Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 7581–7585, August 1992
Microbiology

Statistical analysis of sparse infection data and its implications for retroviral treatment trials in primates

(animal trial/infectivity assay/statistics/computer program)

JOHN L. SPOUGE

National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894

Communicated by Stirling A. Colgate, May 21, 1992

ABSTRACT **Reports on retroviral primate trials rarely publish any statistical analysis. Present statistical methodology lacks appropriate tests for these trials and effectively discourages quantitative assessment. This paper describes the theory behind VACMAN, a user-friendly computer program that calculates statistics for *in vitro* and *in vivo* infectivity data. VACMAN's analysis applies to many retroviral trials using i.v. challenges and is valid whenever the viral dose–response curve has a particular shape. Statistics from actual i.v. retroviral trials illustrate some unappreciated principles of effective animal use: dilutions other than 1:10 can improve titration accuracy; infecting titration animals at the lowest doses possible can lower challenge doses; and finally, challenging test animals in small trials with more virus than controls safeguards against false successes, "reuses" animals, and strengthens experimental conclusions. The theory presented also explains the important concept of viral saturation, a phenomenon that may cause *in vitro* and *in vivo* titrations to agree for some retroviral strains and disagree for others.**

1. Introduction

Vaccine development is an important therapeutic strategy against the human immunodeficiency virus (HIV) (1, 2). Because HIV productively infects only humans, chimpanzees, gibbon apes (3), and *scid*-hu mice (4), HIV vaccines are usually tested in chimpanzees (3, 5–10). Similar retroviral vaccines against the simian immunodeficiency virus are tested in macaques (11–13).

Primate trials have two phases. In the titration phase, aliquots of a frozen viral stock are thawed. The experimenter administers various doses to animals and determines which doses infect. After the titration is complete, treatments (or pretreatments like vaccines) are evaluated in the test phase. More virus is thawed, and test animals and their controls receive a viral challenge. The challenge chosen should be sufficient to ensure infection of any naive animal. On the other hand, excessively generous challenges (3, 9, 11) may overwhelm an otherwise resistant animal (12). Any challenge chosen under these conflicting conditions will be called a minimal challenge dose (MCD). MCD selection is crucial, for a test often is considered successful when a MCD infects all control but not all test animals.

Primate trials are strikingly small (3, 5–13). For example, one titration phase used six chimpanzees (3); another used seven macaques (11). Test phases have been even more sparing. Conclusions have been drawn from as few as four animals, corresponding to three different treatments with one overall control (7).

No statistician could feel comfortable with such sparse data. But the demands of animal care make primate experiments very laborious. Also, because the chimpanzee is an endangered species, trials often reuse animals from hepatitis experiments (3, 5–7, 9). In fact, the price of using a chimpanzee has escalated so much (from $100 in the 1950s to $25,000–$75,000 now) that any statistician protesting sparse primate data will soon do so with extreme trepidation.

Consider an actual primate trial. First, Arthur *et al.* (3) used tissue cultures to estimate the infectious dose 50% (ID_{50}) of an HIV stock. (By definition one ID_{50} infects half its recipients, and the ID_{50} in tissue culture is called the $TCID_{50}$.) Then, in the titration phase, four chimpanzees were given different viral doses intravenously (i.v.): 4, 40, 400, and 4000 $TCID_{50}$. Only 4 $TCID_{50}$ did not infect. Next, in a secondary titration, two more chimpanzees were given either 4 or 40 $TCID_{50}$ i.v. Both doses infected. Finally, after the viral titration, Berman *et al.* (6) tested two separate vaccines, gp120 and gp160. Five chimpanzees, two for each vaccine and one naive control, were challenged with 40 $TCID_{50}$. The control and the two gp160 chimpanzees were infected, but not the two gp120 chimpanzees.

The titration and test phases of this trial together cost a little less than a million dollars. But how confidently can HIV researchers abandon gp160 in favor of a gp120 vaccine? Was the challenge really sufficient, or could the two gp120 chimpanzees have remained uninfected through chance alone?

Comparing the gp120 and control results directly with the Fisher exact test (14) assigns an inappropriately small significance ($p = 0.33$) to the trial. Experimental confidence really depends on estimating how often any chosen challenge dose fails to infect a naive animal and so depends on the viral dose–response curve, which graphs infection probability vs. dose.

The dose–response curve must be estimated from titration data. Classical nonparametric methods like Spearman-Kärber (15) are unsuitable because they estimate only one point on the curve. Classical parametric methods like logit (16) or probit (17) analysis and nonclassical Bayesian methods (18) like Dirichlet curve fitting (19, 20) do estimate the entire curve but fit at least two arbitrary parameters. Typical primate data sensibly determine at most one parameter (e.g., the ID_{50}), so their interpretation requires a single-parameter theory.

