Norma Pugh
October 24, 2008

Integrity Check for the Inflammatory Bowel Disease Genetic Consortium (IBDGC) Phenotype Analysis File

As a partial check of the integrity of the IBDGC phenotype analysis dataset archived in the NIDDK data repository, a set of tabulations was performed to verify that published results can be reproduced using the archived dataset. Analyses were performed to duplicate results for the data published by Nguyen et al [1] in the *American Journal of Gastroenterology* in May 2006. The results of this integrity check are described below. The full text of the *American Journal of Gastroenterology* article can be found in Attachment 1, and the SAS code for our tabulations is included in Attachment 2. Attachment 3 includes confirmatory Stata code provided by the DCC, and Attachment 4 provides a complete description of the genotypic data. The genotypic data have not been replicated, as they are not included in the NIDDK Data Repository at RTI. They are housed in the NIDDK Genetics Repository at Rutgers.

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 on a first (or second) exercise 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.

Background. The NIDDK Inflammatory Bowel Disease Genetics Consortium (IBDGC) consists of investigators from seven sites in the U.S. and Canada who have recruited a large sample of inflammatory bowel disease patients, their relatives, and control subjects. All of the individuals in this sample have been evaluated according to a standardized protocol for clinical traits related to IBD, and have donated blood samples as a source of DNA. The IBDGC investigators are conducting genetic linkage and association studies to identify genes influencing predisposition to IBD [2].

The Nguyen paper examines racial differences in Inflammatory Bowel Disease (IBD) in family history, disease location, and extraintestinal manifestations (EIMs) that may reflect underlying genetic variations and have important implications for diagnosis and management of the disease [1].

The IBDGC analysis file provided to the repository was the most current version at the time of delivery. It is not the version that was used to produce the published results. In order to perform this replication, the repository was provided a supplemental list of the patient IDs analyzed in the publication.

The DCC indicated that one patient requested to be removed from the repository. In addition, two others either had no samples collected (or there was a problem in sample processing), and so are not part of the repository. This accounts for the total of 1,123 patients analyzed in this replication versus the total of 1,126 patients analyzed in the publication. Additionally, per the publication, the replication analysis was restricted to all unrelated index subjects (‘control’ variable = 0) with a confirmed diagnosis of IBD (‘diag’

variable = 1, 2 or 3) and self-identified race of African American, Hispanic, or Non-Hispanic White ('hispanic' and 'race' variables).

Finally, Attachment 2 provides the SAS code that was used by the data repository for the replication analysis. The DCC has reviewed and confirmed that the appropriate variables were used. Attachment 3 documents the DCC-provided Stata code used to clarify the definition of various variables and aid in the resolution of discrepancies. The Stata analysis was run using a copy of the data housed at the Data Repository.

Demographic and Baseline Characteristics. Table 1 in the publication [1] reports on demographic and baseline characteristics. Table A lists the variables we used in our replication.

Table A: Variables Used to Replicate Table 1

Table Variable	Variables Used in Replication
Race distribution	hispanic, race
Age at study entry	yob
Age at diagnosis	diag_yr, yob
Female	sex=2
Diagnosis: CD	diag=1
Diagnosis: UC	diag=3
Diagnosis: Indeterminate colitis	diag=2
Family history of IBD	chld_cd>0 or chld_uc>0 or chld_ibd>0 or fthr_ibd in(1,2,3) or mthr_ibd in(1,2,3) or sib_cd>0 or sib_uc>0 or sib_ibd>0 or fam_hist=1
Family history of IBD among CD patients	Family history (as defined above) and diag=1
Family history of IBD among UC patients	Family history (as defined above) and diag=3
% Siblings affected	sib_cd, sib_uc, sib_ibd, sib_unf
Packs/day at diagnosis	no_cigar/20, where smoking=1
Smoking at diagnosis: Current	smoking=1
Smoking at diagnosis: Former	smoking=2
Smoking at diagnosis: Never	smoking=3
Appendectomy (2 yr prior to IBD)	diag_yr, app_yr
Note: 'age at study entry' is not collected on the phenotype forms, and therefore is not included in the analysis dataset. Per advice from the DCC, an approximate age was calculated by subtracting 'year of birth' from 2005. Total patient enrollment took place from 2003-2007.	

In Table B, we compare the results calculated from the archived dataset to the results published in Table 1, Patient Demographics by Race/Ethnicity of the NIDDK-IBDGC Repository. As Table B shows, the results are similar.

Table B: Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values

Table	Group: All			Group: White			Group: AA			Group: Hispanic		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
Race	1,126	1,123	3	830	829	1	127	126	1	169	168	1
Age at study entry in yr, mean (SD)	36.1 (14.4)	37.1 (14.4)	1 (0)	36.7 (14.9)	37.7 (14.9)	1 (0)	36.1 (14.6)	37.1 (14.5)	1.0 (0.1)	33.3 (10.6)	34.2 (10.6)	0.9 (0)
Age at diagnosis in yr, mean (SD)	26.4 (12.8)	26.3 (12.8)	0.1 (0)	26.6 (13.3)	26.6 (13.3)	0	25.4 (13.2)	25.4 (13.2)	0	25.9 (9.8)	25.9 (9.8)	0
Female, no. (%)	584 (51.9)	584 (52.0)	0 (0.1)	427 (51.5)	428 (51.6)	1 (0)	80 (63.0)	80 (63.5)	0 (0.5)	77 (45.6)	76 (45.2)	1 (0.4)
Diagnosis, no. (%):												
CD	697 (61.9)	695 (61.9)	2 (0)	510 (61.5)	508 (61.3)	2 (0.2)	81 (63.8)	80 (63.5)	1 (0.3)	106 (62.7)	107 (63.7)	1 (1.0)
UC	396 (35.2)	397 (35.4)	1 (0.2)	299 (36.0)	302 (36.4)	3 (0.4)	35 (27.6)	35 (27.8)	0 (0.2)	62 (36.7)	60 (35.7)	2 (1.0)
IC	33 (2.9)	31 (2.8)	2 (0.1)	21 (2.5)	19 (2.3)	2 (0.2)	11 (8.7)	11 (8.7)	0	1 (0.6)	1 (0.6)	0
Fam Hx IBD	288 (25.6)	288 (25.6)	0	237 (28.6)	237 (28.6)	0	23 (18.1)	23 (18.3)	0 (0.2)	28 (16.6)	28 (16.7)	0 (0.1)
CD patients	191 (27.4)	191 (27.5)	0 (0.1)	156 (30.6)	156 (30.7)	0 (0.1)	16 (19.8)	16 (20.0)	0 (0.2)	19 (17.9)	19 (17.8)	0 (0.1)
UC patients	88 (22.2)	88 (22.2)	0	74 (24.8)	74 (24.5)	0 (0.3)	5 (14.3)	5 (14.3)	0	9 (14.5)	9 (15.0)	0 (0.5)
% Siblings	3.8	3.8	0	4.6	4.6	0	2.5	2.5	0	1.3	1.3	0
Packs/day at diagnosis	0.12	0.12	0	0.15	0.15	0	0.07	0.07	0	0.04	0.04	0

Table B: Comparison of Values Computed in Integrity Check to Reference Article Table 1 Values (cont.)

Table	Group: All			Group: White			Group: AA			Group: Hispanic		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
Smoking at diagnosis, no. (%):												
Current	192 (17.2)	193 (17.2)	1 (0)	161 (19.5)	162 (19.5)	1 (0)	20 (16.1)	20 (15.9)	0 (0.2)	11 (6.6)	11 (6.6)	0
Former	140 (12.6)	138 (12.3)	2 (0.3)	117 (14.2)	116 (14.0)	1 (0.2)	15 (12.1)	15 (11.9)	0 (0.2)	8 (4.8)	7 (4.2)	1 (0.6)
Never	783 (70.2)	781 (69.6)	2 (0.6)	547 (66.3)	546 (65.9)	1 (0.4)	89 (71.8)	88 (69.8)	1 (2.0)	147 (88.6)	147 (87.5)	0 (1.1)
Appendectomy (2 yr prior to IBD)	47 (4.4)	47 (4.2)	0 (0.2)	32 (4.0)	32 (3.9)	0 (0.1)	3 (2.9)	3 (2.4)	0 (0.5)	12 (7.1)	12 (7.1)	0

Notes:
(1) AA = African American, CD = Crohn's disease, UC = ulcerative colitis, IC = indeterminate colitis, IBD = inflammatory bowel disease

CD and UC Phenotype. Table 2 in the publication [1] reports on the unadjusted distribution of CD location and behavioral pattern, as well as UC disease extent. Table C lists the variables we used in our replication.

