Notes on the GEMS-1 and GEMS-1A Data sets

1) What are the pn variables?

The pn (for "pathogen") variables were defined for convenience in computer programming for analysis. They are 0-1 variables (0 if the putative pathogen was not isolated, 1 if it was isolated). The following table lists the pn variables and the corresponding pathogens in GEMS-1 (a total of 42 pathogens, pn1 through pn42).

Pathogen number	Pathogen name
pn1	ETEC, any
pn2	ST-producing ETEC (ST only or LTST)
pn3[1]	ETEC LT and ST
pn4[1]	ETEC ST only
pn5	ETEC LT only
pn6	EAEC
pn7[1]	EAEC AATA only
pn8[1]	EAEC AAIC only
pn9[1]	EAEC AATA and AAIC
pn10	Typical EPEC (tEPEC)
pn11	EAE and BFPA
pn12	BFPA only
pn13	EAE positive, BFPA negative
pn14	Shiga toxin
pn15	Atypical EPEC (aEPEC)
pn16	EHEC (EAE positive, stx positive)
pn17	*Shigella*, any
pn18	*Shigella flexneri*
pn19	*Shigella sonnei*
pn20	*Shigella dysenteriae*
pn21	*Shigella boydii*
pn22	Rotavirus
pn23	*Aeromonas*
pn24	Norovirus, any (GI or GII)
pn25	Norovirus, GI
Pn26	Norovirus, GII
pn27	Adenovirus 4041
pn28	Adenovirus non-4041
pn29	*Vibrio cholerae* O1
pn30[1]	*Vibrio cholerae* O1 Inaba
pn31[1]	*Vibrio cholerae* O1 Ogawa
pn32	*Vibrio cholerae* O139
pn33	*Salmonella typhi*
pn34	Non-typhoidal *salmonella*
pn35	*Cryptosporidium*
pn36	*Giardia*

pn37	*Entamoeba histolytica*
pn38[1]	*Campylobacter*, any
pn39	*Campylobacter* jejuni
pn40	*Campylobacter* coli
pn41	Astrovirus
pn42	Sapovirus

[1] - Those pathogens are present only in GEMS-1.

Following is a table of the pn variables in GEMS-1A (a total of 52 pathogens):

Pathogen number	Pathogen name
pn1	ETEC, any
pn2	ST-producing ETEC (ST only or LTST)
pn3	ETEC LT and ST
pn4	ETEC ST only
pn5	ETEC LT only
pn6	EAEC
pn7	EAEC AATA only
pn8	EAEC AAIC only
pn9	EAEC AATA and AAIC
pn10	Typical EPEC (tEPEC)
pn11	EAE and BFPA
pn12	BFPA only
pn13	EAE positive, BFPA negative
pn14	Shiga toxin
pn15	Atypical EPEC (aEPEC)
pn16	EHEC (EAE positive, stx positive)
pn17	*Shigella* any
pn18	*Shigella* flexneri
pn19	*Shigella* sonnei
pn20	*Shigella* dysenteriae
pn21	*Shigella* boydii
pn21a[2]	*Shigella* with non-typeable serogroup
pn22	Rotavirus
pn23	*Aeromonas*
pn24	Norovirus_any (GI+GII)
pn25	Norovirus_GI
pn26	Norovirus_GII
pn27	Adenovirus_4041
pn28	Adenovirus_non4041
pn51[2]	*Vibrio cholerae* O1 or O139
pn29	*Vibrio cholerae* O1
pn32	*Vibrio cholerae* O139

pn33	*Salmonella typhi*
pn34	Non-typhoidal *salmonella*
pn35	*Cryptosporidium*
pn36	*Giardia*
pn37	*Entamoeba histolytica*
pn38	*Campylobacter*, any
pn39	*Campylobacter jejuni*
pn40	*Campylobacter coli*
pn41	Astrovirus
pn42	Sapovirus
pn43[2]	*Helicobacter pylori*
pn44[2]	*Clostridium difficile* no toxin
pn45[2]	*Clostridium difficile* toxin
pn46[2]	No *Clostridium difficile* but toxin
pn47[2]	*Ascaris lumbricoides*
pn48[2]	*Strongyloides stercoralis*
pn49[2]	Hookworm
pn50[2]	Toxin positive *Bacteroides fragilis*
pn61[2]	GDH positive, toxin negative
pn62[2]	GDH negative, toxin positive

[2] - Those pathogens are present only in GEMS-1A.

