Statistical Methods in the Global Enteric Multicenter Study (GEMS)

William C. Blackwelder,[1] Kousick Biswas,[2] Yukun Wu,[1] Karen L. Kotloff,[1] Tamer H. Farag,[1] Dilruba Nasrin,[1] Barry I. Graubard,[3] Halvor Sommerfelt,[4,5] and Myron M. Levine[1]

[1]Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, [2]Veterans Administration, Perry Point, and [3]National Cancer Institute, Bethesda, Maryland; [4]Centre for International Health, University of Bergen, and [5]Division of Infectious Disease Control, Norwegian Institute of Public Health, Oslo, Norway

The Global Enteric Multicenter Study (GEMS) is an investigation of the burden (number of cases and incidence) of moderate-to-severe diarrhea (MSD) in children <60 months of age at 7 sites in sub-Saharan Africa and South Asia. The population attributable fraction for a putative pathogen, either unadjusted or adjusted for other pathogens, is estimated using the proportion of MSD cases from whom the pathogen was isolated and the odds ratio for MSD and the pathogen from conditional logistic regression modeling. The adjusted attributable fraction, proportion of MSD cases taken to a sentinel health center (SHC), number of cases presenting to an SHC, and the site's population are used to estimate the annual number of MSD cases and MSD incidence rate attributable to a pathogen or group of pathogens. Associations with death and nutritional outcomes, ascertained at follow-up visits to case and control households, are evaluated both in MSD cases and in the population.

Diarrheal diseases are one of the top 2 causes of death among children <60 months of age in the developing world [1]. Interventions to diminish this enteric disease burden among the world's most disadvantaged pediatric populations are expected to include pathogen-specific vaccines and diagnostics (followed by specific treatment), as well as expanded use of nonspecific therapeutic regimens such as oral rehydration and zinc. Despite a plethora of individual site studies [2–8], and a few coordinated multicountry studies [9] of both case/control and prospective cohort design, there remains much disagreement over the relative importance of various specific bacterial, viral, and protozoal pathogens as causes of diarrheal illness, particularly of clinically more severe forms. As additional diarrheal pathogens have come to be described in recent years and as diagnostic microbiologic tests have become more sensitive, the need for a definitive study of pediatric diarrheal disease gained widespread support and momentum [10]. This led to initiation of the Global Enteric Multicenter Study (GEMS), a matched case/control study of the burden, in terms of numbers of cases and incidence rates, of moderate-to-severe diarrhea (MSD) in children <60 months of age at 4 sites in sub-Saharan Africa and 3 sites in South Asia (see Levine et al and Kotloff et al in this supplement) [11, 12]. GEMS, which involves the detection of a wide array of etiologic agents in enrolled MSD cases and their matched controls, represents a historic multisite undertaking to apply standardized specific microbiologic methods to detect evidence of infection with 1 or more of a wide array of potential pathogens, and to use the resulting data to estimate the disease burden attributable to specific pathogens.

In this paper we describe and illustrate the major statistical methods used in GEMS. These include methods for assessing associations between the presence of specific pathogens (and other variables of

Correspondence: William C. Blackwelder, PhD, Center for Vaccine Development, University of Maryland School of Medicine, 685 W Baltimore St, Baltimore, MD 21201 (wblackwe@medicine.umaryland.edu).

Clinical Infectious Diseases 2012;55(S4):S246–53

DOI: 10.1093/cid/cis788

interest) and MSD, estimating the proportion and absolute number of cases of MSD due to specific pathogens, estimating the incidence of MSD and the proportion of MSD cases who are taken to one of the health centers designated as a study site (referred to as sentinel health centers [SHCs]), and assessing associations between pathogens and other variables of interest with outcomes other than MSD.

