American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2012.

Vol. 176, No. 11
DOI: 10.1093/aje/kws183
Advance Access publication:
November 8, 2012

Practice of Epidemiology

Revealing the Complexity of Health Determinants in Resource-poor Settings

Fraser I. Lewis* and Benjamin J. J. McCormick

* Correspondence to Dr. Fraser I. Lewis, Section of Epidemiology, Vetsuisse Faculty, University of Zürich, Winterthurerstrasse 270, CH-8057 Zürich, Switzerland (e-mail: fraseriain.lewis@uzh.ch).

Initially submitted November 28, 2011; accepted for publication March 19, 2012.

An epidemiologic systems analysis of diarrhea in children in Pakistan is presented. Application of additive Bayesian network modeling to 2005–2006 data from the Pakistan Social and Living Standards Measurement Survey reveals the complexity of child diarrhea as a disease system. The key distinction between standard analytical approaches, such as multivariable regression, and Bayesian network analyses is that the latter attempt to not only identify statistically associated variables but also, additionally and empirically, separate these into those directly and indirectly dependent upon the outcome variable. Such discrimination is vastly more ambitious but has the potential to reveal far more about key features of complex disease systems. Additive Bayesian network analyses across 41 variables from the Pakistan Social and Living Standards Measurement Survey identified 182 direct dependencies but with only 3 variables: 1) access to a dry pit latrine (protective; odds ratio = 0.67); 2) access to an atypical water source (protective; odds ratio = 0.49); and 3) no formal garbage collection (unprotective; odds ratio = 1.32), supported as directly dependent with the presence of diarrhea. All but 2 of the remaining variables were also, in turn, directly or indirectly dependent upon these 3 key variables. These results are contrasted with the use of a standard approach (multivariable regression).

Bayesian network; diarrhea; epidemiologic determinants; graphical model; socioeconomic factors

Abbreviations: AIC, Akaike Information Criterion; DAG, directed acyclic graph; PSLSM, Pakistan Social and Living Standards Measurement.

Presented here is an epidemiologic systems approach for identifying potential determinants of diarrhea in children under 5 years using data from the Pakistan Social and Living Standards Measurement (PSLSM) Survey. Childhood diarrhea is the second biggest cause of worldwide mortality in children under 5 years of age (1, 2), and health surveys targeting this disease are common (3–6). Although such study designs are not without issue, potentially suffering from data quality and reliability concerns (7), they are widely used as low cost methods of data collection in developing countries.

A major challenge when analyzing data from surveys is that they are typically exploratory in nature, where the precise etiology of health outcomes is not known, and information on a large number of variables is often collected, not all of which are necessarily complete or relevant. It is also typical that many variables of potential interest are interrelated, both to each other and also to health outcomes. Such data may be conceptualized as describing an epidemiologic system (8–11), that is, a collection of mutually interdependent variables some or all of which can predict or affect the health outcomes of interest.

Additive Bayesian networks are introduced as a methodology for identifying statistical dependencies in complex disease systems by using observational data. Ultimately, what is desired in many epidemiologic analyses is the identification of causal pathways (12–14), which can be extremely challenging in systems such as diarrheal disease where many high level casual factors have been postulated (15), and the identification of statistical dependencies can be invaluable for informing such analyses.

The key distinction between standard multivariable regression analyses and Bayesian network-type analyses is that multivariable regressions seek to identify covariates associated with some outcome variable, for example, the presence of diarrhea. Bayesian network analyses go much further and attempt to not only identify associated variables but also, additionally and crucially, empirically, separate these into those directly and indirectly dependent upon the outcome variable. The latter is vastly more ambitious but has the potential to reveal far more about key features of complex disease systems than existing commonly used approaches. This is the central message of the work presented: Additive Bayesian network analyses are superior to standard approaches for inferring statistical dependencies from complex observational data.

IDENTIFYING STATISTICAL DEPENDENCIES BY USING MULTIVARIABLE REGRESSION

When exploratory analyses of data comprising many variables are performed, it is common to utilize some form of multivariable regression in which a variable selection process is then used, the goal being to search for variables that are statistically significantly associated with, for example, an outcome variable such as disease presence. Stepwise regression searches are widely used (16–19) despite being viewed rather negatively in the epidemiologic and biostatistical literature (20–22). Such automated searches are arguably overused or, rather, the results from such analyses are too often presented without sufficient additional checks to ensure the robustness of associations against overfitting (23).

In rapidly developing and increasingly data-rich fields such as genetic epidemiology, computational biology, and bioinformatics, automation in statistical modeling is standard and, indeed, arguably essential when faced with exploring observations from large numbers of potentially interdependent variables. That automated searches tend to overfit is well known. There are, however, well-established techniques for addressing this; 2 of the most commonly utilized are model averaging (24, 25) and parametric bootstrapping (26), both of which are explored in the later case study analyses.

IDENTIFYING STATISTICAL DEPENDENCIES BY USING BAYESIAN NETWORKS

Bayesian network analysis is a form of statistical modeling that derives, from empirical data, a graphical network describing the dependency structure between variables, where this is formally depicted as a directed acyclic graph (DAG). Bayesian networks are widely used in areas such as systems biology (27–29), human immunodeficiency virus (HIV) and influenza research (30–33), and also analyses of complex disease systems (34–37). The origins of Bayesian network modeling lie within the machine-learning and data-mining literature (27, 38) with an accessible nontechnical introduction (28).