Table C: Variables Used to Replicate Table 2

Table Variable	Variables Used in Replication
Race distribution	hispanic, race
Crohn's disease: Disease Involvement: Upper gastrointestinal	diag=1 and (jejunal=1 or gi=1)
Crohn's disease: Disease Involvement: Esophagogastroduodenal	diag=1 and gi=1
Crohn's disease: Disease Involvement: Jejunum	diag=1 and jejunal=1
Crohn's disease: Disease Involvement: Ileum	diag=1 and ileal=1
Crohn's disease: Disease Involvement: Colorectal	diag=1 and colorect=1
Crohn's disease: Disease Involvement: Perianal disease	diag=1 and perianal=1
Best-fit disease site: Ileum	*Note:*
Best-fit disease site: Ileo-colon	*'Best-fit' items have not been replicated. They are*
Best-fit disease site: Colon	*generated by other variables and differ slightly from the*
Best-fit disease site: Upper gastrointestinal	*Montreal classification.*
Disease pattern: Inflammatory	diag=1 and behavior=1
Disease pattern: Stricturing	diag=1 and behavior=2
Disease pattern: Penetrating	diag=1 and behavior=3
UC Disease extent: Proctitis	diag=3 and (proctit=1 & left=2 & extensiv=2)
UC Disease extent: Left-sided colitis	else diag=3 and (left=1 & extensiv=2)
UC Disease extent: Extensive colitis	else diag=3 and extensiv=1

In Table D, we compare the results calculated from the archived dataset to the results published in Table 2, CD and UC Phenotype by Race/Ethnicity of the NIDDK-IBDGC Repository. As Table D shows, the results are similar.

Table D: Comparison of Values Computed in Integrity Check to Reference Article Table 2 Values

Table	**Group: All**			**Group: White**			**Group: AA**			**Group: Hispanic**		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
Crohn's disease												
Disease involvement												
Upper gastrointestinal	96 (15.8)	96 (15.8)	0	72 (16.1)	72 (16.1)	0	17 (25.8)	17 (26.2)	0 (0.4)	7 (7.4)	7 (7.3)	0 (0.1)
Esophagogastroduodenal	54 (8.8)	55 (9.0)	1 (0.2)	38 (8.5)	39 (8.7)	1 (0.2)	14 (20.0)	14 (20.3)	0 (0.3)	2 (2.1)	2 (2.1)	0
Jejunum	57 (9.1)	56 (8.9)	1 (0.2)	44 (9.4)	43 (9.1)	1 (0.3)	8 (12.9)	8 (13.1)	0 (0.2)	5 (5.2)	5 (5.1)	0 (0.1)
Ileum	530 (79.0)	530 (79.2)	0 (0.2)	396 (80.0)	395 (80.1)	1 (0.1)	50 (67.6)	50 (68.5)	0 (0.9)	84 (82.4)	85 (82.5)	1 (0.1)
Colorectal	459 (68.3)	458 (68.2)	1 (0.1)	326 (65.6)	325 (65.4)	1 (0.2)	63 (77.8)	62 (77.5)	1 (0.3)	70 (74.5)	71 (74.7)	1 (0.2)
Perianal disease	231 (33.5)	227 (33.0)	4 (0.5)	146 (28.7)	142 (28.1)	4 (0.6)	32 (40.0)	32 (40.5)	0 (0.5)	53 (52.5)	53 (52.0)	0 (0.5)
Disease pattern												
Inflammatory	332 (48.2)	332 (48.1)	0 (0.1)	242 (48.0)	242 (47.9)	0 (0.1)	42 (53.2)	41 (52.6)	1 (0.6)	48 (45.3)	49 (45.8)	1 (0.5)
Stricturing	155 (22.5)	154 (22.3)	1 (0.2)	108 (21.4)	108 (21.4)	0	22 (27.9)	21 (26.9)	1 (1)	25 (23.6)	25 (23.4)	0 (0.2)
Penetrating	202 (29.3)	204 (29.6)	2 (0.3)	154 (30.6)	155 (30.7)	1 (0.1)	15 (19.0)	16 (20.5)	1 (1.5)	33 (31.1)	33 (30.8)	0 (0.3)

Table D: Comparison of Values Computed in Integrity Check to Reference Article Table 2 Values (cont)

Table	Group: All			Group: White			Group: AA			Group: Hispanic		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
Ulcerative Colitis												
Disease extent												
Proctitis	23 (6.2)	24(6.4)	1 (0.2)	17 (5.9)	18 (6.2)	1 (0.3)	4 (13.8)	4 (13.8)	0	2 (3.6)	2 (3.8)	0 (0.2)
Left-sided colitis	109 (29.4)	109 (29.2)	0 (0.2)	90 (31.4)	90 (30.9)	0 (0.5)	10 (34.5)	10 (34.5)	0	9 (16.4)	9 (17.0)	0 (0.6)
Extensive colitis	239 (64.4)	240 (64.3)	1 (0.1)	180 (62.7)	183 (62.9)	3 (0.2)	15 (51.7)	15 (51.7)	0	44 (80.0)	42 (79.2)	2 (0.8)
Notes: (1) AA = African American												

Surgical History. Table 3 in the publication [1] reports on the surgical history of IBD patients. Table E lists the variables we used in our replication of these variables.

Table E: Variables Used to Replicate Table 3

Table Variable	Variables Used in Replication
Race distribution	hispanic, race
CD: # surgeries: abdominal	op_ad
CD: # surgeries: perianal	op_pd
CD: # surgeries: bowel resection	surg_br=1
CD: # surgeries: bowel diversion	surg_div=1
CD: # surgeries: penetrating disease	surg_af=1
CD: # surgeries: perianal disease	surg_pf=1
Time to first surgery	surgery, diag_yr, review, surg_yr
% surgery free: at 2 years	surgery, diag_yr, review, surg_yr
% surgery free: at 5 years	surgery, diag_yr, review, surg_yr
% surgery free: at 10 years	surgery, diag_yr, review, surg_yr
UC: Colectomy	surgery=1 and diag=3
UC: Dysplasia	surg_dys=1
UC: Chronic refractory disease	surg_chr=1
UC: Fulminant colitis	surg_acu=1 and diag=3

In Table F, we compare the results calculated from the archived dataset to the results published in Table 3, Surgical History of IBD Patients in the NIDDK-IBDGC Repository by Race/Ethnicity. As Table F shows, the results are similar. The DCC provided confirmatory Stata code (Attachment 3) in order to clarify the definition of various variables and aid in the resolution of discrepancies. Specifically, 'time to first surgery' and the '% surgery free at *xx* years' variables have been confirmed by the Stata code.

Additionally, the DCC has noted that published results include approximately 50 observations in which the 'number of surgery' variables (abdominal and perianal) should not have been answered. These observations represent cases where the overall surgery variable is not recorded as 'yes'. Therefore, it is unclear how the 'number of surgery' responses should be interpreted and these observations have been excluded from the archived dataset.

Finally, variables labeled as "# surgeries" in the publication are not always "# surgeries". Yes/No surgery variables are actually summarized for bowel resection, bowel diversion, penetrating disease and perianal disease.

