Disentangling the complexity of infectious diseases: Time is ripe to improve the first-line statistical toolbox for epidemiologists.

Matthieu Hanf, Jean-François Guégan, Ismaïl Ahmed, Mathieu Nacher

Disentangling the complexity of infectious diseases: time is ripe to improve the first-line statistical toolbox for epidemiologists

AUTHORS: Matthieu Hanf [1,2], Jean-François Guégan [3,4], Ismail Ahmed [1,2], Mathieu Nacher [5,6]

AFFILIATION: [1]: Biostatistics, CESP Centre for research in Epidemiology and Population Health, U1018, Inserm; Villejuif, France. [2]: Université of Paris Sud 11, UMRS 1018, Villejuif, France. [3]: UMR MIVEGEC IRD-CNRS-Universities of Montpellier I and II, Centre IRD de Montpellier, BP 64501, F-34394 Montpellier cedex 5 France. [4]: French School of Public Health (EHESP), Interdisciplinary Center on Biodiversity, Climate Change and Infectious Diseases, Centre IRD de Montpellier, BP 64501, F-34394 Montpellier cedex 5 France. [5]: Centre d'Investigation Clinique Epidémiologie Clinique Antilles Guyane CIC-EC INSERM CIE 802, Cayenne General Hospital, Cayenne, French Guiana. [6]: Université des Antilles et de la Guyane, EPaT EA3593, Cayenne, French Guiana.

CORRESPONDING AUTHOR : Matthieu Hanf, INSERM U1018, Centre de Recherche en Epidémiologie et Santé des Populations Equipe 1 Biostatistiques, 16 avenue Paul Vaillant Couturier, 94807 VILLEJUIF CEDEX, France. E-mail address: matthieu@hanf.fr, telephone number: +33 1 45 59 50 32. The corresponding author (MATTHIEU HANF) confirms that he had the final responsibility for the decision to submit for publication.

FUNDING: MH contribution has benefited from a grant managed by Agence Nationale de la Recherche (ANR). JFG and MN' contribution has benefited from an "Investissement d'Avenir" grant managed by ANR (CEBA, ref. ANR-10-LABX-0025). ANR had no role in the writing of the report; and in the decision to submit the paper for publication.

ACKNOWLEDGEMENT: M.H thanks Emeline Laurent, Michel Chavance and Neil Pearce for their advices and comments about the draft. J.F.G thanks IRD, CNRS and EHESP.

CONTRIBUTORS: MH and MN initiated the study. MH drafted the article. JG, IA and MN provided guidance in the writing of the manuscript and revised it critically. All authors approved the final version of the manuscript.

COMPETING INTEREST: None

ETHICAL APPROVAL: No need

WORD COUNT: 4747

ABSTRACT: Because many biological processes related to the dynamics of infectious diseases are caused by complex interactions between the environment, the host(s) and the agent(s), the necessity to address the methodological implications of this inherent complexity has recently emerged in epidemiology. Most epidemiologists now acknowledge that most human infectious diseases are likely to have complex dynamics. However, this knowledge still percolates with difficulty in their statistical "*modus operandi*". Indeed, for the study of complex systems, the traditional first-line statistical toolbox of epidemiologists (mainly built around the Generalized Linear Model family), despite its undeniable practicality and robustness, has structural limitations deprecating its usefulness. Three major sources of complexity neglected or not taken into account by this first-line statistical toolbox and having deep statistical implications are the multi-level organization of data, the non-linear relationships between variables and the complex interactions between variables. Three promising candidates to incorporate along with traditional tools for a new first-line statistical toolbox more suitable to apprehend these sources of complexity are the generalized linear mixed models, the generalized additive models, and the structural equation models. The aforementioned methodologies have the advantage to be generalizations of GLM models and are relatively easy to implement. Their assimilation and implementation would thus be greatly facilitated for epidemiologists. More globally, this text underlines that an improved use of other methods as such described here compared to traditional ones has to be performed to better understand the complexity challenging epidemiologists every day. This is particularly true in the field of infectious diseases for which major public health challenges will have to be addressed in the coming decades.

KEYWORDS: infectious diseases; complexity; statistics; multi-level organization; nonlinearity; interactions;

HIGHLIGHTS

- Most epidemiologists now acknowledge that infectious diseases have complex dynamics.
- Multilevel organization of data, non-linear behaviors and interactions are three major sources of complexity
- The epidemiologist first-line statistical toolbox has structural limitations limiting its ability to capture complexity.
- 3 models more able to deal with these sources of complexity are the GLMM, the GAM and the SEM models.
- An improved use of this kind of methods has to be performed to elucidate the complexity of infectious diseases.

Introduction

It is now well admitted that the emergence and reemergence of infectious diseases and their rapid dissemination worldwide are actually major challenges for national and international epidemiological researches (Jones et al., 2008; McMichael, 2004; Smith and Guégan, 2010). Until now, expectations in new vaccines or drugs and global surveillance to reverse the observed trends have been frustrated by the extreme complexity of the dynamics of infectious diseases (Plowright et al., 2008). Various individual or global determinants, such as genetics, extreme poverty, risky behaviors, urbanization, land-use changes e.g. deforestation, agricultural practices, or climate and its perturbations**,** acting at different spatio-temporal scales, may favor the emergence and resurgence of many infectious diseases and increase their epidemiological complexity (Harrus and Baneth, 2005; Morse, 2004, 1995; Weiss and McMichael, 2004; Woolhouse and Gowtage-Sequeria, 2005). In addition, the huge diversity of viruses, bacteria, fungi and parasites (Woolhouse et al., 2008) entails that it is also not unusual for people to be co-infected with various pathogens that circulate within the global environment (Smith et al., 2007). The resulting symptoms and severity may be due to multi-species co-infections and often cannot be predicted by the simple sum of the effects of each pathogen, as notably revealed by the 2008-2009 H1N1*v* pandemic for which mortality was mostly due to opportunistic bacterial infections (MMWR, 2009; Palacios et al., 2009). Traditional approaches for the study of cause and effect relationships are often not possible when studying emerging infections because study units are large and complex and risk factors have non-linear, hierarchical effects (Karesh et al., 2012; Plowright et al., 2008). Systematic, interdisciplinary approaches are clearly needed for understanding disease outbreaks and spread (Harrus and Baneth, 2005; Kilpatrick and Randolph, 2012; Morse, 2004, 1995; Weiss and McMichael, 2004; Woolhouse and Gowtage-Sequeria, 2005).