In multivariable regression analyses, the goal is to identify statistically significant associations between an outcome variable and one or more covariates. Here, "association" denotes that the variables are statistically dependent; it says nothing on whether the variables are directly or indirectly dependent. To borrow an example from Hernán et al. (14), in multivariable regression analyses with "lung cancer" as the outcome variable and "smoking" and "yellow fingers" as covariates, it may then be expected that one or both of these covariates would be identified as statistically significantly associated with lung cancer. In contrast, in a Bayesian network analysis, it would be expected that smoking and lung cancer be identified as directly dependent, yellow fingers and smoking as directly dependent, but that yellow fingers and lung cancer not be identified as directly dependent. In terms of a DAG, this would describe a model with 2 arcs (one between lung cancer and smoking and a second between yellow fingers and smoking) but with no arc between yellow fingers and lung cancer. Note, we have *not* specified the direction of these arcs. In a Bayesian network analysis, each DAG is formally a factorization of the joint probability distribution of the observed data and, because of likelihood equivalence, it is the presence of arcs between variables and not their direction that is the notable feature.

Consider the joint probability of variables X and Y, $P(X, Y)$. Theory gives $P(X, Y) = P(X|Y)\ P(Y)$ and $P(X, Y) = P(Y|X)\ P(X)$, where the former can be depicted as a DAG with one arc, from Y to X, and the latter with one arc, from X to Y. The practical implication of this is that, by using observed data alone, it is not possible to statistically discriminate between different DAGs from within the same likelihood equivalence class, as these are probabilistically identical. However, determining likelihood equivalence between DAGs is extremely difficult in all but the simplest cases; refer to Web Appendix 1, the first of 13 Web appendices available on the *Journal*'s website (http://aje.oxfordjournals.org/), for more details. Because of these complications, it is typical to ignore arc direction in Bayesian network analyses (30–33, 39), although notable exceptions are analyses of longitudinal data where dynamic Bayesian networks may be utilized (40). Using prior belief to impose explicit arc direction may be of some value in analyses that attempt to combine statistical dependency with causality, and this is returned to later.

MATERIALS AND METHODS

Case study data

The PSLSM Survey is a biennial survey of a large number of social, environmental, and economic indicators, motivated and directed in part by efforts to meet the United Nations Millennium Development Goals. The survey is conducted at the household level, sampling from the majority (approximately 97%) of the population across Pakistan in the data analyzed here from 2005 to 2006. According to the Federal Bureau of Statistics, Islamabad, Pakistan, 15,453 households were surveyed comprising 110,909 individuals, of whom 18,202 were under 5 years of age. The survey includes around 250 questions, from which a broad subset (i.e., 40 questions from the questionnaire) was included in the following analyses based on potential relevance to childhood diarrheal disease determined from their inclusion in previous studies (refer to Web Appendices 2 and 3 for variable descriptions and details of previous

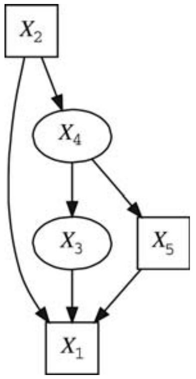

Figure 1. Example of an additive Bayesian network model comprising 3 binary random variables (X_1, X_2, X_5) and 2 continuous (Gaussian) variables (X_3, X_4). The model for each node is a generalized linear regression with identity or logit link function as appropriate. Let π_i for $i = 1, 2, 5$ denote the probability of observing a success: $P(X_i = 1) = 1 - P(X_i = 0)$ and μ_i denote the mean of random variable X_i for $i = 3, 4$. X_2 is independent of the other variables with $\log\{\pi_2/(1 - \pi_2)\} = \beta_{2,0}$; X_4 is conditionally dependent upon X_2 with $\mu_4 = \beta_{4,0} + \beta_{4,1} X_2$; X_5 is conditionally dependent upon X_4 with $\log\{\pi_5/(1 - \pi_5)\} = \beta_{5,0} + \beta_{5,1} X_4$; X_3 is conditionally dependent upon X_4 with $\mu_3 = \beta_{3,0} + \beta_{3,1} X_4$; and X_1 is conditionally dependent upon X_2, X_3, X_5 with $\log\{\pi_1/(1 - \pi_1)\} = \beta_{1,0} + \beta_{1,1} X_2 + \beta_{1,2} X_3 + \beta_{1,3} X_5$.

studies). Diarrheal presence was taken from a binary question of whether children under 5 years of age in the household had experienced diarrheal symptoms within the preceding 30 days.

Bayesian network modeling formulation

Standard multivariable statistical models—linear or generalized linear models or their variants—are additive, that is, they describe the mean value of some response variable conditional on a given covariate pattern, as an additive contribution from each covariate. In contrast, Bayesian network models for categorical data, the most commonly utilized form of Bayesian network, use a parameterization where each and every covariate pattern is modeled by using *independent* sets of parameters; that is, the parameters cannot be interpreted as main effects or interaction effects (Web Appendix 4). This formulation may also be far from parsimonious (41) and does not provide any ready interpretation of the model parameters. The standard formulation of Bayesian networks (27) does facilitate conjugacy; that is, all parameter estimates in a Bayesian network can be computed analytically for the 3 usual types of Bayesian network (categorical, Gaussian, and a special variant of mixed categorical and Gaussian variables (42)).