STUDY DESIGN

For GEMS, MSD is defined by an episode of diarrhea (≥3 loose stools within a 24-hour period) with onset within the past 7 days and at least 7 days after the end of any previous episode, and at least 1 of the following: sunken eyes, more than normal; loss of skin turgor; intravenous rehydration administered or prescribed; visible blood in stool; or hospitalization with diarrhea. At each GEMS study site, children 0–59 months of age who had MSD and were brought to one of the site's SHCs were enrolled into the study, along with 1 or more controls who were matched to the case by age, time (within 7 days of the case enrollment), and geographic location of residence. Up to 8 or 9 cases per 2-week period in each of 3 age strata (0–11, 12–23, and 24–59 months) were typically enrolled during a 3-year enrollment period at each site. These strata represent age groups in which MSD and its clinical presentations (eg, dysentery) are observed with different frequencies and when the etiologies are known to be somewhat different. For example, certain etiologies are relatively more important in infants, while others are more common in toddlers or preschool children with MSD. Most of our analyses have been done within these age strata. Calculation of statistical power was based on comparing 2 independent proportions; for a moderate degree of correlation between presence of a pathogen in a case and in its matched control, the power for a given sample size will be higher for a test designed for matched data. The planned sample size at each site was 600 analyzable case/control pairs. This sample size should be sufficient, for example, for a test at the 2-sided 5% significance level to have 80% power to find a significant difference between proportions of cases and controls for which a specific pathogen is isolated, if the respective true proportions are 5.8% and 2.5%. Stool specimens were collected from each case and control for identification of potential enteric pathogens. Demographic, anthropometric, and other information about the study child and the household was collected at enrollment and also at a home visit approximately 50–90 days after enrollment.

Besides the matched case/control study, a Health Care Utilization and Attitudes Survey (HUAS), based on random sampling from each site's demographic surveillance system (DSS), was conducted before the beginning of the study. Truncated versions of the HUAS, known as HUAS-lite, were conducted several times during the 3 years of case and control enrollment. The HUAS was used to evaluate associations between a variety of demographic factors and characteristics of households (eg, main source of water, main method of disposal of feces) with the presence of diarrhea in the sampled child. The HUAS-lite surveys are used primarily to estimate the proportion of children with MSD who were taken to one of the site's SHCs and to estimate the 1-week incidence of MSD.

The GEMS study design is given in more detail in the article by Kotloff et al in this supplement [12].

ANALYSES OF HUAS AND HUAS-LITE DATA

The primary analyses of the HUAS and HUAS-lite data are (1) estimation of the proportion of children with MSD who are taken to one of the site's SHCs within 7 days of onset of diarrhea, (2) estimation of the 1-week incidence of MSD, and (3) identification of associations between characteristics of a household or primary caretaker and care seeking for diarrhea.

We use "r" to represent the proportion (and its estimate) of MSD cases who were taken to one of our designated SHCs within 7 days of onset of diarrhea; r is calculated from HUAS-lite data, since the HUAS-lite surveys were conducted during the period of case/control enrollment. In calculating r for all sites except Kenya, we use site-specific sampling weights that are defined for each combination of age group (0–11, 12–23, and 24–59 months) and sex. (The entire DSS population is included in HUAS-lite surveys in the Kenya site, so no weighting is necessary.) For each HUAS-lite round, the sampling weight for an age-sex category is the number of children in that category in the DSS population represented by each child in the HUAS-lite sample. Then the weight for each child in an age-sex category is the DSS population total for that category divided by the number of children in the HUAS-lite sample in the category. These weights are used in a time-to-event (life table) analysis using the Kaplan-Meier method to estimate, at each day beginning with the day after onset of diarrhea, the proportion of children with MSD in the population who had been taken to an SHC. Time-to-event analysis is used because for many children who currently had MSD, the HUAS-lite interview was conducted before the child reached the seventh day of the episode, so the child had not had a full 7 days after onset of diarrhea in which to be taken to an SHC. The data for all HUAS-lite rounds conducted during the case/control study are pooled [13], with each child weighted according to the sampling weight assigned for that child's HUAS-lite round. The estimate of r is then the proportion of children who were taken to an SHC by day 7 after onset of diarrhea in the time-to-event analysis.

We also estimate the 1-week incidence of MSD from HUAS-lite data. For this estimate, we pool data from different HUAS-lite rounds and use sampling weights to obtain these estimates, as is done for estimates of the proportion of cases taken to an SHC within 7 days of onset. We count the number of children with MSD whose illness began on the day of the interview or one of the 6 days preceding that day. As for estimating r, time-to-event analysis is used. In this analysis, children with diarrhea that had not progressed to MSD and whose diarrhea began <1 week prior to the HUAS-lite survey were censored after the number of days they had had diarrhea.

Because much more information was collected in the HUAS than in the HUAS-lite rounds, with the HUAS data we study associations between care seeking for diarrhea and a variety of characteristics of a household or primary caretaker. The main analytic method used to assess these associations is logistic regression modeling, in which we use the sampling weights in order to obtain results that relate to the DSS population.