Thus, the elucidation of “complexity” is now at the heart of current epidemiological issues (Leport et al., 2012). Because many biological processes related to the emergence and dynamics of infectious diseases are caused by complex interactions between the natural and socio-economical environment, the host(s) and the agent(s), the necessity to address the methodological implications of such inherent complexity in epidemiology has emerged during the last decade (Karesh et al., 2012; Kilpatrick and Randolph, 2012). This has led to a call for a new paradigm “the theory of complexity” to understand the different mechanisms and drivers underlying pathogen emergence and improve disease prevention (Jayasinghe, 2011; Materia and Baglio, 2005; Morabia, 2007; Pearce and Merletti, 2006). Epidemiologists now acknowledge that most human infectious diseases are likely to have complex, non-linear dynamics, and for some chronic diseases it is now demonstrated that some can have a microbial or an infectious origin, like for Crohn’s disease (Bouskra et al., 2008) or several neurologic diseases (Olival and Daszak, 2005). However, to move forward epidemiologists must not only acknowledge but also directly confront the numerous multi-scale factors that can be involved in

complex infectious disease dynamics (Karesh et al., 2012). The traditional first-line statistical toolbox of epidemiologists, mainly built around the Generalized Linear Models (GLMs) and the general use of risk factors epidemiology (Susser, 1998), have structural limitations limiting its ability to accomplish this task. Significant statistical challenges are thus now facing epidemiologists. Unfortunately, the methodological implications of complexity theory still percolate with difficulty in the field and have difficulties to be routinely applied "statistically" despite the fact that analyses based on this theory are nowadays facilitated by the combined use of relatively new analytical methods and statistical softwares that combine complexity and usability. Our aim in the present paper is thus to explain 1) three major sources of complexity having deep statistical implications in epidemiology of infectious diseases, and 2) why these last ones are neglected or not taken into account by traditional statistical tools for epidemiologists when they are largely used by other fields of research, notably in ecology and evolution of infectious diseases (Plowright et al., 2008). Here, we outline some of the barriers to advancing our understanding of statistical modeling in medical epidemiology. Together with this presentation, three statistical models, relatively easy to implement, and more suitable to apprehend these sources of complexity, are described. The usefulness of these models is also illustrated with recent examples from the literature.

Multilevel analysis with Generalized Linear mixed modeling (GLMM): a well suited tool to apprehend the hierarchical structure of infectious disease dynamics

A first key concept of the complexity theory is that determinants of an infectious disease cannot be conceptualized only as an attribute of a particular level of organization (molecular, cellular, individual or population ones for example). In epidemiology, population and group factors as well as individual factors are all important in understanding the causes of diseases (Pearce and Merletti, 2006; Pearce, 2004, 1999, 1996). Discussions on group and individual factors are often reduced to the idea that on one side, group characteristics are important in understanding the differences between groups and that on another side, individual characteristics are important in understanding differences between individuals. Complexity theory underlines that a set of factors or drivers defined on several levels of organization may be important for understanding the causes of variability within a single level of organization (Pearce and Merletti, 2006). In infectious disease epidemiology, it has long been recognized that factors "independent" of individuals, defined at the population level (Morgenstern, 1995; Rose, 1985; Susser, 1994), influence health. A well known example is the concept of "herd immunity" which implies that the probability of a person to contract an infectious disease agent depends partly on the immunity level of the population to which it belongs (Fine, 1993). It is now acknowledged that this concept of multiple levels of organization can be found in every epidemiological study because they always involve some sort of population (including countries, regions, villages, community, extended families, etc…)(Pearce, 1999). Complexity theory

thus underlines that all levels of organization are of value and that it is particularly valuable to follow an integrative approach which incorporates the various levels whatever the level at which the research is made (Plowright et al., 2008). Interestingly, this approach is called in medicine and public health as ecological studies - which are not used exactly in the same way ecologists are defining this - , and it is criticized by health scientists as being pervasive even fallacious correlation studies only (Pearce, 2000).

A particularly illustrative example is given by the social determinants of HIV/AIDS incidence in populations (Poundstone et al., 2004) (figure 1). The latter are distributed on several levels of organization (individual, community and national) and all of these factors have to be put together to apprehend the dynamic of HIV/AIDS in populations.

Unfortunately, over the past few decades, most epidemiological studies in infectious diseases have taken into account only individual-level risk factors for disease (McMichael, 1999; Susser and Susser, 1996). This approach has led to the intensive use of statistical models developed on a “one level data” spirit. Furthermore, during the same period, epidemiologists acknowledging the hierarchical organization of data were often inhibited from applying the ‘multilevel perspective’ by a lack of understanding of how to analyze such data and by the lack of dedicated statistical tools leading to utilize traditional one-level statistical tools, even when their data and hypotheses were multilevel in nature.

These practices are confronted to at least two problems. First, all of the unmodeled group level or contextual information ends up pooled into the single individual error term of the model (Duncan et al., 1996; Luke, 2004). This is problematic because individuals belonging to the same context will presumably have correlated errors, which violates one of the basic assumptions of classical regression models. The second problem is that by ignoring the context under investigation, the model assumes that the regression coefficients apply equally to all contexts, “thus propagating the notion that processes work out in the same way in different contexts” (Duncan et al., 1996; Luke, 2004).

