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

Prepared by NIDDK-CR
January 2, 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 that can progress to significant fibrosis and cirrhosis. There is an estimated 40-90 million individuals within the United States with NAFLD, 10-30% of whom have nonalcoholic steatohepatitis (NASH) and may develop NASH-related cirrhosis. Identifying through non-invasive means those individuals who are at risk for progressive liver disease remains challenging.

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 Adult Database 2 study continued the longitudinal follow-up of participants enrolled in earlier NASH CRN studies and recruited new participants with recent liver biopsies. Comprehensive data, including demographics, medical history, symptoms, medication use, alcohol use, and routine laboratory results were collected on all participants at entry and at annual visits for up to 10 years after enrollment. A liver biopsy was collected at baseline if not collected in a prior NASH CRN study. Study questionnaires administered at enrollment included Skinner Lifetime Drinking History, Alcohol Use Disorders Identification Test (AUDIT), and Beverage Questionnaire (BEVQ-15).

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 Adult Database 2 folder in the data package. For this replication, variables were taken from the “rg.sas7bdat”, “pe.sas7bdat”, “bg2.sas7bdat”, “lr1.sas7bdat”, “lr2.sas7bdat”, “lr3.sas7bdat”, “fr1.sas7bdat”, “fr2.sas7bdat”, “fr3.sas7bdat”, “fr4.sas7bdat”, and “fr5.sas7bdat” datasets.

4 Statistical Methods

Analyses were performed to replicate results for the data in the publication by Woreta et al. [1]. To verify the integrity of the data, only descriptive statistics were computed. The DCC provided a list of participant IDs used in the Woreta et al. analysis for the purposes of this replication.

5 Results

For Table 1 in the publication [1], Characteristics of the population by NASH activity and fibrosis status, 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 published results.

6 Conclusions

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

7 References

[1] Woreta TA, Van Natta ML, Lazo M, Krishnan A, Neuschwander-Tetri BA, Loomba R, Diehl AM, Abdelmalek MF, Chalasani N, Gawrieh S, Dasarathy S, Vuppalanchi R, Siddiqui MS, Kowdley KV, McCullough A, Terrault NA, Behling C, Kleiner DE, Fishbein M, Hertel P, Wilson LA, Mitchell EP, Miriel LA, Clark JM, Tonascia J, Sanyal AJ. Validation of the Accuracy of the FAST™ Score for Detecting Patients with At-risk Nonalcoholic Steatohepatitis (NASH) in a North American Cohort and Comparison to Other Non-invasive Algorithms. PLoS One, 17(4), e0266859, April 2022. doi: https://doi.org/10.1371/journal.pone.0266859

Table A: Variables used to replicate Table 1 – Characteristics of the population by NASH activity and fibrosis status

Table Variable	**dataset.variable**
Sex	rg.rg111
Race	rg.rg114a rg.rg114b rg.rg114c rg.rg114d rg.rg114e rg.rg114f
Ethnicity	rg.rg112
Age	rg.rg110
BMI	pe.pe108a pe.pe108b pe.pe108c pe.pe109a pe.pe109b pe.pe109c
Diabetes	bg2.bg249a bg2.bg249b
Hyperlipidemia	bg2.bg249am
AST	lr1.lr123 lr2.lr223 lr3.lr323
ALT	lr1.lr124 lr2.lr224 lr3.lr324
AST/ALT ratio	lr1.lr123 lr2.lr223 lr3.lr323 lr1.lr124 lr2.lr224 lr3.lr324
GGT	lr1.lr126 lr2.lr226 lr3.lr326
Alkaline phosphatase	lr1.lr125 lr2.lr225 lr3.lr325
Total bilirubin	lr1.lr121 lr2.lr221 lr3.lr321
Direct bilirubin	lr1.lr122 lr2.lr222 lr3.lr322
Albumin	lr1.lr128