Table F: Comparison of Values Computed in Integrity Check to Reference Article Table 3 Values

Table	Group: All			Group: White			Group: AA			Group: Hispanic		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
CD												
No. surgeries, mean (SD)												
For abdominal CD	1.4 (1.3)	1.5 (1.2)	0.1 (0.1)	1.6 (1.3)	1.6 (1.3)	0	0.8 (1.1)	1.4 (1.2)	0.6 (0.1)	1.1 (1.0)	1.3 (1.0)	0.2 (0)
For perianal CD	0.5 (1.2)	0.6 (1.2)	0.1 (0)	0.5 (1.1)	0.5 (1.1)	0	0.3 (0.7)	0.6 (0.9)	0.3 (0.2)	0.7 (1.7)	0.8 (1.9)	0.1 (0.2)
Bowel resection	335 (71.0)	334 (71.8)	1 (0.8)	263 (78.7)	262 (76.2)	1 (2.5)	27 (35.1)	27 (58.7)	0 (23.6)	45 (73.8)	45 (60.0)	0 (13.8)
Bowel diversion	46 (10.0)	46 (10.1)	0 (0.1)	24 (7.4)	24 (7.1)	0 (0.3)	9 (12.0)	9 (20.0)	0 (8.0)	13 (22.4)	13 (17.6)	0 (4.8)
Surgery for penetrating disease	98 (21.5)	100 (22.1)	2 (0.6)	74 (22.9)	76 (22.7)	2 (0.2)	10 (13.3)	10 (22.2)	0 (8.9)	14 (24.1)	14 (19.2)	0 (4.9)
Surgery for perianal disease	107 (23.0)	104 (22.5)	3 (0.5)	77 (23.4)	76 (22.3)	1 (1.1)	15 (20.0)	15 (33.3)	0 (13.3)	15 (24.6)	13 (17.1)	2 (7.5)
Median time to 1st surgery (yr)	9.5	See Attachment 3		9.8	See Attachment 3		11.4	See Attachment 3		6.6	See Attachment 3	
% Surgery-free												
At 2 yr	86.1%			84.6%			94.8%			86.7%		
At 5 yr	67.7%			67.7%			76.3%			59.8%		
At 10 yr	48.0%			49.6%			58.8%			26.1%		

Table F: Comparison of Values Computed in Integrity Check to Reference Article Table 3 Values (cont.)

Table	Group: All			Group: White			Group: AA			Group: Hispanic		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
Ulcerative colitis												
Colectomy	71 (18.0)	70 (17.6)	1 (0.4)	47 (15.8)	47 (15.6)	0 (0.2)	4 (11.4)	4 (11.4)	0	20 (32.3)	19 (31.7)	1 (0.6)
Dysplasia	4 (3.0)	4 (5.2)	0 (2.2)	3 (3.5)	3 (5.9)	0 (2.4)	1 (3.7)	1 (16.7)	0 (13.0)	0 (0)	0 (0)	0
Chronic refractory disease	54 (40.6)	53 (69.7)	1 (29.1)	35 (40.7)	35 (68.6)	0 (27.9)	1 (3.9)	1 (20.0)	2 (16.1)	18 (85.7)	17 (85.0)	1 (0.7)
Fulminant colitis	8 (5.9)	8 (10.5)	0 (4.6)	6 (6.9)	6 (12.0)	0 (5.1)	0 (0)	0 (0)	0	2 (9.5)	2 (10.0)	0 (0.5)
Notes: (1) AA = African American, CD = Crohn's disease												

Extraintestinal Manifestations. Table 4 in the publication [1] reports on the distribution of EIMs in all patients. Table G lists the variables we used in our replication of these variables.

Table G: Variables Used to Replicate Table 4

Table Variable	Variables Used in Replication
Race distribution	hispanic, race
Sacroiliitis	j_si=1
Ankylosing spondylitis	j_as=1
Uveitis	eye_uv=1
Erythma nodosum	skin_en=1
Pyoderma gangrenosum	skin_py=1
Primary sclerosing cholangitis	liv_psc=1

In Table H, we compare the results calculated from the archived dataset to the results published in Table 4, Extraintestinal Manifestations in the NIDDK-IBDGC Repository by Race/Ethnicity. As Table H shows, the results are similar.

Table H: Comparison of Values Computed in Integrity Check to Reference Article Table 4 Values

Table	Group: All			Group: White			Group: AA			Group: Hispanic		
Variable	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**	**Nguyen et al (2006)**	**Integrity Check**	**Diff**
Sacroiliitis	23 (2.1)	23 (2.1)	0	14 (1.7)	16 (2.0)	2 (0.3)	7 (5.8)	6 (5.3)	1 (0.5)	2 (1.2)	1 (0.6)	1 (0.6)
Ankylosing spondylitis	14 (1.3)	12 (1.1)	2 (0.2)	12 (1.5)	11 (1.4)	1 (0.1)	2 (1.7)	1 (0.9)	1 (0.8)	0 (0)	0 (0)	0
Uveitis	25 (2.2)	24 (2.2)	1 (0)	13 (1.6)	13 (1.6)	0	10 (8.1)	9 (8.0)	1 (0.1)	2 (1.2)	2 (1.2)	0
Erythema nodosum	48 (4.3)	46 (4.3)	2 (0)	28 (3.4)	29 (3.6)	1 (0.2)	2 (1.7)	2 (1.8)	0 (0.1)	18 (10.7)	15 (9.3)	3 (1.4)
Pyoderma gangrenosum	23 (2.1)	22 (2.0)	1 (0.1)	16 (1.9)	16 (2.0)	0 (0.1)	3 (2.5)	3 (2.7)	0 (0.2)	4 (2.4)	3 (1.9)	1 (0.5)
Primary sclerosing cholangitis	20 (1.8)	20 (1.9)	0 (0.1)	11 (1.3)	11 (1.4)	0 (0.1)	4 (3.2)	4 (3.5)	0 (0.3)	5 (3.0)	5 (3.1)	0 (0.1)
Note: AA = African American												

Notes

1. The few discrepancies documented in this report are likely due to database corrections made after the data freeze used for the publication. The analysis dataset housed at the repository is the most current, updated dataset and contains more than twice the number of patients as was analyzed for the publication. The DCC has confirmed all variables used in our replication analysis and, further, has used a copy of the data sent to the repository to conduct their own replication of several variables (see Attachment 3).

2. In addition to the analysis dataset examined in this replication analysis (PHENOTYP), there exist genotypic data from the IBDGC cohort. Attachment 4 provides a complete description of the genotypic data, which is housed in the NIDDK Genetics Repository at Rutgers. These data have not been replicated, as they are not included in the NIDDK Data Repository at RTI.

3. The SAS datasets provided to the NIDDK Data Repository are provided as transport files. In order to use SAS Viewer, limit CPU resources and increase performance when using these datasets, they must be converted back to an un-archived state. One method to do this is via PROC COPY, as follows:

```
/* Location of IBDGC SAS Data Files */
libname indata xport 'R:\IBDGC\Phenotype data\phenotypes.xpt';
libname infmts xport 'R:\IBDGC\Phenotype data\formats.xpf';

/* Location for Un-archived IBDGC SAS Data Files */
libname sasdb 'R:\IBDGC\Phenotype data\Unarchived';

/* Create SAS-Readable File & Formats */
proc copy in=indata out=sasdb; run;
proc copy in=infmts out=sasdb; run;

/* Output Formats */
data fmts; set sasdb.formats;
proc format cntlin=fmts; run;
```

References

1. Geoffrey C. Nguyen, M.D. et al, **Inflammatory Bowel Disease Characteristics Among African Americans, Hispanics, and Non-Hispanic Whites: Characterization of a Large North American Cohort**, American Journal of Gastroenterology 2006; 101:1012-1023.

2. NIDDK Website: IBDGC page. Inflammatory Bowel Disease Genetic Consortium (IBDGC) : NIDDK.

ATTACHMENT 1

"The full text of the article referenced will be provided to approved requestors along with the data archive."

Geoffrey C. Nguyen, M.D. et al, **Inflammatory Bowel Disease Characteristics Among African Americans, Hispanics, and Non-Hispanic Whites: Characterization of a Large North American Cohort**, American Journal of Gastroenterology 2006; 101:1012-1023.