To solve these methodological problems, specific statistical modeling called multilevel modeling was developed during the last two decades. Such models have been created to allow analysis at several levels simultaneously, rather than having to choose at which level to carry out a single level analysis. They were relatively new compared to other common types of modeling, such as GLMs. To avoid previously described pitfalls in the analysis of hierarchical data, these multilevel models incorporate, in parallel to individual factors (commonly referred as fixed effects), group level effects describing the variability associated with particular group levels (commonly referred as random effects). With this ability, these models radically outperform classical regression in term of predictive accuracy. The vast increase in computing power over recent decades has led to the emergence of

these multilevel models as practical and powerful tools to better explain data variability. All statistical programs now have dedicated functions and packages that allow the study of hierarchical structures in data of all kinds in one unified framework called generalized linear mixed models (McCulloch et al., 2008), whatever the design of the study (cross-sectional or longitudinal) and with the distinct advantage to handle unbalanced designs quite well.

However, multilevel modeling also raises some concerns about potential pitfalls and limitations which have to be carefully apprehended during the model construction. Some of them are, for example, the proper specification of the error structure, the model building strategy, the choice of appropriate software and associated options, and the interpretation and reporting of the results (Diez Roux and Aiello, 2005; Greenland, 2000; Nezlek, 2008). Nevertheless, in the area of infectious disease epidemiology, multilevel analysis remains a very pertinent tool, when properly used, to examine how both group- and individual-level factors are related to individual-level disease outcomes and how factors at both levels contribute to group-to-group differences in disease rates (Diez Roux and Aiello, 2005).

The study of Yang et al. (2009), focused on risk factors for Schistosomiasis, perfectly illustrates the advantages of multi-level modeling on traditional ones (J. Yang et al., 2009). They conducted a cross-sectional survey in 16 villages in the Chinese province of Hunan to investigate both individual and group level (villages) risk factors for Schistosomiasis infection. Surprisingly, contrarily to their single level analysis and of those found in the literature, their multi-level analysis did not find a significant, independent effect of density, in particular, of infected snails on Schistosomiasis infection in humans. They concluded that previous studies having ignored the hierarchical structure of the data may have obtained improper results. These findings, obtained by multi-level modeling, may guide the development of Schistosomiasis infection prevention programs, questioning whether massive application of molluscicides to control snails in endemic areas is an effective preventive measure.

More globally, epidemiological studies using multilevel modeling could now be seen in a large spectrum of infectious diseases such as malaria (Yusuf et al., 2010), HIV (Msisha et al., 2008), visceral leishmaniasis (Werneck et al., 2006) or leprosy (Sales et al., 2011) but this type of applications remain however globally sparse (Diez Roux and Aiello, 2005). As rightly said by Diez Roux and Aiello (Diez Roux and Aiello, 2005), the generalization of the use of multilevel analysis in infectious disease epidemiology could only be done if an upstream work, as usually done in ecological sciences notably (Burnham et al., 2002; Grace, 2006), was performed by the community to identify the levels that are relevant to the research question of interest, specifying the relevant constructs or variables at each level, operationalizing the relevant groups, and measuring the relevant group-level variables. The incorporation of group-level data in individual-level studies (if done carefully) can only strengthen the field (Duell, 2006).

Elucidating the non linear behaviors observed in infectious disease dynamics: the great value of the generalized additive model (GAM)

In addition to the difficulty provided by the multilevel organization of data, complexity theory implies a second fundamental source of complexity: the nonlinear relationships between the variables of a system (Pearce and Merletti, 2006; Pearce, 1996).

A non-linear behavior could be roughly defined as a behavior that is not based on a simple proportional relationship between two quantitative variables. Therefore, the induced changes are often sudden, unexpected and difficult (and sometimes impossible) to predict. In these nonlinear systems, a modification of a small amount of one or two parameters can dramatically change the behavior of the entire system (Pearce and Merletti, 2006; Pearce, 1996). Complex biological systems are often characterized by nonlinear behaviors whatever the level of organization (from the activity of an enzyme to the dynamic of infectious diseases in human populations). It is now well acknowledged that the incidence of an infectious disease is a non-linear function of the number of infectious and susceptible individuals within the population (Anderson and May, 1991), or that the relationship between malaria transmission and vector-sources adaptation to temperature is profoundly non-linear (Patz and Olson, 2006).

However, to be able to apprehend natural complex systems, science has always tried to reduce their description to a simpler system. One of the most widely used methods to study and explain such systems was to consider these systems under the assumption of linearity. This paradigm requires that the relationship between two variables X and Y depends on a weight α representing the strength of the relationship. Because of its conceptual usefulness, this assumption is no longer challenged when selecting the methodology to apply. This is particularly true for GLMs used every day. Undoubtedly, this paradigm of linearity helps researchers to better understand phenomena of interest in epidemiology, but its usefulness is inherently limited when the investigator wants to better understand complex systems that involve non linear behaviors (Philippe and Mansi, 1998). There are important non-linearities in nature for which the linear approximation is an uninformative (and possibly misleading) first analysis step especially in the case of threshold, belt-shaped or Gaussian curves, and aggregated functions that are common in nature, and U, or J-shaped, or even more complicated relationships (May and Bigelow, 2005).