Table Variable	dataset.variable
	lr2.lr228 lr3.lr328
Total protein	lr1.lr127 lr2.lr227 lr3.lr327
INR	lr1.lr130 lr2.lr230 lr3.lr330
Platelets	lr1.lr113 lr2.lr213 lr3.lr313
Glucose	lr1.lr134a lr2.lr234a lr3.lr334a
FibroScan LSM	fr1.fr115a fr2.fr215a fr3.fr315a fr4.fr415a fr5.fr515a
FibroScan CAP	fr1.fr116a fr2.fr216a fr3.fr316a fr4.fr416a fr5.fr516a
FibroScan Probe type	fr1.fr112 fr2.fr212 fr3.fr312 fr4.fr412 fr5.fr512

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

Mean (SD) / n (%)	Publication: At-risk NASH = No (n=371)	DSIC: At-risk NASH = No (n=371)	Diff. (n=0)	Publication: At-risk NASH = Yes (n=214)	DSIC: At-risk NASH = Yes (n=214)	Diff. (n=0)	Publication: Total (n=585)	DSIC: Total (n=585)	Diff. (n=0)
Sex (Male)	161 (43.4)	161 (43.4)	0 (0)	63 (29.4)	63 (29.4)	0 (0)	224 (38.3)	224 (38.3)	0 (0)
Race									
White	282 (76.0)	282 (76.0)	0 (0)	178 (83.2)	178 (83.2)	0 (0)	460 (78.6)	460 (78.6)	0 (0)
Black	11 (3.0)	11 (3.0)	0 (0)	11 (5.1)	11 (5.1)	0 (0)	22 (3.8)	22 (3.8)	0 (0)
Asian	42 (11.3)	42 (11.3)	0 (0)	12 (5.6)	12 (5.6)	0 (0)	54 (9.2)	54 (9.2)	0 (0)
Am Indian/Pacific Islander	5 (1.4)	9 (2.4)	4 (1.0)	3 (1.4)	4 (1.9)	1 (0.5)	8 (1.4)	13 (2.2)	5 (0.8)
Refused	27 (7.3)	27 (7.3)	0 (0)	9 (4.2)	9 (4.2)	0 (0)	36 (6.2)	36 (6.2)	0 (0)
Ethnicity (Hispanic)	63 (17.0)	63 (17.0)	0 (0)	19 (8.9)	19 (8.9)	0 (0)	82 (14.0)	82 (14.0)	0 (0)
Age (Years)	50 (12)	50 (12)	0 (0)	54 (12)	54 (12)	0 (0)	51 (12)	51 (12)	0 (0)
BMI (kg/m^2)									
18.5-24.9	17 (4.6)	17 (4.6)	0 (0)	6 (2.8)	6 (2.8)	0 (0)	23 (3.9)	23 (3.9)	0 (0)
25.0-29.9	100 (27.0)	100 (27.0)	0 (0)	33 (15.4)	33 (15.4)	0 (0)	133 (22.7)	133 (22.8)	0 (0.1)
30.0-39.9	116 (31.3)	202 (54.6)	86 (23.3)	72 (33.6)	121 (56.5)	49 (22.9)	188 (32.1)	323 (55.3)	135 (23.2)
≥ 40	138 (37.2)	51 (13.8)	87 (23.4)	103 (48.1)	54 (25.2)	49 (22.9)	241 (41.2)	105 (18.0)	136 (23.2)
Diabetes									
None	235 (63.3)	235 (63.3)	0 (0)	93 (43.5)	93 (43.5)	0 (0)	328 (56.1)	328 (56.1)	0 (0)
Type 1	2 (0.5)	2 (0.5)	0 (0)	3 (1.4)	3 (1.4)	0 (0)	5 (0.8)	5 (0.8)	0 (0)
Type 2	134 (36.1)	134 (36.1)	0 (0)	118 (55.1)	118 (55.1)	0 (0)	252 (43.1)	252 (43.1)	0 (0)
Hyperlipidemia (Yes)	174 (46.9)	174 (46.9)	0 (0)	115 (53.7)	115 (53.7)	0 (0)	289 (49.4)	289 (49.4)	0 (0)
AST (U/L)	40 (31)	40 (31)	0 (0)	65 (41)	65 (41)	0 (0)	50 (37)	50 (37)	0 (0)
ALT (U/L)	58 (41)	58 (41)	0 (0)	78 (50)	78 (50)	0 (0)	66 (45)	66 (45)	0 (0)
AST/ALT ratio	0.78 (0.32)	0.78 (0.32)	0 (0)	0.91 (0.33)	0.91 (0.33)	0 (0)	0.83 (0.33)	0.83 (0.33)	0 (0)
GGT (U/L)	59 (74)	59 (74)	0 (0)	95 (96)	95 (96)	0 (0)	72 (85)	72 (85)	0 (0)
Alkaline phosphatase (U/L)	78 (27)	78 (27)	0 (0)	90 (33)	90 (33)	0 (0)	82 (30)	82 (30)	0 (0)
Total bilirubin (mg/dL)	0.63 (0.38)	0.63 (0.38)	0 (0)	0.63 (0.35)	0.63 (0.35)	0 (0)	0.63 (0.37)	0.63 (0.37)	0 (0)
Direct bilirubin (mg/dL)	0.19 (0.10)	0.19 (0.10)	0 (0)	0.18 (0.10)	0.18 (0.10)	0 (0)	0.19 (0.10)	0.19 (0.10)	0 (0)
Albumin (g/dL)	4.41 (0.34)	4.41 (0.34)	0 (0)	4.35 (0.37)	4.35 (0.37)	0 (0)	4.39 (0.35)	4.39 (0.35)	0 (0)