ATTACHMENT 2

SAS Code for Tabulations from the IBDGC Phenotype Analysis Dataset in the NIDDK Repository

```
/*************************************************************************************/
/*
/* Program: R:\05_Users\Norma\IBDGC\PhenotypeData\NguyenPaper\Update\table1_updated.sas
/* Author:  Norma Pugh
/* Date:    28 July 2008
/* Revised: 03 October 2008 per DCC comments.
/* Revised: 24 October 2008 per DCC comments re: calculation of '% affected siblings'.
/* Purpose: Replicate table 1 results.
/*
/*************************************************************************************/
/* DATA SOURCE */
libname sasdb 'R:\05_Users\Norma\IBDGC\PhenotypeData';

/*******************/
/* INCLUDE FORMATS */
/*******************/
data fmts; set sasdb.formats;
proc format cntlin=fmts; run;
proc format; value yn 1='1=Y' 2='2=N'; run;

/************************/
/* GET STUDY POPULATION */
/************************/
data table1; merge sasdb.phenotyp(in=x1) sasdb.patients(in=x2); by suid; if x1 & x2;
 /* Define age at study entry */
 /* NOTE: Enrollments were from 2003-2005. Use enrollment year of 2005, per P.Schumm e-
mail. */
 ageentry=2005-yob; label ageentry='Age at study entry (yr)';

 /* Define age at diagnosis */
 agediag=diag_yr-yob; label agediag='Age at diagnosis (yr)';

 /* Define racial group */
 if race=1 & hispanic=2 then racegrp='2_White';
 if race=2 then racegrp='3_AA';
 if hispanic=1 and race^=2 then racegrp='4_Hispanic';
 label racegrp='Racial group';

 /* Define family history variables */
 label famhx='1st/2nd degree relative w/ IBD(CD or UC)'
       numsibaff='# Sibling(s) affected'
       numsib='# Sibling(s)';

 if chld_cd>0 or chld_uc>0 or chld_ibd>0 or fthr_ibd in(1,2,3) or mthr_ibd in(1,2,3) or
sib_cd>0 or sib_uc>0 or sib_ibd>0 or fam_hist=1 then famhx=1;
 else famhx=2;

 numsibaff=sib_cd+sib_uc+sib_ibd;
 numsib=sib_cd+sib_uc+sib_ibd+sib_unf;

 /* Define packs/day */
 if smoking=1 then numpacks=no_cigar/20; else numpacks=0;

 /* Define appendectomy 2 yrs prior to IBD */
 app2yr=2;

 if diag_yr-app_yr>=2 then app2yr=1;
```

```
 /* Output patients w/ IBD & appropriate racial group */
 if control=0 & diag in(1,2,3) & racegrp>'';
run;

/************************/
/* CHECK FOR DUPLICATES */
/************************/
data check_dups; set table1; keep suid; run;
proc sort data=check_dups nodup; by suid; run;

/***********************************************/
/* CREATE 'ALL' GROUPING FOR RACE & DENOMINATORS */
/***********************************************/
data table1; set table1; output; racegrp='1_All'; output; run;

proc freq data=table1 noprint; tables racegrp / out=denom(drop=percent
rename=(count=denom)); run;
proc print data=denom; title'Denominators: Racial group'; run;

/******************/
/* GET STATISTICS */
/******************/
proc sort data=table1; by racegrp; run;
proc means data=table1 n mean stddev; by racegrp; var ageentry agediag; title'Means/SDs:
Ages'; run;

proc freq data=table1 noprint; tables racegrp*sex / out=outsex(drop=percent
rename=(count=numer)); format sex sex.; run;
data outsex; merge outsex denom; by racegrp; pct=(numer/denom)*100; run;
proc print data=outsex(where=(sex=2)); title'Frequency counts: Female'; run;

proc freq data=table1 noprint; tables racegrp*diag / out=outdiag(drop=percent
rename=(count=numer)); format diag diag.; run;
data outdiag; merge outdiag denom; by racegrp; pct=(numer/denom)*100; run;
proc print data=outdiag; title'Frequency counts: Diagnosis'; run;

proc freq data=table1 noprint; tables racegrp*famhx / out=outfamhx(drop=percent
rename=(count=numer)); format famhx yn.; run;
data outfamhx; merge outfamhx denom; by racegrp; pct=(numer/denom)*100; run;
proc print data=outfamhx; title'Frequency counts: Family History'; run;

proc freq data=table1(where=(diag=1)); tables racegrp*famhx; format famhx yn.;
 title'Frequency counts: Family History for Patient Diagnosis of CD';
run;

proc freq data=table1(where=(diag=3)); tables racegrp*famhx; format famhx yn.;
 title'Frequency counts: Family History for Patient Diagnosis of UC';
run;

proc means data=table1 n sum; by racegrp; var numsibaff numsib;
 title'Sums: # Sibs Affected, # Sibs Total';
 title'Calculate sample proportion as # affected/# total';
run;

proc means data=table1 n mean; by racegrp; var numpacks; title'Means: Packs of cigarettes
per day AMONG SMOKERS & NON-SMOKERS'; run;
```

```
proc means data=table1(where=(smoking=1 & racegrp='1_All')) n mean; by racegrp; var
numpacks; title'Means: Packs of cigarettes per day AMONG SMOKERS'; run;

proc freq data=table1; tables racegrp*(smoking app2yr); format smoking smoking. app2yr
yn.;
 title'Frequency counts: Smoking at diagnosis, Appendectomy (2 yr prior to IBD)';
run;
```

```
/*******************************************************************************************/
/*
/* Program: R:\05_Users\Norma\IBDGC\PhenotypeData\NguyenPaper\Update\table2.sas
/* Author:  Norma Pugh
/* Date:    29 July 2008
/* Revised: 03 October 2008 per DCC comments.
/* Purpose: Replicate table 2 results.
/*
/*******************************************************************************************/
/* DATA SOURCE */
libname sasdb 'R:\05_Users\Norma\IBDGC\PhenotypeData';

/******************/
/* INCLUDE FORMATS */
/******************/
data fmts; set sasdb.formats;
proc format cntlin=fmts; run;
proc format; value yn 1='1=Y' 2='2=N'; value uc 1='1=Proctitis' 2='2=Left'
3='3=Extensive'; run;

/************************/
/* GET STUDY POPULATION */
/************************/
data table2; merge sasdb.phenotyp(in=x1) sasdb.patients(in=x2); by suid; if x1 & x2;

 /* Define racial group */
 if race=1 & hispanic=2 then racegrp='2_White';
 if race=2 then racegrp='3_AA';
 if hispanic=1 and race^=2 then racegrp='4_Hispanic';
 label racegrp='Racial group';

 /* Define Crohn's dx: Dx involvement: Upper gastro  */
 if jejunal=1 or gi=1 then crohndx_gastro=1; else crohndx_gastro=2;
 label crohndx_gastro='CD: Dx involvement: Upper GI';

 /* Define disease pattern */
 inflam=2; strict=2; penetrate=2;

 if behavior=1 then inflam=1;
 if behavior=2 then strict=1;
 if behavior=3 then penetrate=1;

 label inflam='Dx pattern: inflammatory'
       strict='Dx pattern: stricturing'
       penetrate='Dx pattern: penetrating';

 /* Define UC Disease Extent: Proctitis, Left-sided Colitis or Extensive Colitis */
 if (proctit=1 & left=2 & extensiv=2) then uc_dx_ext=1;
 else if (left=1 & extensiv=2) then uc_dx_ext=2;
 else if (extensiv=1) then uc_dx_ext=3;

 label uc_dx_ext='UC Dx Xtent:1=Proctitis,2=Left,3=Extensv';

 /* Output patients w/ IBD & appropriate racial group */
 if control=0 & diag in(1,2,3) & racegrp>'';
run;
```

```
/************************/
/* CHECK FOR DUPLICATES */
/************************/
data check_dups; set table2; keep suid; run;
proc sort data=check_dups nodup; by suid; run;

/*****************************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE INVOLVEMENT UPPER GI DENOMINATORS */
/*****************************************************************************/
data table2_cd_upper; set table2;
 if diag=1;
 if (jejunal in(1,2) & gi in(1,2)) or (jejunal=1 & gi=3) or (gi=1 & jejunal=3); output;
racegrp='1_All'; output;
run;

proc freq data=table2_cd_upper noprint; tables racegrp / out=denom_cd_upper(drop=percent
rename=(count=denom)); run;
proc print data=denom_cd_upper; title'CD Dx Involvement Upper GI Denominators: Racial
group'; run;

/*****************************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE INVOLVEMENT ESOPH. DENOMINATORS */
/*****************************************************************************/
data table2_cd_esoph; set table2;
 if diag=1 & gi in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table2_cd_esoph noprint; tables racegrp / out=denom_cd_esoph(drop=percent
