Dataset Integrity Check for the Boston Area Community Health (BACH) II Data Files

Prepared by Allyson Mateja
IMS, Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
June 20, 2016

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 Boston Area Community Health cohort is a multi-stage 1:1:1 stratified random sample of self-identified African American, Hispanic, and white adults from three Boston inner city areas. The goal of the study was to test among diabetes-free urban community-dwelling adults the hypothesis that the proportion of African genetic ancestry is positive associated with glycaemia, after accounting for other continental ancestry proportions, BMI, and socioeconomic status. It was found that a greater proportion of African genetic ancestry is independently associated with higher fasting glucose levels in a non-diabetic community-based cohort, even accounting for other ancestry proportions, obesity, and socioeconomic status. The results suggest that differences between African-Americans and whites in type 2 diabetes risk may include genetically mediated differences in glucose homeostasis.

3 Archived Datasets

The SAS data files, as provided by the Data Coordinating Center (DCC), are located in the data package. For this replication, variables were taken from the “bach2publicuse.sas7bdat” data file.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Maserejian, et al. in The Journal of Urology in 2014 [1]. To verify the integrity of the datasets, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Characteristics of BACH-II Participants, Overall and by Sex and Presence of LUTS at Baseline, Table A lists the variables that can be used in the replication. Table B compares the results calculated from the archived data file to the results published in Table 1. The results of the replication are a close match.

6 Conclusions

The NIDDK repository is confident that the BACH II data files to be distributed are a true copy of the manuscript data.

7 References

[1] Maserejian, N.N., Chen, S., Chiu, G.R., Araujo, A.B., Kupelian, V., Hall, S.A., and McKinlay, J.B. Treatment Status and Progression or Regression of Lower Urinary Tract Symptoms among Adults in a General Population Sample. J Urol (2014) 191 (1): 107-113.

Table A: Variables used to replicate Table 1: Characteristics of BACH-II Participants, Overall and by Sex and Presence of LUTS at Baseline

Table Variable	**Dataset Variable**
Sex	gender_1
LUTS Present at Baseline	luts
Age, years	age
Age category	agegrp5
Race/ethnicity	re
BMI	bmi_mi
LUTS Medication, Baseline	S12080800, S92080000, spoab, spui, spbph
LUTS Medication, Follow-up	S12080800_1, S92080000_1, spoab_1, spui_1, spbph_1

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

Characteristic	Total Manuscript N=4,144	Total DSIC N=4,144	Difference N=0
Age (years) mean (SE)	50.3 (0.3)	50.3 (0.2)	0 (0.1)
Age category, %			
<40 y	24.4	24.4	0
40-49 y	27.5	27.5	0
50-59 y	24.7	24.7	0
60-69 y	16.4	16.4	0
70+ y	7	7	0
Race/ethnicity, %			
Black	32	32	0
Hispanic	32.4	32.4	0
White	35.6	35.6	0
BMI, mean (SE) kg/m²	29.9 (0.1)	29.9 (0.1)	0 (0)
LUTS Medication, %			
Both baseline and follow-up	2.5	2.5	0
Baseline only	1.6	1.6	0
Follow-up only	4.2	4.1	0.1
None	91.7	91.8	0.1

Characteristic	Male Manuscript (N=1,610)	Male DSIC (N=1,610)	Difference (N=0)	Female Manuscript (N=2,534)	Female DSIC (N=2,534)	Difference (N=0)
Age (years) mean (SE)	49.7 (0.4)	49.7 (0.3)	0 (0.1)	50.7 (0.3)	50.7 (0.2)	0 (0.1)
Age category, %						
<40 y	25	25	0	24	24	0
40-49 y	29.8	29.8	0	26	26.1	0.1
50-59 y	23.7	23.7	0	25.4	25.4	0
60-69 y	15.2	15.2	0	17.1	17.1	0
70+ y	6.3	6.3	0	7.4	7.4	0
Race/ethnicity, %						
Black	30.2	30.2	0	33.2	33.2	0
Hispanic	30.5	30.5	0	33.5	33.5	0
White	39.3	39.3	0	33.3	33.3	0
BMI, mean (SE) kg/m²	28.9 (0.2)	28.9 (0.1)	0 (0.1)	30.6 (0.2)	30.6 (0.1)	0 (0.1)
LUTS Medication, %						
Both baseline and follow-up	3.7	3.7	0	1.7	1.7	0
Baseline only	2	2.1	0.1	1.3	1.3	0
Follow-up only	5.5	5.5	0	3.4	3.2	0.2
None	88.7	88.7	0	93.6	93.8	0.2