ASSOCIATIONS WITH MSD IN THE CASE/CONTROL STUDY

An analysis that is central to the aims of GEMS is the evaluation from the case/control data of potential risk and protective factors for MSD. Because cases of MSD are matched with 1 or more controls, we use conditional logistic regression (CLR) modeling to estimate associations with MSD [14]. In this type of model, case status (case = 1, control = 0) is the dependent variable. The model differs slightly from the usual (unconditional) logistic regression model in that there is no intercept. Thus, the fitted model is of the form

$$\log_e(\text{odds}) = b_1x_1 + b_2x_2 + \cdots + b_kx_k,$$

where x_1, x_2, … , x_k are independent variables under study for association with MSD and b_1, b_2, … , b_k are estimates of the corresponding coefficients. When a variable x_i is dichotomous (ie, 1 if the factor is present and 0 if the factor is not present), $\exp(b_i)$ is an estimate of the odds ratio for the factor—that is, the ratio of the odds of MSD when the factor is present to the odds when the factor is absent, where the odds of an event that occurs with probability Q are Q/(1-Q). In order to obtain appropriate results when the number of discordant case/control pairs (ie, pairs where the factor is present in the case and absent in the control, or vice versa) is 0, we use a penalized likelihood approach [15]. Typically, we fit CLR models for each site and age category separately. In certain analyses it may be appropriate to combine data for different sites and/or different age groups.

Of special interest are associations of putative enteric pathogens—bacterial, viral, and protozoan—with MSD. To assess and quantify the contribution of a specific pathogen without regard to the presence of other pathogens, we fit a CLR model with a dichotomous variable, representing presence or absence of the pathogen, as the only covariate. In analysis of the contribution of a pathogen adjusted for other specific pathogen (s), we fit models with multiple dichotomous variables, each representing presence or absence of one of the pathogens, as covariates. In developing these models, interactions between the effects of pairs of pathogens are considered (ie, the possibility that the association of a pathogen with MSD depends on whether another pathogen is present).

CLR modeling is also used to evaluate associations of environmental and socioeconomic factors with MSD. Two analyses of particular interest are of associations of water sources, sanitary facilities, and hygiene practices with MSD, and of care-seeking costs with MSD.

THE POPULATION ATTRIBUTABLE FRACTION OF MSD DUE TO 1 OR MORE PATHOGENS

The population attributable fraction (AF) of a disease due to a risk factor is the proportion of disease cases (or the proportion of the risk of disease) that might theoretically be eliminated if the risk factor were eliminated. Other names that have been used for this concept include attributable risk and etiologic fraction. AF can be estimated equivalently [16] from the distribution of the exposure (risk factor) either in the entire population [17] or in cases of disease [18]. Although the concept of AF has been known and applied for decades and there have been scores of case/control and cohort studies that have tested for multiple etiologic agents of diarrheal disease to gather information on the relative importance of different agents in association with diarrhea, we have noted only 2 etiologic studies of diarrhea in which the AF concept was applied [8, 19].

In GEMS we use AF to estimate the fraction of MSD cases due to a specific pathogen or a group of pathogens. We calculate AF for a pathogen, A, as though A were the only risk factor for MSD, and also for A adjusted for other pathogens that might be present. Adjustment for other pathogens is important, since at least 2 of the potential pathogens under study in GEMS were identified in substantial percentages of both cases and controls.

To determine what pathogens are associated with MSD in GEMS, we fit CLR models, as described above, to the data on cases and matched controls. Unadjusted AF for pathogen A is estimated from a model in which the only covariate is an indicator variable y for the presence of A (ie, y = 1 if A is present and y = 0 if A is absent). Given the coefficient b of y in the fitted model, the odds ratio (OR) for MSD and A is estimated as e^b. We assume we have a random sample of MSD cases and represent the proportion of MSD cases for which A is present by Pr (A|MSD). If OR >1, the unadjusted attributable fraction AF_u is then given by

$$AF_u = \Pr(A|MSD)\left(1 - \frac{1}{OR}\right). \tag{1}$$

As is common in case/control studies, we use OR as an approximation to the risk ratio (RR). In GEMS the 1-week incidence of MSD, which is the basis for choosing cases, is small (ranging from <1% to approximately 9%, depending on the study site and age group). Since the controls are closely matched in time to cases, OR is a close approximation to the incidence rate ratio [20], which with these small incidence rates is in turn close to the RR.

