Dataset Integrity Check for Longitudinal Assessment of Transient Elastography in Cystic Fibrosis (ELASTIC-CF)

Prepared by NIDDK-CR
September 23, 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 Longitudinal Assessment of Transient Elastography in Cystic Fibrosis (ELASTIC-CF) study assessed if combining FibroScan® measurement of liver stiffness transient elastography with ultrasound would improve the prediction of the development of a nodular liver on ultrasound and development of portal hypertension over time in children and young adults with CF. Study visits occurred at baseline and for at least two annual follow-up visits.

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 ELASTIC-CF and the CFLD PUSH folders in their respective data packages. For this replication, variables were taken from the "fibroscan_data_setup.sas7bdat" dataset from the ELASTIC-CF package; and "elg.sas7bdat", "usgrade_data_setup.sas7bdat", "lab_data_setup.sas7bdat", "dop_data_setup.sas7bdat", and "demographics_data_setup.sas7bdat" datasets from the CFLD PUSH package.

4 Statistical Methods

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

5 Results

For Table 1 in the publication [1], Characteristics of Study Participants With Valid Baseline VCTE LSM, 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 ELASTIC-CF data files to be distributed are a true copy of the study data.

7 References

[1] Ye W, Leung DH, Molleston JP, Ling SC, Murray KF, Nicholas JL, Huang S, Karmazyn BW, Harned RK, Masand P, Alazraki AL, Navarro OM, Otto RK Palermo JJ, Towbin AJ, Alonso EM, Karnsakul WW, Schwarzenberg SJ, Seidel GF, Siegel M, Magee J, Narkewicz MR, Freeman JA. Association Between Transient Elastography and Controlled Attenuated Parameter and Liver Ultrasound in Children With Cystic Fibrosis. Hepatology Communications, 5(8), 1362-1372, August 2021. doi: https://doi.org/10.1002/hep4.1719

Table A: Variables used to replicate Table 1 – Characteristics of Study Participants With Valid Baseline VCTE LSM

Table Variable	**dataset.variable**
Age	elg.elgb01adt fibroscan_data_setup.visit_date fibroscan_data_setup.id
Male gender	demographics_data_setup.gender fibroscan_data_setup.id
Ethnicity	demographics.ethnicity fibroscan_data_setup.id
Race	demographics.race fibroscan_data_setup.id
Pseudomonas-positive at enrollment	lab_data_setup.pseudomonas fibroscan_data_setup.id
US grade	usgrade_data_setup.congrade fibroscan_data_setup.id
Platelets	lab_data_setup.platelet fibroscan_data_setup.id
APRI	lab_data_setup.apri fibroscan_data_setup.id
FIB-4	lab_data_setup.fib4 fibroscan_data_setup.id
GGT	lab_data_setup.ggtp_ul fibroscan_data_setup.id
Spleen size	dop_data_setup.spleen_size_cm fibroscan_data_setup.id
Spleen size z-score	dop_data_setup.spleen_size_cm_zscore fibroscan_data_setup.id

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

Characteristic	Pub (n=133)	DSIC (n=135)	Diff. (n=2)
Age (years)			
Mean (SD)	14.3 (3.4)	14.5 (3.3)	0.2 (0.1)
Median (IQR)	14.7 (11.9, 17)	14.9 (12.3, 17.3)	0.2 (0.4, 0.3)
Min, Max	7, 21	7.5, 21.1	0.5, 0.1
Male gender, n (%)	70 (52.6)	72 (53.3)	2 (0.7)
Ethnicity, n (%)			
Hispanic	10 (7.5)	10 (7.4)	0 (0.1)
Non-Hispanic	123 (92.5)	125 (92.6)	2 (0.1)
Race, n (%)			
Asian	1 (0.8)	1 (0.7)	0 (0.1)
Black or African American	2 (1.5)	2 (1.5)	0 (0)
Multiracial	1 (0.8)	1 (0.7)	0 (0.1)
White	129 (97)	131 (97)	2 (0)
Pseudomonas-positive at enrollment, n (%)	41 (30.8)	59 (45.4)	18 (14.6)
US grade			
NL	72 (54.1)	70 (57.4)	2 (3.3)
HTG	18 (13.5)	21 (17.2)	3 (3.7)
HMG	17 (12.8)	13 (10.7)	4 (2.1)
NOD	26 (19.5)	18 (14.8)	8 (4.7)
Platelets ($10^3/mm^3$)			
Mean (SD)	288.8 (95.6)	287.2 (99.6)	1.6 (4)
Median (IQR)	281.5 (232, 347)	284 (234.5, 343.5)	2.5 (2.5, 3.5)
Min, Max	51, 628	51, 618	0, 10
APRI			
Mean (SD)	0.6 (0.8)	0.7 (1.0)	0.1 (0.2)
Median (IQR)	0.4 (0.3, 0.7)	0.4 (0.3, 0.7)	0 (0, 0)
Min, Max	0.1, 5.2	0.1, 7.2	0, 2
FIB-4			
Mean (SD)	0.3 (0.3)	0.3 (0.3)	0 (0)
Median (IQR)	0.2 (0.2, 0.3)	0.2 (0.2, 0.3)	0 (0, 0)
Min, Max	0.1, 1.6	0.1, 2.1	0, 0.5
GGT (U/L)			
Mean (SD)	30.8 (37.1)	28.3 (31.6)	2.5 (5.5)
Median (IQR)	17 (13, 40)	17 (12, 28)	0 (1, 12)
Min, Max	4, 274	4, 239	0, 35
Spleen size (cm)			
Mean (SD)	11.1 (2)	11.4 (2.3)	0.3 (0.3)
Median (IQR)	10.8 (9.8, 12.3)	11 (10, 12.6)	0.2 (0.2, 0.3)
Min, Max	7.1, 18	7.1, 23.8	0, 5.8
Spleen size z-score			
Mean (SD)	1.2 (2.1)	1.5 (2.5)	0.3 (0.4)
Median (IQR)	0.9 (-0.1, 2.3)	1.0 (-0.1, 2.5)	0.1 (0, 0.2)
Min, Max	-3.8, 8	-3.8, 14.5	0, 6.5