At the cost of a loss of conjugacy, it is possible to formulate Bayesian network models that are direct analogs of standard multivariable linear and generalized linear models, where each variable in the data is modeled by an additive multivariable regression model, with an appropriate link function (e.g., a logit) if required (Figure 1). As in classical Bayesian network models, additive Bayesian networks are described by a DAG. The price for this considerably greater model flexibility is that the goodness of fit and model parameters now must be estimated numerically rather than analytically. Laplace approximations (43) are used here.

Bayesian inference requires prior distributions, and in a Bayesian network there are 2 possible types of priors: priors on the model parameters and priors on the DAG. In terms of parameter priors, the approach utilized here assumes uninformative Gaussian priors with zero mean and large variance for each of the regression parameters across all parts of the model, as well as diffuse gamma priors for the precision parameter in Gaussian nodes of the model. In terms of structural priors, it is currently assumed that each DAG structure is equally plausible in the absence of any data. Imposing prior causal knowledge onto network structures, for example, by imposing conditions on arc direction, is returned to later.

Model selection in statistical analyses

Statistical model selection comprises 3 parts, denoted "A" (choosing a general form of model); "B" (deciding the scope of the model search space and how to cross it); and "C" (deciding how to summarize the results from "B").

For the PSLSM case study, the first analysis presented comprises a conventional stepwise regression search. Therefore, "A" is standard multivariable logistic regression. "B," a stepwise search, forwards from a null model and backwards from a full model, with comparisons performed within a maximum likelihood framework, as that is what is provided in common statistical software and is most usual in practice, and with the Akaike Information Criterion (AIC) as the goodness-of-fit metric. For "C," the single best model found in "B" was then subjected to parametric bootstrapping to identify any issues of overfitting (Web Appendix 5).

In the Bayesian network analyses, for "A" an additive form of Bayesian network is used. For "B," the 2 most widely utilized "structure discovery" approaches are used. First, the local search heuristic given by Heckerman et al. (27) is analogous to the usual stepwise search in multivariable regression. Second, there is a search over node orderings rather than DAG structures. Order-based approaches were introduced by Friedman and Koller (44) and then substantially extended by Koivisto and Sood (45). The motivation behind local heuristics (including stepwise searches in multivariable regression) is that they will identify high-scoring, well-fitting models when it is not computationally feasible to identify the very best model with any certainty. The second approach for searching for optimal Bayesian network models is to collapse DAGs over node orderings; a node ordering is simply a list of the nodes, for example, as indices 1 through *n*, where a given DAG structure is consistent with an ordering if, and only if, the parents of each node precede their child node in this list. Orderings can be thought of as groups of DAG structures (those structures which are consistent with that particular ordering), and note that each DAG may be consistent with more than one ordering; for example, the empty DAG (no arcs) is consistent with every possible ordering. The basic idea is that, by searching across orders, the dimension of the search space is vastly reduced from $\approx n!2^{\binom{n}{2}}$ unique DAGs down to $n!$ unique orders (45), although the latter may still be computationally impractical. The price for this reduction in size of search space is that searching across orders is biased relative to searching across DAGs.

Finally, consider step "C," how to summarize the results of Bayesian network model searches. Two options are either to construct some form of summary or "average" model by pooling across heuristic search results or else to select a single "best" model. A popular approach for the former is to construct a majority consensus network that builds a DAG comprising all those arcs present in at least a majority (>50%) of the DAGs identified by using heuristic searches (30, 35). Because of likelihood equivalence, it is common to collapse over arc direction to avoid missing important structural features. For example, if arc $X \rightarrow Y$ appears in 50% of heuristic results, and $Y \rightarrow X$ appears in the other 50%, then even although this direct dependency between X and Y features in every search result it will never appear in a (directed) majority consensus network. For this reason, collapsing over arc direction when presenting results of Bayesian network analyses is common (30, 39). The purpose of summarizing over many DAGs is to address concerns of overfitting, and it is directly analogous to the ubiquitous use of majority consensus trees in phylogenetics (46). The second option in "C" is to choose a single best model, with the most obvious concern being overfitting, and parametric bootstrapping is not generally computationally feasible here. An accepted approach for choosing a single best model is to use the exact order-based method (45) that finds the globally most probable posterior DAG.

All modeling results were carried out in R (47) by using an R library called "abn" developed by the authors for the purpose of analyzing epidemiologic data. This software is freely available for download from CRAN (http://cran.r-project.org/).

RESULTS

Multivariable regression

Table 1 shows the variables in the optimal model from the stepwise multivariable regression analysis (Web Appendix 6). There are 12 covariates, many of which have low *P* values. To identify spurious covariates arising from overfitting, we used parametric bootstrapping to generate 10,000 data sets from the optimal model. The parametric bootstrapping results provide convincing evidence that each of the 12 identified covariates is robust in terms of being statistically associated with the presence of diarrhea (Web Appendix 6). This regression model can be represented as a DAG where each of the explanatory variables is a node with an arc directed toward the node for the response variable (Figure 2).

Additive Bayesian network

Three different, although related, sets of results are presented, all with the same goal of identifying those variables directly dependent upon the presence of diarrhea.