Characteristic	LUTS Present at Baseline = No Manuscript (N=3,302)	LUTS Present at Baseline = No DSIC (N=3,302)	Difference (N=0)	LUTS Present at Baseline = Yes Manuscript (N=842)	LUTS Present at Baseline = Yes DSIC (N=842)	Difference (N=0)
Age (years) mean (SE)	49.4 (0.3)	49.4 (0.2)	0 (0.1)	53.7 (0.4)	53.7 (0.4)	0 (0)
Age category, %						
<40 y	27.5	27.5	0	12.3	12.4	0.1
40-49 y	27.6	27.6	0	27.2	27.2	0
50-59 y	23.4	23.4	0	29.8	29.8	0
60-69 y	15.2	15.2	0	21	21	0
70+ y	6.3	6.3	0	9.7	9.6	0.1
Race/ethnicity, %						
Black	31.5	31.5	0	33.9	34	0.1
Hispanic	33.3	33.3	0	28.8	28.7	0.1
White	35.2	35.2	0	37.2	37.3	0.1
BMI, mean (SE) kg/m²	29.5 (0.1)	29.5 (0.1)	0 (0)	31.6 (0.3)	31.6 (0.3)	0 (0)
LUTS Medication, %						
Both baseline and follow-up	1.5	1.5	0	6.5	6.4	0.1
Baseline only	1	1	0	3.8	3.8	0
Follow-up only	2.9	2.9	0	9.3	8.9	0.4
None	94.6	94.6	0	80.4	80.9	0.5