There again, when confronted with these very difficult conceptual problems due to non linear relations between variables, some substitution strategies were developed to better accounts for it than in traditional models. When a relationship between two continuous variables is identified as non linear, a first practical solution is often to categorize one of the studied variables and to estimate for each associated categories the resulting effect on the second variable. Such methods have become very popular due to their easy interpretation and the ensuing intuitiveness of the

communication of the results. Unfortunately, it is now well admitted that the fit of such strategies is often very poor (Altman, 1991; Bennette and Vickers, 2012; Greenland, 1995; Zhao and Kolonel, 1992)**.** Indeed, categorization of continuously distributed variables is associated with three problems: first, it involves multiple hypothesis testing with pair-wise comparisons of groups; second, it requires an unrealistic function of risk that assumes homogeneity of risk within groups, leading to both a loss of power and inaccurate estimation; and third, it leads to difficulty comparing results across studies due to the data-driven cut off points often used to define categories (median, quintiles,…)(Bennette and Vickers, 2012). A second strategy massively adopted by the epidemiological community is to transform one or several exploratory variables to obtain a relationship that is linear (logarithmic, square root, inverse or square transformation …) (Flanders et al., 1992) or to use a parametric function of the original variable (most often quadratic and occasionally cubic or polynomial). Similarly, applying a transformation of the outcome through the use of a non linear link function possibly selected by a model selection procedure is another common strategy to deal with non linearity. There again, with their limited flexibility, the fitting of such models is often quite poor (Royston, 2000).

An important statistical development of the last thirty years has been the advance in regression analysis provided by generalized additive models (GAM) (Hastie and Tibshirani, 1990, 1986). The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relationships between the response and a set of explanatory variables. They are a semi-parametric extension of the GLM in that one or more predictors may be specified using a smooth function. The smoothness for the functions is calculated internally with the goal of optimal balance between the fit to the data and excessive ‘‘tortuosity’’ of the functions. Furthermore, group-level effect can also be taken into account in GAMs by the possible inclusion of random effects. Therefore, the hierarchical structure of explanatory variables can also be modeled with GAMs. Since their development, GAMs have been extensively applied in biological sciences as ecology, as evidenced by the growing number of published papers incorporating these modern tools (Guisan et al., 2002). This is due, in part, to their ability to deal with the multitude of distributions that define data in the same way as GLM, and to the fact that they blend in well with traditional practices used in linear modeling and analysis of variance. Like in ecology, the use of GAMs in epidemiology to handle non-linear data structures could improve the representation of the underlying data, and hence increase our understanding of complex epidemiological systems (Guisan et al., 2002).

A simple but powerful example of GAM usefulness could be found in the study made by Giraudoux et al. (2013) focalizing on Human alveolar echinococcosis (*Echinococcus. multilocularis*). Through the use of a GAM model investigating the non linear effect of a large panel of environmental determinants on the infections status of more than 15,000 Chinese people, they showed and describe precisely the non linear impact of landscape features and climate on Human alveolar

echinococcocosis (Giraudoux et al., 2013) (figure 2). The authors concluded that their study may be a starting point for further research wherein landscape management could be used to predict human disease risk and for controlling this zoonotic helminthic. With the use of traditional statistical models assuming linearity alone, this simple and useful message could not have been elaborated.

In infectious diseases, some research fields such as environmental science or biogeography have already understood the potential benefits of this kind of methods in the understanding of infectious disease dynamics. Applications could now be found for a variety of diseases such as influenza (L. Yang et al., 2009), malaria (Nkurunziza et al., 2011), cholera (Piarroux, 2011) and many others (Dukić et al., 2012; Hens et al., 2007; Schindeler et al., 2009). However, similarly to multilevel modeling, empirical applications of GAM analysis in infectious disease researches remain globally sparse. Yet despite the methodological advancements provided by methods like GAM models and calls for the abandonment of variable categorization, the epidemiologic community continues to rely heavily on the use of linearity hypothesis as a primary means of analyzing and presenting results (Bennette and Vickers, 2012). Possible explanations could be found in the fact that GAMs are more complicated to fit, require a sufficient amount of data to be performed, may lead to over-fitting when improperly used, are criticized to have a "black box" behavior and could provide difficulties in assigning biological meaning to the fitted model due to the flexibility of GAM in allowing different model types. With recent dedicated functions and packages simplifying their use, these problems and limitations become increasingly obsolete. By extending GLM and relaxing the linear assumption, GAMs could thus represent a new kind of "screwdriver" in the first line statistical toolbox of epidemiologists specialized in the study of non linear behaviors. They really offer to epidemiologists a practical methodology for improving on the extensive practice of linearity by default (Beck and Jackman, 1998).

Unraveling the complexity behind the interactions of variables: identification of the web of determinants by structural equation modeling (SEM)

A third major source of complexity in epidemiology remains to be described: the existence of complex interactions between the outputs (or explanatory variables) and inputs (or dependent variables) (Pearce and Merletti, 2006; Pearce, 1996). A first source of interaction which could be described is when a relationship between a predictor and an outcome is weakened or strengthened by a second predictor. Furthermore, in a complex system, a particular determinant could have the ability to impact not directly the disease outcome (proximate determinants) but rather through a complex web of interactions involving others factors (distal determinants). For example, in tuberculosis, HIV status is clearly a proximal determinant of occurrence in individuals and belonging country economic level a distal one (acting for example through capability of health structures management or health intervention implementation). Complexity theory thus underlines that health,

disease and the balance between the two are determined by many interwoven factors, which may reinforce, interact synergistically, mask or inhibit each other in a dynamic web of interactions (Albrecht et al., 1998). Indeed, to understand a natural process, it is critical to know which groups of variables are joined in such complex effects and must be examined together. This “web of determinants” in infectious diseases is illustrated in figure 3 for Hendra virus emergence determinants in Australia. Furthermore, complexity theory also underlines that an infectious disease epidemiologist has to interrogate himself on the principle of “causation” (Joffe et al., 2012; Plowright et al., 2008). Under this theory, the definition of a causal relation between a determinant and a disease is much more than a direct relationship between the two; it is rather a confirmed effect of a determinant on a complex system in which many variables interact and influence the disease dynamic in a form of more or less complex cascade-of-effects structure.

In a reductionist scientific tradition, epidemiology has tried to understand and explain the impact of different factors on outcomes by isolating and studying them separately (Susser, 1998). This philosophy is mainly achieved in epidemiology through traditional multivariate statistical analyses (as GLMs) revealing the impact of each health- or disease-promoting factor by controlling for the effect of all other factors included. These kinds of models are very useful to examine direct relationships between independent and dependent variables but are intrinsically limited to study complex interactions where distal influences could be at stake. Real life may not be so parsimonious; relationships between various variables may be much more complex, more “web-like” (Krieger, 1994).

