Dataset Integrity Check for the Barretts Esophagus Data Files

Prepared by David Ruggieri
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
August 28, 2014

Table of 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

Genetic influences may be discerned in families that have multiple affected members and may manifest as an earlier age of cancer diagnosis. In this study we determine whether cancers develop at an earlier age in multiplex Familial Barrett's Esophagus (FBE) kindreds, defined by 3 or more members affected by Barrett's esophagus (BE) or esophageal adenocarcinoma (EAC).

3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the barretts_esophagus data package. For this replication, variables were taken from the "Comprehensive_V8_forDavid" dataset and modified as per instruction to match the study data used in the publications.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Chak et al [1] in the Cancer Epidemiol Biomarkers journal in February of 2012.

To verify the integrity of the dataset, descriptive statistics were computed (table 1, table 2, and table 4).

5 Results

Table 1 in the publication [1], Individuals with BE and esophageal cancer, by family type. Our Table A lists the variables we used in our replication and Table B compare the results calculated from the archived data file to the results published in Table 1. The results of the replication are identical to those listed in the table.

Table 2 in the publication [1] Age of diagnosis and cancer stage for individuals with esophageal cancer. Our Table C lists the variables we used in our replication and Table D compares the results calculated from the archived data file to the results published in supplemental table 2. The differences in n counts for the average age at cancer diagnosis were a result of our exclusion of patients who were missing an age at cancer diagnosis and therefore did not contribute to the average from that portion of the table. When the patients with a missing age at cancer diagnosis are included, the counts match the manuscript perfectly. The results of the replication are within expected results.

Table 4 in the publication [1] Symptoms and exposures according to family type (includes affected individuals in the non-familial group and all individuals in the duplex and multiplex kindreds). Our Table E lists the variables we used in our replication and Table F compares the results calculated from the archived data file to the results published in supplemental table 4. The results of the replication are identical to those listed in the table.

6 Conclusions

The NIDDK repository is confident that the Barrett's Esophagus data files to be distributed are within expected results.

7 References

1. Amitabh Chak, Yanwen Chen, Jaime Vengoechea, Marcia I. Canto, Robert Elston, Gary W. Falk MD, William M. Grady, Kishore Guda, Margaret Kinnard, Sanford Markowitz, Sumeet Mittal, Ganapathy Prasad, Nicholas Shaheen, Joseph E. Willis, Jill Barnholtz-Sloan

Table A: Variables used to replicate Table 1: Table 1 Frequency of renal insufficiency at baseline and Year 8

Table Variable	Variables used in replication of the Table 1 Dataset
Nonfamilial/ Duplex/ Multiplex Grouping	is_multi
Probands/ Relatives grouping	proband
BE / Esophageal Cancer grouping	diagnosis
Gender	sex
Race	race

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

	Nonfamilial (mansucript)	Nomfamilial (calculated)	Duplex (manusript)	Duplex (calculated)	Mutliplex (manuscript)	Multiplex (calculated)
Probands	830	830	275	275	41	41
Relatives	N/a	N/a	146	146	105	105
Total Individuals	830	830	421	421	146	146
BE	556	556	288	288	103	103
Esophageal Cancer	274	274	133	133	43	43
Gender						
Men	693/830	693/830	326/421	326/421	110/146	110/146
Women	137/830	137/830	95/421	95/421	36/146	36/146
White Race	803/828	803/828	390/392	390/392	137/137	137/137

Table C: Variables used to replicate table 2: Age of diagnosis and cancer stage for individuals with esophageal cancer

Table Variable	Variables used in replication of the Table 2 Dataset
Nonfamilial/ Duplex/ Multiplex Grouping	is_multi
Age at cancer diagnosis	ageatcadiagnosis
Cancer Stage	stage

Table D: Comparison of values computed in integrity check to reference article table 2.

	Nonfamilial (mansucript)	Nomfamilial (calculated)	Duplex (manusript)	Duplex (calculated)	Mutliplex (manuscript)	Multiplex (calculated)
Average Age at Cancer Diagnosis	62.9 (n=274)	62.9 (n=211)	62.6 (n= 133)	62.6 (n= 104)	57.6 (n=43)	57.6 (n=38)
Cancer Stage						
I	55/195	55/195	13/57	13/57	2/16	2/16
II	49/195	49/195	15/57	15/57	0	0
III	66/195	66/195	19/57	19/57	10/16	10/16
IV	25/195	25/195	10/57	10/57	4/16	4/16

Table E: Variables used to replicate table 4: Symptoms and exposures according to family type (includes affected individuals in the non-familial group and all individuals in the duplex and multiplex kindreds)

Table Variable	Variables used in replication of the Table 4 Dataset
Nonfamilial/ Duplex/ Multiplex Grouping	is_multi
BE/EAC/EGJAC	diagnosis
Heartburn Present	hadheartburn
Regurgiation Present	hadacidregurgitation
Smoking	smokecigarettes
Alcohol	alcoholicdrinksperweek

Table F: Comparison of values computed in integrity check to reference article supplemental Table 4 values

