169.25, 317.34]	234.2 [169.2, 317]	0.12 [0.05, 0.34]	N/A	N/A	N/A
SHBG, nmol/L	38.8 [26.4, 55.8]	38.9 [26.4, 55.8]	0.1 [0, 0]	44.1 [30.4, 68.3]	43.0 [30.1, 66.9]	1.1 [0.3, 1.4]

[1] The publication has this value represented as “10”. This is a typographical error and the Data Coordinating Center (DCC) has been notified. Per the DCC, the correct value is “103.5” and has been updated in the table above.

Table B2: Comparison of values computed in integrity check to reference article Table 1 values for Postmenopausal Women on hormone replacement (HRT) and Postmenopausal Women not on hormone replacement

Variable	Manuscript Postmenopausal Women on HRT (n=431)	DSIC Postmenopausal Women on HRT (n=435)	Diff. (n=4)	Manuscript Postmenopausal Women Not on HRT (n=551)	DSIC Postmenopausal Women Not on HRT (n=537)	Diff. (n=14)
DHEA, pg/mL	1390 [920, 2140]	1410 [920, 2160]	20 [0, 20]	1640 [1050, 2429]	1662.7 [1030, 2429.7]	22.7 [20, 0.7]
DHEA-S, pg/mL	515 [305, 794]	508.6 [303.3, 789.4]	6.4 [1.7, 4.6]	594 [365, 914]	598.5 [370.1, 923.1]	4.5 [5.1, 9.1]
Estrone, pg/mL	N/A	N/A	N/A	N/A	N/A	N/A
Estrone-S, pg/mL	1359 [640, 2577]	1352.4 [620, 2595.8]	6.6 [20, 18.8]	419 [209, 761]	421.3 [210.4, 762.4]	2.3 [1.4, 1.4]
Estradiol, pg/mL	17.5 [11.9, 27.2]	17.2 [11.7, 26.8]	0.3 [0.2, 0.4]	9.4 [6.0, 17.8]	9.6 [6.2, 18.3]	0.2 [0.2, 0.5]
T, pg/mL	520 [347, 728]	555.2 [416.4, 763.4]	35.2 [69.4, 35.4]	514 [347, 729]	555.2 [416.4, 798.1]	41.2 [69.4, 69.1]
DHT, pg/mL	N/A	N/A	N/A	N/A	N/A	N/A
SHBG, nmol/L	75.3 [43.9, 121.3]	74.4 [42.7, 119.2]	0.9 [1.2, 2.1]	35.1 [25.7, 46.5]	35.0 [25.5, 45.8]	0.1 [0.2, 0.7]

Attachment A: SAS Code

```
libname dpp_a "X:\NIDDK\niddk-
dr_studies1\DPP\private_orig_data\dpp_ancillary_2021 0303";
libname dpp "X:\NIDDK\niddk-
dr_data_curation2\DPP_V5\Data\DPP_Data_2008\Non-Form_Data\Data";
libname dppf "X:\NIDDK\niddk-
dr_data_curation2\DPP_V6\Data\DPP_Data_2008\Form_Data\Data";

proc format;
      value sexf 1='Male'
                        2='Female';

      value groupf 1='Men'
                        2='Premenopausal'
                        3='Postmenopausal on HRT'
                        4='Postmenopausal not on HRT';
run;

*define datasets;

data dpp_a_hormone;
      set dpp_a.hormone;
      if visit='BAS';
run;

data dpp_s07; set dppf.s07;
run;

data dpp_s03;
      set dppf.s03;
run;

data dpp_sex;
      set dpp.basedata;
run;

data dpp_menopause;
      set dppf.s05; *note: s05 is a screening form, so no need to
restrict by visit type;
run;

data dpp_menopausee;
      set dppf.r04;
      if visit='BAS';
run;

proc sort data=dpp_s07;
      by release_id;
```

```
run;

proc sort data= dpp_s03;
     by release_id;
run;

proc sort data=dpp_a_hormone;
     by release_id;
run;

proc sort data=dpp_sex;
     by release_id;
run;

proc sort data=dpp_menopause;
     by release_id;
run;

proc sort data=dpp_menopausee;
     by release_id;
run;

*merge selected datasets;
data dpp_a_combined;