Such a theory is feasible. Every titration has a smallest infecting dose (SID). For example, after Arthur's primary titration, it is 40 $TCID_{50}$; after his secondary, it is 4 $TCID_{50}$. Now, for the sake of argument, if the infection probability were to switch from 0 to 1 over a dose factor of 20, challenging with 20 times the SID would always infect—with progressively steeper dose–response curves allowing progressively smaller MCDs. Since five trials chose MCDs between 10 and 30 times the SID in the corresponding titration (6, 7, 10, 12, 13), many virologists (perhaps unconsciously) agree approx-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: HIV, human immunodeficiency virus; MCD, minimal challenge dose; MCD_{99}, minimal challenge dose 99%; SID, smallest infecting dose; $TCID_{50}$, tissue culture 50% infectious dose.

imately on a standard shape for the dose–response curve of i.v. retroviral challenges.

Mathematics can quantify the standard shape explicitly. Examine the following assumptions:

ASSUMPTION 1. *Interchangeability of animals. Consider any single infectious particle in a viral dose. The particle's chance of infecting is the same in all naive animals.*

ASSUMPTION 2. *Independence of infectious particles. Infectious particles in a viral dose act independently of one another in producing infection.*

Let D be the ID_{50}, and d be any viral dose. Given the assumptions, a standard mathematical derivation (21) shows that the probability of not infecting a naive animal is

$$q(d|D) = 2^{-d/D}. \quad [1]$$

Hence doses $d = D, 2D, 3D, \ldots$ fail to infect with probabilities 1/2, 1/4, 1/8, ...

Fig. 1 is the dose–response curve from Eq. **1**. Since Fig. 1 would be remarkably useful even if its predictions erred by a dose factor of 2, the assumptions above should be regarded as approximations, not exactitudes.

In a trial monitoring for HIV infection after i.v. challenge, the assumptions are plausible as approximations, even before the data in section 3 are examined. In well-mixed fluids, infection rates depend only on average concentrations [e.g., of viral targets or blockers (22)], so if an i.v. challenge with HIV initially infects in a well-mixed compartment (e.g., blood and/or lymph) and if average concentrations in this compartment are similar in all naive animals, animal interchangeability is credible.

The independence assumption is also credible, since *in vitro* experiments (23, 24) suggest that HIV particles do not cooperate when infecting cells. Because HIV escapes immune surveillance as a dormant DNA provirus and also replicates in bursts (25), HIV particles in a viral dose have no obvious need to cooperate in producing detectable infection.

Fig. 1 is not specifically limited to i.v. retroviral challenges, but its assumptions should be reexamined if trial conditions change. In vaginal HIV challenge, for example, variability due to lesions, diseases, or hormonal cycles may cause significant animal variations. Also, in i.v. challenge with a virus not infecting blood cells, anatomical and physiological variations may affect viral delivery to a target compartment (e.g., the liver or central nervous system). In general, interchangeability is suspect when animal variations affect the accessibility or susceptibility of viral targets.

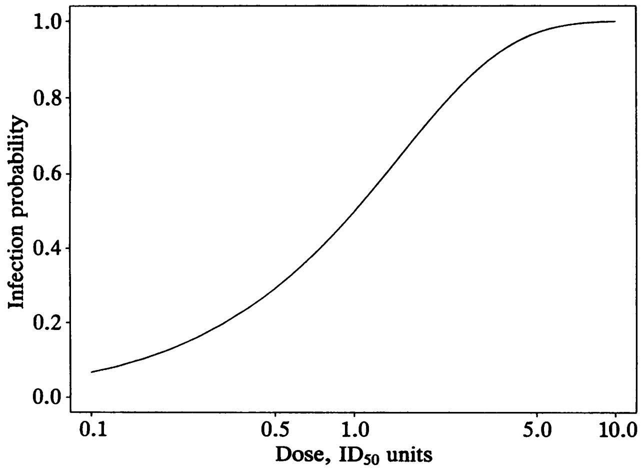

FIG. 1. Plot of viral infection probability vs. log dose. The doses are in ID_{50} units. This dose–response curve, derived from Eq. **1**, plots the infection probability $[1 - q(d|D)]$ vs. $\log(d|D)$, where D is the ID_{50}, and d is the viral dose.

