Dataset Integrity Check for the Prevalence and Impact of Hepatitis E Virus (HEV) Infection in the HBRN Cohort Ancillary Study (HBRN HEV)

Prepared by NIDDK-CR
August 31, 2023

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

An ancillary study from the Hepatitis B Research Network (HBRN) Adult Cohort Study, the Prevalence and Impact of Hepatitis E Virus (HEV) Infection, aimed to determine the prevalence of prior and possible acute HEV infection among persons with chronic hepatitis B virus (HBV), and identify if an association exists between HEV infection and liver disease flares among persons living with chronic HBV.

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 HBRN Adult Cohort folder in the data package. For this replication, variables were taken from datasets in the HBRN Adult Cohort study (“baselinechar.sas7bdat”, “bc.sas7bdat”, and “bp.sas7bdat”), and merged with the HEV ancillary study dataset (“hev.sas7bdat”).

4 Statistical Methods

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

5 Results

For Table 1 in the publication [1], Participants’ Characteristics by Anti-HEV Positivity, 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 to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the HBRN HEV data files to be distributed are a true copy of the study data.

7 References

[1] McGivern DR, Lin HS, Wang J, Benzine T, Janssen HLA, Khalili M, Lisker-Melman M, Fontana RJ, Belle SH, Fried MW. Prevalence and Impact of Hepatitis E Virus Infection Among Persons With Chronic Hepatitis B Living in the US and Canada. Open Forum Infectious Diseases, 6(5), 1-9, April 2019. doi: https://doi.org/10.1093/ofid/ofz175

Table A: Variables used to replicate Table 1 – Participants' Characteristics by Anti-HEV Positivity

Table Variable	**dataset.variable**
Age	baselinechar.age_erl hev.igg_positive hev.igm_positive
Sex	bp.sex hev.igg_positive hev.igm_positive
Race	baselinechar.asian hev.igg_positive hev.igm_positive
Time since migration	bc.cborns hev.igg_positive hev.igm_positive
Education level	bc.educ hev.igg_positive hev.igm_positive
Employment status	bc.work hev.igg_positive hev.igm_positive
ALT	bc.alt hev.igg_positive hev.igm_positive
HBV DNA	bc.bdna hev.igg_positive hev.igm_positive
Genotype	baselinechar.gen_cat hev.igg_positive hev.igm_positive
HBeAg	bc.hbeag hev.igg_positive hev.igm_positive

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

Characteristics	**Pub: Total (n=600)**	**DSIC: Total (n=600)**	**Diff. (n=0)**	**Pub: Anti-HEV (-) (n=426)**	**DSIC: Anti-HEV (-) (n=425)**	**Diff. (n=1)**	**Pub: Anti-HEV (+) (n=174)**	**DSIC: Anti-HEV (+) (n=175)**	**Diff. (n=1)**
Age at current study visit, years	42.1	41.9	0.2	39.8	39.7	0.1	47.7	47.9	0.2
Median (IQR)	(32.8:53.0)	(32.5:52.6)	(0.3:0.4)	(31.4:50.6)	(31.3:50.0)	(0.1:0.6)	(38.1:58.1)	(38.1:57.9)	(0:0.2)
Sex, No. (%)									
Female	304 (50.7)	304 (50.7)	0 (0)	233 (76.6)	233 (76.6)	0 (0)	71 (23.4)	71 (23.4)	0 (0)
Male	296 (49.3)	295 (49.2)	1 (0.1)	193 (65.2)	193 (65.1)	0 (0.1)	103 (34.8)	103 (34.9)	0 (0.1)
Race, No. (%)									
Non-Asian	165 (27.5)	150 (26.6)	15 (0.9)	130 (78.8)	118 (78.7)	12 (0.1)	35 (21.2)	32 (21.3)	3 (0.1)
Asian	434 (72.5)	413 (73.4)	21 (0.9)	296 (68.2)	280 (67.8)	16 (0.4)	138 (31.8)	133 (32.2)	5 (0.4)
Time since migration, No. (%)									
Born in U.S./Canada	106 (19.4)	106 (17.7)	0 (1.7)	89 (84.0)	89 (84.0)	0 (0)	17 (16.0)	17 (16.0)	0 (0)
Born outside U.S./Canada	439 (80.6)	494 (82.3)	55 (1.7)	298 (67.9)	336 (68.0)	38 (0.1)	141 (32.1)	158 (32.0)	17 (0.1)
Education level, No. (%)									
Bachelor's or higher	282 (47.5)	283 (47.2)	1 (0.3)	224 (79.4)	223 (78.8)	1 (0.6)	58 (20.6)	60 (21.2)	2 (0.6)
Less than Bachelor's	312 (52.5)	317 (52.8)	5 (0.3)	202 (64.7)	202 (63.7)	0 (1.0)	110 (35.3)	115 (36.3)	5 (1.0)
Employment status, No. (%)									
Employed, full-time or part-time	447 (74.9)	447 (74.9)	0 (0)	326 (72.9)	325 (72.7)	1 (0.2)	121 (27.1)	122 (27.3)	1 (0.2)
Homemaker, not currently working for pay	25 (4.2)	25 (4.2)	0 (0)	18 (72.0)	18 (72.0)	0 (0)	7 (28.0)	7 (28.0)	0 (0)
Not currently employed	125 (20.9)	125 (20.9)	0 (0)	82 (65.6)	82 (65.6)	0 (0)	43 (34.4)	43 (34.4)	0 (0)
ALT (U/L) Median (IQR)	33 (22:51)	33 (22:52)	0 (0:1)	31 (22:51)	32 (22:52)	1 (0:1)	35 (22:52)	36 (23:56)	1 (1:4)
HBV DNA, $\log_{10}$ IU/mL	3.5	3.9	0.4	3.5	3.8	0.3	3.7	4.1	0.4
Median (IQR)	(2.3:5.3)	(2.6:5.4)	(0.3:0.1)	(2.3:5.1)	(2.8:5.3)	(0.5:0.2)	(2.6:5.6)	(2.9:5.8)	(0.3:0.2)
Genotype, No. (%)									
A	90 (16.2)	85 (16.2)	5 (0)	65 (72.2)	62 (72.9)	3 (0.7)	25 (27.8)	23 (27.1)	2 (0.7)
B	224 (40.4)	215 (41.0)	9 (0.6)	154 (68.8)	146 (67.9)	8 (0.9)	70 (31.3)	69 (32.1)	1 (0.8)
C	176 (31.7)	169 (32.2)	7 (0.5)	123 (69.9)	118 (69.8)	5 (0.1)	53 (30.1)	51 (30.2)	2 (0.1)
D	44 (7.9)	38 (7.2)	6 (0.7)	32 (72.7)	28 (73.7)	4 (1.0)	12 (27.3)	10 (26.3)	2 (1.0)
Other: E, F, or multiple genotypes	21 (3.8)	15 (2.9)	6 (0.9)	17 (81.0)	12 (80.0)	5 (1.0)	4 (19.0)	3 (20.0)	1 (1.0)
HBeAg, No. (%)									
Negative	464 (77.3)	427 (76.1)	37 (1.2)	329 (70.9)	301 (70.5)	28 (0.4)	135 (29.1)	126 (29.5)	9 (0.4)
Positive	136 (22.7)	134 (23.9)	2 (1.2)	97 (71.3)	97 (72.4)	0 (1.1)	39 (28.7)	37 (27.6)	2 (1.1)

