American Journal of Epidemiology

Vol. 170, No. 12
DOI: 10.1093/aje/kwp366
Advance Access publication: November 12, 2009

Invited Commentary

Invited Commentary: Evaluating Vaccination Programs Using Genetic Sequence Data

M. Elizabeth Halloran* and Edward C. Holmes

* Correspondence to Dr. M. Elizabeth Halloran, Center for Statistics and Quantitative Infectious Diseases, 1100 Fairview Avenue North, M2-C200, Seattle, WA 98109 (e-mail: betz@u.washington.edu).

Initially submitted July 22, 2009; accepted for publication August 21, 2009.

Genomic data will become an increasingly important component of epidemiologic studies in coming years. The authors of the accompanying *Journal* article, van Ballegooijen et al. (*Am J Epidemiol*. 2009;170(12):1455–1463), are to be commended for attempting to use the coalescent analysis of viral sequence data to evaluate a hepatitis B vaccination program. Coalescent theory attempts to link the phylogenetic history of populations with rates of population growth and decline. In particular, under certain assumptions, a reduction in genetic diversity can be interpreted as a reduction in disease incidence. However, the authors of this commentary contend that van Ballegooijen et al.'s interpretation of changes in viral genetic diversity as a measure of hepatitis B vaccine effectiveness has major limitations. Because of the potential use of these methods in future vaccination studies, the authors discuss the utility of these methods and the data requirements needed for them to be convincing. First, data sets should be large enough to provide sufficient epidemiologic-scale resolution. Second, data need to reflect sufficiently fine-grained temporal sampling. Third, other processes that can potentially influence genetic diversity and confuse demographic inferences should be considered.

communicable diseases; disease notification; disease transmission, infectious; genetic variation; hepatitis B virus; molecular sequence data; vaccination

Abbreviation: HBV, hepatitis B virus.

In their article, van Ballegooijen et al. (1) claim that viral gene sequence data analyzed with coalescent-based techniques offer a complementary tool to assess epidemiologic dynamics. In particular, they claim to have assessed the added value of using sequence data to monitor the effects of a vaccination program against hepatitis B virus (HBV). Genomic data will undoubtedly become an increasingly important component of epidemiologic studies in coming years, and the authors are to be commended for attempting to use this approach. However, we contend that their interpretation of changes in viral genetic diversity as a measure of vaccine effectiveness has major limitations. Because of the potential use of these methods in future vaccination studies, we discuss their utility and the data requirements and assumptions needed for them to be convincing. We also highlight the limitations of the particular analysis undertaken by van Ballegooijen et al.

The coalescent forms a bridge between evolutionary and epidemiologic processes: given a particular rate of population growth or decline, expected values can be obtained for the distribution of coalescent (branching) events on phylogenetic trees linking gene sequences (2). In the case of rapidly evolving viruses, these coalescent events can also occur within the timescale of sampling. In some simple cases, such as constant population size or exponential and logistic population growth, the relation between demography and phylogeny can be described in more precise terms. However, rather than basing parameter estimates on a single estimate of phylogeny, which may obviously contain error, the coalescent approaches used for viruses are commonly set within a Bayesian Markov chain Monte Carlo framework, incorporating millions of plausible representations of phylogenetic history. Such a sampling process provides a built-in measure of statistical uncertainty.

In other cases, most notably for viruses with more complex, fluctuating population dynamics, such as seasonal epidemic influenza and the periodic epidemics of measles, it is necessary to depict changes in population size by using a temporally piecewise Bayesian skyline plot (3). Such a plot is central to van Ballegooijen et al.'s study (1, Figure 4). Therefore, it is important to understand exactly what such a figure describes. Although a major aim of coalescent approaches is to make a precise statement about population growth, the Bayesian skyline plot simply represents an estimate of genetic diversity for different time segments from the root to the tip of the phylogeny. The exact measure of genetic diversity used in this context is $N_e\tau$, where N_e is the effective population size and τ the viral generation time (i.e., the time from host to host). Under an entirely neutral evolutionary system, such that the only process controlling the spread of mutations is genetic drift, a measure of $N_e\tau$ is also a measure of effective population size, which in simple terms can be thought of as the number of individuals in the population who contribute offspring to the next generation. However, $N_e\tau$ can also be strongly influenced by nonneutral evolutionary processes. For example, a selective sweep, in which a genetic variant of enhanced fitness outcompetes all other variants in the population to become the dominant form, will lead to a reduction in $N_e\tau$ although population size itself may have remained constant.