rename=(count=denom)); run;
proc print data=denom_cd_esoph; title'CD Dx Involvement Esoph. Denominators: Racial
group'; run;

/*****************************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE INVOLVEMENT JEJUNUM DENOMINATORS */
/*****************************************************************************/
data table2_cd_jej; set table2;
 if diag=1 & jejunal in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table2_cd_jej noprint; tables racegrp / out=denom_cd_jej(drop=percent
rename=(count=denom)); run;
proc print data=denom_cd_jej; title'CD Dx Involvement Jejunal Denominators: Racial
group'; run;


/*****************************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE INVOLVEMENT ILEUM DENOMINATORS */
/*****************************************************************************/
data table2_cd_il; set table2;
 if diag=1 & ileal in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table2_cd_il noprint; tables racegrp / out=denom_cd_il(drop=percent
rename=(count=denom)); run;
proc print data=denom_cd_il; title'CD Dx Involvement Ileum Denominators: Racial group';
run;
```

```
/*****************************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE INVOLVEMENT COLORECTAL DENOMINATORS */
/*****************************************************************************/
data table2_cd_colo; set table2;
 if diag=1 & colorect in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table2_cd_colo noprint; tables racegrp / out=denom_cd_colo(drop=percent
rename=(count=denom)); run;
proc print data=denom_cd_colo; title'CD Dx Involvement Colorectal Denominators: Racial
group'; run;

/*****************************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE INVOLVEMENT PERIANAL DX DENOMINATORS */
/*****************************************************************************/
data table2_cd_peri; set table2;
 if diag=1 & perianal in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table2_cd_peri noprint; tables racegrp / out=denom_cd_peri(drop=percent
rename=(count=denom)); run;
proc print data=denom_cd_peri; title'CD Dx Involvement Perianal Dx Denominators: Racial
group'; run;

/*****************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CD DISEASE PATTERN DENOMINATORS */
/*****************************************************************/
data table2_cd_dxpatt; set table2;
 if diag=1 & behavior in(1,2,3); output; racegrp='1_All'; output;
run;

proc freq data=table2_cd_dxpatt noprint; tables racegrp /
out=denom_cd_dxpatt(drop=percent rename=(count=denom)); run;
proc print data=denom_cd_dxpatt; title'CD Dx Pattern Denominators: Racial group'; run;

/*************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & UC DENOMINATORS */
/*************************************************/
data table2_uc; set table2;
 if diag=3 & proctit in(1,2) & left in(1,2) & extensiv in(1,2); output; racegrp='1_All';
output;
run;

proc freq data=table2_uc noprint; tables racegrp / out=denom_uc(drop=percent
rename=(count=denom)); run;
proc print data=denom_uc; title'UC Denominators: Racial group'; run;

/******************/
/* GET STATISTICS */
/******************/
%macro table2(indata,var,denom,title);
 proc freq data=&indata noprint; tables racegrp*&var / out=outdx(drop=percent
rename=(count=numer)); format &var yn.; run;
 data outdx; merge outdx &denom; by racegrp; pct=(numer/denom)*100; run;
 proc print data=outdx(where=(&var=1)); title"Frequency counts: &title"; run;
%mend table2;
```

```
%table2(table2_cd_upper,crohndx_gastro,denom_cd_upper,%str(CD Disease Involvement - Upper
gastrointestinal));
%table2(table2_cd_esoph,gi,denom_cd_esoph,%str(CD Disease Involvement -
Esophagogastroduodenal));
%table2(table2_cd_jej,jejunal,denom_cd_jej,%str(CD Disease Involvement - Jejunum));
%table2(table2_cd_il,ileal,denom_cd_il,%str(CD Disease Involvement - Ileum));
%table2(table2_cd_colo,colorect,denom_cd_colo,%str(CD Disease Involvement - Colorectal));
%table2(table2_cd_peri,perianal,denom_cd_peri,%str(CD Disease Involvement - Perianal
disease));

%table2(table2_cd_dxpatt,inflam,denom_cd_dxpatt,%str(CD Disease Pattern - Inflammatory));
%table2(table2_cd_dxpatt,strict,denom_cd_dxpatt,%str(CD Disease Pattern - Stricturing));
%table2(table2_cd_dxpatt,penetrate,denom_cd_dxpatt,%str(CD Disease Pattern -
Penetrating));

proc freq data=table2_uc noprint; tables racegrp*uc_dx_ext / out=outdx(drop=percent
rename=(count=numer)); format uc_dx_ext uc.; run;
data outdx; merge outdx denom_uc; by racegrp; pct=(numer/denom)*100; run;
proc print data=outdx; title"Frequency counts: UC Disease Extent"; run;
```

```
/*************************************************************************************/
/*
/* Program: R:\05_Users\Norma\IBDGC\PhenotypeData\NguyenPaper\Update\table3.sas
/* Author:  Norma Pugh
/* Date:    29 July 2008
/* Revised: 03 October 2008 per DCC comments.
/* Purpose: Replicate table 3 results.
/*
/*************************************************************************************/
/* DATA SOURCE */
libname sasdb 'R:\05_Users\Norma\IBDGC\PhenotypeData';

/******************/
/* INCLUDE FORMATS */
/******************/
data fmts; set sasdb.formats;
proc format cntlin=fmts; run;
proc format; value yn 1='1=Y' 2='2=N'; run;

/************************/
/* GET STUDY POPULATION */
/************************/
data table3; merge sasdb.phenotyp(in=x1) sasdb.patients(in=x2); by suid; if x1 & x2;
 /* Define racial group */
 if race=1 & hispanic=2 then racegrp='2_White';
 if race=2 then racegrp='3_AA';
 if hispanic=1 and race^=2 then racegrp='4_Hispanic';
 label racegrp='Racial group';

 /* Define surgery variables */
 uc_colectomy=2;

 if surgery=1 then surgtime=surg_yr-diag_yr;
  else if surgery=2 then surgtime=review-diag_yr;

 if surgery=1 & diag=3 then uc_colectomy=1;

 label surgtime='Time to 1st surgery (yrs)'
       uc_colectomy='UC: Colectomy surgery';

 /* Output patients w/ IBD & appropriate racial group */
 if control=0 & diag in(1,2,3) & racegrp>'';
run;

/************************/
/* CHECK FOR DUPLICATES */
/************************/
data check_dups; set table3; keep suid; run;
proc sort data=check_dups nodup; by suid; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & ABDOMINAL CD SURGERY DENOMINATORS */
/*********************************************************************/
data table3_abcd; set table3;
 if surgery=1 & diag=1 & op_ad>.; output; racegrp='1_All'; output;
run;
```

```
proc freq data=table3_abcd noprint; tables racegrp / out=denom_abcd(drop=percent
rename=(count=denom)); run;
proc print data=denom_abcd; title'Abdominal CD Surgery: Racial group'; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & PERIANAL CD SURGERY DENOMINATORS */
/*********************************************************************/
data table3_percd; set table3;
 if surgery=1 & diag=1 & op_pd>.; output; racegrp='1_All'; output;
run;

proc freq data=table3_percd noprint; tables racegrp / out=denom_percd(drop=percent
rename=(count=denom)); run;
proc print data=denom_percd; title'Perianal CD Surgery: Racial group'; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & BOWEL RESECTION SURGERY DENOMINATORS */
/*********************************************************************/
data table3_br; set table3;
 if surgery=1 & surg_br in(.,1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_br noprint; tables racegrp / out=denom_br(drop=percent
rename=(count=denom)); run;
proc print data=denom_br; title'Bowel Resection Surgery: Racial group'; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & BOWEL DIVERSION SURGERY DENOMINATORS */
/*********************************************************************/
data table3_div; set table3;