     merge dpp_a_hormone (in=a)
             dpp_sex (keep=release_id sex agegroup in=b)
             dpp_menopause (keep=release_id simens simenst siovar
sihyst siestrn sibcpn in=c)
             dpp_menopausee (keep=release_id chsex in=d)
             dpp_s03 (keep= release_id sorxdq sorxex sorxda1 sorxdb1
sorxdc1 sorxdd1 sorxde1 sorxdf1 sorxdg1
             sorxdh1 sorxdi1 sorxdj1 in=e)
             dpp_s07 (keep= release_id srrxdq srrxda1 srrxdb1 srrxdc1
srrxdd1 srrxde1 srrxdf1 srrxdg1
             srrxdh1 srrxdi1 srrxdj1 in=f);
     by release_id;
     if a=1 and b=1 and c=1 and d=1 and e=1 and f=1;
run;

data dpp_a_combined_2; set dpp_a_combined;
     if sorxex = 1 then delete;
run;



*creating flag variable for hormone medications;
proc freq data=dpp_a_combined_2;
     tables sorxdq sorxex sorxda1 sorxdb1 sorxdc1 sorxdd1 sorxde1
sorxdf1 sorxdg1
             sorxdh1 sorxdi1 sorxdj1; *from screening for s03;
             where sex = 2;
             where sorxdq = 1;
```

```
run;


proc freq data= dpp_a_combined_2;
     tables srrxdq srrxda1 srrxdb1 srrxdc1 srrxdd1 srrxde1 srrxdf1
srrxdg1
             srrxdh1 srrxdi1 srrxdj1; *from screening form s07;
     where sex = 2;
     where sorxdq = 1;
run;

*creating medication flags for HRT medications from s03 and s07
screening forms;
data exhorm; set dpp_a_combined_2;

flag = 0;

array horm {10} sorxda1 sorxdb1 sorxdc1 sorxdd1 sorxde1 sorxdf1
sorxdg1
             sorxdh1 sorxdi1 sorxdj1;

             do i = 1 to 10;

             if horm{i} = "ESTRADIOL" OR horm{i} = "PROGESTERONE" OR
horm{i} = "ESTROGEN" OR
             horm{i} = "ESTROGEN,CON/M-PROGEST ACET" OR horm{i} =
"ESTROGENS,CONJUGATED" OR  horm{i} = "ESTROPIPATE" OR
             horm{i} = "ESTRATAB" OR horm{i} = "MEDROXYPROGESTERONE
ACET" OR horm{i} = "MEDROXYPROGESTERONE ACETATE" OR
           horm{i} = "ESTRACE" OR horm{i} = "ESTRADERM" OR horm{i} =
"ORTHO-EST" OR horm{i} = "MENEST" OR  horm{i} = "PREMARIN" OR
         horm{i} = "PREMARIN W/METHYLTESTOSTERONE" OR horm{i} =
"PREMPRO" OR horm{i} = "DEPO-PROVERA" OR
             horm{i} = "PREMPHASE" OR horm{i} = "PROVERA" OR horm{i} =
"AYGESTIN" OR horm{i} = "OGEN" OR
         horm{i} = "VIVELLE" OR horm{i} = "CLIMARA" OR horm{i} =
"CYCRIN" OR horm{i} = "PROGESTERONE-50 IN OIL"

             then flag = 1;
             end;
     run;

proc freq data=exhorm;
tables flag;
run;

data exhorm2; set exhorm;

flag2 = 0;

array horm2 {10} srrxda1 srrxdb1 srrxdc1 srrxdd1 srrxde1 srrxdf1
srrxdg1
```

```
            srrxdh1 srrxdi1 srrxdj1;

            do i = 1 to 10;

            if horm2{i} = "ESTRADIOL" OR horm2{i} = "PROGESTERONE" OR
horm2{i} = "ESTROGEN" OR
            horm2{i} = "ESTROGEN,CON/M-PROGEST ACET" OR horm2{i} =
"ESTROGENS,CONJUGATED" OR  horm2{i} = "ESTROPIPATE" OR
          horm2{i} = "ESTRATAB" OR horm2{i} = "MEDROXYPROGESTERONE
ACET" OR horm2{i} = "MEDROXYPROGESTERONE ACETATE" OR
