Kinetics methods for clinical epidemiology problems

Alexandru Dan Corlan[a] and John Ross[b,1]

[a]Research Department, University and Emergency Hospital of Bucharest, 5 Bucharest, Romania; and [b]Chemistry Department, Stanford University, Stanford, CA 94305

Edited by R. Stephen Berry, University of Chicago, Chicago, IL, and approved October 1, 2015 (received for review June 4, 2015)

Calculating the probability of each possible outcome for a patient at any time in the future is currently possible only in the simplest cases: short-term prediction in acute diseases of otherwise healthy persons. This problem is to some extent analogous to predicting the concentrations of species in a reactor when knowing initial concentrations and after examining reaction rates at the individual molecule level. The existing theoretical framework behind predicting contagion and the immediate outcome of acute diseases in previously healthy individuals is largely analogous to deterministic kinetics of chemical systems consisting of one or a few reactions. We show that current statistical models commonly used in chronic disease epidemiology correspond to simple stochastic treatment of single reaction systems. The general problem corresponds to stochastic kinetics of complex reaction systems. We attempt to formulate epidemiologic problems related to chronic diseases in chemical kinetics terms. We review methods that may be adapted for use in epidemiology. We show that some reactions cannot fit into the mass-action law paradigm and solutions to these systems would frequently exhibit an antiportfolio effect. We provide a complete example application of stochastic kinetics modeling for a deductive meta-analysis of two papers on atrial fibrillation incidence, prevalence, and mortality.

chemical kinetics | paradigm | noncommunicable disease | epidemiology | stochastic model

Much of the medical progress over the last century can be attributed to the objective assessment of the effect of treatments on the evolution of specific diseases. Treatment effect is measured as the rate of an event such as recovery in a sample of the patient population. Relatively immediate results were obtained from studies involving acute diseases, occurring in previously healthy individuals, in which recovery could be clearly identified. This resulted in the development of effective treatments for most acute diseases affecting children and younger adults and a substantial prolongation of life expectancy (1). Consequently, many acute diseases were treated effectively. This led to the current, more complex situation, in which an elderly population suffers from a combination of chronic conditions. Few older people are strictly healthy, and besides the evolution of the chronic conditions themselves, acute diseases occurring in this setting do not always evolve as in a young, healthy population. This combination of chronic conditions and risk factors amounts to the presence of more heterogenous populations. Thus, samples need to be larger to allow reproducible predictions, compared with those for acute diseases occurring in a young and previously healthy population. Predictions that are also more complex (there is no strict "recovery") apply to a limited range of cases.

The concepts used by clinicians and epidemiologists to describe the health status of individuals and their prevalence in the population, as well as the rates of change in this status and the general predictive laws, are quite analogous to the concepts used by chemists for predicting the future concentrations of species in a reactor. Early works on the spread of epidemics of communicable diseases (2, 3) made reference to this analogy, unlike later developments in mathematical epidemiology (4, 5). The purpose of current mathematical modeling of epidemiology is mostly the kinetics (or "dynamics" as it is frequently called) of the spread of a communicable disease in an acute epidemic (6), an important and pressing problem when such epidemics occur. The primary phenomenon represented in these models is contagion. The models used are typically deterministic, self-catalytic kinetic models of the whole population, with mass-action law assumption.

The focus of clinical studies in chronic diseases is the risk for various possible outcomes for the patient that are usually distinct from a complete recovery and sometimes of a quantitative nature—for example, how much of the function of an organ is preserved. Kinetics-type mathematical support for this purpose is rarely available, other than a simple statement of risk or relative risk that is directly inferred from a study in a sample.

Deterministic event rates are the basis of virtually all clinical judgement. A typical example is, What is the yearly risk of stroke in patients with atrial fibrillation on either of two treatments, such as warfarin or aspirin? (More individual parameters are usually taken into account when classifying each patient). A lower yearly risk rate in one of the treatment categories is an argument to choose that treatment for a particular patient. This risk rate is usually the rate of stroke that has been directly measured in a study where a sample of patients belonging to certain classes (for example, middle-aged males with atrial fibrillation and no history of stroke) has been followed for some time. The observed event rates are taken as the best estimations for the population sharing the same characteristics as the patients in the sample, a population that is presumed infinite. The event rates are inferred, however, starting from observations made in finite, small, ensembles of individuals (case series). Thus, inference is always probabilistic, as event rates in the population can be estimated only with some uncertainty, even in homogenous populations.