Attachment A: SAS Code

```
libname hev "X:\NIDDK\niddk-dr_studies2\HBRN\private_orig_data\HBRN Ancillary Studies\HBRN
Ancillary Studies\HEV";
libname hbrn "X:\NIDDK\niddk-dr_studies2\HBRN\private_created_data\Adult Cohort\Redacted Data";

/***************************/
/* HBRN Ancillary HEV Study */
/* McGivern et al. */
/***************************/

*identifying the cohort;
data dem; set hbrn.baselinechar;
run;

data work.hev; set hev.hev;
orig_id = ID;
if hevselect = 1;
run;

data bp; set hbrn.bp;
run;

*merging Adult cohort data with the ancillary study;
proc contents data=hbrn.baselinechar;
run;

proc contents data=hev;
run;

Proc sort data=dem;
by orig_id;
run;

proc sort data=work.hev;
by orig_id;
run;

data one; merge
hev (in=a)
dem (in=b);
by orig_id;
if a=1;
run;

*age;
proc means data=one n median q1 q3;
```

```
var age_erl;
run;

data two; set one;
pos = 0;
if igg_positive = 1 OR igm_positive = 1 then pos = 1;
run;

proc freq data=two;
tables pos;
run;

proc means data=two n median q1 q3;
var age_erl;
class pos;
run;

*age category;
data three; set two;
age_4cat = 0;
if age_erl < 30 then age_4cat = 1;
if age_erl >= 30 AND age_erl < 40 then age_4cat = 2;
if age_erl >= 40 AND age_erl < 50 then age_4cat = 3;
if age_erl >= 50 then age_4cat = 4;
run;

proc freq data=three;
tables age_4cat*pos/norow;
run;

proc freq data=three;
tables age_cat age_cat2 age_cat3;
run;

*sex;
proc sort data=bp;
by id;
run;

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

data sex; merge
hev (in=a)
bp (in=b);
by id;
if a=b;
```

```
run;

data sex_one; set sex;
pos = 0;
if igg_positive = 1 OR igm_positive = 1 then pos = 1;
run;

proc freq data=sex_one;
tables sex*pos;
run;

*race;
proc freq data=three;
tables asian*pos;
run;

*country of birth;
proc freq data=three;
tables cborn*pos;
run;

data bc; set hbrn.bc;
run;

proc sort data=bc;
by id;
run;

data country; merge
bc (in=a)
hev (in=b);
by id;
if a=b;
run;

data country_one; set country;
pos = 0;
if igg_positive = 1 OR igm_positive = 1 then pos = 1;
run;

proc freq data=country_one;
tables cborns/missing;
run;

data country_two; set country_one;
US_canada = 0;
if cborns = "UNITED STATES OF AMERICA" OR cborns = "CANADA" then US_canada = 1;
run;
```

```
proc freq data=country_two;
tables US_canada*pos;
run;

*education level;
data educ; set country_one;
education = .;
if educ < 8 then education = 0;
if educ >=8 then education = 1;
run;

proc freq data=educ;
tables education*pos;
run;

*employment status;
data work; set educ;
work_status = .;
if work = 1 OR work = 2 then work_status = 1;
if work = 3 then work_status = 2;
if work = 4 or work = 5 then work_status = 3;
run;

proc freq data=work;
tables work_status*pos;
run;

*ALT;
proc means data=work n median q1 q3;
var alt;
class pos;
run;

*HBV DNA;
data bdna; set work;
log_bdna = log10(bdna);
run;

proc means data=bdna n median q1 q3;
var log_bdna;
class pos;
run;

*genotype;
proc freq data=two;
tables gen_cat*pos;
run;
```

```
*HbeAg;
proc freq data=work;
tables hbeag*pos;
run;
```