Another important limitation that needs to be considered in any coalescent analysis is that of biased sampling. First, it is obviously important that sequences be sampled randomly from the population under study, without a bias toward a particular group (such as those with a particular disease manifestation). In addition, a tacit assumption of the coalescent is that sequences are also drawn from a randomly mating population, such that all sequences have the same probability of having an ancestor in the previous generation. If this assumption of panmixis is broken—for example, because of population subdivision—drawing demographic inferences can become a perilous exercise. An example would be if the authors (1) had combined in the coalescent analysis the HBV genotype A sequences from the Amsterdam population of men who have sex with men with the HBV genotype B and C sequences, generally associated with Asian populations (4).

The study by van Ballegooijen et al. (1) opens important new doors on ways to assess the effectiveness of intervention strategies, and their expectations of how vaccination will affect genetic diversity are correct. However, the study suffers from a chronically small sample size: only 43 sequences (and just 21 different alleles) covering a 654-nucleotide fragment of the *S* (surface antigen) gene and sampled over a 14-year period (1992–2006) were available for analysis (i.e., an average of 3 sequences per year). The effect of a small sample is clearly apparent in a number of results represented. First, the Bayesian skyline plot (1, Figure 4) provides no strong evidence for a stepwise decline in genetic diversity (population size) coincident with the onset of vaccination. Although the median value of $N_e\tau$ undoubtedly declines, the 95% credible values cover a multitude of other demographic scenarios from a constant population size through time to even a step increase in genetic diversity. The adverse effect of small-scale sampling is also apparent in the parameter estimates made under the constant and stepwise-change demographic models. Although the latter has (slightly) better log likelihood, estimates of genetic diversity strongly overlap in both cases. In sum, although the authors may be right in their assertion that vaccination has effectively limited viral population size, their analysis needs to be based on a greatly expanded database of sequences before it can be considered conclusive.

The authors (1) expect that incidence of new infections would decrease because of reduced transmission. However, such a decrease was not observed, so they posit an increase in risk behavior to explain the lack of decreased incidence. Vaccine coverage is low (<20% in men who have sex with men based on their numbers), however, and, as the authors themselves state, the study lacks power to detect a change in incidence because incidence of reported HBV cases in the population of men who have sex with men is too low.

Other evolutionary processes could be also responsible for any decline in genetic diversity through time. As noted above, the predicted effect of positive natural selection is to cause a transient decrease in genetic diversity. The authors analyzed the *S* gene, the main antigenic determinant of HBV and where vaccine escape mutations have been described previously (5). Thus, it is theoretically possible that the decline in $N_e\tau$ is due to the recent spread of a vaccine escape mutation. Although perhaps less likely than the scenario conceived by van Ballegooijen et al. (1), at the very least it would be important to exclude the possibility that vaccine escape has contributed to the patterns of genetic diversity observed.

It is important to recall that genetic diversity is being used as a marker of population size. Hence, a reduction in genetic diversity is assumed to mean that there has also been an effect on viral population size, which in turn means that vaccination has reduced transmission. The reduction in genetic diversity is a general approach that could be used to measure an overall effect of a vaccination program (6). For strongly immunizing vaccines, reduced transmission is entirely to be expected. However, for a vaccine that is not completely cross-protective, which may often be the case for viruses that exhibit antigenic diversity, this is more of an open question, so the approach taken by the authors (1) would be useful. As such, it is of interest to determine how efficacy of vaccine varies across genetic variants of the pathogen; that is, what is the level of cross-protection across genetic variants? Methods have recently been developed to assess strain variations in vaccine efficacy based on the distance of genetic divergence of the infecting strain and the vaccine strain (7, 8). For this approach, it would be useful to examine the viral sequences of the vaccinated cases and unvaccinated cases separately. If many serotypes naturally occur, as with pneumococcal bacteria, widespread vaccination could enable expansion of nonvaccine serotypes that had been less important before vaccination (9, 10).

