Dataset Integrity Check for Diabetes Prevention Program/Diabetes Prevention Program Outcomes Study (DPP/DPPOS) PFAS Secondary Data

Prepared by NIDDK-CR
June 4, 2024

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 multicenter trial examining the ability of an intensive lifestyle program or treatment with metformin to prevent or delay the development of type 2 diabetes in high-risk individuals with prediabetes. The DPP study showed that both interventions reduced the incidence of diabetes in participants, compared with placebo; the lifestyle intervention proved more effective than metformin in preventing the onset of diabetes. The Diabetes Prevention Program Outcomes Study (DPPOS) was the long-term follow-up of the original DPP study. The DPPOS sought to evaluate the effects of the interventions on the further development of diabetes and diabetes complications, including retinopathy, microangiopathy, and cardiovascular disease.

The PFAS (per- and polyfluoroalkyl substances) Secondary Data was an ancillary study to determine trends and rate of change of plasma PFAS concentrations in overweight or obese U.S. adults and evaluate variation by sex, race/ethnicity, and age.

3 Archived Datasets

A full listing of the archived datasets included in the package can be found in the Roadmap document. All data files, as provided by the Data Coordinating Center (DCC), are located in the DPP and DPPOS folders in the respective data packages. For this replication, variables were taken from the “basedata.sas7bdat” dataset from the DPP package, and the “dpp_dppos_plasma_pfas.sas7bdat” dataset found in both DPP and DPPOS packages.

4 Statistical Methods

Analyses were performed to replicate results for the data in the publication by Lin et al. [1]. To verify the integrity of the data, only descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Baseline characteristics of the 957 participants with plasma PFAS measurements in the Diabetes Prevention Program (DPP), 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 in the publication. The results of the replication are within expected variation to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the DPP/DPPOS PFAS Secondary Data files to be distributed are a true copy of the study data.

7 References

[1] Lin PD, Cardenas A, Hauser R, Gold DR, Kleinman KP, Hivert MF, Calafat AM, Webster TF, Horton ES, Oken E. Temporal Trends of Concentrations of Per- and Polyfluoroalkyl Substances among Adults with Overweight and Obesity in the United States: Results from the Diabetes Prevention Program and NHANES. Environment International, 157, 106789, December 2021. doi: https://doi.org/10.1016/j.envint.2021.106789

Table A: Variables used to replicate results – Baseline characteristics of the 957 participants with plasma PFAS measurements in the Diabetes Prevention Program (DPP)

Table Variable	**dataset.variable**
Treatment arm	basedata.assign dpp_dppos_plasma_pfas.release_id
Sex	basedata.sex dpp_dppos_plasma_pfas.release_id
Race/ethnicity	basedata.race_eth dpp_dppos_plasma_pfas.release_id
Age	basedata.agegroup dpp_dppos_plasma_pfas.release_id
BMI	basedata.bmi_cat dpp_dppos_plasma_pfas.release_id

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

Characteristic	Publication: N (%)	DSIC: N (%)	Diff.
Treatment arm			
Lifestyle	481 (50.3)	481 (50.3)	0 (0)
Placebo	476 (49.7)	476 (49.7)	0 (0)
Sex			
Male	332 (34.7)	332 (34.7)	0 (0)
Female	625 (65.3)	625 (65.3)	0 (0)
Race/ethnicity			
White	552 (57.7)	552 (57.7)	0 (0)
Black	184 (19.2)	184 (19.2)	0 (0)
Hispanic	179 (18.7)	179 (18.7)	0 (0)
All other races	42 (4.4)	42 (4.4)	0 (0)
Age (years)			
25-39	112 (11.7)	112 (11.7)	0 (0)
40-44	107 (11.2)	107 (11.2)	0 (0)
45-49	213 (22.3)	213 (22.3)	0 (0)
50-54	167 (17.5)	167 (17.5)	0 (0)
55-59	137 (14.3)	137 (14.3)	0 (0)
60-64	107 (11.2)	107 (11.2)	0 (0)
65+	114 (11.9)	114 (11.9)	0 (0)
BMI (kg/m^2)*			
< 25	9 (0.9)	74 (7.7)	65 (6.8)
25 to 29.9	308 (32.2)	242 (25.3)	66 (6.9)
30 to 34.9	238 (24.9)	239 (25.0)	1 (0.1)
35 to 39.9	251 (26.2)	251 (26.2)	0 (0)
40+	151 (15.8)	151 (15.8)	0 (0)

*Note: The BMI categories in the DPP data are not the BMI categories in the publication. There were 10 BMI categories in the DPP data: 1) < 26 kg/m^2, 2) $\geqslant$ 26 to < 28 kg/m^2, 3) $\geqslant$ 28 to < 30 kg/m^2, 4) $\geqslant$ 30 to < 32 kg/m^2, 5) $\geqslant$ 32 to < 34 kg/m^2, 6) $\geqslant$ 34 to < 36 kg/m^2, 7) $\geqslant$ 36 to < 38 kg/m^2, 8) $\geqslant$ 38 to < 40 kg/m^2, 9) $\geqslant$ 40 to < 42 kg/m^2, and 10) $\geqslant$ 42 kg/m^2.

Attachment A: SAS Code

```
libname pfas "X:\NIDDK\niddk-dr_studies1\DPPOS\private_created_data\pfas";
libname dpp "X:\NIDDK\niddk-
dr_studies1\DPP\private_created_data\DPP_V7R\Data\DPP_Data_2008\Non-Form_Data\Data";

proc means data=work.dpp_dppos_plasma_pfas n median q1 q3;
var total_PFOS;
class Visit;
run;

/**************************/
/* DPP/DPPOS PFAS DSIC */
/* Lin et al. */
/**************************/

data dem; set dpp.basedata;
id = input(release_id, 8.);
drop release_id;
run;

data pfa; set work.dpp_dppos_plasma_pfas;
if visit = "BAS";
id = release_id;
drop release_id;
run;

*Merging baseline DPP data with the PFAS data;
proc sort data=work.pfa;
by id;
run;

proc sort data=work.dem;
by id;
run;

data one; merge
pfa (in=a)
dem (in=b);
by id;
if a=b;
run;

*Table 1 from pub Lin et al. ;
proc freq data=one;
tables assign sex race_eth agegroup b;
run;
```

```
*Fixing BMI categories;
data two; set one;
if bmi_cat = 1 then bmi = 1;
if bmi_cat = 2 OR bmi_cat = 3 then bmi = 2;
if bmi_cat = 4 OR bmi_Cat = 5 then bmi = 3;
if bmi_cat = 6 or bmi_cat = 7 or bmi_cat = 8 then bmi = 4;
if bmi_cat > 8 then bmi = 5;
run;

proc freq data=two;
tables bmi;
run;
```