Table 1. Results of Stepwise Regression Search and Additive Bayesian Network Analyses (With All 3 Variables Directly Dependent Upon Diarrhea), Pakistan Social and Living Standards Measurement Survey, 2005–2006

Indicator	Odds Ratio	*P* Value[a]	Bayesian Odds Ratio
Child's age	1.05	0.012	
No. of rooms	0.95	0.007	
Sex, male	1.14	0.012	
Dwelling type, part of compound	0.70	0.020	
Drinking water source			
Piped	0.85	0.011	
Canal/river/stream	0.65	0.0087	
Spring	0.76	0.13	
Other	0.46	0.0016	0.49
Type of toilet			
Flush, connected to open drain	0.84	0.046	
Dry pit latrine	0.66	<0.0001	0.67
Connection to sewerage, yes, covered drains	0.72	0.064	
Organizer of garbage collection from house, no formal system	1.23	0.0048	1.32

[a] *P* values are from type III chi-squared tests. The odds ratios are marginal; for example, for dry pit latrine, the odds ratio = 0.66, which is relative to not having a dry pit latrine (ignoring all other covariates). For the continuous variables (age and number of rooms), the odds ratios are in respect of a 1-unit increase.

Heuristic search across 13 variables. The standard local heuristic search (27) was applied to the subset of 12 covariates identified in the optimal multiple regression model. A (directed) majority consensus additive Bayesian network model was constructed by pooling results across 20,000 heuristic searches; it was sufficient for robust results (Web Appendix 7). This summary network (Web Appendix 7) identifies "dry pit latrine" and access to an atypical "other water source" as directly dependent upon diarrhea. Figure 3 shows an undirected majority consensus additive Bayesian network constructed from the same 20,000 heuristic searches. This now additionally has "no formal garbage collection" as directly dependent upon diarrhea, although its structural support is relatively weaker in terms of how often it was chosen for inclusion in each locally optimal DAG (Web Appendix 7), compared with the other 2 variables. Posterior density estimates for the 3 variables directly dependent upon diarrhea can be found in Web Appendix 8. Note that the odds ratio estimates in Table 1 and posterior densities will be identical in any additive Bayesian network that has only these 3 variables with arcs to diarrhea.

Exact search for most probable DAG across 13 variables. The most probable posterior DAG was identified by using the exact method (45), again on the reduced set of 13

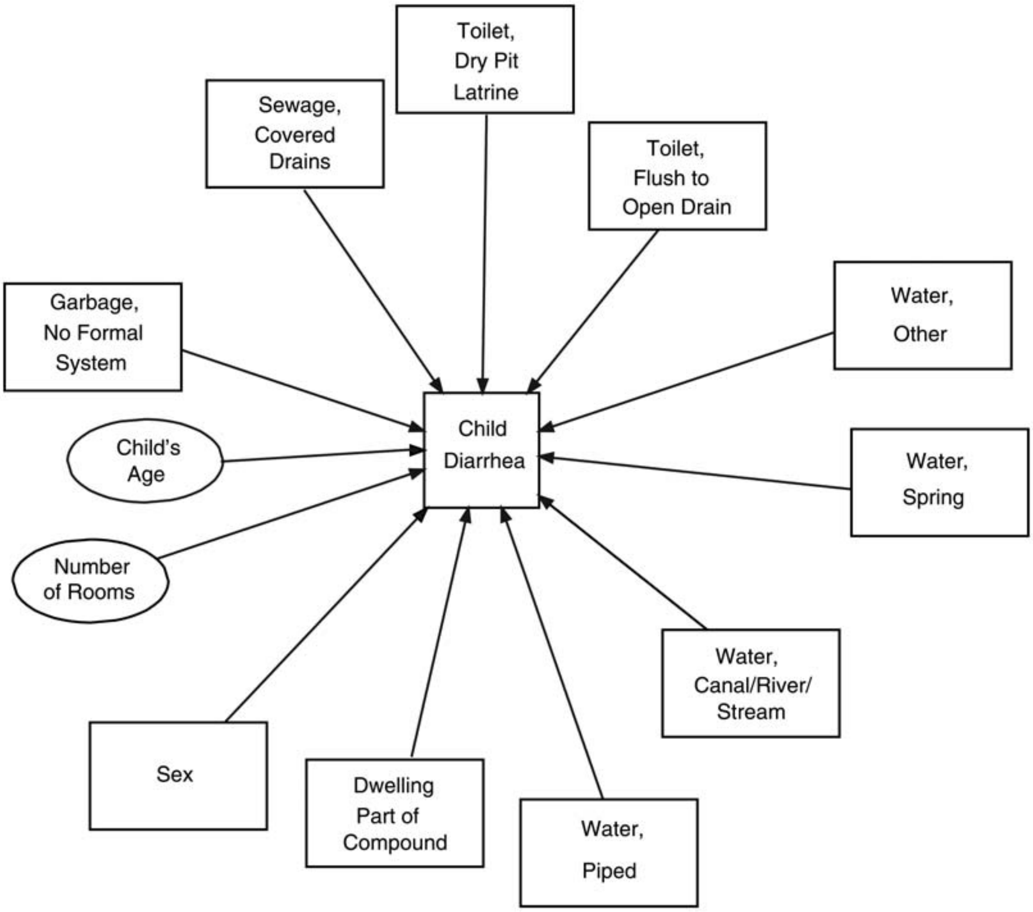

Figure 2. Final model in stepwise (forwards and backwards) multivariable regression search depicted as a directed acyclic graph. Parametric bootstrapping statistically supports all 12 covariates in this model. Ovals are continuous variables; squares, discrete.

variables (specific details are given in Web Appendix 9). This exact search identifies a maximal additive Bayesian network that has "dry pit latrine" and access to an atypical "other water source" as directly dependent upon diarrhea but not "no formal garbage collection." The goodness of fit of this model is −105,028.4 (log marginal likelihood), and it has 32 arcs in total. During the previous 20,000 search heuristics across DAGs, a number of models with improved goodness of fit were identified (e.g., −105,025.9 with 34 arcs), which demonstrates the bias toward parsimony in order-based searches, as the fewer arcs a DAG has, the more orders it will be consistent with.