In populations in which individuals have various combinations of underlying pathologies that may each influence the future event rates, this approach may frequently lead to unreproducible

Significance

Medical predictions rely on probabilistic models that link the risk of future events to diagnostic and therapeutic variables. These models result from epidemiologic studies based on direct observation of events and variables in population samples. They work well for acute diseases in otherwise healthy individuals. For chronic, multiple conditions and diseases, more complex predictive models are needed. Such models have been developed for chemical kinetics problems that are analogous to epidemiologic problems if we view molecules as individuals and macroscopic systems as populations: Deterministic kinetics with one reaction are like most direct epidemiologic studies; the Arrhenius model is analogous to survival analysis. Stochastic systems with time-dependent reaction coefficients are worth exploring for deriving new epidemiologic methods.

Author contributions: A.D.C. and J.R. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: john.ross@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1510927112/-/DCSupplemental.

results (7, 8). Aiming to predict events that would occur in the more distant future, as needed with chronic disease, further complicates the problem: The longer the prediction time is, the higher the number of other events that may intervene and invalidate the prediction.

Both epidemiology and chemical kinetics have evolved independently over the previous decades, each developing its own stochastic methods with a specific terminology that frequently refers to somewhat similar concepts. Prediction of event rates from relatively small samples, using probabilistic models, is of primary importance for epidemiology. Half a century ago, the factors influencing the recovery from acute diseases in otherwise healthy (that is, homogenous) populations were the main concern. Thus, the problem was to estimate event rates in otherwise simple systems. Models of the epidemiologic equivalent of a single reaction, with a few other parameters, were adequate for this. Probabilistic issues were mostly related to the errors associated with the limited sizes of the samples used, but the inferred rates were typically deterministic. In chemistry, at that time, model development focused on identifying the relatively complex reaction mechanisms that occur, even when only a few initial species are involved, and on describing their kinetics, typically with systems of deterministic differential equations adjusted using macroscopic measurements of species concentrations. Uncertainty due to small molecule numbers was not usually involved. Over the last 70 y, however, chemical kinetics developed new methods, such as models of more complex systems that do not rely on mass-action law (9) or models of single-molecule kinetics that might be closer to the problem of predicting clinical evolution in individual patients. Also, issues that occur in the biochemical kinetics of more complex systems, such as crowding (10), are to some extent analogous to event prediction in heterogenous populations.

The development of numerical methods allows the practical approach to problems that involve systems that are both complex and stochastic and the exploration of uncertain phenomena at both individual and population levels (11). An important clinical problem that cannot, in general, be solved without such a systematic approach is to compute, for an individual, the risks for each possible disease over the next time interval (such as 1 y), given what we know about his or her health status and history and based on currently available epidemiologic data. Solving this problem would allow much more accurate planning of clinical interventions than is possible today.

In this paper, we attempt to compare concepts, methods, and models that have been developed in the two fields, by reformulating the epidemiologic approaches in chemical kinetics terms, to identify chemical kinetics methods that might be adaptable for epidemiologic use. In *Supporting Information*, we show an example of a deductive meta-analysis of two epidemiology papers, using a stochastic kinetic system.

Unified Terminology

We use the following unique terms to describe either chemical systems or populations of patients: individual, either an individual molecule or a patient; ensemble, a set of individuals; population, a real or imaginary ensemble large enough for deterministic mass laws to occur; sample, a small ensemble in which measurements are performed; species, a binary, yes/no criterion, for classifying individuals and also the subsystem comprising individuals with "yes" for the criterion [it could be a covalent structure for molecules such as warfarin or a diagnostic class (such as "diabetic") for a patient]; subspecies, a species that is a subset of another species, for example "insulin-dependent diabetic" as a subspecies of diabetic for a patient or a warfarin enantiomer (R or S) as a subspecies of "warfarin" in chemistry; elementary species, a species that contains no subspecies; coverage, a selection of species such that each individual in a population belongs to exactly one species; prevalence is the equivalent of relative concentration, the proportion of individuals of a species in an ensemble; parameter, another measurable term characterizing either the system or each individual, such as temperature or pressure for chemical systems or age, weight, and fasting glycemia for patients; transition, the probabilistic transformation of an individual from one species to another; reaction, the macroscopic process of transition of individuals between two species, possibly going through a number of intermediate species; elementary reaction, a reaction that consists of a single transition, without intermediates; rate is the derivative of the prevalence in time, due to one reaction; rate coefficient is a number, or a function, that characterizes the rate after removing its dependency on the prevalence (concentration); evolution, the succession of species through which an individual progresses in time (for example: susceptible → infected → recovered); system, a theoretical construction, comprising individuals, species, and reactions; and process, the successive states undergone by a system between two time instants.