	Nonfamilial (mansucript)	Nomfamilial (calculated)	Duplex (manusript)	Duplex (calculated)	Mutliplex (manuscript)	Multiplex (calculated)
BE/EAC/EGJAC						
Affected	830	830	421	421	146	146
Unaffected	N/A	N/A	3071	3071	1065	1065
Total	830	830	3438	3438	1211	1211
Heartburn Present						
Yes	507/724	507/724	339/528	339/528	178/263	178/263
No	211/724	211/724	183/528	183/528	84/263	84/263
Unknown	6/724	6/724	6/528	6/528	1/263	1/263
Regurgiation Present						
Yes	475/717	475/717	301/525	301/525	114/164	114/164
No	233/717	233/717	216/525	216/525	118/264	118/264
Unknown	9/717	9/717	8/525	8/525	2/264	2/264
Smoking						
Yes	466/725	466/725	271/529	271/529	120/263	120/263
No	251/725	251/725	253/529	253/529	140/263	140/263
Unknown	8/725	8/725	5/259	5/259	3/263	3/263
Alcohol						
Yes	581/723	581/723	446/527	446/527	238/261	238/261
No	126/723	126/723	74/527	74/527	22/261	22/261
Unknown	16/723	16/723	7/527	7/527	1/261	1/261

```
%let flnm = %sysfunc(getoption(sysin));
title "Program saved as: &FLNM.";
title2 "Checks tables 1, 2, and 4 on nihms345321.";

/*************************************************************************************************************************************
Programmer: Dave Ruggieri
Date: 21 July 2014

Function/Notes: This program checks tables 1, 2, and 4 on nihms345321.
*************************************************************************************************************************************/
*  Input file  *;
***************;
libname libin "/prj/niddk/ims_analysis/barrett_esophagus/private_created_data/sasfile/";

*************;
*  Formats  *;
*************;
proc format;
  value famtyp
    0 = 'None'
    1 = 'Non-Familial'
    2 = 'Duplex'
    3-high = 'Multiplex'
    ;

  value $ dx1fmt
    'BE'   = 'Cancer'
    'ECA'  = 'Cancer'
    'JCA'  = 'Cancer'
    'SSBE' = 'Cancer'
    other  = 'None'
    ;

  value $ dx2fmt
    'BE'   = 'BE'
    'ECA'  = 'ECA'
    'JCA'  = 'ECA'
    'SSBE' = 'BE'
    other  = 'None'
    ;

  value $ dx3fmt
    'BE','ECA','JCA','SSBE'   = 'Affected'
    other  = 'Unaffected'
    ;

  value missfmt
    . = 'Missing'
    0 = '0'
    other = 'Other'
    ;

  value $ racefmt
    'White' = 'White'
    other = 'Other'
    ;
```

```
  value $ drinkfmt
    'Unknown' = 'Unknown'
    'None' = 'No'
    other = 'Yes'
    ;

*************;
*  Table 1  *;
*************;
data newdata_tab;
  set libin.barretts_analysis_nihms345321;

proc freq data = newdata_tab;
  where diagnosis ^= '';
  tables proband*is_multi
         diagnosis*is_multi
         sex*is_multi
         race*is_multi
         /missing;
  format diagnosis $dx2fmt. race $racefmt.;
  title4 'Table 1';

proc sort data = newdata_tab;
  by is_multi;

proc means data = newdata_tab;
  where diagnosis in('ECA', 'JCA');
  by is_multi;
  var AgeAtCADiagnosis;
  title4 'Table 2: age at CA diagnosis with N counts equal to non-missing contributors';
  title5 'Subset to ECA, and JCA';

proc means data=newdata_tab mean median;
  where diagnosis in ("ECA","JCA");
  class is_multi;
  var ageatcadiagnosis;
  title4 'Table 2: age at CA diagnosis with N counts that include patients with missing ageatcadiagnosis to match the counts in the paper';
  title5 'Subset to ECA, and JCA';

proc freq data = newdata_tab;
  where diagnosis in('ECA', 'JCA','BE','SSBE');
  tables is_multi*stage
         is_multi
         /missing list;
  title4 'Table 2: Cancer stage by familial categories';
  title5 'Subset to ECA, JCA, BE, and SSBE';

proc freq data = newdata_tab;
  tables diagnosis*is_multi
         /missing;
  format diagnosis $dx3fmt.;
  title4 'Table 4: Affected vs. unaffected';
  title5 'NOTE: There is a typo in the unaffected duplex number: it should be 3017 in the paper';

proc freq data = newdata_tab;
```

```
  where hadheartburn ^= '';
  tables hadheartburn*is_multi
         /missing;
  title4 'Table 4';
  title5 'Subset to non-missing heartburn results';

proc freq data = newdata_tab;
  where hadacidregurgitation ^= '';
  tables hadacidregurgitation*is_multi
         /missing;
  title4 'Table 4';
  title5 'Subset to non-missing hadacidregurgitation results';

proc freq data = newdata_tab;
  where smokecigarettes ^= '';
  tables smokecigarettes*is_multi
         /missing;
  title4 'Table 4';
  title5 'Subset to non-missing smokecigarettes results';

proc freq data = newdata_tab;
  where alcoholicdrinksperweek ^= '';
  tables alcoholicdrinksperweek*is_multi
         /missing;
  format alcoholicdrinksperweek $drinkfmt.;
  title4 'Table 4';
  title5 'Subset to non-missing alcoholicdrinksperweek results';
```