Some adjustments are however possible. In traditional tools, interactions terms could be included in models to correct all deviations due to strong interactions between inputs. Nevertheless, these terms only represent statistical corrections and do not take explicitly into account the complex structural relations existing between variables. This traditional approach, which emphasizes single causes and bivariate associations, has dominated epidemiological researches until recently.

Structural-equation models (SEMs) were developed in the mid-late 1980's to model more efficiently complex relationships between factors (Bollen, 1989; Kaplan, 2000). Statistically, they represent an extension of path analyses and GLM procedures. They are applicable to both experimental and non-experimental data, as well as cross-sectional and longitudinal data. Traditional SEMs are multiple-equation regression models in which the response variable in one regression equation can appear as an explanatory variable in another equation. Indeed, two variables in a SEM can even effect one-another reciprocally, either directly, or indirectly through a feedback loop. SEMs can also include variables that are not measured directly (latent variables). The goal of SEM is to determine whether a hypothesized theoretical model is consistent with the data collected. The consistency is evaluated through model-data fit, which indicates the extent to which the postulated network of relations among variables is plausible. Indeed, on the contrary to traditional methods

such as regression, SEM is able to yield unique information about the complex nature of disease and health behaviors when used within good research design. Nevertheless, like any procedure in data analysis, this methodology is also subject to misspecifications, and the researcher must be aware of several considerations to develop a legitimate model. These include the steps in model development, testing for reliability and validity, sample size requirements and interpretation of fitting measures (Beran and Violato, 2010).

With the advent of SEM computer programs and the development of methods such as causal diagrams helping to structure the statistical analysis of the hypothesized pathways (Joffe et al., 2012; Plowright et al., 2008), SEM has now become a well-established and respected methodology. Important contributions to SEM have come out of the behavioral and social sciences. Currently, the potential of such techniques are just beginning to be appreciated in epidemiologic and clinical studies (Amorim et al., 2010; Beran and Violato, 2010).

The advantages of SEM approaches compared to traditional analyses were perfectly illustrated by the study of Calis et al. (Calis et al., 2008). Little is known about the causes of severe anemia in African children. Among them, iron deficiency and infectious diseases are widely held to be some of the most common causes. To test this assertion, Calis et al. conducted a SEM analysis to finely model the complex relations existing between potential determinants and severe anemia. Retrieved significant associations were shown in figure 4. One of their counterintuitive results is that iron deficiency, due to complex relations with other determinants (as hookworm and bacteria load), could be in fact a protective factor of severe anemia. They concluded that treatment recommendations for severe anemia that promote iron and ignore bacteremia or hookworm infections appear to be of limited applicability. These important results could not have been obtained when developing classical analyzes.

More globally, special uses of SEM are now emerging in fields as diverse as exposure assessment (Davis, 2011), nutritional epidemiology (Chavance et al., 2010) or human genetics (Li et al., 2006) but still percolate difficultly in the infectious disease area. Apart from the behavioral studies linked to infectious diseases (Rao et al., 2011), applications in infectious diseases remain quite rare (Guan et al., 2009; Obel et al., 2010).

By permitting the study of the complex web of interactions exiting in every infectious disease dynamic, SEM could however be a promising tool to complement or an alternative to traditional ones. Incorporating SEMs in their statistical *modus operandi* could give infectious disease epidemiologists a real opportunity to better apprehend the inherent complexity of infectious diseases challenging them every day.

Concluding remarks

We have seen that the traditional first-line statistical toolbox (mainly built around the GLM family), despite its undeniable practicality, has structural limitations limiting its ability to capture the complexity provided by the multilevel organization of data and the potential non-linear behaviors and/or complex interactions at stake in infectious diseases. As pointed in other research areas (Thornton-Wells et al., 2004), there is currently a crucial need for an extensive reevaluation of existing methodologies to study the infectious diseases. This discussion tries to make a move in this direction. Three additional candidates for this new statistical toolbox have been described here: the GLMM models (taking into account the multi-level organization of data), the GAM models (able to manage deep non-linear relationships between variables), and the SEM models (allowing the modeling of complex interactions between variables). We are convinced that a more systematic use of these of these kinds of models could help epidemiologists to better elucidate the inherent complexity of infectious diseases and fill the gap between acknowledgement of limitations and action to overcome them.

The simultaneous application of these three models on every epidemiological datasets with which the GLM family is an adequate strategy of analysis cannot obviously be done systematically. Everything depends on the question under investigation, the collected data and the particular dynamic of the studied phenomenon. However, we think that, in a non-negligible proportion of these epidemiological studies, at least one of these models is applicable and can be used to investigate the underlying complexity of the epidemiological phenomenon more accurately. Furthermore, an upstream reflection must also be performed by the community to enable these kinds of models to be applied as often as possible. This reflection should primarily focus on formalizing assumptions on the complexity of the studied phenomena, the type of study to conduct to efficiently investigate this complexity and the nature of the data which have to be collected to accomplish this task.