Part of the usefulness of the coalescent is that it allows epidemiologic dynamics to be dissected for individual viral lineages rather than relying on broader-scale measures such as seroprevalence. By focusing on individual lineages, more analytical precision can be achieved. As a case in point,

although simple serologic tests provide vital information on overall viral prevalence or incidence, they cannot provide information on the population size changes in individual lineages, such as the specific genotypes of HBV. Second, because viral lineages will usually circulate among all case types in a population, including those with both mild and severe disease manifestations, coalescent analyses consider the average evolutionary behavior of the virus in the population (i.e., it is reasonably assumed that the asymptomatic infections and unreported cases harbor the same viral lineages as those associated with reported cases). As such, inferences based on individual lineages effectively act as powerful markers for transmission through the whole population.

It is the power of the coalescent that forms the foundation of the work by van Ballegooijen et al. (1). However, the HBV data set they used is far from ideal, leading to substantial uncertainty. What would an idealized data set look like? As noted before, first, it is crucial that analyses be based on data sets large enough to provide sufficient epidemiologic-scale resolution. How large this data set will need to be will depend on the particular pathogen under consideration and is influenced by such factors as the length of the nucleotide sequence and the rate of evolutionary change. Second, it is critical that the data reflect sufficiently fine-grained temporal sampling to accurately inform on population growth rates. Although yearly sampling may be adequate for viruses with slow epidemiologic dynamics such as HBV, it is evidently inadequate for viruses that experience more rapid population dynamics such as influenza virus (11). Third, because $N_e\tau$ is used as a marker for population size, it is always important to consider what other processes can potentially influence genetic diversity and confuse demographic inferences. Natural selection and population subdivision are perhaps the most important. Finally, in the longer term, it is important that a broader range of population growth models be developed and that they enable precise parameter estimates. The Bayesian skyline plot is essentially a graphic method. The stepwise-change model of van Ballegooijen et al. is an important first step in this direction, and we hope that others will follow.

ACKNOWLEDGMENTS

Author affiliations: Center for Statistics and Quantitative Infectious Diseases, Fred Hutchinson Cancer Research Center, Seattle, Washington (M. Elizabeth Halloran); Department of Biostatistics, The University of Washington, Seattle, Washington (M. Elizabeth Halloran); Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, Pennsylvania (Edward C. Holmes); and Fogarty International Center, National Institutes of Health, Bethesda, Maryland (Edward C. Holmes).

The work of M. E. H. was partially supported by National Institute of Allergy and Infectious Diseases grant R01-AI32042.

Conflict of interest: none declared.

REFERENCES

1. van Ballegooijen MW, van Houdt R, Bruisten SM, et al. Molecular sequence data of hepatitis B virus and genetic diversity after vaccination. *Am J Epidemiol*. 2009;170(12):1455–1463.
2. Kingman JF. The coalescent. *Stoch Proc Appl.* 1982;13: 235–248.
3. Drummond AJ, Rambaut A, Shapiro B, et al. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 2005;22(5):1185–1192.
4. Magnius LO, Norder H. Subtypes, genotypes and molecular epidemiology of the hepatitis B virus as reflected by sequence variability of the *S*-gene. *Intervirology.* 1995;38(1-2): 24–34.
5. Torresi J. The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. *J Clin Virol*. 2002;25(2):97–106.
6. Halloran ME, Struchiner CJ, Longini IM Jr. Study designs for different efficacy and effectiveness aspects of vaccination. *Am J Epidemiol*. 1997;146(10):789–803.
7. Gilbert P, Self S, Rao M, et al. Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *J Clin Epidemiol*. 2001;54(1):68–85.
8. Sun Y, Gilbert PB, McKeague IW. Proportional hazards models with continuous marks. *Ann Stat*. 2009;37:394–426.
9. Lipsitch M. Interpreting results from trials of pneumococcal conjugate vaccines: a statistical test for detecting vaccine-induced increases in carriage of nonvaccine serotypes. *Am J Epidemiol*. 2000;154(1):85–92.
10. Peters TR, Poehling KA. Invasive pneumococcal disease: the target is moving. *JAMA*. 2007;297(16):1825–1826.
11. Rambaut A, Pybus OG, Nelson MI, et al. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 2008;453(7195):615–619.