Mean (SD) / n (%)	Publication: At-risk NASH = No (n=371)	DSIC: At-risk NASH = No (n=371)	Diff. (n=0)	Publication: At-risk NASH = Yes (n=214)	DSIC: At-risk NASH = Yes (n=214)	Diff. (n=0)	Publication: Total (n=585)	DSIC: Total (n=585)	Diff. (n=0)
Total protein (g/dL)	7.39 (0.45)	7.39 (0.45)	0 (0)	7.44 (0.55)	7.44 (0.55)	0 (0)	7.41 (0.49)	7.41 (0.49)	0 (0)
INR	1.03 (0.08)	1.03 (0.08)	0 (0)	1.06 (0.09)	1.06 (0.09)	0 (0)	1.04 (0.09)	1.04 (0.09)	0 (0)
Platelets (1000/µL)	243 (76)	243 (76)	0 (0)	227 (71)	227 (71)	0 (0)	237 (74)	237 (74)	0 (0)
Glucose (mg/dL)	110 (35)	110 (35)	0 (0)	121 (39)	121 (39)	0 (0)	114 (37)	114 (37)	0 (0)
FibroScan LSM (kPa)	9.3 (9.6)	9.3 (9.6)	0 (0)	15.2 (12.4)	15.2 (12.4)	0 (0)	11.4 (11.1)	11.4 (11.1)	0 (0)
FibroScan CAP (dB/m)	317 (51)	317 (51)	0 (0)	327 (51)	327 (51)	0 (0)	321 (52)	321 (52)	0 (0)
FibroScan Probe type									
M	181 (48.8)	181 (48.8)	0 (0)	80 (37.4)	80 (37.4)	0 (0)	261 (44.6)	261 (44.6)	0 (0)
XL	190 (51.2)	190 (51.2)	0 (0)	134 (62.6)	134 (62.6)	0 (0)	324 (55.4)	324 (55.4)	0 (0)