These three models are only examples of new interesting statistical methods; many others, also able to meet these challenges, already exist or are under development. These include among others, decision trees, neural networks, projection pursuit regression, boosting, bayesian hierarchical models, penalized regressions, generalized method of moments or quantile regression (Hastie et al., 2009). Nevertheless the aforementioned methodologies have the unique advantage to be generalizations of GLM models: their assimilation and implementation would thus be greatly facilitated for epidemiologists. They also have the decisive advantages of being applicable to a wide variety of data and of having been tested and validated in many other scientific areas. They thus could be rapidly assimilated and used by the infectious disease community. Nevertheless, what we propose here only represents a preliminary part of the in depth introspection that the community should perform. Indeed, Similarly to GLM models, the use of more well-suited models instead of others statistical tools commonly used in infectious disease epidemiology (as survival analysis or spatial analysis for example) are needed to better apprehend the complexity provided by the

multilevel organization of data and associated potential non-linear behaviors and/or complex interactions. Furthermore, recent progresses in advanced statistics as in contact networks, spatial point processes, or transmission tree reconstruction to name a few have also to be tested and assimilated by the community to help to the definition of a new statistical toolbox plenty able to study the complexity of infectious diseases (Lawson, 2006; Mollison, 1995; Waller, 2004). This text underlines that the techniques necessary to answer current infectious diseases questions are quite different from the standard statistical techniques that are taught in most epidemiological textbooks and courses today (Pearce and Merletti, 2006). A sound reflection on what to teach in statistics and/or on how to better expose the future epidemiologists to new statistical methods must also be performed. In addition, recent advances outside the scope of statistical modeling in numerical simulation and mathematical models (as for example agent-based modeling or SEIR models) have shown their great utility in studying the complexity of infectious diseases and critically reinforce this need. This task appears to us as a necessity if the community wants to equip future epidemiologists for the study of the complex dynamics provided by infectious diseases in the next decades.

This discussion is far from a plea against the traditional models used in epidemiology. Due to their simplicity, functionality and robustness, they must continue to be implemented in view to provide a first picture of studied phenomena. But the epidemiological community must now be aware about the fact that the former are necessary but not sufficient and that the implementation of more refined methodologies has to be performed concomitantly to go further in the understanding of the dynamic of infectious diseases. However to be largely used, selected new methodologies must themselves not fall in the trap of complexity. They have to be designed keeping in mind both "simplicity" in use/interpretation and "complexity" of potential phenomena under investigation. "Simplexity" paradigm argued that simple interfaces tend to improve the usability and understanding of complex systems (Kluger, 2008). Its full application is a challenging task for the statistician community in the next decades.

To conclude, Neil Pearce and Franco Merletti asked the question to know if we are going to continue to use the epidemiological methods of the 20th century to address the scientific and public health problems of the 21st century (Pearce and Merletti, 2006). Our response is "yes" but a concomitant improved use and development of other methods, such as those described here, also have to be performed to entirely and efficiently address this task. This is particularly true in the field of infectious diseases for which major public health challenges operating at different spatial and temporal scales, from the local to upper scale and vice versa will have to be addressed in the coming decades.

REFERENCES

Albrecht, G., Freeman, S., Higginbotham, N., 1998. Complexity and human health: the case for a transdisciplinary paradigm. Cult Med Psychiatry 22, 55–92.

Altman, D.G., 1991. Categorising continuous variables. Br. J. Cancer 64, 975.

Amorim, L.D.A.F., Fiaccone, R.L., Santos, C.A.S.T., Santos, T.N. dos, Moraes, L.T.L.P. de, Oliveira, N.F., Barbosa, S.O., Santos, D.N. dos, Santos, L.M. dos, Matos, S.M.A., Barreto, M.L., 2010. Structural equation modeling in epidemiology. Cadernos de Saúde Pública 26, 2251–2262.

Anderson, R.M., May, R.M., 1991. Infectious diseases of humans : dynamics and control. Oxford University Press, Oxford; New York.

Beck, N., Jackman, S., 1998. Beyond Linearity by Default: Generalized Additive Models. American Journal of Political Science 42, 596.

Bennette, C., Vickers, A., 2012. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol 12, 21.

Beran, T.N., Violato, C., 2010. Structural equation modeling in medical research: a primer. BMC Research Notes 3, 267.

Bollen, K., 1989. Structural equations with latent variables. New York: Wiley; 1989. Wiley, New York.

Bouskra, D., Brézillon, C., Bérard, M., Werts, C., Varona, R., Boneca, I.G., Eberl, G., 2008. Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. Nature 456, 507–510.

Burnham, K.P., Anderson, D.R., Burnham, K.P., 2002. Model selection and multi-model inference : a practical information-theoretic approach. Springer, New York.

Calis, J.C.J., Phiri, K.S., Faragher, E.B., Brabin, B.J., Bates, I., Cuevas, L.E., de Haan, R.J., Phiri, A.I., Malange, P., Khoka, M., Hulshof, P.J.M., van Lieshout, L., Beld, M.G.H.M., Teo, Y.Y., Rockett, K.A., Richardson, A., Kwiatkowski, D.P., Molyneux, M.E., van Hensbroek, M.B., 2008. Severe anemia in Malawian children. N. Engl. J. Med. 358, 888–899.

Chavance, M., Escolano, S., Romon, M., Basdevant, A., de Lauzon-Guillain, B., Charles, M., 2010. Latent variables and structural equation models for longitudinal relationships: an illustration in nutritional epidemiology. BMC Medical Research Methodology 10, 37.

Davis, M.E., 2011. Structural equation models in occupational health: an application to exposure modelling. Occupational and Environmental Medicine 69, 184–190.

Diez Roux, A.V., Aiello, A.E., 2005. Multilevel analysis of infectious diseases. J. Infect. Dis. 191 Suppl 1, S25–33.

Duell, E.J., 2006. The future of epidemiology: methodological challenges and multilevel inference. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 49, 622–627.

Dukić, V., Hayden, M., Forgor, A.A., Hopson, T., Akweongo, P., Hodgson, A., Monaghan, A., Wiedinmyer, C., Yoksas, T., Thomson, M.C., Trzaska, S., Pandya, R., 2012. The Role of Weather in Meningitis Outbreaks in Navrongo, Ghana: A Generalized Additive Modeling Approach. Journal of Agricultural, Biological, and Environmental Statistics 17, 442–460.

Duncan, C., Jones, K., Moon, G., 1996. Health-related behaviour in context: A multilevel modelling approach. Social Science & Medicine 42, 817–830.

Fine, P.E., 1993. Herd immunity: history, theory, practice. Epidemiol Rev 15, 265–302.

