Progressive Diseases: Interpretation of Genetic Data

LARRY BAUM and ERIC BAUM

*a PhD, Departments of Anatomical and Cellular Pathology and of Chemical Pathology, Chinese University of Hong Kong, Shatin, Hong Kong; b PhD, TRW, Redondo Beach, CA

(Received 2 April 1998; In final form 11 September 1998)

Simple modeling is proposed to represent the screening of gene polymorphisms for association with a progressive disease of insidious onset such as Alzheimer’s disease. The modeling demonstrates that when a polymorphism affects the rate of progression as well as the risk of disease, the correct interpretation of DNA data requires an accurate sampling of the living, diseased population. Furthermore, in this population, the effect of the polymorphism on disease risk cannot be distinguished from a corresponding effect on the rate of progression of the disease, and a null result does not preclude a significant effect of the gene on the disease. By contrast, when the population is sampled either at time of diagnosis or at autopsy, the effect of the polymorphism on disease frequency can be directly related to the frequency of the polymorphism in the sample, but evaluating the rate of disease progression requires additional data. When the only available data are obtained from a live patient population, substantial differences in interpretation can result from subtle differences in the patient selection protocol. When existing DNA databases are used in which this protocol is not well characterized, there is a corresponding uncertainty introduced into the deduced effect of the polymorphism on disease risk and rate of progression.

Keywords: Gene polymorphism Alzheimer modeling

INTRODUCTION

There is a broad class of degenerative diseases which are characterized by an insidious onset and progressive loss of function leading to a lingering debilitated terminal stage. We will use Alzheimer’s disease (AD) as a model. When it is suspected the disease has a genetic component, epidemiological data may be used to relate a gene polymorphism to a change in the prevalence of the disease. The association of Apolipoprotein E (ApoE) alleles to AD is an example. The sample size must be sufficient to produce statistically significant results, and this reduces the usefulness of the approach for rare mutations. There are other problems which have not been routinely considered which can lead to large uncertainties in the interpretation of the data. Since these problems are not intuitively obvious and some of the possible conclusions are indeed counter-intuitive, it is useful in a tutorial sense to define a
simple mathematical model of the disease and of the associated sampling process.

**DISEASE MODEL**

Assume that the progression of the idealized disease is measured by a test having scores ranging from $S = 1$ (normal function) to $S = 0$ (total loss of function). The progression of the disease with time is represented (see Figure 1) by the function

$$S = \frac{1}{1 + \exp(4(t - t_0)/\tau - 2)}$$

where $t =$ age at which noticeable disease progression begins, and $t_0 + \tau =$ age at which the terminal phase of disease begins. This function has the appropriate asymptotic behavior; it approaches unity for $t \ll t_0$ and approaches zero for $t \gg t_0 + \tau$, with a mid-point $S = 1/2$ at $t = t_0 + \tau/2$. The slope at the mid-point is $dS/dt = -1/\tau$. An approximation to this function is given by the three straight line segments:

- $S = 1 \quad t < t_0$
- $S = 1 - (t - t_0)/\tau \quad t_0 < t < t_0 + \tau$
- $S = 0 \quad t > t_0 + \tau$

The rate of disease progression (in the linear approximation) is represented by $r = 1/\tau$.

**GENETIC EFFECTS MODEL**

Assume that there is a single gene polymorphism, of prevalence $f$ in the general population, which changes the probability of being diagnosed with the disease in any given year from $p$ to $p^*$ and changes the rate of progression from $r = 1/\tau$ to $r^* = 1/\tau^*$.

---

**FIGURE 1** Functions representing the degree of disease progression.
For simplicity, assume that \( p, p^*, r \) and \( r^* \) are constants, neglecting their change with age. Out of a total population of size \( n \), \( n_d \) subjects without the gene polymorphism have been diagnosed with the disease but are not yet in the terminal phase. For a slowly changing population, this changes at a rate of approximately

\[
d n_d / dt = n p (1 - f) - n_d / \tau
\]

since \( n p (1 - f) \) is the number of people being diagnosed and \( n_d / \tau \) is the number of people entering the terminal phase each year. In a steady-state population, \( d n_d / dt = 0 \), so

\[
n_d / n = p r (1 - f) = p (1 - f) / r
\]

for subjects with the normal gene. By similar reasoning

\[
n_d^* / n = p^* r^* f = p^* f / r^*
\]

for subjects with the polymorphism. The fraction \( f^* \) of those diagnosed but not yet in the terminal phase who carry the polymorphism is

\[
f^* = n_d^* / (n_d + n_d^*) = p^* r^* f / [p r (1 - f) + p^* r^* f]
\]