          horm2{i} = "ESTRACE" OR horm2{i} = "ESTRADERM" OR horm2{i}
= "ORTHO-EST" OR horm2{i} = "MENEST" OR
          horm2{i} = "PREMARIN" OR horm2{i} = "PREMARIN
W/METHYLTESTOSTERONE" OR horm2{i} = "PREMPRO" OR
          horm2{i} = "DEPO-PROVERA" OR horm2{i} = "PREMPHASE" OR
horm2{i} = "PROVERA" OR horm2{i} = "AYGESTIN" OR
          horm2{i} = "OGEN" OR horm2{i} = "VIVELLE" OR horm2{i} =
"CLIMARA" OR horm2{i} = "CYCRIN" OR
            horm2{i} = "PROGESTERONE-50 IN OIL"

            then flag2 = 1;
            end;
     run;

*cross checking flags;
proc freq data=exhorm2;
tables flag2*flag;
run;


*define participant groups (males and menopause status);
data parts; set exhorm2;
     if sex = 1 then male = 1; *males;
     if sex = 2 then do; *women;
          if simens = 1 then meno = 3; *post-menopausal;
               else if simens = 2 then meno = 1; *premenopausal;
               else if simens = 3 AND 1<=simenst<=2 then meno = 2;
*peri-menopausal;
               else if simens = 3 AND 3<=simenst<=4 then meno = 4;
*uncertain;
          if siovar = 3 OR sihyst = 1 then meno = 3; *post-
menopausal;
          if agegroup>=5 then meno = 3; *post-menopausal;
     end;
     label meno = "Menopause Status";
run;

*checking meno status variable;
proc freq data=parts;
tables meno;
run;
```

```
*Assign post-menopause status;
data parts_2; set parts;
     if meno = 3 then postmeno = 1;
          else if meno = . then postmeno = .;
          else postmeno = 0;
label postemeno = 'Post-Menopausal';
run;

*cross checking meno and postmeno status variables;
proc freq data=parts_2;
tables meno*postmeno;
run;


*identify those taking HRT (as best we can);
data parts_3; set parts_2;
     if flag = 1 or flag2 = 1 then exhorm = 1;
     else exhorm = 0;
run;

proc freq data= parts_3;
tables exhorm;
run;

*Create four groups from publication;
data parts_4; set parts_3;
     if postmeno = 0 and exhorm = 1 then delete; *exclude
premenopausal women with hormone use*;
     if sex = 1 then analgp = 1; *men;
          else if (sex = 2 and postmeno = 0) then analgp = 2;
*premenopausal;
          else if (sex = 2 and postmeno = 1 and exhorm = 0) then
analgp = 3; *postmenopausal non hormone users;
          else if (sex = 2 and postmeno = 1 and exhorm = 1) then
analgp = 4; *postmenopausal hormone users;
run;

proc freq data=parts_4;
tables sex;
run;



proc format;
     value analgpf 1 = 'Men'
                   2 = 'Premenopausal'
                   3 = 'Post-Menopausal non-hormone users'
                   4 = 'Post-menopausal hormone users';
run;

data parts_4; set parts_4;
format analgp analgpf. sex sexf.;
```

```
run;


*converting Testosterone to match publication unit;
data parts_5; set parts_4;
testosterone_1 = (testosterone*3.47);
run;


*Table 1 values;
proc means data=parts_5 n median q1 q3 maxdec=1;
     class analgp;
     var DHEA dhea_s e1 e1_s estradiol testosterone_1 dht shbg;
run;
```