The Clinician's Problem

Restated in these terms, the purpose of clinical research is the prediction of the evolution (state or species transition) of individuals, given variable values at the initial time. The purpose of epidemiological research is the prediction of the prevalence of each species in time, given an initial variable value distribution. An example problem is, Given what we know about an individual, what is the probability of each possible health outcome (status belonging to a species from a coverage) for him or her after a definite period? For example, given a 45-y-old man with uncomplicated diabetes, free of any other disease, on adequate treatment, what is the probability of each disease classification from a coverage 1 y in the future? A simple example of coverage could be (*i*) alive, healthy, and diabetes free; (*ii*) the same status as now; (*iii*) having a new disease; or (*iv*) dead. The clinical problem, for an individual, is an aspect of the epidemiological problem, for a population, just as chemistry at the macroscopic level is another view of molecular collisions and transformations at the molecular level. The answer to the general clinical problem stated above is typically a probabilistic translation (a "risk" law) derived from an epidemiologic law that was, in its turn, directly inferred from patient sample measurements.

What one cannot usually do, using the current epidemiologic methods, is the equivalent of the prediction of future species concentrations in a chemical reactor: Compute all of the disease and complication risks of the individual, at any time over the next years, such that the sum of the risks is 1 and the result is consistent with experimental data. Such a prediction is usually possible only for the simplest cases: acute diseases in otherwise healthy people and only for the short term.

Types of Data Available from Epidemiology Studies

In chemistry, both measured data and models deal mostly with macroscopic variables (concentrations in time from which rates over finite durations are computed). In epidemiology, models are also at the population level (prevalences, incidences, risk rates) but the data are available only for small samples, starting from measurements in individuals. Measurements consist of diagnostic classifications (such as "atrial fibrillation") and continuous parameters (such as "serum glucose level"). Sometimes, diagnostic classifications refer to ranges of the continuous parameters, for example "bradycardia," "normocardia," and "tachycardia" may refer to a heart rate under 50/min, between 50/min and 100/min, and over 100/min. We denote by x the vector of these n measurements for an individual and by S the set of possible values of x, which is a contiguous subset of $\mathbf{R}^n$, $x \in S$. In a typical experiment, individuals are selected to belong to some subset of S, for example having a certain disease with certain parameters. They are further divided into subgroups, for example by randomly subjecting them to various treatments, say S_0, S_1. Each of these subsets corresponds to a species in the unified terminology. Species dealt with in one study may be subspecies of species from other studies or may overlap in various ways. In time t, each individual undergoes changes in his or her status, $x(t)$. Each individual is followed from an initial instant, T, over a finite time interval Δt, to see whether he or she changed into some other subset, named endpoint S_e—for example, if the individual became free of one of the diseases he or she originally had or if he or she died or developed a specific new disease.

The observed frequencies of transition to new states are inferred as rates of transition for the broader population of individuals belonging to S_0 or S_1 into S_e over the same finite time interval; that is,

$$p_i = p(x(T+\Delta t)\in S_e | x(T)\in S_i), \quad i=0,1. \qquad [1]$$

In other words, the vast majority of clinical studies that assess the p_i for every S_i and S_e aim in fact to document the master equation that governs the $x(t)$ stochastic function. Due to small numbers in samples, the inferred p_i are uncertain, with a confidence interval computed using the binomial distribution that corresponds to the process of random sampling. Alternatively, the probability of the null hypothesis, $P(p_0 = p_1)$ is reported; a P below a threshold, such as 0.05, results in the conclusion that, for example, $p_0 > p_1$. The accuracy of estimation of p_i depends on the product of Δt and the number of cases considered in each study.