Heuristic search across all 41 variables. The standard heuristic search (27) over 41 variables was not computationally feasible, nor was the exact order-based method. By necessity, an ad-hoc approach was instead utilized by adding several constraints to the standard heuristic search (Web Appendix 10). A majority consensus additive Bayesian network model was constructed by pooling results across ≈500,000 separate searches. This model identifies the same 3 variables as directly dependent upon diarrhea as in the undirected majority consensus network with 13 variables. The additive Bayesian network model supports 182 interdependencies among the 41 variables, where 179 are dependencies indirectly related to the presence of diarrhea, that is, between variables which can potentially affect disease presence, but only through their associations with other variables (refer to Web Appendix 11 for a detailed description of the model).

DISCUSSION

The objective of the analyses presented was to identify potential determinants of the presence of diarrhea and, in particular, to contrast results by using standard multivariable regression with those of an epidemiologic systems approach utilizing additive Bayesian networks.

Comparison of methods

Table 1, along with Figures 2 and 3, shows clearly that the 2 approaches provide very different, though overlapping, results. The additive Bayesian network results suggest that most—9 out of 12—of the covariates identified in the

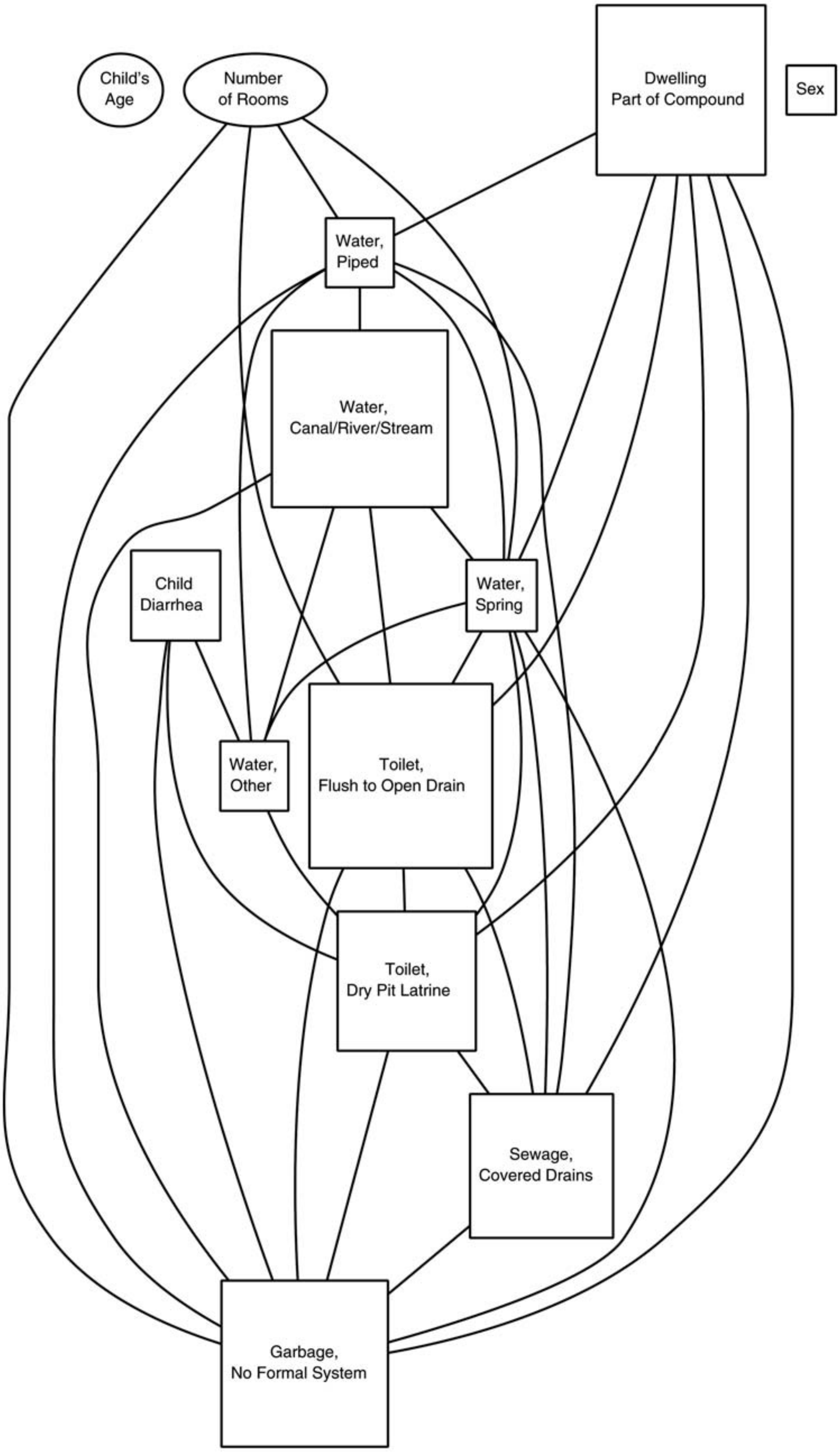

Figure 3. Undirected majority consensus additive Bayesian network model constructed by pooling results across 20,000 heuristic searches. Only 3 variables (no formal garbage collection, access to a dry pit latrine, and access to an atypical other water source) are supported as directly dependent upon the presence of diarrhea.

multivariable regression analyses, while associated with the presence of diarrhea, are only indirectly rather than directly related to this outcome.

