Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2006;35:513–514
doi:10.1093/ije/dyl101

EDITORIAL

Everything should be made as simple as possible but not simpler

Rodolfo Saracci

Can complexity theory throw some different light on the aetiology of complex diseases,[1] currently explored mostly by the probe of molecular genetics?[2] Epidemiology, particularly as developing in the last half a century, has been dealing with disease aetiology—'*a priori*' not known whether simple or complex—with quite simple tools from an epistemological viewpoint. Observational epidemiology studies of aetiological factors are conceived, and whenever feasible carried out, as association studies at the individual level. An association and its nature, causal or non-causal, is researched within the same individuals of one or more exposures of interest with an outcome adjusting for other exposures, which may distort (confound) the association. This basic and invariant study concept has been developed into a vast and sound array of methods of study design and statistical analysis aimed at (i) making it applicable within a variety of purely observational circumstances and (ii) approaching the same study validity attainable by an experiment in which exposures are selected and assigned at will by the investigator and randomization is employed to control confounding and biasing factors unknown or known but not controlled by systematic arrangement.

This approach is in principle universally applicable and robust, as it does not depend on the specific nature of the exposures and of the disease under study or on any formal and quantitative model of disease causation. Only a very limited number of such models have in fact been produced, the best known being the multi-stage models of carcinogenesis initially proposed by Armitage and Doll,[3] useful in refining but not crucial for the interpretation of the data from the association studies. Ironically for a discipline defined as 'The study of the distribution and determinants of health-related states and events in specified populations…'[4] the research approach operates in fact at the level of individuals, whose disease risk produced by an exposure is estimated using the collection of conceptually identical and independent subjects called a 'population' (often a sub-group of the real population of interest). The distributions of the exposures and outcomes in the population and their evolution in time are regarded as just descriptive data providing only suggestive or ancillary evidence on causation, always infiltrated by the possibility of an ecological fallacy. In a paradigmatic study the real population is stripped of all its history to extract, as closely as possible to a controlled experiment, selected exposure values for individuals and to find out what frequencies of outcome correspond to them. Two domains escape this restriction: infectious disease epidemiology and population genetics, in which models of transmission of putative causal agents contribute substantially to their identification as actual agents. Still in both fields the final word goes to the association method at the individual level, as shown by the studies on HPV viruses and cervical cancer[5] or *Helicobacter pylori*[6] and gastric cancer, and by the recent major shift in emphasis towards association studies, both exploratory and confirmatory, in genetic epidemiology. A second stripping of possibly available information may occur when moving from the population level, above the individual, to the levels below the individual as a unit, namely when biological markers are the exposures under investigation. Common models of data analysis, as logistic, Poisson or proportional hazard models, do not take into account (or only partially in Cox's model with time-dependent covariates) the time sequence, the interconnections and the hierarchical level within the body physiology of the various markers potentially involved in disease pathogenesis. Basically they are 'omnibus' models, in which the interrelationships between exposures collapse into first order (two factors) or at the maximum second order (three factors) interaction terms in a linear model (higher-order interactions are very hard to assess even in balanced randomized factorial designs).[7]

It is tempting to speculate that incorporating in a comprehensive analysis whatever information may be available (and usually wasted) both at population level and at physiological level may improve the ability to elucidate the aetiology of complex diseases: and complexity theory may be a place to look for this purpose. Complexity theory[8,9] means different things in different scientific fields, sometimes the word being loosely used. In biology it refers to the approach dealing with the 'emergence' of new measurable properties of living organisms as one moves up from the molecular to the cellular, organ, system, organism, and population levels. In the simplest formulation each level, although formed by elements of the level immediately below, exhibits properties that cannot be explained by the properties of such elements *taken independently*: it has instead to be treated as a 'system' of multiple interacting elements (hence also the label of 'system biology'). Paradoxically for a theory of 'complexity' the approach derives its quantitative modelling power from the methodologically—as opposed to substantively—reductionist insights of the statistical mechanics founders Maxwell and Boltzmann in the second half of the 19th century, developed and systematized at the start of the 20th century by Gibbs (interestingly his theoretical statistical treatment is at the origin of 'Gibbs

Division of Epidemiology, IFC-National Research Council, Pisa, Italy. E-mail: saracci@hotmail.com