Also, different trial endpoints reflect different viral processes. Consider, for example, a virus that causes disease only if many particles from an i.v. challenge infect cells. If a trial uses disease to detect infection, it is examining the outcome of a cooperative viral process. By contrast, retroviral trials usually seek sterilizing immunity [i.e., the absence of even a single successful cellular infection (6, 7, 10, 12, 13)], so the viral processes they examine are probably not cooperative. This paper's statistics assume (as stated precisely above) that infectious particles behave independently and usually require a trial endpoint that corresponds to cellular infection, not disease.

Hence each new application requires that the assumptions be reexamined.

The next section explains the statistical theory and is important but not essential to the rest of the paper. Section 3 exemplifies practical analysis with Arthur's titration (3), Berman's gp120–gp160 test (6), and some *in vitro* data. Section 4 describes the practical use of the VACMAN computer program. Finally, the *Discussion* examines the implications for primate retroviral trials.

2. Theory

Virologists starting a titration usually do 1:10 dilutions (3, 11–13). Hence in the chosen dose range and before collecting data, virologists judge all ID_{50} values to be about equally probable on a logarithmic scale. In the terminology of Bayesian statistics, their "prior probability" for $X = \log_{10}D$ is roughly uniform (26).

A logarithmic scale, $x = \log_{10}d$, therefore, is convenient. Rewriting Eq. **1** gives

$$q(x|X) = 2^{-10^{x-X}}. \quad [2]$$

The hypothesis in all subsequent equations implicitly assumes Eq. 2.

The data $\boldsymbol{r} = \{r_i$ of n_i animals infected at a viral dose d_i, $i = 1, \ldots, N\}$ change an experimenter's judgment about X. The "posterior probability" $p(X|\boldsymbol{r})$ is the probability after collecting the data $\boldsymbol{r}$ that a particular X is the $\log_{10}ID_{50}$. Given a uniform prior probability for X, Bayes Theorem (27) yields

$$p(X|\boldsymbol{r}) \propto p(\boldsymbol{r}|X) = \prod_{i=1}^{N} \binom{n_i}{r_i}[1 - q(x_i|X)]^{r_i}[q(x_i|X)]^{n_i - r_i}, \quad [3]$$

where $p(\boldsymbol{r}|X)$ is the probability of the data $\boldsymbol{r}$, given the true $\log_{10}ID_{50}$ equals X.

There are two trivial cases: all animals are infected $(r_i \equiv n_i)$ and all animals are uninfected $(r_i \equiv 0)$. In other cases, $p(X|\boldsymbol{r})$ has a unique maximum, and choosing an appropriate constant of proportionality in Eq. **3** makes $p(X|\boldsymbol{r})$ a proper probability distribution. The rest of the paper will assume this choice has been made.

In the limit of large data, the mode of $p(X|\boldsymbol{r})$ enjoys all the advantages accruing to a maximum likelihood estimator, including asymptotic normality and efficiency (28). More importantly for primate trials, $p(X|\boldsymbol{r})$ also extrapolates from the data $\boldsymbol{r}$ to give the posterior probability that a dose fails to infect:

$$q(x|\boldsymbol{r}) = \int_{-\infty}^{\infty} q(x|X)p(X|\boldsymbol{r})dX. \quad [4]$$

Eq. **4** can, therefore, help to select the challenge.

Consider now a trial with titration data $\boldsymbol{r}$ and test data $\boldsymbol{s}$. Assume for the moment that the viral titration conforms to

Eq. 2 and consider the null hypothesis H_0, that the conditions producing the titration and test data are the same. The experimenter wishes to find arguments against H_0.

Two types of arguments are feasible. The first treats r and s symmetrically. Let X_r equal $\log_{10}ID_{50}$ for the titration animals; let X_s equal $\log_{10}ID_{50}$ for the test animals. Under H_0, $X_r = X_s$. Data r generate a posterior distribution for X_r; data s generate a posterior distribution for X_s. The posterior distribution for $X_s - X_r$ can be calculated, so Bayesian procedures based on $X_s - X_r = \log_{10}ID_{50}$ ratio can decide on H_0.

Such a procedure is flawed, however, when s contains no infections (e.g., the gp120 test data in the *Introduction*). The posterior distribution for X_s, and therefore for $X_s - X_r$, is then improper toward positive infinity. Without an artificial "fix," any procedure based on $X_s - X_r$ will always conclude that the treatment succeeded. For extremely small challenges, this conclusion is clearly fallacious.

A different argument accounts for the challenge correctly, however. It acknowledges the logical asymmetry between r and s by viewing the titration animals as a "standard population." It then examines the test animals, whose status is uncertain, for conformity to this standard.