We estimate AF for pathogen A1, adjusted for the presence of other pathogens, as in Bruzzi et al [21]. For example, suppose we adjust A1 for another pathogen A2. We fit a (multiple) conditional logistic regression model that in its most general form includes variables y_1 and y_2, indicating presence or absence of A1 and A2, respectively; and the product of y_1 and y_2, which represents the interaction of the effects of A1 and A2. In this model an interaction indicates that the OR for MSD and A1 depends on whether or not A2 is present. The model will have estimated coefficients b_{10} for y_1, b_{01} for y_2, and b_{11} for the product of y_1 and y_2. We let ρ_{ij} be the proportion of cases with $y_1 = i$ and $y_2 = j$, for i and j = 0 or 1; for example, ρ_{10} is the proportion of cases with A1 present and A2 absent. Then AF_a, the adjusted attributable fraction estimate for A1, can be written

$$AF_a = \rho_{10}\left(1 - \frac{1}{T_{10}}\right) + \rho_{11}\left(1 - \frac{1}{T_{11}}\right), \tag{2}$$

where $T_{10} = \exp(b_{10})$ and $T_{11} = \exp(b_{10} + b_{11})$. Note that the coefficient b_{01} does not appear in formula (2); only coefficients corresponding to the presence of A1 are included. T_{1j} is the ratio of the OR for the combination (1j), in which A1 is present, to the OR for (0j), the same combination of pathogens except that A1 is absent. If there is no interaction term in the model, $AF_a = (\rho_{10} + \rho_{11})\ (1–1/T_{10}) = \Pr\ (A1|MSD)\ (1–1/T_{10})$, where T_{10} is now simply the odds ratio for A1 when A2 is absent; in this case AF_a has the same form as AF_u in equation (1), the only difference being that the OR T_{10} is estimated from a model that includes both y_1 and y_2.

In general, we can estimate the combined attributable fraction for a set of 1 or more pathogens, each with AF >0, possibly adjusted for 1 or more other pathogens—that is, in the above description of AF for A1 adjusted for A2, both pathogens A1 and A2 can be replaced by multiple pathogens. For example, suppose we want estimate to AF_a for 2 pathogens A1 and A2 (set I, the pathogens for which a summary AF is desired), adjusted for pathogens A3 and A4 (set II, the pathogens included only for conditioning). In this case we consider proportions ρ_{ijkl}, representing all combinations of presence or absence of the 4 pathogens; in the GEMS data, some of these proportions will be 0, since specific combinations with more than 2 pathogens occur infrequently. In its most general form, the model will include indicator variables y_1, y_2, y_3, and y_4, and all possible products (interactions) of these variables. The formula for AF_a is

$$AF_a = 1 - \frac{\Sigma\rho_{ijkl}}{T_{ijkl}}. \tag{3}$$

In formula (3), the summation is over all i, j, k, l = 0 or 1; T_{ijkl}, the term for combination (ijkl), is a ratio of the OR for the combination and an OR when the variables representing pathogens in set I (pathogens A1 and A2, for which a summary AF is to be estimated) are all set to 0. T_{ijkl} thus includes coefficients that correspond to pathogens in set I that are present in the combination, as well as interactions between any of them and pathogens in set II (pathogens A3 and A4, the conditioning set) that are present in the combination. T_{ijkl} will not include any coefficients of "main effects" for pathogens in set II or interactions between pathogens in set II, because these appear in both the numerator and denominator ORs that determine T_{ijkl} [21]. For a combination of the form (00kl), all pathogens in set I are absent, and $T_{ijkl} = 1$.

Suppose the CLR model for 4 pathogens, the first 2 in the set for which AF is to be estimated and the last two in the conditioning set, includes terms for all main effects and 2-way interactions (but not higher-order interactions), with coefficients b_{ijkl} for combination (ijkl). Then, for example, the combination of pathogens 1, 2, and 4 corresponds to $T_{1101} = \exp(b_{1000} + b_{0100} + b_{1100} + b_{1001} + b_{0101})$. Note that a coefficient with >1 of the subscripts equal to 1 is the coefficient of an interaction term; b_{1100} is the coefficient of the product y_1y_2, etc. For this example Table 1 gives all possible values of the natural logarithm of T, $\log_e(T_{ijkl})$, in terms of the estimates b_{ijkl} from the CLR model. Note that any combination (ijkl) for which no pathogen from set I is present will have $\log_e(T_{ijkl}) = 0$ (ie, $T_{ijkl} = 1$) in formula (3).

In the GEMS data we have occasionally seen evidence of 2-way interactions, but we have seen no evidence of 3-way or higher interactions; thus, only interactions involving 2 pathogens need be considered in the GEMS analysis.