Attachment A: SAS Code

```
libname id "X:\NIDDK\niddk-dr_studies6\NAFLD\private_orig_data\DB2 Program from DCC";
libname db "X:\NIDDK\niddk-dr_studies6\NAFLD\private_created_data\ADULT DB2\1. SAS Datasets
Orig";

/******************/
/* NAFLD DB2 DSIC */
/* Woreta et al. */
/******************/

*Connecting linking file to NASH id list provided by DCC;
data pub; set id.id_list;
run;

data id; set db.id;
run;

proc sort data=pub;
by nash;
run;

proc sort data=id;
by nash;
run;

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

*585 participants used in the publication;

/*************************/
*Identifying datasets and variables needed for Table 1 and merging;

*registration dataset (RG)
*physical examination (PE)
*baseline history (BG1)
*laboratory results (LR1)
*Fibroscan report (FR1)

/*************************************/
*Demographics from registration dataset;
data demo; set db.rg;
```

```
run;

proc sort data=demo;
by id;
run;

proc sort data=one;
by id;
run;

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

*Sex;
proc freq data=two;
tables rg111*atriskdefnash/norow;
run;

*Race and Ethnicity;
data two_1; set two;
if rg114d = 1 OR rg114a = 1 then race_cat = "AmInd/PI";
if rg114c = 1 then race_cat = "Black/";
if rg114f = 1 then race_cat = "Refused";
if rg114b = 1 then race_cat = "Asian";
if rg114e = 1 then race_cat = "White";
run;

proc freq data=two_1 ;
tables (rg112 race_cat)*atriskdefnash/norow;
run;

*Age (convert character age to numeric);
proc freq data=two_1;
tables rg110;
run;

data two_2; set two_1;
age = input(rg110, 3.);
run;

proc means data=two_2 mean std maxdec=0;
var age;
/*class atriskdefnash; */
run;
```

```
/***************************************/
*Anthropometrics;

*BMI;
data work.pe; set db.pe;
run;

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

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

proc freq data=three;
tables visit;
run;

data three_1;
set three;
if visit = "t0";
run;

proc freq data=three_1;
tables pe108c pe109c; *units of measurement 1 = imperial, 2 = metric;
run;

data three_2; set three_1;
*height converions;
if pe108c = 1 then height_1 = (pe108a*2.54)/100; else height_1 = pe108a/100;
if pe108c = 1 then height_2 = (pe108b*2.54)/100; else height_2 = pe108b/100;
*weight conversions;
if pe109c = 1 then weight_1 = (pe109a/2.20462); else weight_1 = pe109a;
if pe109c = 1 then weight_2 = (pe109b/2.20462); else weight_2 = pe109b;
/*height_1 = round(height_1, 0.1); */
/*height_2 = round(height_2, 0.1); */
/*weight_1 = round(weight_1, 0.1); */
/*weight_2 = round(weight_2, 0.1); */
run;

*Checking conversions;
proc means data=three_2 n min max mean median;
var height_1 height_2 pe108a pe108b;
```

```
class pe108c;
run;

proc means data=three_2 n min max mean median;
var weight_1 weight_2 pe109a pe109b;
class pe109c;
run;

*BMI;
data three_3; set three_2;
heightsq1 = (height_1*height_1);
heightsq2 = (height_2*height_2);
bmi1 = weight_1/heightsq1;
bmi2 = weight_2/heightsq2;
avg_bmi = (bmi1+bmi2)/2;
/*bmi1 = round(bmi1, .1); */
/*bmi2 = round(bmi2, .1); */
/*avg_bmi = round(avg_bmi, .1); */
run;

proc means data=three_3 n min max mean median;
var bmi1 bmi2 avg_bmi;
run;

data three_4; set three_3;
if bmi1 < 25 then bmicat1 = 1;
if bmi1 >=25 AND bmi1 <30 then bmicat1 = 2;
if bmi1 >=30 AND bmi1 <39.9 then bmicat1 = 3;
if bmi1 >=40 then bmicat1 = 4;
if bmi2 < 25 then bmicat2 = 1;
if bmi2 >=25 AND bmi2 <30 then bmicat2 = 2;