Flanders, W.D., DerSimonian, R., Freedman, D.S., 1992. Interpretation of linear regression models that include transformations or interaction terms. Ann Epidemiol 2, 735–744.

Giraudoux, P., Raoul, F., Pleydell, D., Li, T., Han, X., Qiu, J., Xie, Y., Wang, H., Ito, A., Craig, P.S., 2013. Drivers of Echinococcus multilocularis Transmission in China: Small Mammal Diversity, Landscape or Climate? PLoS Neglected Tropical Diseases 7, e2045.

Grace, J.B., 2006. Structural equation modeling. Cambridge University Press, Cambridge, UK; New York.

Greenland, S., 1995. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 6, 450–454.

Greenland, S., 2000. Principles of multilevel modelling. Int J Epidemiol 29, 158–167.

Guan, P., Huang, D., He, M., Shen, T., Guo, J., Zhou, B., 2009. Investigating the effects of climatic variables and reservoir on the incidence of hemorrhagic fever with renal syndrome in Huludao City, China: a 17-year data analysis based on structure equation model. BMC Infectious Diseases 9, 109.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100.

Harrus, S., Baneth, G., 2005. Drivers for the emergence and re-emergence of vector-borne protozoal and bacterial diseases. Int. J. Parasitol. 35, 1309–1318.

Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. Statistical Science 1, 297–310.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall Ltd, London, United Kingdom.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning data mining, inference, and prediction. Springer, New York.

Hens, N., Aerts, M., Shkedy, Z., Kung'U Kimani, P., Kojouhorova, M., Van Damme, P., Beutels, P., 2007. Estimating the impact of vaccination using age–time-dependent incidence rates of hepatitis B. Epidemiology and Infection 136.

Jayasinghe, S., 2011. Conceptualising population health: from mechanistic thinking to complexity science. Emerg Themes Epidemiol 8, 2.

Joffe, M., Gambhir, M., Chadeau-Hyam, M., Vineis, P., 2012. Causal diagrams in systems epidemiology. Emerging Themes in Epidemiology 9, 1.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990–993.

Kaplan, D., 2000. Structural equation modeling. Sage, Thousand Oakes.

Karesh, W.B., Dobson, A., Lloyd-Smith, J.O., Lubroth, J., Dixon, M.A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E.H., Machalaba, C.C., Thomas, M.J., Heymann, D.L., 2012. Ecology of zoonoses: natural and unnatural histories. The Lancet 380, 1936–1945.

Kilpatrick, A.M., Randolph, S.E., 2012. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. The Lancet 380, 1946–1955.

Kluger, J., 2008. Simplexity : why simple things become complex (and how complex things can be made simple). Hyperion, New York.

Krieger, N., 1994. Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med 39, 887–903.

Lawson, A., 2006. Statistical methods in spatial epidemiology, 2nd ed. ed, Wiley series in probability and statistics. Wiley, Chichester, England ; Hoboken, NJ.

Leport, C., Guégan, J.., Zylberman, P., Bitar, D., Bricaire, F., Cavallo, J.., Eliaszewicz, M., Moatti, J.., 2012. Proceedings of the seminar on Emerging Infectious Diseases,November 9, 2011: Current trends and proposals. Médecine et Maladies Infectieuses (In press).

Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I.M., Wergedal, J., Paigen, B., Churchill, G.A., 2006. Structural Model Analysis of Multiple Quantitative Traits. PLoS Genetics 2, e114.

Luke, D.A., 2004. Multilevel modeling. Sage, Thousand Oaks, CA.

Materia, E., Baglio, G., 2005. Health, science, and complexity. J Epidemiol Community Health 59, 534–535.

May, S., Bigelow, C., 2005. Modeling Nonlinear Dose-Response Relationships in Epidemiologic Studies: Statistical Approaches and Practical Challenges. Dose-Response 3, 474–490.

McCulloch, C.E., Searle, S.R., Neuhaus, J.M., 2008. Generalized, linear, and mixed models, 2nd ed. ed, Wiley series in probability and statistics. Wiley, Hoboken, N.J.

McMichael, A.J., 1999. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am. J. Epidemiol. 149, 887–897.

McMichael, A.J., 2004. Environmental and social influences on emerging infectious diseases: past, present and future. Philosophical Transactions of the Royal Society B: Biological Sciences 359, 1049–1058.

MMWR, 2009. Bacterial Coinfections in Lung Tissue Specimens from Fatal Cases of 2009 Pandemic Influenza A (H1N1) United States, May-August 2009 (No. 58(Early Release);1-4). CDC, Atlanta.

Mollison, D., 1995. Epidemic models: their structure and relation to data, Publications of the Newton Institute. New York, NY : Cambridge University Press, Cambridge [England].

Morabia, A., 2007. Epidemiologic interactions, complexity, and the lonesome death of Max von Pettenkofer. Am. J. Epidemiol. 166, 1233–1238.

Morgenstern, H., 1995. Ecologic studies in epidemiology: concepts, principles, and methods. Annu Rev Public Health 16, 61–81.

Morse, S.S., 1995. Factors in the emergence of infectious diseases. Emerging Infect. Dis. 1, 7–15.

Morse, S.S., 2004. Factors and determinants of disease emergence. Rev. - Off. Int. Epizoot. 23, 443–451.

Msisha, W.M., Kapiga, S.H., Earls, F.J., Subramanian, S., 2008. Place matters: multilevel investigation of HIV distribution in Tanzania. AIDS 22, 741–748.

Nezlek, J.B., 2008. An Introduction to Multilevel Modeling for Social and Personality Psychology. Social and Personality Psychology Compass 2, 842–860.

Nkurunziza, H., Gebhardt, A., Pilz, J., 2011. Geo-additive modelling of malaria in Burundi. Malaria Journal 10, 234.