 if surgery=1 & surg_div in(.,1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_div noprint; tables racegrp / out=denom_div(drop=percent
rename=(count=denom)); run;
proc print data=denom_div; title'Bowel Diversion Surgery: Racial group'; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & PENETRATING DX SURGERY DENOMINATORS */
/*********************************************************************/
data table3_af; set table3;
 if surgery=1 & surg_af in(.,1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_af noprint; tables racegrp / out=denom_af(drop=percent
rename=(count=denom)); run;
proc print data=denom_af; title'Penetrating Dx Surgery: Racial group'; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & PERIANAL DX SURGERY DENOMINATORS */
/*********************************************************************/
data table3_pf; set table3;
 if surgery=1 & surg_pf in(.,1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_pf noprint; tables racegrp / out=denom_pf(drop=percent
rename=(count=denom)); run;
```

```
proc print data=denom_pf; title'Perianal Dx Surgery: Racial group'; run;

/*********************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & SURGERY DENOMINATORS */
/*********************************************************/
data table3_surg; set table3;
 if surgery=1; output; racegrp='1_All'; output;
run;

proc freq data=table3_surg noprint; tables racegrp / out=denom_surg(drop=percent
rename=(count=denom)); run;
proc print data=denom_surg; title'Surgery Denominators: Racial group'; run;

/*********************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & OVERALL DENOMINATORS */
/*********************************************************/
data table3_all; set table3; output; racegrp='1_All'; output; run;

proc freq data=table3_all noprint; tables racegrp / out=denom_all(drop=percent
rename=(count=denom)); run;
proc print data=denom_all; title'Overall Denominators: Racial group'; run;

/*********************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & COLECTOMY DENOMINATORS */
/*********************************************************/
data table3_col; set table3;
 if diag=3; output; racegrp='1_All'; output;
run;

proc freq data=table3_col noprint; tables racegrp / out=denom_col(drop=percent
rename=(count=denom)); run;
proc print data=denom_col; title'Colectomy Denominators: Racial group'; run;

/*********************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & DYSPLASIA DENOMINATORS */
/*********************************************************/
data table3_dys; set table3;
 if surgery=1 & surg_dys in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_dys noprint; tables racegrp / out=denom_dys(drop=percent
rename=(count=denom)); run;
proc print data=denom_dys; title'Dysplasia Denominators: Racial group'; run;

/*********************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & CHRONIC REFRACTORY DX DENOMINATORS */
/*********************************************************************/
data table3_chr; set table3;
 if surgery=1 & surg_chr in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_chr noprint; tables racegrp / out=denom_chr(drop=percent
rename=(count=denom)); run;
proc print data=denom_chr; title'Chronic Refractory Dx Denominators: Racial group'; run;
```

```
/*****************************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & FULMINANT COLITIS DENOMINATORS */
/*****************************************************************/
data table3_acu; set table3;
 if surgery=1 & surg_acu in(1,2); output; racegrp='1_All'; output;
run;

proc freq data=table3_acu noprint; tables racegrp / out=denom_acu(drop=percent
rename=(count=denom)); run;
proc print data=denom_acu; title'Fulminant colitis Denominators: Racial group'; run;


/******************/
/* GET STATISTICS */
/******************/
proc sort data=table3_abcd; by racegrp; run;
proc sort data=table3_percd; by racegrp; run;
proc sort data=table3_surg; by racegrp; run;
proc sort data=table3_all; by racegrp; run;

proc means data=table3_abcd n mean stddev; by racegrp; var op_ad;
 title'Means/SDs: # Surgeries: abdominal';
run;

proc means data=table3_percd n mean stddev; by racegrp; var op_pd;
 title'Means/SDs: # Surgeries: perianal';
run;

%macro table3(indata,var,denom,title);
 proc freq data=&indata noprint; tables racegrp*&var / out=outsurg(drop=percent
rename=(count=numer)); format &var yn.; run;
 data outsurg; merge outsurg &denom; by racegrp; pct=(numer/denom)*100; run;
 proc print data=outsurg(where=(&var=1)); title"Frequency counts: &title"; run;
%mend table3;

%table3(table3_br,surg_br,denom_br,%str(Bowel resection surgery));
%table3(table3_div,surg_div,denom_div,%str(Bowel diversion surgery));
%table3(table3_af,surg_af,denom_af,%str(Penetrating disease surgery));
%table3(table3_pf,surg_pf,denom_pf,%str(Perianal disease surgery));

%table3(table3_col(where=(diag=3)),uc_colectomy,denom_col,%str(UC Surgery: Colectomy));
%table3(table3_dys(where=(diag=3)),surg_dys,denom_dys,%str(UC Surgery: Dysplasia));
%table3(table3_chr(where=(diag=3)),surg_chr,denom_chr,%str(UC Surgery: Chronic refractory
disease));
%table3(table3_acu(where=(diag=3)),surg_acu,denom_acu,%str(UC Surgery: Fulminant
colitis));
```

```
/*************************************************************************************/
/*
/* Program: R:\05_Users\Norma\IBDGC\PhenotypeData\NguyenPaper\Update\table4.sas
/* Author:  Norma Pugh
/* Date:    29 July 2008
/* Revised: 03 October 2008 per DCC comments.
/* Purpose: Replicate table 4 results.
/*
/*************************************************************************************/
/* DATA SOURCE */
libname sasdb 'R:\05_Users\Norma\IBDGC\PhenotypeData';

/******************/
/* INCLUDE FORMATS */
/******************/
data fmts; set sasdb.formats;
proc format cntlin=fmts; run;

/************************/
/* GET STUDY POPULATION */
/************************/
data table4; merge sasdb.phenotyp(in=x1) sasdb.patients(in=x2); by suid; if x1 & x2;
 /* Define racial group */
 if race=1 & hispanic=2 then racegrp='2_White';
 if race=2 then racegrp='3_AA';
 if hispanic=1 and race^=2 then racegrp='4_Hispanic';
 label racegrp='Racial group';

 /* Output patients w/ IBD & appropriate racial group */
 if control=0 & diag in(1,2,3) & racegrp>'';

 /* Output patients w/ complete data, per table note */
 if j_si in(1,2) & j_as in(1,2) & eye_uv in(1,2) & skin_en in(1,2) & skin_py in(1,2) &
liv_psc in(1,2);
run;

/************************/
/* CHECK FOR DUPLICATES */
/************************/
data check_dups; set table4; keep suid; run;
proc sort data=check_dups nodup; by suid; run;

/*************************************************/
/* CREATE 'ALL' GROUPING FOR RACE & DENOMINATORS */
/*************************************************/
data table4; set table4; output; racegrp='1_All'; output; run;

proc freq data=table4 noprint; tables racegrp / out=denom(drop=percent
rename=(count=denom)); run;
proc print data=denom; title'Denominators: Racial group'; run;

/******************/
/* GET STATISTICS */
/******************/
%macro table4(var,out,title);
 proc freq data=table4 noprint; tables racegrp*&var / out=&out(drop=percent
rename=(count=numer)); format &var &var.; run;
```

```
 data &out; merge &out denom; by racegrp; pct=(numer/denom)*100; run;
 proc print data=&out(where=(&var=1)); title"Frequency counts: &title"; run;
%mend table4;

%table4(j_si,out_j_si,%str(Sacroiliitis (j_si)));
%table4(j_as,out_j_as,%str(Ankylosing spondylitis (j_as)));
%table4(eye_uv,out_eye_uv,%str(Uveitis (eye_uv)));
%table4(skin_en,out_skin_en,%str(Erythema nodosum (skin_en)));
%table4(skin_py,out_skin_py,%str(Pyoderma gangrenosum (skin_py)));
%table4(liv_psc,out_liv_psc,%str(Primary sclerosing cholangitis (liv_psc)));
```