Assumptions of Clinical-Epidemiologic Models

Some common assumptions are necessary to extrapolate the observed frequencies to the entire population: (i) that the components of x represent all of the relevant determinants of future values of $x(t)$, this naturally implying that they are all of the known relevant determinants; as a special case, that the previous history of the individual, that is, $x(t), t < T$, does not influence p_i; (ii) that $S_{0,1}$ are properly sampled by the cases reported in the study, and thus p_i refers to any future individual from $S_{0,1}$; (iii) that p_i do not change over the next years; that is, the current sample is also relevant to future populations (to which the conclusions of the study will be applied); (iv) that for at least some relevant transitions, the transition rate is constant in time or at least that the rate differences observed over Δt are relevant over different time ranges; for example, if a drug is observed to be effective and safe for a disease at exactly 1 mo of treatment, then recommending it for that disease implies the assumption that it will also be safe and effective after 2 wk or 2 mo; although $p_0(\Delta t)$ and $p_1(\Delta t)$ may not be strictly constant, an assumption is made that, for example, at least $p_0(\Delta t) > p_1(\Delta t)$ for a meaningful range of Δt; and (v) that the population from which the samples are drawn is infinite, or large enough to be treated as infinite, and thus state transition or reaction rates make sense as macroscopic properties.

The assumption that, for individuals in a region S_0 of the parameter space, the probability of transition to another region, S_e, is constant over successive Δt intervals is equivalent to the assumption of mass-action law at the population level. This assumption is explicit in most mathematical epidemiology models, and it is also present in many common judgments that involve, for example, comparing risk rates over specific intervals of time. Many patient evolutions are, however, inconsistent with this assumption. One example is progression through age groups: 40-y-old individuals do not have a constant daily, or monthly, risk of progressing to being 41 y old. Instead, the progression rate is 0 except for the 41st birthday when it is 1. (The same is true irrespective of what time interval we choose for an age group, when we choose age groups with fixed limits as species).

Another example is the cure of an acute infectious disease. The duration of the illness can be given, for example, as a median of 7 d, with an interquartile range of 3–10 d (12). This distribution is inconsistent with a constant rate coefficient, because the rate of cure would then have to be the highest in the first days of the disease. Indeed, the rate of cure is not constant, but given by the time needed for the specific immune response to occur, that is, around 1 wk in most healthy individuals.

Other examples not consistent with constant rates include the time-dependent risks following an acute event. For example, following an acute myocardial infarction, the risk of death is much higher in the first days than in the following days or months.

Thus, the type of kinetics that are suitable for modeling many situations needs to consider reaction coefficients that are functions of time or, at the individual level, stochastic chains that have memory, rather than the mass-action law framework. In some cases, a process that does not conform to mass-action law can still be modeled using mass-action law by assuming that there is more than one species, each with its own constant reaction coefficient. For example, following an acute event, the species C suffering the event E may be split, with a specific probability, in "high-risk" (C_h) and "low-risk" (C_l) strata (subspecies) having constant reaction coefficients. The kinetics of $C \to E$ will not obey mass-action law, but the components, $C_h \to E$ and $C_l \to E$ will.

Methods such as hidden Markov or, more generally, inhomogenous stochastic chains have been used occasionally to approach some epidemiologic problems (13–16).

The General Problem

In general, we need to develop theoretical kinetic systems that are solved in the form of computed results of various hypothetical studies. Parameters of the kinetic systems need to be tuned so that computed results match existing studies. A proposed kinetic system is valid as long as it predicts the results of future studies that were not used for tuning it. In the most general case, these systems would be stochastic reaction–diffusion systems in which reaction (and diffusion) coefficients are stochastic functions of time. Because the history of each individual from the system may influence its current transition probabilities, the solution needs to involve simulation at the individual level. Matching with experimental data must, however, be done at the population (macroscopic) level, as experimental results are available as population models. An example application, in which we check, quantitatively, the consistency of some experimental results (17–19) and some epidemiologic hypotheses (17), using a kinetic system, is shown in *Supporting Information*.

Description of Time-Dependent Reaction Coefficients

A reaction

$$\mathrm{A} \xrightarrow{q(t)} \mathrm{X}, \qquad [2]$$

where $q(t)$ is not constant, can be written at the individual level as a set of transitions

$$\mathrm{A_i} \xrightarrow{k_i} \mathrm{X} \qquad [3]$$

$$\mathrm{A_i} \xrightarrow{\delta_i} \mathrm{A_{i+1}} \qquad [4]$$

$$\ldots \mathrm{A_n} \xrightarrow{k_n} \mathrm{X}, \qquad [5]$$

where $\mathrm{A_i}$ are subspecies of A, k_i are different transition coefficients that are, each, constant in time, and thus $\mathrm{A_i} \xrightarrow{k_i} \mathrm{X}$ are memoryless transitions, corresponding to constant reaction coefficients. δ_i represents a special type of reaction, representing only the memory phenomenon, in which the transition rate is 0 until a specific instant, Δt_i, when it becomes 1. The time interval Δt_i, for each individual in species $\mathrm{A_i}$, flows from the instant when it transitioned into species $\mathrm{A_i}$.