Additive Bayesian network models are simply multivariate extensions of standard multivariable regression and nothing more. The single key conceptual difference is that additive Bayesian network models are multidimensional and consider all associations among all variables simultaneously. It is therefore intuitively reasonable to expect both approaches—additive Bayesian network and multivariable regression—to both identify in common those variables with the strongest degree of statistical support (and irrespective of whether different goodness-of-fit metrics or inferential framework is used, e.g., AIC or marginal likelihood, Bayesian or non-Bayesian). This is exactly what was found in the analyses presented (Table 1): The 3 variables with lowest *P* values in the multivariable regression are also those supported as directly dependent upon diarrhea in the additive Bayesian network results.

In the multivariable regression analyses, "number of rooms" was identified as associated with diarrhea and, with $P = 0.007$ sufficiently low to be typically considered strong statistical evidence, this was further supported through the bootstrapping results. This variable has many direct dependencies in the additive Bayesian network (Figure 3; Web Appendix 11) but only with variables other than diarrhea. Biologically, "number of rooms" cannot be directly dependent upon diarrhea as, although it is likely to be related to the living and environmental conditions in a household and the additive Bayesian network model provides empirical evidence of this, intuitively, it cannot contribute directly to diarrheal infection. This suggests that "number of rooms" has been identified in the multivariable regression model as a result of association induced with diarrhea through a network of interdependencies across the disease system. This highlights the difficulty of interpretation in the traditional multivariable regression model where it is possible for variables with low *P* values to be identified as associated with disease but that are likely to be only indirectly related to the outcome variable.

Biologic interpretation of results

The decreased risk of diarrhea from dry pit latrines suggests that infectious enteric pathogens are efficiently removed from fecal-oral transmission cycles. Using a univariate regression model, we found that the absence of a toilet in a household was a substantially greater risk factor for diarrhea (odds ratio = 5.7, 95% confidence interval: 5.07, 6.50) than the presence of any type of toilet. The absence of arcs connecting "no toilet in the household" to childhood diarrhea suggests, however, that those houses lacking toilets have a network of confounding factors that modify the risk of enteric infection. For example, a fuller description of the disease system (Web Appendix 11) shows that the absence of a toilet is dependent upon other descriptions of the household, such as the absence of an electrical connection (as an indicator of socioeconomic status), with certain water sources, especially those that do not require infrastructure, such as ponds, streams, and springs, and also no formal garbage collection. What might be deduced from this is that the lack of a toilet is an index of lower socioeconomic status and therefore living conditions. Socioeconomic status is itself associated with educational levels and certain behaviors, for example, unhygienic practices that increase the risk of diarrhea.

Using "other" sources of water is comparatively unusual in the data (2.1% of households). Given the breadth of options covered across the other 5 categories of water sources (Web Appendix 2), "other water source" is somewhat vague but includes buying water from a local seller and collecting water from a tanker-supplied public standpipe. Water that is further removed from either natural water courses or pipe/well systems may be less susceptible to leakage between sewerage systems and the water table and thereby at risk of contamination with water-borne pathogens. Examination of the 41-variable additive Bayesian network model shows connections from each of the water sources to at least the alternative sewerage connections or to the type of toilet, demonstrating the tight interlinkage among these 3 risk categories in determining the exposure of children to enteric pathogens. The relatively large uncertainty in the log odds estimate between "other" water sources and diarrhea (Web Appendix 8) is likely due to both the rarity of this water source in the data and also the ambiguous definition of this variable. Web Appendix 3 contains a list of additional references and a summary of previous variables associated with diarrhea, including age, which is also briefly discussed.

In the language of Hernán et al. (14), the components of a disease system are the disease outcome, an exposure (that directly results in disease), and a series of confounding variables that act on either the exposure or both exposure and disease outcome. Survey data, as in the PSLSM Survey, do not necessarily contain the sort of proximal exposure that results in infection (contact with enteric pathogens), so much as a collection of variables that might be considered common to both exposure (infection) and disease (pathogenesis). It is, therefore, arguably less useful to discuss causality with respect to such an exploratory model; however, what is both possible and useful is the partitioning of factors into those that are directly or indirectly dependent upon disease outcome. To illustrate this point, consider the type of toilet that is present in a house: There are several types of toilet in the PSLSM data, but these are all variants of the same underlying theme of how the household disposes of human excreta. It is not practical to question whether a flushing toilet is causal of diarrheal disease so much as the relative risk of disease given the different types of toilet present; the actual exposure is still the contact with enteric pathogens, which can then be stratified (assuming such data exist) by type of toilet. In this context, "toilet" is a confounder (according to Hernán et al.) that can be used to stratify the more proximal exposure.