Obel, N., Christensen, K., Petersen, I., Sorensen, T.I.A., Skytthe, A., 2010. Genetic and Environmental Influences on Risk of Death due to Infections Assessed in Danish Twins, 1943-2001. American Journal of Epidemiology 171, 1007–1013.

Olival, K.J., Daszak, P., 2005. The ecology of emerging neurotropic viruses. Journal of Neurovirology 11, 441–446.

Palacios, G., Hornig, M., Cisterna, D., Savji, N., Bussetti, A.V., Kapoor, V., Hui, J., Tokarz, R., Briese, T., Baumeister, E., Lipkin, W.I., 2009. Streptococcus pneumoniae Coinfection Is Correlated with the Severity of H1N1 Pandemic Influenza. PLoS ONE 4, e8540.

Patz, J.A., Olson, S.H., 2006. Malaria risk and temperature: influences from global climate change and local land use practices. Proc. Natl. Acad. Sci. U.S.A. 103, 5635–5636.

Pearce, N., 1996. Traditional epidemiology, modern epidemiology, and public health. Am J Public Health 86, 678–683.

Pearce, N., 1999. Epidemiology as a population science. Int J Epidemiol 28, S1015–1018.

Pearce, N., 2000. The ecological fallacy strikes back. Journal of Epidemiology & Community Health 54, 326–327.

Pearce, N., 2004. Epidemiology: populations, methods and theories. Eur. J. Epidemiol. 19, 729–731.

Pearce, N., Merletti, F., 2006. Complexity, simplicity, and epidemiology. Int J Epidemiol 35, 515–519.

Philippe, P., Mansi, O., 1998. Nonlinearity in the epidemiology of complex health and disease processes. Theor Med Bioeth 19, 591–607.

Piarroux, R., 2011. Understanding the Cholera Epidemic, Haiti. Emerging Infectious Diseases 17, 1161–1168.

Plowright, R.K., Sokolow, S.H., Gorman, M.E., Daszak, P., Foley, J.E., 2008. Causal inference in disease ecology: investigating ecological drivers of disease emergence. Frontiers in Ecology and the Environment 6, 420–429.

Poundstone, K.E., Strathdee, S.A., Celentano, D.D., 2004. The social epidemiology of human immunodeficiency virus/acquired immunodeficiency syndrome. Epidemiol Rev 26, 22–35.

Rao, D., Feldman, B.J., Fredericksen, R.J., Crane, P.K., Simoni, J.M., Kitahata, M.M., Crane, H.M., 2011. A Structural Equation Model of HIV-Related Stigma, Depressive Symptoms, and Medication Adherence. AIDS and Behavior 16, 711–716.

Rose, G., 1985. Sick individuals and sick populations. Int J Epidemiol 14, 32–38.

Royston, P., 2000. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med 19, 1831–1847.

Sales, A.M., Ponce de Leon, A., Düppre, N.C., Hacker, M.A., Nery, J.A.C., Sarno, E.N., Penna, M.L.F., 2011. Leprosy among Patient Contacts: A Multilevel Study of Risk Factors. PLoS Neglected Tropical Diseases 5, e1013.

Schindeler, S.K., Muscatello, D.J., Ferson, M.J., Rogers, K.D., Grant, P., Churches, T., 2009. Evaluation of alternative respiratory syndromes for specific syndromic surveillance of influenza and respiratory syncytial virus: a time series analysis. BMC Infectious Diseases 9, 190.

Smith, K.F., Guégan, J.-F., 2010. Changing Geographic Distributions of Human Pathogens. Annual Review of Ecology, Evolution, and Systematics 41, 231–250.

Smith, K.F., Sax, D.F., Gaines, S.D., Guernier, V., Guégan, J.-F., 2007. Globalization of human infectious disease. Ecology 88, 1903–1910.

Susser, M., 1994. The logic in ecological: I. The logic of analysis. Am J Public Health 84, 825–829.

Susser, M., 1998. Does risk factor epidemiology put epidemiology at risk? Peering into the future. Journal of Epidemiology & Community Health 52, 608–611.

Susser, M., Susser, E., 1996. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. Am J Public Health 86, 674–677.

Thornton-Wells, T.A., Moore, J.H., Haines, J.L., 2004. Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. 20, 640–647.

Waller, L.A., 2004. Applied spatial statistics for public health data, Wiley series in probability and statistics. John Wiley & Sons, Hoboken, N.J.

Weiss, R.A., McMichael, A.J., 2004. Social and environmental risk factors in the emergence of infectious diseases. Nat. Med. 10, S70–76.

Werneck, G.L., Costa, C.H.N., Walker, A.M., David, J.R., Wand, M., Maguire, J.H., 2006. Multilevel modelling of the incidence of visceral leishmaniasis in Teresina, Brazil. Epidemiology and Infection 135, 195.

Woolhouse, M.E.., Howey, R., Gaunt, E., Reilly, L., Chase-Topping, M., Savill, N., 2008. Temporal trends in the discovery of human viruses. Proceedings of the Royal Society B: Biological Sciences 275, 2111–2115.

Woolhouse, M.E.J., Gowtage-Sequeria, S., 2005. Host range and emerging and reemerging pathogens. Emerging Infect. Dis. 11, 1842–1847.

Yang, J., Zhao, Z., Li, Y., Krewski, D., Wen, S.W., 2009. A multi-level analysis of risk factors for Schistosoma japonicum infection in China. International Journal of Infectious Diseases 13, e407–e412.

Yang, L., Wong, C., Chan, King, Chau, P., Ou, C., Chan, Kwok, Peiris, J.M., 2009. Seasonal effects of influenza on mortality in a subtropical city. BMC Infectious Diseases 9, 133.

Yusuf, O.B., Adeoye, B.W., Oladepo, O.O., Peters, D.H., Bishai, D., 2010. Poverty and fever vulnerability in Nigeria: a multilevel analysis. Malar. J. 9, 235.

Zhao, L.P., Kolonel, L.N., 1992. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. Am. J. Epidemiol. 136, 464–474.