Attachment A: SAS Code

```
libname elastic "X:\NIDDK\niddk-dr_studies1\ELASTIC\private_created_data\DEID";
libname push "X:\NIDDK\niddk-dr_studies1\CFLD-PUSH\private_created_data\R_script\derived";
libname pushraw "X:\NIDDK\niddk-dr_studies1\CFLD-PUSH\private_created_data\DEID\raw data\raw";
libname ellink "X:\NIDDK\niddk-dr_studies1\ELASTIC\private_created_data\DEID\ELASTIC_Linking";

/***********************************/
/* ELASTIC - CF DSIC */
/* Ye et al. */
/***********************************/

*keeping only the baseline or first observation for each participant from elastic;
data one; set elastic.fibroscan_data_setup; *creating a temp dataset for manipulation;
run;

*sorting based on ID and visit;
proc sort data=one;
by id ctxvisno;
run;

data elone; set elastic.fibroscan_data_setup; *keeping only the first observation for each ID (135
participants);
by ID;
if first.ID;
keep ID visit_date ;
run;

*creating variables to capture windows of either 6 months or 1 year from visit data in order to find the
closest labs in the PUSH data;
data eltwo; set elone;
mo6_window_start = visit_date - 182;
mo6_window_end = visit_date + 182;
yr_window_start= visit_date - 365;
yr_window_end = visit_date + 365;
drop visit_date;
run;

*merging the new date window dataset with the linking file;
data link; set elastic.link;
run;

proc sort data=link;
by id;
run;

data elthree; merge
```

```
eltwo (in=a)
link (in=b);
by id;
if a = b;
run;

*need to create full file of necessary variables from PUSH with the linking file back to the common PID;
data pushgrade; set push.usgrade_data_setup;
run;

data pushlab; set push.lab_data_setup;
run;

data push_dop; set push.dop_data_setup;
run;

data pushdem; set push.demographics_data_setup;
run;

*combining the grade and labs data;
proc sort data=pushgrade;
by id;
run;

proc sort data=pushlab;
by id;
run;

proc sort data=push_dop;
by id;
run;

proc sort data=pushdem;
by id;
run;

data push_one; merge
pushgrade (in=a)
pushlab (in=b)
push_dop (in=c)
pushdem (in=d);
by ID;
if a=b=c=d;
run;

*connecting the linking file to the new push dataset;
data pushlink; set push.link;
run;
```

```
proc sort data=pushlink;
by id;
run;

data push_two; merge
push_one (in=a)
pushlink (in=b);
by id;
if a=b;
run;

*combining the push data and the elastic data in order to keep the specific time window;
proc sort data=push_two;
by ctxsbjid;
run;

proc sort data=elthree;
by ctxsbjid;
run;

data combined_one; merge
push_two (in=a)
elthree (in=b);
by ctxsbjid;
if a=b;
run;

*creating a dataset with the time window set at 6 months and a dataset with the window set at 1 yr;
data mo6window; set combined_one;

if visit_date <= mo6_window_end AND visit_date >= mo6_window_start;
run;

proc sort data=mo6window;
by ctxsbjid visit_date;
run;

proc freq data=mo6window;
tables ctxsbjid;
run;

data mo6window1; set mo6window; *keeping only the earliest observation if there are multiple obs for
any subject iD;
by ctxsbjid visit_date;

if first.ctxsbjid;
run;
```

```
*109 observations;
*creating the 1yr window dataset;
data yrwindow; set combined_one;

if visit_date <= yr_window_end AND visit_date >= yr_window_start;
run;

proc sort data=yrwindow;
by ctxsbjid visit_date;
run;

data yrwindow1; set yrwindow;
by ctxsbjid visit_date;

if first.ctxsbjid;
run;
*130 observations;
*will uses yr window instead of 6mo window as there are more observations;

/**********************/
/* Labs replication */
/**********************/

*pseudomonas;
proc freq data=combined_one;
tables pseudomonas*ctxvisno;
run;

proc freq data=yrwindow1;
tables pseudomonas;
run;

proc freq data=mo6window1;
tables pseudomonas;
run;

*us grade;
proc freq data=mo6window; *6month was closer;
tables congrade;
run;

*platelets, apri;
proc means data=yrwindow1 n mean std median q1 q3 min max;
var platelet apri fib4 ggtp_ul;
run;

*spleen;
proc means data=yrwindow1 n mean std median q1 q3 min max;