A second population, defined to be composed of individuals identified with the disease at autopsy, clearly has a different genetic composition. This population is representative of everyone who gets the disease, so the fraction \( f_a^* \) of those identified with the disease at autopsy having the polymorphism is

\[
f_a^* = p^* f / [p (1 - f) + p^* f]
\]

since \( p^* f \) is the fraction of those people with the polymorphism who get (and the number who die of) the disease each year, and the denominator is the fraction of the entire population who get (and die of) the disease each year. We assume that \( f, f^* \) and \( f_a^* \) are measured and we would like to infer the resulting values of \( p^* / p \) and of \( r^* / r \). From the definition of \( f_a^* \),

\[
p^* / p = [1 / f - 1] / [1 / f_a^* - 1]
\]

From a combination of (1) and (2),

\[
r^* / r = [1 / f^* - 1] / [1 / f_a^* - 1]
\]

The result is that knowledge only of the frequency of the polymorphism in the general population \( f \) and in a living population diagnosed with the disease \( f^* \) cannot separate the effect of the polymorphism on disease frequency \( (p^* / p) \) from the effect on the rate of disease progression \( (r^* / r) \), as indicated by equation (1). A consequence is that without additional data, the effect of a decrease in disease frequency cannot be distinguished from that of an increase in rate of progression. A corollary is that significant dual effects which can reasonably be expected to coexist (an increase in disease frequency and an increase in rate of progression, for instance) have counterbalancing effects on the data. As a consequence, a null result (no significant difference between the frequency of the polymorphism in a living population diagnosed with the disease and that in the general population, or \( f \sim f^* \)) does not, without additional information, justify the conclusion that the polymorphism has no association with the disease. Measurements of the frequency of the polymorphism in the general population \( f \) and in a population in which the disease is diagnosed at autopsy \( f_a^* \) yield the effect of the polymorphism on disease risk, but no information on the rate of progression, as indicated by equation (2). Both types of measurement together yield both \( (p^* / p) \) and \( (r^* / r) \).

For case-control studies, \( f \) may not be known. However, for rare diseases, \( f \) approximately equals the frequency of the polymorphism among normal controls. Equation (1) then gives the odds ratio of the polymorphism for the living population, and equation (2) for the autopsy population.

**SAMPLING PROBLEMS**

Some of the difficulties in sampling the population of living patients are illustrated by the following
idealized example. Suppose that one in three in the
general population has at least one copy of a poly-
morphism \((f = 1/3)\), and that this doubles the dis-
ease risk \((p^*/p = 2)\) and doubles the rate of disease
progression (duration decrease from \(\tau = 2\) years to
\(\tau^* = 1\) year). Patients are diagnosed at a threshold
level of function \(S_t\), and two patients are entered
into a clinical database and a separate steady-state
longitudinal study database per day (one with the
polymorphism). After two years, the longitudinal
database population has reached a steady state, and
contains 365 patients with the polymorphism (365
reached the terminal stage and were removed from
the database) and twice that number (730) without
it. From the longitudinal database, \(f^* = f^*_L = 1/3\).
An estimate of the value of \(f^*\) from the clinical
database (which reflects only the patients whose
level of function has passed through \(S_t\)), would yield
\(f^*_C = 1/2\). Starting in the third year, a steady-state is
reached where pathological examination of patients
who have entered the terminal stage would yield
\(f^*_o = 1/2\). Both the clinical and the pathological
sampling procedures select their samples from the
stream of patients passing through the study, rather
than from the pool of patients in the population.

Cohort studies, which genotype and follow a
healthy population for several years to observe the
proportion with a polymorphism who develop dis-
case, can determine the disease risk for a poly-
morphism \((f_o^*)\) without interference from effects of
the polymorphism on duration of disease. But these
studies are large, time-consuming, and expensive
compared to case-control studies, and are only prac-
tical for common diseases. In addition, duration of
disease cannot be calculated from their data, unless
it is measured directly.

Community-based sampling which is not per-
fectly random can be expected to have a different
distribution of \(S\) in new enrollees than is present
in the population, but for any narrow range of \(S\)
(and therefore for the entire range), the polymor-
phism prevalence in the sample is the same as that in
the population \((f^*)\) provided that enrollees of each
genotype are attracted in proportion to the number
in the population. This requirement is approximately
satisfied for some sampling procedures, i.e. when
advertisements in general-circulation media are used
to solicit enrollment. A sampling procedure for
which the requirement is not satisfied is to post
the advertisement on a bulletin board in the diagnos-
ing clinic, in which case the sampling is done
on the stream of patients entering the population.
Physicians’ referrals may consist of newly diag-
nosed patients or may consist of the established
patients of a physician who has recently been made
aware of the study.

When the sampling protocol is not well defined,
or when mixed sources of data are used, as is often
the case when studies use preexisting databases,
there is an associated uncertainty in the evaluation of
\(f^*\). The importance of this patient selection proto-
col has been generally disregarded in the literature,
though Khoury et al. [1] do emphasize the impor-
tance of patient selection in case-control studies.
However, they recommend sampling the stream of
new patients rather than the pool of existing cases
in order to remove any effect of polymorphisms on
duration of disease.