This description is equivalent to averaging $q(t)$ on intervals.

Typical Epidemiological Studies and Their Chemical Equivalents

Transversal Studies. Studies called "transversal," meaning "performed across a population or population sample at a given instant in time," correspond to the determination, at a specific instant, of the concentrations of a species in a reactor or of the instantaneous rate of production or consumption of that species. Sometimes transversal studies consider the whole population (this type of study is sometimes called "screening") that is usually large, such as the population of a country, but of course finite. Frequently, the study is performed in a representative sample and the observed prevalences are extrapolated with confidence limits; that is, prevalences in the population are determined as probability distributions. Transversal studies aim to estimate the concentrations $[A](t_0), [B](t_0), \ldots$ at instant t_0. Sometimes an assumption of stationarity is made, so measurements of prevalences (concentrations) performed at one time are supposed to directly predict the prevalences a few years, or decades, later or $[A](t) = [A](t_0)$ for any t.

Longitudinal Studies. Longitudinal studies are repeated transversal studies. Simpler studies are just verifications of the fact that the prevalences are, or are not, stationary. A null hypothesis that they are stationary in the population may be rejected with a statistical test on sample data.

Some long-term surveys monitor the prevalences and incidences of important disease classifications (species) yearly over long periods of time in the whole population of a country. In our notation, this type of study attempts to determine the probability $p([A](t) > [A](t_0))$ for certain values t.

Cohort Studies. Cohort studies are longitudinal studies in which the same ensemble, consisting typically of individuals belonging to a single species, is followed for a longer time, and the rates of transition to other species, in time, are recorded. They correspond to observing reactions in reactors where the initial concentrations are well known, for example a single, pure substance, the decomposition of which is monitored in time. In other words, a reaction $A \rightarrow E$ is monitored, and the result of the study is an estimation of $[A](t) - [A](t_0)$ or $[E](t) - [E](t_0)$ or, if successive measurements are made at regular intervals Δt, they are determinations of

$$\overline{k}_A(t) = \frac{\Delta[A]}{\Delta t}(t) \quad [6]$$

or, rather, the probability distribution $p(\overline{k}_A(t))$, where the $\overline{k}_A$ is the integral rate over some time interval.

Controlled Studies. Control studies consist of comparing rates in two or more cohort studies in parallel. The differences in rates may be attributed to differences in the initial parameters of the populations that may be introduced (for example, as treatments) by the researchers. Prevalences are usually measured at specific time instants, for example every month, or just at the start and the end of the study. The purpose of a control study, put in kinetic terms, is to compare two integral rates (differences in prevalences) over a specific interval of time. For example, we assume the reactions $A \xrightarrow{k_A} E; B \xrightarrow{k_B} E$. We try to estimate the probability that $\overline{k}_A(t) > \overline{k}_B(t)$ from measurements of $[A], [B], [E]$ at various time instants.

Survival Studies. Survival studies are control studies in which the timing of the transition from a species to another is recorded for each individual. In a typical approach, using the Cox proportional hazards model (20), an assumption is made that, with the notation above, $k_B(t) = uk_A(t)$, where u is a constant. The purpose is to calculate the probability $p(u < 1)$, thus inferring that $k_A(t) > k_B(t)$, and also to estimate $p(u)$, sometimes called "relative risk," that is typically given as an estimate with confidence intervals.

In more sophisticated survival studies, u is given as an exponential of a linear combination of constant population parameters, to estimate the relationship between rate coefficients of numerous population subgroups, for a specific reaction. This is apparently analogous to the Arrhenius model (law). The Arrhenius model relates the rate coefficient to the (inverse of) the temperature with a formula like

$$k = Ae^{-E_a/RT}, \quad [7]$$

where k is the rate coefficient, E_a is the activation energy (specific to the reactant species), R is the gas constant, T is the temperature, and A is a fitting constant.