```

```
var spleen_size_cm spleen_size_cm_zscore;
run;

/************/
/* age calc */
/************/
*calculating age from the eligibility raw dataset and the fibroscan data;
*reviewing needed variables from eligibility raw dataset from PUSH;
proc contents data=pushraw.elg;
run;

proc freq data=pushraw.elg;
tables ctxvisdt elgb01adt;
run;

proc freq data=elastic.fibroscan_data_setup;
tables visit_date;
run;

*creating a new PUSH datset;
proc contents data=pushraw.elg;
run;

proc freq data=pushraw.elg;
tables ctxvisdt elgb01adt;
run;

proc freq data=elastic.fibroscan_data_setup;
tables visit_date;
run;

data elg; set pushraw.elg;
keep id elgb01adt;
run;

data pulink; set push.link;
keep ctxsbjid id;
run;

*merging push data with PUSH linking file;
proc sort data=elg;
by id;
run;

proc sort data=pulink;
by id;
run;
```

```
data puage; merge
elg (in=a)
pulink (in=b);
by id;
drop id;
run;

*creating a specific dataset for age from ELASTIC fibroscan dataset;
data age; set elastic.fibroscan_data_setup;
keep id visit_date;
run;

data ellink; set elastic.link;
keep ctxsbjid id;
run;

*sorting linking dataset and elastic dataset;
proc sort data=age;
by id;
run;

proc sort data=ellink;
by id;
run;

data elage; merge
age (in=a)
ellink (in=b);
by id;
if a=b;
drop id;
run;

*merging PUSH and Elastic age datasets;
proc sort data=puage;
by ctxsbjid;
run;

proc sort data=elage;
by ctxsbjid;
run;

data finage; merge
puage (in=a)
elage (in=b);
by ctxsbjid;
if a = b;
run;
```

```
*calculating age;
data finage1; set finage;
age = (visit_date - elgb01adt)/365.25;
run;

*Sort the dataset by ID and AGE;
proc sort data=finage1;
 by ctxsbjid age;
run;

*Create a new dataset with the first and last observations for each ID and flag the first observation;
data finage2; set finage1;
by ctxsbjid;
if first.ctxsbjid then flag_first = 1;
else flag_first = 0;
if first.ctxsbjid or last.ctxsbjid;
run;

*assessing whether the first observation per ID or last is the correct age;
proc means data=finage2 n mean std median q1 q3 min max;
var age;
class flag_first;
run;

*age at the first obs per ID is correct;

/**********************************/
*the rest of the demographics;
/**********************************/

/* PUSH Combined datasets for Table 1, excluding age*/
*creating the PUSH datasets with ID linkage;
proc freq data=push.lab_data_setup;
tables ctxvisno;
run;

proc freq data=push.usgrade_data_setup;
tables ctxvisno;
run;

proc freq data=push.dop_data_setup;
tables ctxvisno;
run;

proc sort data=push.link;
by ID;
run;
```

```
data dem; set push.demographics_data_setup;
run;

data lab; set push.lab_data_setup;
if ctxvisno = "BASELINE";
drop ctxvisno;
run;

data usgrade; set push.usgrade_data_setup;
if ctxvisno = "SCREENING";
drop ctxvisno;
run;

proc freq data=push.dop_data_setup;
tables ctxvisno;
run;

data dop; set push.dop_data_setup;
if ctxvisno = "SCREENING";
drop ctxvisno;
run;

proc sort data=dem;
by id ;
run;

proc sort data=lab;
by id;
run;

proc sort data=usgrade;
by id;
run;

proc sort data=dop;
by id;
run;

data push1; merge
push.link (in=a)
dem (in=b)
lab (in=c)
usgrade (in=d)
dop (in=e);
by ID;
run;
```

```
/* Linkning ELASTIC and PUSH*/
*linking elastic to push demographics;
data elasticlink; set elastic.link;
run;

data fibro; set elastic.fibroscan_data_setup;
keep id ctxvisno;
run;

proc sort data=elasticlink;
by id;
run;

proc sort data=fibro;
by id;
run;

data elasticlink2; merge
elasticlink (in=a)
fibro (in=b);
by id;
if a=b;
run;

proc sort data=work.elasticlink2;
by ctxsbjid ctxvisno;
run;

data elast1; set work.elasticlink2;
by ctxsbjid;
if first.ctxsbjid;
drop id ctxvisno consentdt;
run;

proc sort data=elast1;
by ctxsbjid;
run;

proc sort data=push1;
by ctxsbjid;
run;

data dem2; merge
push1 (in=a)
elast1 (in=b);
by ctxsbjid;
if a = b;
run;
```

```
*sex, ethnicity, race;
proc freq data=dem2;
tables gender ethnicity race;
run;
```