EXAMPLES

Many of the examples in the literature deal with
the effect of ApoE alleles on Alzheimer’s disease.
By virtue of the fact that independent measurements
using longitudinal studies have determined that the
value of \(r^*/r\) is about unity for all genotypes [2–8],
the issue becomes moot. This was not known a
priori, however, and even earlier papers [9–12] fail
to address the influence of \(r^*/r\) on the interpreta-
tion of this type of measurement. However, van Duijn
et al. [13] do point out that the larger \(r^*\) they
observed for ApoE \(\varepsilon2\) than for ApoE \(\varepsilon3\) patients
may have led other studies to erroneously conclude
that \(\varepsilon2\) reduced the risk for AD.

The association between ApoE alleles and other
degenerative diseases has also been studied using
the same kinds of DNA databases. Both positive
[14] and null [15] associations of the \(\varepsilon4\) genotype
with vascular dementia have been reported, as have
both positive [16] and negative [17] associations of the e4 genotype with sporadic frontal lobe dementia.

Reports of the association of alleles of other genes with Alzheimer’s disease include the positive or null association of the I allele of the presenilin-1 gene with the late-onset form of the disease [18–21] and the null association of the A1 allele of the D2 dopamine receptor with both sporadic and familial forms of the disease [22]. Parkinson’s disease represents another slowly progressive disease. Positive associations of CYP2D6 polymorphisms with Parkinson’s disease have been reported [23–25].

The issues which we have pointed out are gene-rally disregarded in these studies, even in qualitative terms. When the reported association of a gene allele with a disease is positive, other studies (either measuring $r^*/r$ directly or using a different type of population) are motivated which will eventually resolve these issues. When a null association with the disease is reported (as would be the case in the idealized example of the previous section, where $f = f^*$), there is no strong motivation to validate the results with additional studies and a possibly strong association could be missed.

APPLICATION

Baum et al [26] studied the associations of two common mutations (Ser447Ter and Asn291Ser) of lipoprotein lipase (LPL) with Alzheimer’s disease (AD). Their results displayed some apparently anomalous features which could be partially explained by the effects under discussion here. They studied a predominantly white European-American population represented by a DNA database gathered by the Alzheimer Disease Research Center at the University of California, San Diego (the UCSD data). There are two subgroups of data: clinically diagnosed (live subjects) and pathologically confirmed (autopsy subjects). The prevalence of the mutation in the reference population ($f$) was estimated from that in a control group not showing evidence of AD.

The data showed

\[
\begin{align*}
\text{Ser447Ter:} & \quad f^* = 0.076(9/119) \\
& \quad f = 0.185(63/340) \\
& \quad f^* = 0.089(7/79) \\
& \quad f = 0.022(4/182) \\
& \quad f^* = 0.020(5/252)
\end{align*}
\]

It would be instructive to interpret the results using the idealized models introduced here. This yields

\[
\begin{align*}
\text{Ser447Ter:} & \quad p^*/p = 1.0 \quad r^*/r = 2.7 \\
\text{Asn291Ser:} & \quad p^*/p = 1.1 \quad r^*/r = 0.23
\end{align*}
\]

The indicated increase in the rate of disease progression associated with the Ser447Ter polymorphism and the indicated decrease in the rate of disease progression associated with the Asn291Ser polymorphism are large enough to be directly measurable through longitudinal studies such as those used to quantify the effect of ApoE e4 on the rate of progression of AD [2–8].

In order to justify the predictions of the modeling with respect to the UCSD data, we must show that the sampling process produces a realistic representation of the population. The database was pre-existing and the uncertainties in the sampling protocol are sufficient that these remain open questions. These uncertainties notwithstanding, the differences between $f^*$ and $f^*_a$ in the case of both polymorphisms are difficult to explain without invoking the effects of the rate-of-progression of the disease.

CONCLUSIONS

When screening gene polymorphisms for association with a progressive disease of insidious onset, DNA data obtained from a population diagnosed at autopsy can be directly interpreted to yield the effect of the polymorphism on disease frequency. We have demonstrated that DNA data which represents a living population diagnosed with the disease is more difficult to interpret. An accurate sampling of that population is required. This is difficult to obtain prospectively and associated uncertainties are
introduced when poorly characterized existing data libraries are used. In a properly sampled population, the effect of the gene polymorphism on disease frequency cannot be distinguished from a corresponding effect on the rate of progression of the disease, and a null result does not necessarily preclude a significant effect of the gene on the disease.

Acknowledgments

We wish to thank Dr. CP Pang, Dr. HK Ng, and Dr. CW Lam for helpful comments and support.

References