Following through a series of DAGs as subsets of the more complete system may offer a closer parallel to causal models that are based on expert opinion (Web Appendix 12). Assuming that the absence of formal garbage collection is the exposure leading to diarrheal disease, 4 alternative routes involving 4 additional confounding factors were compared. The reason for selecting "no formal garbage collection" was because this variable was identified as dependent upon diarrhea in all but the most probable DAG. There was little change in the odds of diarrhea when no formal garbage collection was combined with different combinations of other confounders in the alternative DAGs. The interesting exception is the combination of "other" sources of water and no formal garbage collection. With the more defined water sources, the odds of diarrhea in houses lacking formal garbage collection are approximately 1.2; however, the odds rise to 1.3 when water sources are "other." The explanation for this may be the results of lack of running water to remove the buildup of refuse (and potentially excreta) when garbage is not regularly removed. The accumulation of refuse that has no formal disposal and its removal that is presumably irregular provide a permissive breeding ground for enteric pathogens. This hypothesis requires more targetted studies; however, neither garbage nor water is of itself the "cause" of diarrheal disease despite both being likely sources of enteric pathogens.

Introducing prior causal knowledge

Bayesian network modeling is typically concerned with automated structure discovery (searching for a DAG which best describes the statistical relations in observational data), in this case, the PSLSM. In causal inference, on the other hand, the focus is typically on testing whether a given set of assumptions is sufficient for quantifying causal effects from observational data, conditional on a causal diagram that encodes all the relevant domain-specific assumptions (12, 14). The former is objective (empirically derived DAGs) but lacks a causal component, while the latter provides causal insight but whose weakness is the potentially subjective justification of the causal diagram. An obvious question is, therefore, how can prior causal knowledge be integrated into automated structure discovery?

A very rudimentary approach that repeats the previous heuristic additive Bayesian network analyses (across 13 variables) by introducing some simple commonsense prior causal constraints is given in Web Appendix 13. This modeling prohibits arcs emanating from the diarrheal node and prevents "number of rooms" being directly dependent upon diarrhea. This simple informative structural prior now gives an undirected majority consensus DAG with the same 3 variables as previously identified as directly dependent upon diarrhea, but where support for an arc connecting "no formal garbage collection" to diarrhea is now 100% (appears in all 20,000 searches), whereas using the previous uninformative structural prior, this was only 58% and it also didn't appear in the previous directed majority consensus network (Web Appendix 7). The use of directional constraints alters the model search space (as the data can discriminate between DAGs where the arcs in these constraints are reversed provided these are in different equivalence classes) (Web Appendix 1) and so may provide different results. The key question is whether imposing such directional constraints/informative prior—motivated by causal considerations—is conceptually reasonable. This is an open question.

An alternative approach—and one that seems preferable given the complications of likelihood equivalence—is suggested by Heckerman et al. (27, p. 224). Rather than use an informative structural prior, it is proposed to append onto the observed data additional and likely highly incomplete synthetic observations that reflect causal beliefs. The structure learning process is then applied to all of the data as usual, except with the additional functionality necessary to marginalize over missing data (48). This is an elegant approach, but its feasibility and practicality with respect to additive Bayesian network modeling are an open question and an exciting area of future work.

ACKNOWLEDGMENTS

Author affiliations: Section of Epidemiology, Vetsuisse Faculty, University of Zürich, Zürich, Switzerland (Fraser I. Lewis); and Fogarty International Center, National Institutes of Health, Bethesda, Maryland (Benjamin J. J. McCormick).

This research was partly funded (B. J. J. M.) by the Bill and Melinda Gates Foundation and the Fogarty International Center of the National Institutes of Health as part of the MAL-ED (pronounced "mal a dee") project network (i.e., the Interactions of Malnutrition and Enteric Infections: Consequences for Child Health and Development).

The authors thank Safdar Parvez, Asian Development Bank, based in Manila, Philippines, for assisting in gaining access to the PSLSM data.

Conflict of interest: none declared.

REFERENCES

1. Black RE, Cousens S, Johnson HL, et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. Child Health Epidemiology Reference Group of WHO and UNICEF. *Lancet*. 2010;375(9730):1969–1987.
2. Bryce J, Boschi-Pinto C, Shibuya K, et al. WHO estimates of the causes of death in children. WHO Child Health Epidemiology Reference Group. *Lancet*. 2005;365(9465): 1147–1152.
3. Ahiadeke C. Breast-feeding, diarrhoea and sanitation as components of infant and child health: a study of large scale survey data from Ghana and Nigeria. *J Biosoc Sci*. 2000; 32(1):47–61.
4. de Souza AC, Peterson KE, Cufino E, et al.. Relationship between health services, socioeconomic variables and inadequate weight gain among Brazilian children. *Bull World Health Organ*. 1999;77(11):895–905.
5. Hatt LE, Waters HR. Determinants of child morbidity in Latin America: a pooled analysis of interactions between parental education and economic status. *Soc Sci Med*. 2006;62(2):375–386.
6. Jones LL, Griffiths PL, Adair LS, et al. A comparison of the socio-economic determinants of growth retardation in South African and Filipino infants. *Public Health Nutr*. 2008; 11(12):1220–1228.
7. Boerma JT, Black RE, Sommerfelt AE, et al. Accuracy and completeness of mothers' recall of diarrhoea occurrence in pre-school children in demographic and health surveys. *Int J Epidemiol*. 1991;20(4):1073–1080.