ATTACHMENT 3

Confirmatory Stata Code Provided by the DCC for Selected Tabulations from the IBDGC Phenotype Analysis Dataset in the NIDDK Repository

```
  ___  ____  ____  ____  ____ tm
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   10.1   Copyright 1984-2008
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

60-user Stata for Macintosh (network) license expires 30 Jun 2009:
       Serial number:  81910043852
         Licensed to:  Phil Schumm
                       Department of Health Studies

Notes:
      1.  (-m# option or -set memory-) 10.00 MB allocated to data
      2.  (-v# option or -set maxvar-) 5000 maximum variables
      3.  Command line editing disabled
      4.  Stata running in batch mode


running /usr/local/bin/profile.do ...
--------------------------------------------------------------------------------
----------
       log:  /tmp/statalog_092908_162121.log
  log type:  text
 opened on:  29 Sep 2008, 16:21:21

. do reports/nguyen-replication

. // $Id: nguyen-replication.do 453 2008-09-29 21:18:59Z pschumm $
.
. // replicate results from Geoff Nguyen's paper, using dataset submitted to RTI
.
. version 10

. clear

.
. // grab data submitted to RTI, excluding those subjects that were not part of
. // Geoff's analysis
. insheet using raw/nguyen-sample.txt
(1 var, 1123 obs)

. isid suid, so
(data now sorted by suid)

. tempfile nguyen_sample

. save `"`nguyen_sample'"'
file /var/folders/h6/h6Fm5HMPHWmZr-ei8VIDEk+++TI/-Tmp-//St02364.000002 saved

.
. insheet using tmp/ibdgc-rti-1.0/phenotypes.txt, clear
(74 vars, 4761 obs)
```

```
. isid suid, so
(data now sorted by suid)

. merge suid using `"`nguyen_sample'"'

. keep if _merge==3
(3638 observations deleted)

. drop _merge

.
. // generate race/ethnicity variable for use below
. gen ethgroup = 1 if race==1 & hispanic==2
(294 missing values generated)

. replace ethgroup = 2 if race==2
(126 real changes made)

. replace ethgroup = 3 if hispanic==1 & race!=2
(168 real changes made)

. lab def ethgroup 1 "White" 2 "Black" 3 "non-Black Hispanic"

. lab val ethgroup ethgroup

. tab ethgroup

          ethgroup |      Freq.     Percent        Cum.
-------------------+-----------------------------------
             White |        829       73.82       73.82
             Black |        126       11.22       85.04
non-Black Hispanic |        168       14.96      100.00
-------------------+-----------------------------------
             Total |      1,123      100.00

.
.
. // proportion of affected siblings (Table B)
. // note: here we ignore sibs with unknown affection status, as Geoff did
. gen no_affected_sibs = sib_cd+sib_uc+sib_ibd

. gen no_sibs = sib_cd+sib_uc+sib_ibd+sib_unf

. ratio no_affected_sibs / no_sibs if ethgroup=="Black":ethgroup

Ratio estimation                    Number of obs    =     126

     _ratio_1: no_affected_sibs/no_sibs

------------------------------------------------------------
             |             Linearized
             |      Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
    _ratio_1 |   .0252101   .0119009      .0016567    .0487634
------------------------------------------------------------

.
```

```
.
. // UC disease extent (Table D)
. // switch to maximal-extent coding
. gen uc_dis_extent = 1 if proctit==1 & left==2 & extensiv==2
(1098 missing values generated)

. replace uc_dis_extent = 2 if left==1 & extensiv==2
(114 real changes made)

. replace uc_dis_extent = 3 if extensiv==1
(263 real changes made)

. lab def uc_dis_extent 1 "proctitis" 2 "left-sided colitis" 3 "extensive colitis"

. lab val uc_dis_extent uc_dis_extent

. tab uc_dis_extent ethgroup if diag==3, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

                   |             ethgroup
     uc_dis_extent |     White      Black  non-Black |     Total
-------------------+---------------------------------+----------
         proctitis |        18          4          2 |        24
                   |      6.19      13.79       3.77 |      6.43
-------------------+---------------------------------+----------
left-sided colitis |        90         10          9 |       109
                   |     30.93      34.48      16.98 |     29.22
-------------------+---------------------------------+----------
 extensive colitis |       183         15         42 |       240
                   |     62.89      51.72      79.25 |     64.34
-------------------+---------------------------------+----------
             Total |       291         29         53 |       373
                   |    100.00     100.00     100.00 |    100.00


.
.
. // CD surgery (Table F)
. replace op_ad = 0 if surgery==2
(655 real changes made)

. replace op_pd = 0 if surgery==2
(655 real changes made)

. sum op_ad if diag==1 & ethgroup=="Black":ethgroup

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       op_ad |        79     .721519     1.13156          0          5

. sum op_pd if diag==1 & ethgroup=="Black":ethgroup
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       op_pd |        79    .3164557    .7079096          0          4

.
.
. // compute time till first surgery (Table F)
. gen time_to_surgery = (surg_yr - diag_yr) if surgery==1
(674 missing values generated)

. replace time_to_surgery = (review - diag_yr) if surgery==2
(654 real changes made)

. stset time_to_surgery, fail(surgery==1) if(diag==1)

     failure event:  surgery == 1
obs. time interval:  (0, time_to_surgery]
 exit on or before:  failure
               if:  diag==1

--------------------------------------------------------------------------------
     1123  total obs.
       20  event time missing (time_to_surgery>=.)              PROBABLE ERROR
      427  ignored per request (if(), etc.)
      133  obs. end on or before enter()
--------------------------------------------------------------------------------
      543  obs. remaining, representing
      263  failures in single record/single failure data
     3571  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        35

.
. // median as estimated by Weibull model
. streg, dist(weibull)

         failure _d:  surgery == 1
   analysis time _t:  time_to_surgery

Fitting constant-only model:

Iteration 0:   log likelihood = -572.49241
Iteration 1:   log likelihood = -569.20507
Iteration 2:   log likelihood = -569.20227
Iteration 3:   log likelihood = -569.20227

Fitting full model:
Iteration 0:   log likelihood = -569.20227

Weibull regression -- log relative-hazard form

No. of subjects =          543                     Number of obs   =       543
No. of failures =          263
Time at risk    =         3571
                                                   LR chi2(0)      =      0.00
Log likelihood  =   -569.20227                     Prob > chi2     =         .
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-------------+----------------------------------------------------------------
       /ln_p |   .1249559   .0471767     2.65   0.008     .0324913    .2174205
-------------+----------------------------------------------------------------
           p |   1.133098   .0534558                      1.033025    1.242867
         1/p |   .8825358   .0416351                      .8045916    .9680309
------------------------------------------------------------------------------

. predict median, median
(option median time assumed; predicted median time)
(580 missing values generated)

. sum median

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      median |       543    9.449808           0   9.449808   9.449808

.
. // survivor function at 2, 5, and 10 years
. sts list, at(2 5 10)

         failure _d:  surgery == 1
   analysis time _t:  time_to_surgery

              Beg.                       Survivor      Std.
    Time     Total     Fail             Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
       2       454       74              0.8604    0.0151     0.8278    0.8873
       5       269       78              0.6746    0.0222     0.6290    0.7159
      10       132       59              0.4801    0.0269     0.4266    0.5316
-------------------------------------------------------------------------------
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.

.
. // summarize time till first surgery separately by ethgroup
. xi: streg i.ethgroup, dist(weibull)
i.ethgroup        _Iethgroup_1-3      (naturally coded; _Iethgroup_1 omitted)

         failure _d:  surgery == 1
   analysis time _t:  time_to_surgery

Fitting constant-only model:

Iteration 0:   log likelihood = -572.49241
Iteration 1:   log likelihood = -569.20507
Iteration 2:   log likelihood = -569.20227
Iteration 3:   log likelihood = -569.20227

Fitting full model:

Iteration 0:   log likelihood = -569.20227
Iteration 1:   log likelihood = -566.49388
```

```
Iteration 2:   log likelihood = -566.39486
Iteration 3:   log likelihood = -566.39475
Iteration 4:   log likelihood = -566.39475

Weibull regression -- log relative-hazard form

No. of subjects =          543                     Number of obs   =       543
No. of failures =          263
Time at risk    =         3571
                                                   LR chi2(2)      =      5.62
Log likelihood  =   -566.39475                     Prob > chi2     =    0.0604

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iethgroup_2 |   .8198009   .1739502    -0.94   0.349     .5408731    1.242571
_Iethgroup_3 |    1.46634   .2658532     2.11   0.035     1.027799    2.091996
-------------+----------------------------------------------------------------
       /ln_p |   .1382062   .0474798     2.91   0.004     .0451475    .2312648
-------------+----------------------------------------------------------------
           p |   1.148212   .0545169                      1.046182    1.260193
         1/p |   .8709191    .041351                      .7935293    .9558565
------------------------------------------------------------------------------

. predict median_by_ethgroup, median
(option median time assumed; predicted median time)
(580 missing values generated)

. tabstat median_by_ethgroup, by(ethgroup)

Summary for variables: median_by_ethgroup
     by categories of: ethgroup

        ethgroup |      mean
-----------------+----------
           White |  9.659752
           Black |  11.48468
non-Black Hispan |  6.921323
-----------------+----------
           Total |  9.447864
----------------------------

. sts list, at(2 5 10) by(ethgroup)

         failure _d:  surgery == 1
   analysis time _t:  time_to_surgery

              Beg.                      Survivor      Std.
    Time     Total     Fail             Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
White
       2       331       61              0.8452    0.0183     0.8054    0.8774
       5       201       54              0.6734    0.0255     0.6206    0.7206
      10       108       42              0.4931    0.0305     0.4320    0.5512
Black
       2        55        3              0.9476    0.0295     0.8461    0.9828
       5        37        9              0.7641    0.0602     0.6201    0.8594
```

```
      10        18         6              0.5889    0.0796     0.4175    0.7254
non-Black Hispanic
       2        68        10              0.8700    0.0385     0.7713    0.9280
       5        31        15              0.6053    0.0636     0.4692    0.7168
      10         6        11              0.2826    0.0754     0.1479    0.4337
-------------------------------------------------------------------------------
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.

.
.
. // fulminant colitis (Table F)
. tab ethgroup surg_acu if diag==3, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

                   |            surg_acu
          ethgroup |         1          2          3 |     Total
-------------------+---------------------------------+----------
             White |         6         40          1 |        47
                   |     12.77      85.11       2.13 |    100.00
-------------------+---------------------------------+----------
             Black |         0          4          0 |         4
                   |      0.00     100.00       0.00 |    100.00
-------------------+---------------------------------+----------
non-Black Hispanic |         2         17          0 |        19
                   |     10.53      89.47       0.00 |    100.00
-------------------+---------------------------------+----------
             Total |         8         61          1 |        70
                   |     11.43      87.14       1.43 |    100.00


.
end of do-file
```