Attachment A: SAS Code

```
*** BACH II DSIC;
*** Programmer: Allyson Mateja;
*** Date: 6/14/16;

title 'Bach 2 DSIC';
title2 ' ';

proc format;
       value sexf 1 = 'M'
                  2 = 'F';

       value yesnof 1 = 'Yes'
                   0,2 = 'No';

       value agegrpf 1 = '<40 y'
                     2 = '40-49 y'
                     3 = '50-59 y'
                     4 = '60-69 y'
                     5 = '70+ y';

libname bach2 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK';
libname m01 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_1/';
libname m02 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_2/';
libname m03 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_3/';
libname m04 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_4/';
libname m05 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_5/';
libname m06 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_6/';
libname m07 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_7/';
libname m08 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_8/';
libname m09 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_9/';
libname m10 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_10/';
libname m11 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_11/';
libname m12 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_12/';
libname m13 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_13/';
libname m14 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_14/';
libname m15 '/prj/niddk/ims_analysis/BACH2/BACH 2 NIDDK/To NIDDK/m001_15/';

options nofmterr;

data bach;
       set bach2.bach2publicuse;

data m01;  set m01.m001_1;
data m02;  set m02.m001_2;
data m03;  set m03.m001_3;
data m04;  set m04.m001_4;
data m05;  set m05.m001_5;
data m06;  set m06.m001_6;
data m07;  set m07.m001_7;
data m08;  set m08.m001_8;
data m09;  set m09.m001_9;
```

```
data m10;  set m10.m001_10;
data m11;  set m11.m001_11;
data m12;  set m12.m001_12;
data m13;  set m13.m001_13;
data m14;  set m14.m001_14;
data m15;  set m15.m001_15;

proc contents data = bach;
proc contents data = m01;

data mi_data;
       set m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12 m13 m14 m15;

proc sort data = mi_data;
       by blinded_master_id _imputation_;

proc sort data = bach;
       by blinded_master_id _imputation_;

data bach;
       merge bach    (drop = spui spui_1 spbph spbph_1)
             mi_data (keep=blinded_master_id _imputation_ S12080800 S92080000 spoab spui spbph   S12080800_1 S92080000_1 spoab_1 spui_1 spbph_1);
       by blinded_master_id _imputation_;

data bach2.bach2_use;
       set bach;

data bach;
       set bach;
       if S12080800=1 or S92080000=1 or spoab=1 or spui=1 or spbph=1 then trt_base=1;
  else if S12080800=2 and S92080000=2 and spoab =2 and spui=2 and spbph=2 then trt_base=2;
       if S12080800_1=1 or S92080000_1=1 or spoab_1=1 or spui_1=1 or spbph_1=1 then trt_fu=1;
  else if S12080800_1=2 and S92080000_1=2 and spoab_1 =2 and spui_1=2 and spbph_1=2 then trt_fu=2;

data bach_subjects;
       set bach;
       by blinded_master_id;
       retain total_bmi bmi_count total_age age_count total_luts total_trt_base trt_base_count total_trt_fu trt_fu_count 0;
       if first.blinded_master_id then do;
              total_bmi = 0;
              bmi_count = 0;
              total_age = 0;
              age_count = 0;
              total_luts = 0;
              total_trt_base = 0;
              trt_base_count = 0;
              total_trt_fu = 0;
              trt_fu_count = 0;
       end;
       if bmi_mi ne . then do;
              total_bmi = total_bmi + bmi_mi;
              bmi_count = bmi_count + 1;
       end;
       if age ne . then do;
              total_age = total_age+age;
              age_count = age_count + 1;
```

```
        end;
        total_luts = total_luts + luts;
        if trt_base ne . then do;
                total_trt_base = total_trt_base + trt_base;
                trt_base_count = trt_base_count + 1;
        end;
        if trt_fu ne . then do;
                total_trt_fu = total_trt_fu + trt_fu;
                trt_fu_count = trt_fu_count + 1;
        end;
        if last.blinded_master_id then do;
                avg_bmi = total_bmi/bmi_count;
                avg_age = total_age/age_count;
                avg_trt_fu = total_trt_fu/trt_fu_count;
                avg_trt_base = total_trt_base/trt_base_count;
                luts_imp = round(total_luts/15,1);
                output;
        end;

data bach_subjects;
        set bach_subjects;
        if avg_trt_fu in (2, .) then luts_followup = 0;
        if avg_trt_base in (2, .) then luts_base = 0;
        if avg_trt_fu = 1 then luts_followup = 1;
        if avg_trt_base = 1 then luts_base = 1;

proc means data = bach_subjects n mean stderr;
        var avg_age;
        title3 'Table 1 - Total, Age';

proc freq data = bach_subjects;
        tables agegrp5;
        format agegrp5 agegrpf.;
        title3 'Table 1 - Total, Age Category';

proc freq data = bach_subjects;
        tables re;
        title3 'Table 1 - Total, Race';

proc means data = bach_subjects n mean stderr;
        var avg_bmi;
        title3 'Table 1 - Total, BMI';

proc freq data = bach_subjects;
        tables luts_base*luts_followup /list missing;
        format luts_base luts_followup yesnof.;
        title3 'Table 1 - Total, LUTS Medication';

proc freq data = bach_subjects;
        tables gender_1;
        format gender_1 sexf.;
        title3 'Table 1 - Sex';

proc sort data = bach_subjects;
        by gender_1;
```

```
proc means data = bach_subjects n mean stderr;
       var avg_age;
       class gender_1;
       format gender_1 sexf.;
       title3 'Table 1 - Age, by Sex';

proc freq data = bach_subjects;
       tables agegrp5;
       by gender_1;
       format gender_1 sexf.
              agegrp5 agegrpf.;
       title3 'Table 1 - Age Category, by Sex';

proc freq data = bach_subjects;
       tables re;
       by gender_1;
       format gender_1 sexf.;
       title3 'Table 1 - Race, by Sex';

proc means data = bach_subjects n mean stderr;
       var avg_bmi;
       class gender_1;
       format gender_1 sexf.;
       title3 'Table 1 - BMI, by Sex';

proc freq data = bach_subjects;
       tables luts_base*luts_followup /list missing;
       by gender_1;
       format gender_1 sexf.
              luts_base luts_followup yesnof.;
       title3 'Table 1 - LUTS Medication, by Sex';

proc freq data = bach_subjects;
       tables luts_imp;
       format luts_imp yesnof.;
       title3 'Table 1 - LUTS Present at Baseline';

proc sort data = bach_subjects;
       by luts_imp;

proc means data = bach_subjects n mean stderr;
       var avg_age;
       class luts_imp;
       format luts_imp yesnof.;
       title3 'Table 1 - Age, by LUTS Present at Baseline';

proc freq data = bach_subjects;
       tables agegrp5;
       by luts_imp;
       format luts_imp yesnof.
              agegrp5 agegrpf;
       title3 'Table 1 - Age Category, by LUTS Present at Baseline';

proc freq data = bach_subjects;
       tables re;
       by luts_imp;
```

```
        format luts_imp yesnof.;
        title3 'Table 1 - Race, by LUTS Present at Baseline';

proc means data = bach_subjects n mean stderr;
        var avg_bmi;
        class luts_imp;
        format luts_imp yesnof.;
        title3 'Table 1 - BMI, by LUTS Present at Baseline';

proc freq data = bach_subjects;
        tables luts_base*luts_followup /list missing;
        by luts_imp;
        format luts_imp luts_base luts_followup yesnof.;
        title3 'Table 1 - LUTS Medication, by LUTS Present at Baseline';
```