Proportional hazard models are formulated as

$$k = k_A e^{\beta X}, \quad [8]$$

where k_A from survival models corresponds to the prefactor A in the Arrhenius equation. The influencing variable X corresponds to the (inverse of the) temperature T^{-1} and the reaction-specific constant β corresponds to the activation energy $-E_a/R$. As mentioned above, in some studies, X may be a linear combination of some measurable parameters that are supposed constant for a population subgroup (such as gender, exposure to some environmental factor, genetic variant, etc.).

Theoretical Integration of Experimental Results. The types of studies described in the previous section are experimental studies for which simple, empirical models are fitted to datasets, equivalent to straightforward chemical or chemometric studies.

The main contribution that might be achieved through analogy with chemical kinetics refers to theoretical integration of such results. Theoretical integration of experimental data in epidemiology can be divided into two categories: meta-analysis and mathematical epidemiology studies. Meta-analysis refers in essence to reducing the uncertainty regarding the estimation of a rate coefficient of a single reaction by pooling the results from multiple studies. Mathematical epidemiology studies usually deal with the phenomenon of contagion and use rate estimates (without confidence intervals) resulting from a number of single-reaction studies to build systems of a few reactions and predict their behavior at the population level.

Predicting the risk profile of a specific individual at a given instant, the prototypical problem we specified in the Introduction, requires a system of many different reactions, corresponding to many combinations of the numerous types of pathologies and evolutions. As the available data are always in relatively small samples, any rates are probabilistic estimations rather than deterministic values observed in a macroscopic setting. This uncertainty needs to be taken into account.

Issues Resulting from System Complexity. The example below illustrates some of the predictive limitations that arise from not taking into account sufficient aspects of system complexity in the design of a predictive model.

Consider a cohort of initially healthy individuals with two subgroups: low-risk rate k_0, for example younger, $\mathrm{A_0}$, and high-risk rate $k_1 > k_0$, for example older, $\mathrm{A_1}$. A pathologic condition, denoted as species F, may occur in subjects from either group, but with a higher rate in A_1:

$$\mathrm{A}_0 \xrightarrow{k_A} \mathrm{A}_1 \qquad [9]$$

$$\mathrm{A}_0 \xrightarrow{k_0} \mathrm{F} \qquad [10]$$

$$\mathrm{A}_1 \xrightarrow{k_1} \mathrm{F} \qquad [11]$$

Suppose an investigator measures the initial ($t_0 = 0$) prevalences of A_0 and A_1. We denote them as $a_0(t_0), a_1(t_0)$. At time $t_1 > t_0$ the prevalence of condition F is $f(t_1)$.

He or she draws the conclusion that the incidence of condition F over $t_1 - t_0$ y is

$$\phi_1 = \frac{f(t_1)}{a_0(t_0) + a_1(t_0)}. \qquad [12]$$

If the sampling has been adequate, and the study properly conducted, this conclusion would be directly extrapolated to the evolution of similar populations [with the same $a_1(t_0)/a_0(t_0)$ initial ratio] over time interval t_1.

However, the incidence of F in this cohort increases in time, as subjects move into the high-risk category. Consequently, the estimation over interval t_1 cannot be extrapolated to shorter or longer intervals or to populations in which the initial a_1/a_0 ratio is different. For example, if the estimated incidence is $\phi_1 = 0.2$ over 10 y, then it will be lower than 0.1 over 5 y and higher than 0.3 over 15 y.

In a common setup, we would also know $a_0(t_1)$ and $a_1(t_1)$. However, this is sometimes not the case, as the investigations needed to make the initial A_0 vs. A_1 classification may be expensive or otherwise impractical at time t_1.

Assuming first-order kinetics, we can write the kinetic equations for the system,

$$\frac{da_0(t)}{dt} = -k_A a_0(t) - k_0 a_0(t) \qquad [13]$$

$$\frac{da_1(t)}{dt} = k_A a_0(t) - k_1 a_1(t) \qquad [14]$$

$$\frac{df(t)}{dt} = k_0 a_0(t) + k_1 a_1(t) \qquad [15]$$

$$a_0(0) = \alpha_0, \quad a_1(0) = \alpha_1, \quad f(0) = 0, \quad f(t_1) = \phi_1, \qquad [16]$$

where $\alpha_0, \alpha_1, \phi_1, t_1$ are given constants. A simplistic linear extrapolation of f would be

$$f*(t) = \phi_1 \frac{t}{t_1}. \qquad [17]$$

The error in using a simplistic linear extrapolation is

$$E(t; k_A, k_0, k_1) = \frac{f(t) - f*(t)}{f*(t)} = \frac{f(t)}{\phi_1(t/t_1)} - 1. \qquad [18]$$