Cases were sampled for GEMS in approximately equal numbers during each 2-week period, regardless of the number of MSD cases appearing at the SHCs. We estimate AF both unweighted and using weights defined as (number of eligible cases/number of enrolled cases), ie, as the inverse of the

Table 1. Natural Logarithm of Factors (Ratios of Odds Ratios) Corresponding to Combinations of Pathogens in Example of Adjusted Attributable Fraction (AF) Calculation: AF Is Calculated for Pathogens A1 and A2, Adjusted for Pathogens A3 and A4

Pathogen(s) Present	ijkl	$\log_e(T_{ijkl})$
A1	1 0 0 0	b_{1000}
A2	0 1 0 0	b_{0100}
A1, A2	1 1 0 0	$b_{1000} + b_{0100} + b_{1100}$
A1, A3	1 0 1 0	$b_{1000} + b_{1010}$
A2, A3	0 1 1 0	$b_{0100} + b_{0110}$
A1, A2, A3	1 1 1 0	$b_{1000} + b_{0100} + b_{1100} + b_{1010} + b_{0110}$
A1, A4	1 0 0 1	$b_{1000} + b_{1001}$
A2, A4	0 1 0 1	$b_{0100} + b_{0101}$
A1, A2, A4	1 1 0 1	$b_{1000} + b_{0100} + b_{1100} + b_{1001} + b_{0101}$
A1, A3, A4	1 0 1 1	$b_{1000} + b_{1010} + b_{1001}$
A2, A3, A4	0 1 1 1	$b_{0100} + b_{0110} + b_{0101}$
A1, A2, A3, A4	1 1 1 1	$b_{1000} + b_{0100} + b_{1100} + b_{1010} + b_{0110} + b_{1001} + b_{0101}$
A3	0 0 1 0	0
A4	0 0 0 1	0
A3, A4	0 0 1 1	0
None	0 0 0 0	0

sampling fraction for MSD cases. Data for adjacent 2-week periods are combined when there are no enrolled cases in a period. AF estimation is done separately for the 3 age strata (0–11, 12–23, and 24–59 months) within which MSD cases were sampled.

Table 2 shows unadjusted and adjusted AF estimates, unweighted, for children aged 12–23 months in India for the first 2 years of the 3-year GEMS case/control study. These results are typical of the results for other sites and age groups, in that there are few important interactions and the adjusted estimates are not very different from the unadjusted estimates. In particular, with few exceptions, the adjusted AF and number of attributable cases for the major pathogens change only modestly, compared to the unadjusted estimates.

CALCULATION OF ATTRIBUTABLE MSD CASES AND MSD INCIDENCE

Let M_{CC}, M_{SHC}, and M_{pop} represent the total number of MSD cases enrolled in the study, the total numbers of MSD cases seen at the site's SHCs, and the total number of MSD cases in the population in 3 years, respectively. Then for each site and age category, the respective numbers of cases attributable to pathogen A in the study, in the site's SHCs, and in the population are given by $AF \times M_{CC}$, $AF \times M_{SHC}$, and $AF \times M_{pop}$, respectively.

The numbers of cases attributable to A in the study and the SHCs are calculated directly from AF and the numbers of cases, since we observe M_{CC} and take M_{SHC} as the number of cases presenting at the SHCs who are eligible for the study. However, we do not observe M_{pop} directly, but rather estimate it from M_{SHC} and the estimated proportion, r, of MSD cases taken to one of the study site's SHCs. As indicated above, r is estimated from the HUAS-lite rounds conducted during the study.

The estimated annual number of MSD cases in the population during the 3-year case/control study period is $M_{pop} = M_{SHC}/(3r)$, and the estimated annual number of cases attributable to A is $AF \times M_{pop} = AF \times M_{SHC}/(3r)$. If N is the average population at the site over the study period, the annual incidence rate of MSD attributable to A during the study is

Table 2. Crude and Adjusted Attributable Fraction and Attributable Number of Cases in First 2 Years of the Global Enteric Multicenter Study: India, Ages 12–23 Months (364 Cases, 374 Controls)

		Unadjusted Analysis				Adjusted Analysis			
Pathogen	Cases With Pathogen	OR[a]	*P* Value[b]	AF	Attributable Cases	OR[a]	*P* Value[b]	AF	Attributable Cases
Rotavirus	104	22.5	<.0001	0.273	99	36.4	<.0001	0.278	101
Shigella	30	11.4	.0003	0.075	27	38.9	<.0001	0.080	29
ETEC LT/ST or ST	34	2.6	.004	0.057	21	4.3	.0006	0.072	26
Cryptosporidium	45	1.7	.031	0.052	19	2.4	.006	0.073	26
Vibrio cholerae O1	19	8.8	.002	0.046	17	9.1	.002	0.046	17
Adenovirus 40/41	18	6.0	.003	0.041	15	9.5	.002	0.044	16
Entamoeba histolytica	7	3.0	.15	0.013	5	10.3	.038	0.017	6

Abbreviations: AF, attributable fraction; ETEC, enterotoxigenic *Escherichia coli*; LT, heat-labile enterotoxin; OR, odds ratio; ST, heat-stable enterotoxin.