Under H_0, the predictive distribution of s given r is (29).

$$p(s|r) = \int_{-\infty}^{\infty} p(s|X)p(X|r)dX. \qquad [5]$$

Define the P value of test data s to be

$$P = \sum_{p(t|r) \leq p(s|r)} p(t|r). \qquad [6]$$

Given r as a standard, Eq. **6** quantifies how much one expects a result that is at least as unlikely as s. Hence, in analogy with accepted methods in classical statistics, Eq. **6** can decide on treatment efficacy.

Similarly, when s is a second titration, Eq. **6** can decide on the consistency of s with a "standard" titration r. Testing the ID_{50} ratio (defined above) is usually computationally more practical, however, because Eq. **6** may have many terms, each corresponding to an alternative outcome of the second titration s.

Eq. **6** can also test the internal consistency of a titration r. Replace s in Eq. **6** by the data in r at any single dose d and delete that observation from r. A small P value in Eq. **6** then suggests that either the conditions producing the data at dose d were anomalous with respect to the rest of r (reject H_0) or the model implicit in Eq. **2** is wrong. This outlier test is like a t test of externally Studentized residuals in linear regression (30) that evaluates the fit of individual data points to a model.

Finally, Bayesian methods can help refine viral titrations. Sometimes a primary titration is followed by a secondary titration (3) to give extra information and lower the challenge ensuring infection. The standard deviation of $p(X|r)$ is a measure of the information in the titration data r. Let the plan s' for the secondary titration be $\{n_i'$ animals will be given a viral dose d_i', $i = 1, \ldots, N'\}$. For each data set s corresponding to a possible outcome of the plan s', the standard deviation of $p(X|r, s)$ measures the information in r plus s. The rms standard deviation (weighted by the probabilities of the outcomes s) provides a figure of merit for the plan s'.

Theoretically, the rms standard deviation can be optimized as a function of the doses d_i', but the computation is usually prohibitively long. Experimenters always select doses at convenient dilutions anyway, so comparing the rms standard deviation of specific plans is preferable.

3. Practice

The first part of this section analyzes Arthur's titration (3) by finding the ID_{50} distribution, determining MCDs, and planning the secondary titration. It also analyzes Berman's gp120–gp160 test (6). The final part shows how to test whether a titration conforms to the dose–response curve in Fig. 1.

Fig. 2 shows likelihoods for Arthur's titration data (3). The primary titration used doses of 4 $TCID_{50}$ (not infecting) and 40, 400, and 4000 $TCID_{50}$ (all infecting). The corresponding likelihood gives a 95% Bayesian confidence interval (or "credible interval") of ID_{50} values between 1.3 and 170 $TCID_{50}$. The secondary titration used doses of 4 and 40 $TCID_{50}$ (both infecting), revising the confidence interval to ID_{50} values between 0.87 and 26 $TCID_{50}$.

The confidence intervals allow disconcertingly large ID_{50} values because infections do not delimit the ID_{50} as sharply as noninfections. For example, infecting with 0.25 ID_{50} is more likely than not infecting with 4 ID_{50} (from Eq. **1**: $1 - 2^{-0.25} = 0.159 > 0.063 = 2^{-4}$). Hence, an ID_{50} is more likely to be 4 times an infecting dose than 0.25 times a noninfecting dose.

Consequently, if a sparse primary titration already contains an uninfected animal, any secondary titration attempting to delimit the ID_{50} further should infect as many animals as possible, at the lowest doses possible.

The standard deviation of the ID_{50} measures how well a titration or titration plan delimits the ID_{50}. For example, Arthur planned a secondary titration using doses of 4 and 40 $TCID_{50}$. The VACMAN program computed the rms standard deviation for Arthur's plan (weighted over the plan's four possible outcomes) as 0.401. Both animals actually became infected, the most favorable outcome, yielding a lower standard deviation of 0.382.

The VACMAN program also evaluated several other plans. The lowest rms standard deviation among them was 0.378 for a plan using 16 and 32 $TCID_{50}$, doses that are relatively low but still likely to infect. Doses between 1:10 dilutions often improve the effectiveness of secondary titrations.