ATTACHMENT 4

Description of the Genotypic Data Housed in the NIDDK Genetics Repository at Rutgers

NIDDK IBDGC Crohn's Disease Genome-Wide Association Study

:Study name: NIDDK IBDGC Crohn's Disease Genome-Wide Association Study
:Study report name: NIDDK IBD Genetics Consortium Crohn's Disease Genome-Wide Association Study
:Version: $Revision: 432 $
:Date: $Date: 2008-03-28 17:46:54 -0500 (Fri, 28 Mar 2008) $

Description

This dataset contains data from a genome-wide association study performed with 968 IBD-affected cases and 995 unrelated controls using the Illumina HumanHap300 Genotyping BeadChip. Cases were selected to have Crohn's disease with ileal involvement, and controls were matched to cases based on sex and year of birth. Subjects were drawn from two cohorts: (1) persons with non-Jewish, European ancestry (561 cases and 563 controls), and (2) persons with Jewish ancestry (407 cases and 432 controls). Genotyping was performed at the Feinstein Institute for Medical Research.

Seven-hundred and fifty-four of the samples (468 cases and 286 controls) were taken from the NIDDK IBD Genetics Consortium cell line repository. These samples are identified in the file dbGaP_consent.txt. The subject IDs for these individuals may be used to request corresponding samples for follow-up research through the repository. In addition, complete phenotype data for these individuals are available, together with the Consortium's phenotyping manual and the forms used to collect the data. The remaining 1,209 samples were obtained from pre-existing collections ascertained through Cedars-Sinai Medical Center, Johns Hopkins University, University of Chicago, University of Montreal, University of Pittsburgh, University of Toronto, and the New York Health project (controls only). For these samples, only sex, cohort (Jewish vs. non-Jewish), and age at diagnosis (cases only) are available.

Two-hundred and three individuals from among the pre-existing samples did not provide consent to release their genotype data (designated as consent group 2 in the file dbGaP_consent.txt). Thus, individual genotype data are only provided for 1,760 samples. To compensate for this, we have provided summary results for each SNP. These are based on a stratified analysis testing case/control association. Fifty-one samples had a call rate less than 93% and were therefore excluded from this analysis, leaving an overall sample size of 1,963 - 51 = 1,912.

X Chromosome Heterozygosity

Nine samples have X chromosome heterozygosity that is neither consistent nor inconsistent with their phenotypic sex. One of these (3019572) was found to have Turner Syndrome. The remaining 8 (3001651, 3002191, 3003302, 3003339, 3005474, 3006537, 3014051, 3017976) have heterozygosity ranging from 35-76%.

Type

Case-control study, stratified by ancestry (European non-Jewish vs. Jewish).

Disease Name(s)

- Crohn's disease
- Ulcerative colitis
- Inflammatory bowel disease

Inclusion/Exclusion Criteria

Cases were selected to have Crohn's disease with ileal involvement. After genotyping was completed and updated phenotype information was obtained, two (non-repository) cases were identified as having Indeterminate Colitis instead of Crohn's disease, and one Crohn's case was found not to have ileal involvement (these three samples are nonetheless included in the dataset).

Diagnosis of IBD required (i) one or more of the following symptoms: diarrhea, rectal bleeding, abdominal pain, fever or complicated perianal disease; (ii) occurrence of symptoms on two or more occasions separated by at least 8 weeks or ongoing symptoms of at least 6 weeks' duration and (iii) objective evidence of inflammation from radiologic, endoscopic and histologic evaluation. Ileal Crohn's disease involvement was defined as mucosal ulceration, cobblestoning, stricturing or bowel wall thickening from endoscopy reports, barium X-rays, operative reports and/or pathology resection specimen reports. Individuals with either 'ileal only' or 'ileocolonic' were included.

Within each of the non-Jewish and Jewish cohorts, controls were matched to cases based on sex and year of birth. In addition, controls were required to meet the following inclusion criteria: (1) no history of IBD among 1st and 2nd degree relatives, (2) never been diagnosed with IBD, and (3) never experienced chronic diarrhea, unexplained rectal bleeding, or unexplained weight loss.

Relevant Publications

[1] R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, T. Dassopoulos, A. Bitton, H. Yang, S. Targan, L. W. Datta, E. O. Kistner, L. P. Schumm, A. T. Lee, P. K. Gregersen, M. M. Barmada, J. I. Rotter, D. L. Nicolae, and J. H. Cho. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. Science, 314(5804):1461╤1463, 2006. PMID: 17068223.

[2] J. D. Rioux, R. J. Xavier, K. D. Taylor, M. S. Silverberg, P. Goyette, A. Huett, T. Green, P. Kuballa, M. M. Barmada, L. W. Datta, Y. Y. Shugart, A. M. Griffiths, S. R. Targan, A. F. Ippoliti, E.-J. Bernard, L. Mei, D. L. Nicolae, M. Regueiro, L. P. Schumm, A. H. Steinhart, J. I. Rotter, R. H. Duerr, J. H. Cho, M. J. Daly, and S. R. Brant. Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. Nat Genet, 39(5):596╤604, 2007. PMID: 17435756.

Attribution
==========

:Principal Investigator: Judy H. Cho, MD
:Affiliation: Department of Medicine, Yale University, New Haven
:Funding Source:

    NIDDK DK62431 (Steven R. Brant), DK62422 (Judy H. Cho), DK62420 (Richard H. Duerr),
    DK62432 (John D. Rioux), DK62423 (Mark S. Silverberg), DK62413 (Kent D. Taylor),
    and DK62429 (Judy H. Cho)