[a] OR: ratio of the odds of moderate-to-severe diarrhea when the putative pathogen is present to the odds when it is absent. OR >1 indicates a positive association.

[b] *P* value from logistic regression.

Table 3. Annual Attributable Moderate-to-Severe Diarrhea Cases in Population and Incidence per 100 Child-Years in First 2 Years of the Global Enteric Multicenter Study: The Gambia, Ages 0–11 Months (312 Cases, 398 Controls)

Pathogen	AF	Annual Attributable Cases[a]	Attributable Incidence Rate[b]
Rotavirus	0.211	135	2.3
Cryptosporidium	0.095	61	1.0

M_{SHC} = No. of eligible MSD cases at SHCs (observed) = 625; r = proportion of MSD cases seen at SHC = 0.487; M_{pop} = annual MSD cases in population = $M_{SHC}/(2r)$ = 642; N = No. of children in population = 5922.

Abbreviations: AF, attributable fraction; MSD, moderate-to-severe diarrhea; SHC, sentinel health center.

[a] Calculated as AF × $M_{SHC}/(2r)$.

[b] Per 100 child-years; calculated as 100 × AF × $M_{SHC}/(2rN)$.

approximately AF × $M_{SHC}/(3rN)$. N is estimated as the median of population estimates from several DSS rounds performed during the study. Table 3 illustrates these calculations for data from the first 2 years of the study on infants aged 0–11 months in The Gambia.

The variance of the incidence rate is approximated by Taylor series to first derivative terms (delta method). The variance of AF is estimated using a jackknife procedure [13], the variance of r as the variance of the probability of an event in a weighted Kaplan-Meier analysis, the variance of $M_{SHC}/3$ as the variance of the mean of 3 yearly totals of cases coming to an SHC, and the variance of N as the variance of the median of several observations from a normal distribution [22].

ASSOCIATIONS WITH OUTCOMES IN CASES OF MSD

Because the GEMS case/control study includes follow-up visits at approximately 60 days (range, 50–90 days), we have information on certain outcomes. Among these are death and, among cases and controls who survive to the follow-up visit, linear and ponderal growth. It is thus natural to investigate risk or protective factors for these outcomes in MSD cases. For a dichotomous outcome such as death, we use logistic regression modeling or, in order to use the actual follow-up times, Cox proportional hazards regression. For a continuous outcome we use linear regression analysis. This analysis is especially relevant for prioritizing the development of point-of-care diagnostics and therapeutic interventions. For a pathogen that is associated with a high case-fatality rate, it might also suggest a need for a prophylactic intervention, such as a vaccine.

ASSOCIATIONS WITH OUTCOMES IN THE POPULATION

Besides evaluating associations between outcomes and risk or protective factors in cases, we are interested in such associations in the general population. This type of analysis is particularly well suited for evaluating the need for preventive interventions, such as vaccines, that target specific pathogens or environmental conditions and would be widely applied in the population. To evaluate this type of association, we use a weighted analysis of cases and controls, which is described in detail by Sommerfelt et al in this supplement [23]. In this approach, weights are chosen so as to make the proportion of cases in the analysis approximately the same as in the population from which the cases were drawn.

DISCUSSION

Of the various statistical analyses in GEMS, it is the analysis of etiology that is the most important and demanding. The reason is that one of the driving rationales for initiating the GEMS was to be able, on the basis of the results, to prioritize the allocation of financial and other resources toward the implementation of existing interventions (such as vaccines and therapeutics) and to prioritize investments in research aimed at developing new interventions, based on the relative contributions of different pathogens to the overall burden of MSD in young children. From this perspective, one sees clearly the potential utility of the AF, defined as the proportion of MSD that would be eliminated if the target population were no longer exposed to a specific risk factor (such as a specific pathogen). AF allows us to estimate the number of MSD cases at one of our sites that can be attributed to a specific pathogen, adjusted for other pathogens that might also be present. Thus, we can distinguish between a pathogen that is responsible for a large number of cases and another pathogen that might be associated with MSD but for which the number of attributable cases is considerably smaller. This is crucial in allowing policy makers to set priorities for interventions. Further, it can help us to identify locations where a specific intervention might make a large impact and other locations where its impact might be relatively minor.