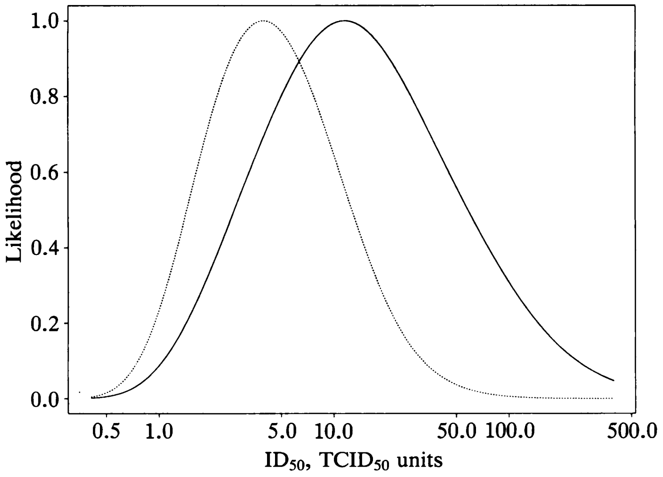

FIG. 2. Likelihood vs. log ID_{50} for titration data from Arthur (3). The accompanying text gives the titration data. The ID_{50} values are in $TCID_{50}$ units and the curves are Eq. 3 normalized with maximum ordinate 1. The likelihood is proportional to the probability that the corresponding dose is the ID_{50}. The curve on the right gives the distribution $p(X|r)$ for the primary titration data, with a most likely ID_{50} value of 12 $TCID_{50}$. After the secondary titration, it shifts to the curve on the left, with a most likely ID_{50} value of 4.0 $TCID_{50}$. The curves are skewed right but eventually approach a Gaussian curve as data accumulate.

The main aim of a titration is not, however, to delimit the ID_{50} but to ensure the infection of naive animals at the lowest challenges possible. Fortunately, infecting titration animals not only delimits the ID_{50} but also lowers challenges as well. For example, define the MCD_{99} from Eq. 4 as the MCD infecting with probability $1 - q(x|r) = 0.99$. After Arthur's primary titration, the MCD_{99} computed is 622 $TCID_{50}$; after his secondary, it is 89 $TCID_{50}$. By infecting two extra animals, Arthur's secondary titration profoundly lowered the MCD_{99}.

Statistics should aid, and not replace, experimental judgment, however. Animal interchangeability is a useful approximation, but animal ID_{50} values do vary. Thawed HIV-1 (HXB3) is labile and can undergo a 2-fold loss of infectivity in an hour (24); dilution error can also be a factor of 2. The statistics do not include these extrinsic errors. Many practical decisions require an experimenter to estimate extrinsic experimental error.

For example, after Arthur's primary titration, the posterior ID_{50} distribution has a mean ± SD of 1.15 ± 0.55 in $\log_{10}$ $TCID_{50}$. If Arthur had believed the extrinsic error to be ±0.60 in $\log_{10}$, his secondary titration would have had little point. If another experimenter believed the extrinsic error to be ±0.30 in $\log_{10}$ (a factor of 2), doubling computed MCDs would provide a conservative strategy for ensuring infection of controls.

Analysis of the gp120–gp160 trial (6) is next and adheres to three principles. (*i*) Random chances of success increase when a trial tests $N > 1$ treatments. If p is the P value of a treatment, the usual test at significance level α, $p \leq \alpha$, should be made stricter and replaced with Bonferroni's inequality (31), $Np \leq \alpha$. Because both gp120 and gp160 were tested, $N = 2$, so that any P value p really has significance $2p$. (*ii*) Extrinsic error may make the actual challenge smaller than the presumed 40 $TCID_{50}$. Since personal estimates of the extrinsic error differ, three "adjusted" challenges of 40 $TCID_{50}$, 20 $TCID_{50}$, and 10 $TCID_{50}$ are analyzed. (*iii*) Historical controls provide collateral titration information, but trials failing to infect a control are probably underreported. Hence, only the single simultaneous gp120–gp160 control is added to Arthur's titration data. The titration dose attributed to the simultaneous control is always the adjusted challenge under analysis.

The P values computed for infecting zero of two gp120 chimpanzees at adjusted challenges of 40 $TCID_{50}$, 20 $TCID_{50}$, and 10 $TCID_{50}$ are 0.008, 0.027, and 0.073, corresponding to significances of 0.016, 0.054, and 0.146, respectively. Hence if an experimenter believed that viral lability and other extrinsic errors had reduced the actual challenge from 40 $TCID_{50}$ to, e.g., 20 $TCID_{50}$, the trial significance would be 0.054. Interpretation at significance level 0.05 is, therefore, sensitive to an experimenter's estimate of extrinsic error.

A control safeguards against lost viral infectivity (7), and its infection supposedly demonstrates that actual challenges were sufficient. If lost infectivity reduced the control's actual challenge to 1 ID_{50}, however, a false treatment success (infecting the control, but not infecting at least one of the four test animals) had probability 0.47. If, however, test animals had been challenged at double the control's dose, the same lost infectivity gives a lower probability of false success. Trials using only one or two control animals should always consider doubled test challenges (or more).