Some of the above pathogens were not assayed in GEMS-1, and the values of the corresponding "pn" variables are missing in the GEMS-1 data (variables are pn43 through pn62 in the above table). Also, in both the GEMS-1 and GEMS-1A data sets, some pn variables and other variables were used only as intermediate variables when deriving certain pn variables; those variables were not used directly in data analysis (for example, pn11-pn14 in the above tables and tEPEC_bfpA_Roy1, tEPEC_bfpA_Roy2, and stec_c1 and stec_c2 in GEMS-1).

2) The pn variables were used in the main GEMS-1 and GEMS-1A analyses of potential pathogens. It should be noted, however, that the current pn variables for any *Shigella* and *Shigella* serotypes (pn17-pn21) are in some instances different from the original pn variables. *Shigella* strains were re-assayed at UMB and CDC, and these re-assayed results are now given in the pn variables. It was decided to keep both the original and corrected results for strains originally coded as being positive for *Shigella*. Thus, in both the GEMS-1 and GEMS-1A data sets, variable names that begin with F16 (Stool Culture) or F17 (E. Coli Polymerase) have "CORR" after the F16 or F17 prefix if they contain the corrected values and "CRF" after the F16 or F17 if they contain the original data as recorded on the case report form. The F16CORR and F17 CORR variables agree with the current pn variables.

3) In the GEMS-1 data set, one child from Mozambique (CHILDID # and CASEID # 303000207, LAB_SPECIMEN_ID # 300220) had no lab data, and therefore this child and the matching control (CHILDID # 303900141, CASEID # 303000207, LAB_SPECIMEN_ID # 300331) were deleted from the GEMS-1 data set.

4) Continuous variables in the SPSS data sets will sometimes need to have their decimal allowance increased. This can be done from the Variable View tab of the data.

5) Additions made to GEMS-1 and GEMS-1A data sets on May 9, 2016:

a) In the GEMS-1 data set, a variable **base_age** was added. (This variable was already present in the GEMS-1A data set.) The variable **base_age** is the maximum value of (**F3_AGE**, **F6_AGE**). In addition, a new **agegroup** variable was redefined based on the maximum value of (**F3_AGE**, **F6_CASE_AGE**), rather than on **base_age**; thus, the age group (0-11, 12-23, or 24-59 months) of a control is now defined to be the same as the age group of the corresponding case. The old **CASE_AGE_CAT** variable (based on **base_age**) was deleted. (In the GEMS-1A data set, **CASE_AGE_CAT** was already correctly defined.)

b) In both the GEMS-1 and GEMS-1A data sets, a new **death** variable was created. If a case died in a hospital/health center, or a case or control died within the follow-up period, this **death** variable was assigned a value of 1. If the vital status at the end of the follow-up period was unknown, the **death** variable was coded as missing. If the child was known to be alive at follow-up, the **death** variable was coded 0. In the GEMS-1A data set, the old **deathind** variable was deleted, and the **death** variable as described above was added.

6) Changes made to GEMS-1 data set on July 12, 2016 and to GEMS-1A data set on July 26, 2016:

In some instances the same CHILDID appears in both the GEMS-1 and GEMS-1A data sets; however, they represent different individuals. To avoid any confusion in an analysis of the combined GEMS-1 and GEMS-1A, a digit was added at the right for CHILDID and CASEID, thus increasing the length of CHILDID and CASEID by one digit. The new rightmost digit is 1 in the GEMS-1 data and 2 in the GEMS-1A data.