sampling' already employed in some advanced analyses of biomedical and epidemiological data).[10] The thermodynamic problem at stake was to find out how temperature affects the individual trajectories and collisions of the huge number of invisible molecules of a gas that hit the walls of a container and give origin to a measurable pressure (being a well-established law that at the macroscopic level the pressure increases with temperature at a constant volume of the container). The problem was solved by discarding the hopeless task of modelling the individual behaviour of each molecule, assumed to move at random in all directions, and by finding out how their collective behaviour, synthesized in their velocities (hence their kinetic energy generating the pressure), depends on temperature. Maxwell modelled such collective behaviour as probability distributions of molecular velocities, which increase both in mean value and scatter with increasing temperature, and Boltzmann put their derivation on a firm theoretical basis. In these models the gas molecules were considered as perfectly elastic on collision and in practice dimensionless. van der Waals went one step forward and by introducing molecule size and an attractive force active at close intermolecular distance extended a similar statistical treatment to liquids. Even more crucially he was able to show that—within a range of the relevant parameters—the change from gas to liquid status that a population of molecules undergoes with increasing pressure or decreasing temperature takes place in an abrupt way, namely with a minimal change of pressure or temperature once a given critical level of these parameters has been reached. This non-linear response to a minimal, perhaps accidental, change of the stimulus that results in a complete change of status when the system is in the neighbourhood of a threshold is a central process in complexity theory. It goes under the name of 'equilibrium transition phase' when it takes place in equilibrium static systems, like a gas or a liquid in a container, or 'non-equilibrium transition phase' when taking place, for instance, in living systems that are not static but steady state and energy consuming systems. Multiple bifurcations between alternative states, some precipitated by mere random fluctuations around threshold values of key parameters, may occur in the history of a system (physical, biological or social) making its time evolution hardly predictable. As Borges wrote: 'Time perpetually bifurcates towards innumerable futures'.[11]

At each level of aggregative organization—molecules, bacterial cells, multi-cellular organisms including the human body, human populations—phenomena have been observed or simulated,[9] which appear, as the macroscopic properties of a gas, to derive from the collective behaviour of the elements at the lower level(s), as dictated by the ground rules of interaction of such elements: perfect elasticity, size, van der Waals attractive forces for molecules; ability to self-propel for bacteria; directionality and obstacle avoidance in a schematic population of humans walking in a street. The resulting phenomena range from the characteristic shape taken by expanding bacterial cultures[12] to metabolic networks[13] in several species to preferential walking patterns.[14] Crime rates have been simulated in relation to population deprivation indexes, exhibiting a behaviour in which a small change in deprivation level may induce a jump from a high to a low level of crime rates or vice-versa.[15]

In aetiological epidemiology a complexity theory perspective on metabolic networks may help in building a biologically rational selection of exposures and biomarkers for association studies. The same perspective on population may help to understand to what extent differences in disease occurrence between populations may be explained not only by the distribution of individual risk exposures and group exposures with contextual effects at a point in time but also by the history of these exposures including critical (transition phase type) changes intervened in the distributions of the group exposures. To be able to examine and reconcile formally and quantitatively the findings at the different levels of observation is an important step to strengthen the credibility of the causal role of putative aetiological factors. At the very least exploring these perspectives will allow charting more accurately the gaps in our dominant research approaches in terms of data needed, measuring techniques, and methods of analyses. Aetiological epidemiology has been by and large reliant on a robust and epistemologically simple investigative approach; however, in the words attributed to Einstein,[16] 'Everything should be made as simple as possible, but not simpler'.

References

1 Pearce N, Merletti F. Complexity, simplicity and epidemiology. *Int J Epidemiol* 2006;**35**:515–19.

2 Buchanan AV, Weiss KM, Fullerton SM. Dissecting complex disease: the quest for the philosopher's stone? *Int J Epidemiol* 2006;**35**:562–71.

3 Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* 1954;**8:**1–12.

4 Last J. *A Dictionary of Epidemiology*. 4th edn. Oxford: Oxford University Press, 2001, p. 62.

5 International Agency for Research on Cancer. *IARC Monographs on the evaluation of the carcinogenic risks to humans. Human Papilloma Viruses.* Vol. 64. Lyon: International Agency for Research on Cancer, 1995.

6 International Agency for Research on Cancer. *IARC Monographs on the evaluation of the carcinogenic risks to humans. Schistosomes, Liver flukes and Helicobacter pylori.* Vol. 61. Lyon: International Agency for Research on Cancer, 1991.

7 Cochran WG, Cox GM. *Experimental Designs*. 2nd edn. New York: J Wiley & Sons, 1966.

8 Ruelle D. *Chance and Chaos*. Princeton: Princeton University Press, 1991.

9 Ball P. *Critical Mass*. London: Arrow Books, 2004.

10 Gilks WR, Clayton DG, Spiegelhalter DJ *et al*. Modelling complexity: application of Gibbs sampling in medicine. *J R Stat Soc B* 1993;**55:** 39–52.

11 Borges JL. *The garden of bifurcating paths.* Italian translation in: Borges JL. Tutte le opere. Vol. 1. Mondadori: Milano, 1991, p. 701.

12 Matsuyama T, Matsushita M. Fractal morphogenesis by a bacterial cell population. *Crit Rev Microbiol* 1993;**19:**117–35.

13 Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000;**407:**851–54.

14 Helbing D, Keltsch J, Molnar P. Modelling the evolution of human trail systems. *Nature* 1997;**388:**47–49.

15 Campbell M, Ormerod P. Social interaction and the dynamics of crime. Available at: http://scholar.google.com/scholar?g=campbell+m+AND+crime (Accessed April 11, 2006).

16 Calaprice A (ed.). *The Expanded Quotable Einstein*. Princeton: Princeton University Press, 2000, p. 314.