The rest of this section shows when not to use this paper's statistics. The statistics are valid, however, whenever Fig. 1 approximates the shape of dose–response curve.

If primatologists disagree with computed MCDs, the true dose–response curve probably deviates from Fig. 1. Consider the MCD_{99} values above. The SID of Arthur's primary titration is 40 $TCID_{50}$, giving 622 $TCID_{50}$ = MCD_{99} = 15 SID. The secondary SID is 4 $TCID_{50}$, giving 89 $TCID_{50}$ = MCD_{99} = 22 SID. Hence the MCD_{99} values are between 10 and 30 SID, agreeing with the MCDs selected in i.v. retroviral trials (see *Introduction*). Larger trial MCDs would have indicated a dose–response curve conspicuously flatter than Fig. 1, making the analysis in this paper inapplicable.

The outlier P values that the VACMAN program computes from titration data also detect deviations from the dose–response curve in Fig. 1. Deviations are probable if any outlier P value is significant in Bonferroni's inequality (examples are given below). Outlier P values computed from published i.v. retroviral titrations (3, 11–13) are all 1.00, however, confirming that these titrations are completely consistent with the dose–response curve in Fig. 1.

Because *in vivo* titrations are consistent with Fig. 1, the final part of the section uses *in vitro* data and compares the outlier test to the classical tests (32, 33) of deviation from Fig. 1.

Stevens (33) gives 2 data sets demonstrating what he politely terms "faulty technique." Both data sets use 1:10 dilutions, with five wells at each dilution. Results will always be presented as the number of infected wells at increasing viral doses. The first data set has two anomalous infected wells: 0, 0, 0, 0, 2, 0, 0, 3, 5, 5, 5, 5. The outlier test correctly singles out the anomalous 2 wells with a P value of <0.0001. The P value for the 3 well is 0.16 and all other P values are 1.00. The second data set has anomalous sterile wells: 0, 0, 0, 2, 4, 5, 5, 5, 1, 4, 5, 5. The P value for the initial three 0 wells and the final two 5 wells is 1.00, for the 1 well is 0.004, and for the other counts is <0.0001. The anomalous noninfections, the 1 and the second 4 wells, delimit the ID_{50} more sharply than any infection. Hence ID_{50} distributions peak near them, causing infected wells at lower doses to appear anomalous. Eliminating the 1 and the 4 wells from the data removes all anomalous P values. The patterns presented are typical of anomalous wells, and the outlier test accords well with Stevens' test.

Armitage (32) gives two data sets, the first with a wide transition from sterility to infection, typical of ID_{50} variation from well to well. Both of Armitage's data sets use 1:2 dilutions, with 40 wells at each dilution. Outlier P values in parentheses accompany each of the $N = 9$ observations in the first set: 4(0.02), 5(0.07), 8(0.07), 10(0.45), 19(0.16), 25(0.78), 32(0.43), 35(0.0008), and 40(1.00). The ID_{50} distributions peak near the 25, causing the P value pattern. By using Bonferroni's inequality (31) (defined in the analysis of the gp120–gp160 trial above) with $N = 9$ to test the outlier P values, $p = 0.0008$ is significant at level $Np = 0.0072$, so the test detects the wide transition. Armitage's second data set conforms to Fig. 1: 2, 4, 8, 10, 19, 25, 33, 39, 40. The outlier test accords with Armitage's test: all outlier P values exceed 0.15.

Hence, the outlier test agrees well with classical tests of deviation.

4. The VACMAN Computer Program

This section describes how to use VACMAN, the program that computes this paper's statistics. VACMAN's use has two prerequisites: (*i*) as approximations, the assumptions above Eq. 1 should not be unreasonable and (*ii*) the titration data should be screened for anomalies with VACMAN's outlier test (see examples in section 3). An absence of significant outlier P values supports animal interchangeability or equivalently assay reproducibility (see VACMAN's interchangeability assumption, given in the *Introduction*).

Significant outlier P values invalidate VACMAN's analysis, unless experimental judgment can justify discarding the corresponding anomalous data. Certain patterns of significance (discussed at the end of section 3) suggest conspicuous variations between wells or animals, in which case reducing assay variability should be considered.

With sparse *in vivo* data, VACMAN computes the following: an estimate and a standard deviation for $\log_{10}ID_{50}$; an rms standard deviation, which ranks by merit secondary titrations designed to refine an ID_{50}; the noninfection probability at different viral doses, which permits MCD estimation; and test *P* values, which decide whether a treatment was efficacious.