There are, of course, limitations of our study and analysis. The most important limitation regarding statistical analysis is probably in the estimation from HUAS-lite surveys of the proportion of MSD cases taken to an SHC, which we call “r.” This proportion is important in our estimation of total MSD burden, as well as the burden attributable to specific pathogens. The HUAS-lite surveys were based on random samples from the DSS population. However, for various reasons (eg, in Kolkata many MSD cases were taken to private healthcare providers) in most sites and age groups r is smaller than we had hoped it would be (<40% in all except 1 site). Thus, there is the potential for bias in the proportion of cases enrolled with a specific manifestation of MSD or in which a specific pathogen was isolated. A minor limitation is that our requirement that eligible cases should have had onset of diarrhea within the past 7 days could produce a slight underestimate of the true incidence of MSD. Data from the HUAS and HUAS-lite surveys, including estimates of r, will be presented in papers that are in preparation.

Various bacterial, viral, and protozoal pathogens can each cause MSD in children, and some set of these enteropathogens are collectively responsible for most of the MSD that occurs among young children in developing countries. These diarrheal pathogens are transmitted to susceptible children in 1 or more ways, depending on the pathogen, including via contaminated water or food, direct contact with fecally contaminated hands, flies acting as mechanical vectors, and contaminated fomites. There are 2 broad approaches to diminish MSD by active interventions. One approach aims to diminish transmission by instituting broad, cross-cutting water/sanitation/hygiene interventions to reduce the risk factors that result in fecally contaminated hands, food, and water and in allowing house flies to serve as mechanical vectors to carry enteric pathogens that can cause illness with small inocula (eg, *Shigella*). Examples of these interventions include household-based methods of treating water, refrigeration (to prevent pathogens in food and drink from growing to become potentially large inocula), washing hands with soap at critical points during the day (following defecation, prior to handling food, and before and after holding infants). Each of these interventions is estimated to diminish the incidence of diarrhea illness by 12%–25% [24–27], and each is presumed to be cross-cutting (ie, to diminish all enteric pathogens transmitted by a particular mechanism against which the intervention is directed).

The alternative strategy whereby the burden of MSD may be diminished, even without water, sanitation, and hygiene interventions that diminish the overall fecal burden in the environment, is to modify the immunologic status of the host from susceptible to immune by means of vaccination against specific pathogens. To pursue this strategy, one must first know the major agents responsible for MSD and their relative contribution to the overall MSD burden to prioritize what existing vaccines need to be implemented and what others need investments to be developed. It is in this context that the concept of AF is so potentially useful and important. While this might be straightforward in a study of diarrheal disease in an industrialized country setting, deciphering the data in a developing country project such as GEMS is daunting because approximately 85% of MSD cases can yield 1 or more enteropathogens, as can >70% of healthy controls, and a substantial proportion can yield multiple pathogens. Use of adjusted AF takes into account not only the prevalence of a pathogen of interest in controls as well as in cases, but also the presence of other pathogens besides the pathogen of interest, in both cases and controls.

Surprisingly, despite the fact that a large number of case/control studies have been carried out in developing countries to look for the predominant pathogens of MSD, the statistical analyses have only rarely utilized the concept of AF. In part this may have been due to lack of clear understanding of methods for adjustment for the presence of multiple pathogens [28]. However, the AF methodology, including the calculation of adjusted AFs, has undergone considerable development in recent decades. Several recent reviews have addressed subtleties in both the mathematical models and assumptions that underlie the use of AF [16, 28, 29]. One fundamental point to consider is that AF for an enteropathogen as a cause of MSD can be calculated either based on the distribution of exposure to the pathogen in the population [17] or the distribution of exposure in the cases [18]. In the GEMS analyses we employ the latter approach. We believe that calculation of AF, with adjustment for the presence of multiple pathogens among cases and controls, provides an appropriate approach for identifying the relative burden of diarrheal disease that could be eliminated through interventions against specific pathogens. We propose that this be adopted as a standard methodology (among others) for studies similar to GEMS, so that it will be possible to compare results of studies across time and geography, if other relevant case definition, selection, and laboratory methods are similar.

Notes

Acknowledgments. We are indebted to the following individuals who participated in 1 or more meetings of the GEMS Steering Committee on Biostatistical Issues and contributed their experience, wisdom, creativity, and advice: Lawrence Moulton, Janet Wittes, William Pan, and Peter Smith. We also thank Robert Black for valuable suggestions concerning GEMS statistical methodology.

Financial support. This work was supported by the Bill & Melinda Gates Foundation (grant number 38874).

Supplement sponsorship. This article was published as part of the supplement entitled "The Global Enteric Multicenter Study (GEMS)," sponsored by the Bill & Melinda Gates Foundation.