Usually, one type of transversal study (S1) would consist of examining a large number of patients that just suffered the event F and classify them by the a criterion (were they high or low risk, for example). In another type (S2) the prevalence of the risk class (a_0 vs. a_1) in the general population could be assessed. These frequencies would allow the direct estimation of k_1 and k_2. In yet a different kind of study (S3), the integral incidence ϕ_1 is measured for a given value of t_1; initial $a_0(t_0)$ and $a_1(t_0)$ are also known. The problem of the kinetic modeling of such a case could be to estimate k_A given the above frequency estimations. $\phi(k_A)$ is a monotonous function, and thus the equation $\phi(k_A) = \phi_1$ can be solved numerically for k_A using, for example, Newton's method.

In an imaginary example of the above with $a_0(0) = 0.9$, $a_1(0) = 0.1$, $t_1 = 10$ (y), $\phi_1 = f(t_1) = 0.2$, $k_0 = 0.01$ (/y), $k_1 = 0.05$ (/y) we found $k_A = 0.064$, $f(5) = 0.089$, and $f(15) = 0.314$, which means an error of about 10% would have been made for $f(5)$ with the linear approximation.

In summary, measurement of an event rate after a specific interval (such as 10 y) cannot, in general, be extrapolated to other time intervals or to populations having different distributions of initial parameters than the one in which the measurement was performed, if any of these parameters are correlated with the event rate. For example, an event rate estimated based on direct measurements in a sample of patients that are both younger and older, say between 50 y old and 70 y old, will erroneously estimate the risk for younger patients, say between 50 y old and 55 y old, and indeed it will imply a prediction error for a single individual. To avoid this error, a kinetic model of sufficient complexity needs to be developed.

Determinability of Reaction Coefficients. Prevalences in epidemiologic studies are usually inferred from (small) samples and are thus known with approximation. For simple mass-action systems, in which we take this uncertainty into account, the rate coefficients have probability distributions that may be calculated analytically. Consider the elementary reaction

$$\mathrm{A} \xrightarrow{k} \mathrm{F}. \qquad [19]$$

Let $a(t)$ be the prevalence (concentration) of species A and $f(t)$ be the prevalence of species F. We assume that $f(t_0 = 0) = 0$ and the prevalence $f(t)$ is estimated at time t_1. If we use a large number of cases, the binomial distribution of the real prevalence can be approximated as normally distributed, with the mean f_1 and the dispersion σ_1. As $f(t) = a_0 e^{-kt}$,

$$k = -\frac{\ln(f(t)/a_0)}{t}. \qquad [20]$$

If we take $a_0 = 1$ and f_1 is normally distributed, then $\ln f(t_1)$ has a lognormal (ln) distribution with mean

$$\langle \ln f_1 \rangle = e^{f_1 + \sigma_1^2/2}. \qquad [21]$$

Thus, k also has a lognormal distribution with mean

$$\langle k \rangle = \frac{e^{f_1 + \sigma_1^2/2}}{t}. \qquad [22]$$

Successive determinations of the rate of $f(t)$ at later time instants allow an iterative improvement of the distribution of k, using a procedure such as that described in ref. 21.

However, if we do not take a_0 as being known exactly, but rather as having a probability distribution, then the fraction $f(t)/a_0$ will exhibit an antiportfolio effect (22) that may reduce the possibility to estimate the value of k very much (the information content of the resulting distribution of k will become very low). A portfolio effect occurs when random variables are combined so that their result has higher information content, such as when we take the average of two measurements of the same noisy event. An antiportfolio effect occurs when combinations, such as multiplication of random variables, result in reduced information, as in solving the above stochastic equation.

The presence of an antiportfolio effect provides a previously unidentified explanation for the lack of reproducibility of studies estimating event rates, beyond the reasons stated in ref. 7.

Summary

The computation of the risks of pathological events facing an individual at any future instant in time is not possible with the currently widespread epidemiology theory except for the simplest problems. In general, it requires complex models of his or her possible evolution that are analogous with complex stochastic kinetics systems used in chemistry.

We explored the analogy between common types of clinical/epidemiologic studies and corresponding kinetic models.

Deterministic models that involve only a few reactions/species can be expected to be effective predictors only for the simplest cases, such as acute diseases occurring in otherwise healthy populations. Stochastic descriptions that correspond to a single reaction, directly inferred from sample measurements, that are common in current studies of epidemiologic processes are, by themselves, limited to prediction over a fixed time interval and for populations with a specific composition.

