Dataset Integrity Check for Nonalcoholic Fatty Liver Disease Pediatric Database 2 (NAFLD Pediatric Database 2)

Prepared by NIDDK-CR
March 25, 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

Nonalcoholic fatty liver disease (NAFLD) is a spectrum of liver conditions associated with fat accumulation that range from benign, non-progressive liver fat accumulation to severe liver injury, cirrhosis, and liver failure. NAFLD is highly prevalent within the United States and is most common in adults who are overweight or have diabetes, insulin resistance, or hyperlipidemia. However, the disease also occurs in children and in persons who are not obese or diabetic. The Nonalcoholic Steatohepatitis Clinical Research Network (NASH CRN) was initiated in 2002 to conduct multicenter, collaborative studies on the etiology, contributing factors, natural history, complications, and treatment of NASH.

The NAFLD Pediatric Database 2 was a multicenter, prospective follow-up study of participants with NAFLD or NASH which aimed to investigate the etiology, pathogenesis, natural history, diagnosis, treatment, and prevention of NAFLD and NASH. The study included longitudinal follow-up of participants enrolled in earlier NASH CRN studies and recruited new participants. The study population included pediatric participants in the United States that were 2-17 years old at the time of enrollment with histologically confirmed NAFLD or NASH. Comprehensive data, including demographics, medical history, symptoms, medication use, alcohol use, and routine laboratory studies were collected on all participants at entry and at follow-up visits every 48 weeks from enrollment. A standard of care liver biopsy was collected at baseline if not previously collected, and specimens were collected every 48 weeks during follow-up.

3 Archived Datasets

A full listing of 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 NAFLD Pediatric Database 2 folder in the data package. For this replication, variables were taken from the “rg.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to replicate demographic descriptive statistics from an unpublished report provided by the DCC. To verify the integrity of the data, only demographic descriptive statistics were computed. The DCC provided a list of participant IDs used in the unpublished report for the purposes of this replication.

5 Results

For Table 1 in the unpublished report, Baseline characteristics of the study population, according to NASH study into which they were first enrolled with a liver biopsy performed within six months before enrollment, 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 Table 1. The results of the replication are within expected variation of the results in the unpublished report.

6 Conclusions

The NIDDK Central Repository is confident that the NAFLD Pediatric Database 2 data files to be distributed are a true copy of the study data.

Table A: Variables used to replicate report Table 1 – Baseline characteristics of the study population, according to NASH study into which they were first enrolled with a liver biopsy performed within six months before enrollment

Table Variable	**dataset.variable**
Age	rg.formdate rg.rg109
Sex	rg.rg111
Hispanic ethnicity	rg.rg112
Race	rg.rg114a rg.rg114b rg.rg114c rg.rg114d rg.rg114e rg.rg114f
Education	rg.rg116
Income	rg.rg122

Table B: Comparison of values computed in integrity check to report Table 1 provided by the DCC

Demographics	**Unpublished Report: NAFLD DB2 (n=711)**	**DSIC: NAFLD DB2 (n =711)**	**Diff. (n=0)**
Age, years – mean (SD)	12.9 (2.8)	12.8 (2.8)	0.1 (0)
Age, years – median (range)	12.7 (4.8, 18)	12.6 (4.7, 17.9)	0.1 (0.1, 0.1)
Sex – no. (%)			
Male	495 (69.6)	495 (69.6)	0 (0)
Female	216 (30.4)	216 (30.4)	0 (0)
Hispanic Ethnicity – no. (%)	520 (73.2)	520 (73.2)	0 (0)
Race – no. (%)			
White	464 (65.3)	464 (65.3)	0 (0)
Black or African American	26 (3.7)	26 (3.7)	0 (0)
Native Hawaiian or Pacific Islander	8 (1.1)	8 (1.1)	0 (0)
Asian	18 (2.5)	18 (2.5)	0 (0)
American Indian/Alaska Native	24 (3.4)	24 (3.4)	0 (0)
Refused or not stated	193 (27.1)	193 (27.1)	0 (0)
Education – no. (%)			
Did not answer or not in school	1 (0.1)	1 (0.1)	0 (0)
Preschool to Grade 6	255 (35.9)	255 (35.9)	0 (0)
Grade 6 to 8	236 (33.2)	236 (33.2)	0 (0)
Grade 9 or Higher	219 (30.8)	219 (30.8)	0 (0)
Income – no. (%)			
< $15,000	213 (30.4)	213 (30.4)	0 (0)
$15,000 - $29,000	216 (30.9)	216 (30.9)	0 (0)
$30,000 - $49,000	132 (18.9)	132 (18.9)	0 (0)
$50,000 or more	139 (19.9)	139 (19.9)	0 (0)
Did not answer	11	11	0 (0)

Attachment A: SAS Code

```
libname ped "X:\NIDDK\niddk-dr_studies6\NAFLD\private_created_data\PED DB2\Redacted Data";
libname id "X:\NIDDK\niddk-dr_studies6\NAFLD\private_created_data\PED DB2";

*use the registration dataset (rg);
*use the DCC provided list of IDs for participants;
data reg; set ped.rg;
run;

proc freq data=id.table1_ids;
tables visit;
run;

data id; set id.table1_ids;
keep id;
run;

proc freq data=reg;
tables id;
run;

*limiting registration dataset to only those participants from the ID list;
proc sort data=reg;
by id;
run;

proc sort data=id;
by id;
run;

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

*demographics;
*age;
data two; set one;
age = (formdate - rg109)/365.25;
run;

proc means data=two n mean std median min max;
var age;
run;
```

```
*sex;
proc freq data=two;
tables rg111;
run;

*ethnicity;
proc freq data=two;
tables rg112;
where rg112 ^= "d";
run;

*race;
proc freq data=two;
tables rg114a rg114b rg114c rg114d rg114e rg114f/missing;
run;

*education;
data three; set two;
if rg116 = "0" then school = 0;
if rg116 = "1" or rg116 = "2" then school = 1;
if rg116 = "3" then school = 2;
if rg116 = "4" or rg116 = "5" or rg116 = "6" or rg116 = "7" then school = 3;
run;

proc freq data=three;
tables school;
run;

*income;
data four; set three;
if rg122 = "1" then income = 1;
if rg122 = "2" then income = 2;
if rg122 = "3" then income = 3;
if rg122 = "4" then income = 4;
if rg122 = "m" or rg122 = "r" or rg122 = "d" then income = 5;
run;

proc freq data=four;
tables income;
where income ^= 5;
run;
```