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Black RE, Cousens S, Johnson HL, et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. Lancet **2010**; 375:1969–87.
2. Black RE, Lopez de Romana G, Brown KH, Bravo N, Bazalar OG, Kanashiro HC. Incidence and etiology of infantile diarrhea and major

routes of transmission in Huascar, Peru. Am J Epidemiol **1989**; 129:785–99.

3. Albert MJ, Faruque AS, Faruque SM, Sack RB, Mahalanabis D. Case control study of enteropathogens associated with childhood diarrhea in Dhaka, Bangladesh. J Clin Microbiol **1999**; 37:3458–64.
4. Tin A, Mar MN, Kyi KK, et al. Epidemiology and aetiology of acute childhood diarrhoea in Burma: a rural community survey. Trans R Soc Trop Med Hyg **1989**; 83:827–30.
5. De MP, Brasseur D, Hemelhof W, Kalala T, Butzler JP, Vis HL. Enteropathogenic agents in children with diarrhoea in rural Zaire. Lancet **1983**; 1:516–8.
6. Gascon J, Vargas M, Schellenberg D, et al. Diarrhea in children under 5 years of age from Ifakara, Tanzania: a case-control study. J Clin Microbiol **2000**; 38:4459–62.
7. Echeverria P, Taylor DN, Lexsomboon U, et al. Case-control study of endemic diarrheal disease in Thai children. J Infect Dis **1989**; 159:543–8.
8. Valentiner-Branth P, Steinsland H, Fischer TK, et al. Cohort study of Guinean children: incidence, pathogenicity, conferred protection, and attributable risk for enteropathogens during the first 2 years of life. J Clin Microbiol **2003**; 41:4238–45.
9. Huilan S, Zhen LG, Mathan MM, et al. Etiology of acute diarrhoea among children in developing countries: a multicentre study in five countries. Bull World Health Organ **1991**; 69:549–55.
10. Levine MM. Enteric infections and the vaccines to counter them: future directions. Vaccine **2006**; 24:3865–73.
11. Levine MM, Kotloff KL, Nataro JP, Muhsen K. Impetus and rationale for the Global Enteric Multicenter Study (GEMS). Clin Infect Dis **2012**; 55(Suppl 4):S215–24.
12. Kotloff KL, Blackwelder WC, Nasrin D, et al. The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study. Clin Infect Dis **2012**; 55(Suppl 4):S232–45.
13. Korn EL, Graubard BI. Analysis of health surveys. New York: Wiley, **1999**:124–5.
14. Breslow NE, Day NE. Statistical methods in cancer research. Vol. I—The analysis of case-control studies. IARC Sci Publ **1980**; 202–5, 248–79.
15. Firth D. Bias reduction of maximum likelihood estimates. Biometrika **1993**; 80:27–38.
16. Hanley JA. A heuristic approach to the formulas for population attributable fraction. J Epidemiol Community Health **2001**; 55:508–14.
17. Levin ML. The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum **1953**; 9:531–41.
18. Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. Am J Epidemiol **1974**; 99:325–32.
19. Sobel J, Gomes TA, Ramos RT, et al. Pathogen-specific risk factors and protective factors for acute diarrheal illness in children aged 12–59 months in Sao Paulo, Brazil. Clin Infect Dis **2004**; 38:1545–51.
20. Rothman KJ, Greenland S. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, **2008**.
21. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol **1985**; 122:904–14.
22. Rider PR. Variance of the median of small samples from several populations. J Am Stat Assoc **1960**; 55:148–50.
23. Sommerfelt H, Steinsland H, van der Merwe L, et al. Case/control studies with follow-up—constructing the source population to estimate effects of risk factors on development, disease and survival. Clin Infect Dis **2012**; 55(Suppl 4):S262–70.
24. Clasen T, Schmidt WP, Rabie T, Roberts I, Cairncross S. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. BMJ **2007**; 334:782.
25. Clasen TF, Bostoen K, Schmidt WP, et al. Interventions to improve disposal of human excreta for preventing diarrhoea. Cochrane Database Syst Rev **2010**; CD007180.
26. Curtis V, Cairncross S. Effect of washing hands with soap on diarrhoea risk in the community: a systematic review. Lancet Infect Dis **2003**; 3:275–81.
27. Curtis V, Schmidt W, Luby S, Florez R, Toure O, Biran A. Hygiene: new hopes, new horizons. Lancet Infect Dis **2011**; 11:312–21.
28. Benichou J. A review of adjusted estimators of attributable risk. Stat Methods Med Res **2001**; 10:195–216.
29. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health **1998**; 88:15–9.