8. Galea S, Riddle M, Kaplan GA. Causal thinking and complex system approaches in epidemiology. *Int J Epidemiol*. 2010; 39(1):97–106.
9. Fenner L, Egger M, Gagneux S. Annie Darwin's death, the evolution of tuberculosis and the need for systems epidemiology. *Int J Epidemiol*. 2009;38(6):1425–1428.
10. Lusis AJ, Attie AD, Reue K. Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet*. 2008; 9(11):819–830.
11. Diez Roux AV. Integrating social and biologic factors in health research: a systems view. *Ann Epidemiol*. 2007;17(7):569–574.
12. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669–688.
13. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578–586.
14. Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155(2):176–184.
15. Eisenberg JN, Desai MA, Levy K, et al. Environmental determinants of infectious disease: a framework for tracking causal links and guiding public health research. *Environ Health Perspect*. 2007;115(8):1216–1223.
16. Howard G, McClure LA, Moy CS, et al. Imputation of incident events in longitudinal cohort studies. *Am J Epidemiol*. 2011;174(6):718–726.
17. Johnson HL, Liu L, Fischer-Walker C, et al. Estimating the distribution of causes of death among children age 1–59 months in high-mortality countries with incomplete death certification. *Int J Epidemiol*. 2010;39(4):1103–1114.
18. Shultz A, Omollo JO, Burke H, et al. Cholera outbreak in Kenyan refugee camp: risk factors for illness and importance of sanitation. *Am J Trop Med Hyg*. 2009;80(4): 640–645.
19. Phillips G, Lopman B, Rodrigues LC, et al. Asymptomatic rotavirus infections in England: prevalence, characteristics, and risk factors. *Am J Epidemiol*. 2010;171(9):1023–1030.
20. Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001;16(3):199–215.
21. Cox DR, Efron B, Hoadley B, et al. Statistical modeling: the two cultures—comments and rejoinders. *Stat Sci*. 2001; 16(3):216–231.
22. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *Am J Epidemiol*. 2011;174(5):613–620.
23. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*. 2004;66(3):411–421.
24. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Stat Assoc*. 1997;92(437):179–191.
25. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol*. 2004;53(5):793–808.
26. Friedman N, Goldszmidt M, Wyner A. Data analysis with Bayesian networks: a bootstrap approach. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann; 1999:206–215.
27. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks—the combination of knowledge and statistical data. *Mach Learn*. 1995;20(3):197–243.
28. Needham CJ, Bradford JR, Bulpitt AJ, et al. A primer on learning in Bayesian networks for computational biology. *PLoS Comput Biol*. 2007;3(8):pe129. (doi:10.1371/journal.pcbi.0030129).
29. Djebbari A, Quackenbush J. Seeded Bayesian networks: constructing genetic networks from microarray data. *BMC Syst Biol*. 2008;2:p57. (doi:10.1186/1752-0509-2-57).
30. Poon AF, Lewis FI, Pond SL, et al. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. *PLoS Comput Biol*. 2007;3(1):pe11. (doi:10.1371/journal.pcbi.0030011).
31. Poon AF, Lewis FI, Pond SL, et al. An evolutionary-network model reveals stratified interactions in the V_3 loop of the HIV_{-1} envelope. *PLoS Comput Biol*. 2007;3(11):pe231. (doi:10.1371/journal.pcbi.0030231).
32. Poon AF, Lewis FI, Frost SD, et al. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics*. 2008;24(17):1949–1950.
33. Lycett SJ, Ward MJ, Lewis FI, et al. Detection of mammalian virulence determinants in highly pathogenic avian influenza H5N1 viruses: multivariate analysis of published data. *J Virol*. 2009;83(19):9901–9910.
34. Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*. 2003;302(5644):449–453.
35. Hodges AP, Dai D, Xiang Z, et al. Bayesian network expansion identifies new ROS and biofilm regulators. *PLoS One*. 2010;5(3):pe9513. (doi:10.1371/journal.pone.0009513).
36. Dojer N, Gambin A, Mizera A, et al. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*. 2006;7:p249. (doi:10.1186/1471-2105-7-249).
37. Lewis FI, Brülisauer F, Gunn GJ. Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data. *Prev Vet Med*. 2011;100(2): 109–115.
38. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn*. 1992;9(4):309–347.
39. Milns I, Beale CM, Smith VA. Revealing ecological networks using Bayesian network inference algorithms. *Ecology*. 2010;91(7):1892–1899.
40. Kim SY, Imoto S, Miyano S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief Bioinform*. 2003;4(3):228–235.
41. Rijmen F. Bayesian networks with a logistic regression model for the conditional probabilities. *Int J Approx Reason*. 2008;48(2):659–666.
42. Boettcher SG, Dethlefsen C. Deal: a package for learning Bayesian networks. *J Stat Softw*. 2003;8(20):1–40.
43. Smith AFM. Bayesian computational methods. *Philos Trans R Soc A*. 1991;337(1647):369–386.
44. Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Mach Learn*. 2003;50(1-2):95–125.
45. Koivisto M, Sood K. Exact Bayesian structure discovery in Bayesian networks. *J Mach Learn Res*. 2004;5:549–573.
46. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003;19(12):1572–1574.
47. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2006.
48. Spiegelhalter DJ, Lauritzen SL. Sequential updating of conditional probabilities on directed graphical structures. *Networks*. 1990;20(5):579–605.