With abundant *in vitro* data, the distribution of $\log_{10}ID_{50}$ approaches normal, and VACMAN and Fisher's dilution method (34) produce similar values for the mean and standard deviation. To detect ID_{50} shifts between two data sets, both sets must be screened for outliers. If no outlier *P* values are significant, VACMAN can give the mean and standard deviation of $\log_{10}ID_{50}$ ratio, which is approximately normally distributed.

5. Discussion

Without statistical interpretation of animal trials, HIV research could be giving disproportionate credence to marginal data. The resulting misdirection will be costly in time, effort, and money.

This paper presents statistics specifically designed for analyzing sparse primate infection data. The statistics are also useful for *in vitro* data (see section 4 and the end of section 3). I am distributing free a user-friendly computer program, VACMAN, that calculates the statistics. When requesting VACMAN 2.0, the current version, please specify your machine type, Macintosh or IBM. VACMAN requires a math coprocessor under Macintosh or Windows under IBM.

VACMAN's statistics are valid whenever the viral dose–response curve has a shape like Fig. 1. By the arguments presented in the *Introduction* and the data analyzed in section 3, Fig. 1 is clearly applicable to many, if not all, i.v. retroviral trials. Fig. 1 does not, however, include the effect of experimental errors like dilution or viral lability. Because errors influence the actual challenge administered, several different "adjusted" challenges should always be analyzed to appraise the sensitivity of conclusions to errors (cf., gp120–gp160 *P* values in section 3). Given this caveat, however, the statistics in this paper provide minimum standards for designing and analyzing retroviral trials using i.v. challenges.

By clarifying the consequences of experimental design (see section 3), the statistics also elucidate the principles of effective animal use. The rms standard deviation shows that in titrations, dilutions other than 1:10 can improve the accuracy of ID_{50} estimates. Computed MCD_{99} values demonstrate that infecting titration animals at the lowest doses possible can lower challenges.

The statistics also demonstrate the benefits of challenging test animals with more virus than controls (data not shown). In small trials incremental losses in viral infectivity can seriously affect conclusions (cf., gp120–gp160 *P* values in section 3). A doubled test challenge safeguards a trial with only one or two controls against false successes, "reuses" animals, and strengthens experimental conclusions.

Finally, the hypotheses after Eq. 1 postulate that in i.v. challenge trials, retroviral particles initially infect in a well-mixed compartment (e.g., blood and/or lymph), where only average concentrations influence each particle's chances of infecting. Matching test-tube assays to i.v. animal trials would then require *in vitro* concentrations to mimic blood and lymph mixed in some unknown ratio.

Some viral strains will be more sensitive to errors in the mixture ratio than others. At very low target-cell concentrations, target availability limits infection. Virions are "unsaturated" (22), and the number of infections is proportional to target-cell concentration (24). At higher target-cell concentrations, however, so many virions infect that increasing target concentration has little effect, and the virions are then "saturated." Different viral strains saturate at different target concentrations (23). Hence when *in vitro* target concentrations mimic the wrong mixture ratio, *in vitro* and *in vivo* titrations agree only for saturated retroviral strains, the ones insensitive to changes in target concentrations. Changing *in vitro* target concentrations may improve titration agreement for unsaturated retroviral strains as well, enhancing the possibilities for mimicking *in vivo* trials *in vitro*.

Scott Layne first pointed out the statistical omissions in primate trials to me. Jonathan Kans, Warren Gish, and other staff at the National Center for Biotechnology Information contributed to the VACMAN program. This paper benefitted greatly by suggestions from Alan Schultz, Natalya Merezhinskaya, and other informal referees, whom I thank.