if bmi2 >=30 AND bmi2 <40 then bmicat2 = 3;
if bmi2 >=40 then bmicat2 = 4;
if avg_bmi < 25 then bmicat = 1;
if avg_bmi >=25 AND avg_bmi <30 then bmicat = 2;
if avg_bmi >=30 AND avg_bmi <40 then bmicat = 3;
if avg_bmi >=40 then bmicat = 4;
run;

proc freq data=three_4;
tables (bmicat1 bmicat2 bmicat)*atriskdefnash/norow;
run;

proc freq data=three_4;
tables bmi1;
where bmicat1 = 3;
run;
```

```
*Diabetes;
data base1; set db.bg1;
run;

data base2; set db.bg2;
run;

proc sort data=base2;
by id;
run;

data base; merge
one (in=a)
base2 (in=b);
by id;
if a=b;
run;

proc freq data=base;
tables bg249a bg249b;
run;

data base_1; set base;
diab = 0;
if bg249a = 1 then diab = 1;
if bg249b = 1 then diab = 2;
run;

proc freq data=base_1;
tables diab*atriskdefnash/norow;
run;

*Hyperlipidemia;
proc freq data=base_1;
tables bg249am*atriskdefnash/norow missing;
run;

*Lab values;
proc freq data=db.lr1;
tables visit;
run;

data lab2; set db.lr2;
where visit = "t0";
ast = lr223;
alt = lr224;
ggt = lr226;
alk = lr225;
```

```
tbil = lr221;
dbil = lr222;
alb = lr228;
tpro = lr227;
inr = lr230;
plat = lr213;
gluc = lr234a;
keep id visit ast alt ggt alk tbil dbil alb tpro inr plat gluc;
run;

data lab1; set db.lr1;
where visit = "t0";
ast = lr123;
alt = lr124;
ggt = lr126;
alk = lr125;
tbil = lr121;
dbil = lr122;
alb = lr128;
tpro = lr127;
inr = lr130;
plat = lr113;
gluc = lr134a;
keep id visit ast alt ggt alk tbil dbil alb tpro inr plat gluc;
run;

data lab3; set db.lr3;
where visit = "t0";
ast = lr323;
alt = lr324;
ggt = lr326;
alk = lr325;
tbil = lr321;
dbil = lr322;
alb = lr328;
tpro = lr327;
inr = lr330;
plat = lr313;
gluc = lr334a;
keep id visit ast alt ggt alk tbil dbil alb tpro inr plat gluc;
run;

data labs; set lab1 lab2 lab3;
run;

proc sort data=labs nodupkey;
by id;
run;
```

```
data labs_1; merge
one (in=a)
labs (in=b);
by id;
if a=b;
run;

*Creating AST/ALT ratio and converting platelet count;
data labs_2; set labs_1;
ast_alt_ratio = ast/alt;
platelet = plat/1000;
run;

proc means data=labs_2 n mean std maxdec=2;
var ast alt ast_alt_ratio ggt alk tbil dbil alb tpro inr platelet
gluc;
class atriskdefnash;
run;

proc means data=labs_2 n mean std maxdec=2;
var ast alt ast_alt_ratio ggt alk tbil dbil alb tpro inr platelet
gluc;
run;

*Fibroscan;
data fr1; set db.fr1;
where visit = "t0";
probe = fr112;
lsm1 = fr115a;
cap1 = fr116a;
lsm2 = fr119a;
cap2 = fr120a;
keep id visit probe lsm1 cap1 lsm2 cap2;
run;

data fr2; set db.fr2;
where visit = "t0";
probe = fr212;
lsm1 = fr215a;
cap1 = fr216a;
lsm2 = fr219a;
cap2 = fr220a;
keep id visit probe lsm1 cap1 lsm2 cap2;
run;

data fr3; set db.fr3;
where visit = "t0";
```

```
probe = fr312;
lsm1 = fr315a;
cap1 = fr316a;
lsm2 = fr319a;
cap2 = fr320a;
keep id visit probe lsm1 cap1 lsm2 cap2;
run;

data fr4; set db.fr4;
where visit = "t0";
probe = fr412;
lsm1 = fr415a;
cap1 = fr416a;
lsm2 = fr419a;
cap2 = fr420a;
keep id visit probe lsm1 cap1 lsm2 cap2;
run;

data fr5; set db.fr5;
where visit = "t0";
probe = fr512;
lsm1 = fr515a;
cap1 = fr516a;
lsm2 = fr519a;
cap2 = fr520a;
keep id visit probe lsm1 cap1 lsm2 cap2;
run;

data fr; set fr1 fr2 fr3 fr4 fr5;
run;

proc sort data=fr nodupkey;
by id;
run;

data fr_1; merge
one (in=a)
fr (in=b);
by id;
if a=b;
run;

*Fibroscan LSM;
proc means data=fr_1 mean std;
var lsm1 cap1;
/*class atriskdefnash; */
run;
```

```
*Fiboscan probe type;
proc freq data=fr_1;
tables probe*atriskdefnash/norow;
run;
```