We showed that many common kinetic phenomena that correspond to epidemiologic processes cannot be expected to fit into the mass-action law paradigm. As reaction coefficients are necessarily inferred from small sample measurements, they are uncertainly known. Consequently, some of the more complex systems may exhibit an antiportfolio effect, with a spread of the probability of some solutions on a broad range. Thus, such systems may actually prove to be poor predictors for some problems, due only to this effect. Still, besides stochastic kinetics, no other well-developed theoretical framework is available to approach the essential problem of risk estimation for patients already suffering from multiple diseases and conditions.

ACKNOWLEDGMENTS. This work was supported, in part, by the National Science Foundation (CHE 0847073).

1. De Flora S, Quaglia A, Bennicelli C, Vercelli M (2005) The epidemiological revolution of the 20th century. *FASEB J* 19(8):892–897.
2. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc R Soc Lond A Contain Pap Math Phys Character* 115(772):700–721.
3. Muench H (1959) *Catalytic Models in Epidemiology* (Harvard Univ Press, Cambridge, MA).
4. Hairston NG (1965) An analysis of age-prevalence data by catalytic models. A contribution to the study of bilharziasis. *Bull World Health Organ* 33(2):163–175.
5. Cohen JE (1973) Selective host mortality in a catalytic model applied to schistosomiasis. *Am Nat* 107:199–212.
6. Anderson RM, May R (1992) *Infectious Diseases of Humans* (Oxford Univ Press, Oxford, UK).
7. Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8): e124.
8. Moonesinghe R, Khoury MJ, Janssens ACJW (2007) Most published research findings are false-but a little replication goes a long way. *PLoS Med* 4(2):e28.
9. Ross J, Mazur P (1961) Some deduction from a formal statistical mechanical theory of chemical kinetics. *J Chem Phys* 35(1):19–28.
10. Ryan TA, Myers J, Holowka D, Baird B, Webb WW (1988) Molecular crowding on the cell surface. *Science* 239(4835):61–64.
11. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys* 22:403–434.
12. Crum-Cianflone NF, et al. (2009) Clinical and epidemiologic characteristics of an outbreak of novel H1N1 (swine origin) influenza A virus among United States military beneficiaries. *Clin Infect Dis* 49(12):1801–1810.
13. Ross JV (2012) On parameter estimation in population models III: Time-inhomogeneous processes and observation error. *Theor Popul Biol* 82(1):1–17.
14. Gil J, et al. (2007) Disease progression and survival in ALS: First multi-state model approach. *Amyotroph Lateral Scler* 8(4):224–229.
15. Sweeting MJ, Farewell VT, De Angelis D (2010) Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. *Stat Med* 29(11):1161–1174.
16. Bureau A, Shiboski S, Hughes JP (2003) Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Stat Med* 22(3): 441–462.
17. Piccini JP, et al.; ROCKET AF Steering Committee and Investigators (2013) Renal dysfunction as a predictor of stroke and systemic embolism in patients with nonvalvular atrial fibrillation: Validation of the R(2)CHADS(2) index in the ROCKET AF (Rivaroxaban Once-daily, oral, direct factor Xa inhibition Compared with vitamin K antagonism for prevention of stroke and Embolism Trial in Atrial Fibrillation) and ATRIA (AnTicoagulation and Risk factors In Atrial fibrillation) study cohorts. *Circulation* 127(2):224–232.
18. Benjamin EJ, et al. (1998) Impact of atrial fibrillation on the risk of death: The Framingham Heart Study. *Circulation* 98(10):946–952.
19. Center for Disease Control and Prevention (1998) *Mortality Data* (Center for Disease Control and Prevention, Atlanta).
20. Cox DR (1972) Regression models and life tables. *J R Stat Soc B* 34(2):187–220.
21. Vlad MO, Corlan AD, Morán F, Oefner P, Ross J (2008) Incremental parameter evaluation from incomplete data with application to the population pharmacology of anticoagulants. *Proc Natl Acad Sci USA* 105(12):4627–4632.
22. Vlad MO, Corlan AD, Popa VT, Ross J (2007) On anti-portfolio effects in science and technology with application to reaction kinetics, chemical synthesis, and molecular biology. *Proc Natl Acad Sci USA* 104(47):18398–18403.