1. Barnes, D. M. (1988) *Science* **240**, 719–721.
2. Koff, W. C. & Hoth, D. F. (1988) *Science* **241**, 426–432.
3. Arthur, L. O., Bess, J. W., Waters, D. J., Pyle, S. W., Kelliher, J. C., Nara, P. L., Krohn, K., Robey, W. G., Langlois, A. J., Gallo, R. C. & Fischinger, P. J. (1989) *J. Virol.* **63**, 5046–5053.
4. Namikawa, R., Kaneshima, H., Lieberman, M., Weissman, I. L. & McCune, J. M. (1988) *Science* **242**, 1684–1686.
5. Arthur, L. O., Pyle, S. W., Nara, P. L., Bess, J. W., Gonda, M. A., Kelliher, J. C., Gilden, R. V., Robey, W. G., Bolognesi, D. P., Gallo, R. C. & Fischinger, P. J. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 8583–8587.
6. Berman, P. W., Gregory, T. J., Riddle, L., Nakamura, G. R., Champe, M. A., Porter, J. P., Wurm, F. M., Hershberg, R. D., Cobb, E. K. & Eichberg, J. W. (1990) *Nature (London)* **345**, 622–625.
7. Girard, M., Kieny, M.-P., Pinter, A., Barre-Sinoussi, F., Nara, P. L., Kolbe, H., Kusumi, K., Chaput, A., Reinhard, T., Muchmore, E., Ronco, J., Kaczorek, M., Gomard, E., Gluckman, J.-C. & Fultz, P. N. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 542–546.
8. Hu, S.-L., Fultz, P. N., McClure, H. M., Eichberg, J. W., Thomas, E. K., Zarling, J., Singhal, M. C., Kosowski, S. G., Swenson, R. B., Anderson, D. C. & Todaro, G. (1987) *Nature (London)* **328**, 721–723.
9. Prince, A. M., Horowitz, B., Baker, L., Shulman, R. W., Ralph, H., Valinsky, J., Cundell, A., Brotman, B., Boehle, W., Rey, F., Piet, M., Reesink, H., Lelie, N., Tersmette, M., Miedema, F., Barbosa, L., Nemo, G., Nastala, C. L., Allen, J. S., Lee, D. R. & Eichberg, J. W. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6944–6948.
10. Ward, R. H. R., Capon, D. J., Jett, C. M., Murthy, K. M., Mordenti, J., Lucas, C., Frie, S. W., Prince, A. M., Green, J. D. & Eichberg, J. W. (1991) *Nature (London)* **352**, 434–436.
11. Desrosiers, R. C., Wyand, M. S., Kodama, T., Ringler, D. J., Arthur, L. O., Sehgal, P. K., Letvin, N. L., King, N. W. & Daniel, M. D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6353–6357.
12. Murphey-Corb, M., Martin, J. N., Davison-Fairburn, B., Montelaro, R. C., Miller, M., West, M., Baskin, S. O. G. B., Zhang, J. Y., Putney, S. D., Allison, A. C. & Eppstein, D. A. (1989) *Science* **246**, 1293–1297.
13. Putkonen, P., Thorstenson, R., Ghavamzadeh, L., Albert, J., Hild, K., Biberfeld, G. & Norrby, E. (1991) *Nature (London)* **352**, 436–438.
14. Siegel, S. (1956) *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York), pp. 96–104.
15. Miller, R. G. (1973) *Biometrika* **60**, 535–542.
16. Cox, D. R. & Snell, E. J. (1989) *Analysis of Binary Data* (Chapman & Hall, London), pp. 18–20.
17. Finney, D. J. (1971) *Probit Analysis* (Cambridge Univ. Press, Cambridge, U.K.).
18. Kraft, C. H. & Van Eeden, C. (1964) *Ann. Math. Stat.* **35**, 886–890.
19. Ferguson, T. S. (1973) *Ann. Stat.* **1**, 209–230.
20. Ramsey, F. L. (1972) *Biometrics* **28**, 841–858.
21. Armitage, P. & Bartsch, G. E. (1960) *Biometrics* **16**, 582–592.
22. Layne, S. P., Spouge, J. L. & Dembo, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4644–4648.
23. Layne, S. P., Merges, M. J., Dembo, M., Spouge, J. L. & Nara, P. L. (1990) *Nature (London)* **346**, 277–279.
24. Layne, S. P., Merges, M. J., Spouge, J. L., Dembo, M. & Nara, P. L. (1991) *J. Virol.* **65**, 3293–3300.
25. Fauci, A. S. (1988) *Science* **239**, 617–622.
26. Jeffreys, H. (1967) *Theory of Probability* (Oxford Univ. Press, Oxford), p. 117.
27. Lee, P. M. (1989) *Bayesian Statistics: An Introduction* (Oxford Univ. Press, Oxford), p. 20.
28. Kendall, M. & Stuart, A. (1979) *The Advanced Theory of Statistics* (Griffin, London), Vol. 2, pp. 46–47.
29. Aitchison, J. & Dunsmore, I. R. (1975) *Statistical Prediction Analysis* (Cambridge Univ. Press, Cambridge, U.K.), pp. 19–23.
30. Weisberg, S. (1985) *Applied Linear Regression* (Wiley, New York), pp. 113–116.
31. Weisberg, S. (1985) *Applied Linear Regression* (Wiley, New York), p. 116.
32. Armitage, P. (1959) *Biometrics* **15**, 1–9.
33. Stevens, W. L. (1958) *J. R. Stat. Soc.* **B20**, 205–214.
34. Fisher, R. A. (1921) *Philos. Trans. R. Soc. London A* **22**, 363–366.