Research Paper

Statistical considerations for the design and analysis of the ELISpot assay in HIV-1 vaccine trials

Michael G. Hudgens\textsuperscript{a,*}, Steven G. Self\textsuperscript{a}, Ya-Lin Chiu\textsuperscript{a}, Nina D. Russell\textsuperscript{b}, Helen Horton\textsuperscript{b}, M. Juliana McElrath\textsuperscript{b,c}

\textsuperscript{a}Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, NW-500, PO Box 19024, Seattle 98109-1024, USA
\textsuperscript{b}Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, USA
\textsuperscript{c}Department of Medicine, University of Washington, Seattle, USA

Received 11 September 2003; received in revised form 23 December 2003; accepted 24 January 2004
Available online 5 March 2004

Abstract

Effector T lymphocyte responses are considered critical for controlling human immunodeficiency virus type-1 (HIV-1) infection. The enzyme-linked immunospot (ELISpot) assay has emerged as a primary means of assessing HIV-specific T cell responses, and the development of objective methods that distinguish positive and negative ELISpot responses while properly controlling the rate of false positives is critical. In this paper, we consider several statistical methods that are helpful in defining such a positive criterion. Simulation results under a variety of scenarios suggest that a permutation-based criterion using a resampling adjustment for multiple comparisons yields the desired false positive rate while remaining competitive with other potential criteria in terms of sensitivity. These results also provide guidance on the effect of the number of experimental and negative control replicate wells on assay sensitivity. Application of different potential positive criteria using ELISpot assay results from IFN-\(\gamma\)-secreting T cells of HIV-1 seropositive and seronegative donors confirmed several of the results obtained under simulation. Our findings support the application of statistically-based positive criteria such as the permutation-based resampling approach in assessing HIV vaccine-induced T cell responses. Moreover, the proposed methods have potential utility in related HIV immunopathogenesis studies and in non-HIV clinical vaccine trials.

\(2004\) Elsevier B.V. All rights reserved.

Keywords: AIDS; Cytokines; Human; T lymphocytes; Vaccination

\textit{Abbreviations:} CTL, cytolytic T lymphocyte; ELISpot, enzyme-linked immunospot; FDR, false discovery rate; FWE, family-wise error rate; HIV-1, human immunodeficiency virus type-1; HVTN, HIV Vaccine Trials Network; MCPs, multiple comparison procedures; PBMC, peripheral blood mononuclear cells; PHA, phytohemagglutinin; SFCs, spot forming cells; SIV, simian immunodeficiency virus.

* Corresponding author. Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, 3107E McGavran-Greenberg Hall Chapel Hill, NC 27599. Tel.: +1-919-966-8872; fax: +1-919-966-3804.
E-mail address: m hudgens@bios.unc.edu (M.G. Hudgens).

0022-1759/ - see front matter \(2004\) Elsevier B.V. All rights reserved.
doi:10.1016/j.jim.2004.01.018
1. Introduction

Induction of strong and durable T lymphocytes capable of recognizing HIV-1-infected cells has become an important immunogenicity outcome in vaccine trials, given the increasing evidence that HIV-1-specific cell-mediated immune responses are at least partially protective against disease in macaques challenged with SHIV (Barouch et al., 2000; Amara et al., 2001), and against HIV-1 disease (Koup et al., 1994; Borrow et al., 1994; Rowland-Jones et al., 1995; Mazzoli et al., 1997; Musey et al., 1997; Ogg et al., 1998; Goh et al., 1999) in humans. With a considerable number of vaccine candidates being tested in non-human primate models and in phase I and II clinical trials, it is imperative to utilize a standardized, sensitive assay to measure HIV-1-specific T cell responses. The need to design and conduct phase III trials to establish correlates of immunity provides further motivation for the development of a sensitive, quantitative, and reproducible assay.

Both CD4+ and CD8+ virus-specific T cells are likely to contribute to control of viral replication. Anti-viral activities of HIV-1-specific CD8+ T cells include cytolysis of infected cells and release of chemokines and cytokines, such as IFN-\(\gamma\). HIV-specific CD4+ T cells are likely to be critical for maintenance of effective virus-specific CD8 T cell responses (Matloubian et al., 1994; Zajac et al., 1998). While CD8+ T cells constitute the majority of HIV-specific T cell responses in HIV-infected individuals and SIV-infected rhesus macaques (Vogel et al., 2002; Cao et al., 2003), the contributions of virus-specific CD4+ versus CD8+ T cells elicited by vaccines are likely to be dependent on the type of vaccine administered. Protein subunit vaccines are more likely to induce CD4+ T cell responses since they are processed and presented through the MHC class II processing pathway. Vaccine modalities that are based on viral vectors are, however, likely to induce CD8+ and CD4+ T cell responses since they will be processed and presented through both the MHC class I and II pathways. Our main interests lie in measurement of HIV-specific CD8+ T cell responses since these have been shown to directly contribute to control of viral replication in rhesus macaque models of simian immunodeficiency virus (SIV) infection (Jin et al., 1999).

Traditionally in phase I and II HIV-1 vaccine studies, CD8+ T cell responses have been identified as cytolytic T lymphocytes (CTLs) utilizing \(^{51}\text{Cr}\) release cytotoxicity assays (Evans et al., 1999; AIDS Vaccine Evaluation Group 022 Protocol Team, 2001; Belshe et al., 2001). Unfortunately, this assay requires fresh peripheral blood mononuclear cells (PBMC), is extremely labor intensive, and does not afford a reliable quantitative response. As a result, many researchers have turned to alternative techniques for measuring ex vivo viral-specific CD8+ and CD4+ T cells in cryopreserved PBMC such as flow cytometric analysis of intracellular cytokines, MHC-peptide tetramer staining, and IFN-\(\gamma\) ELISpot. Here we consider how statistical methodology can be employed in defining an ELISpot positive criterion. While the ELISpot assay does provide quantitative information, assessing the ability to induce a detectable qualitative response is a natural first step in initial evaluation of a candidate vaccine.

Based on principles similar to the ELISA assay, the ELISpot allows enumeration of both CD4+ and CD8+ T cells that secrete cytokines in response to an antigenic stimulus at the single cell level. Using 96-well plates coated with anti-cytokine antibodies and an appropriate detection system, each cytokine-producing cell can be identified by an individual spot within a well. Thus, the raw form of the data arising from the ELISpot assay is the number of spot forming cells (SFCs) for each experimental and control (negative and positive) well, each of which are often performed in duplicate or triplicate. An assay is subsequently categorized as positive or negative depending on whether or not the number of SFCs in the experimental wells is significantly greater than in the control wells.

Current methods for making a positive determination often rely on reasonable but ad hoc approaches. For example, responses to a particular antigen stimulation might be considered positive if on average there are (i) at least 10 more SFCs/2\(\times\)10\(^5\) PBMC in the experimental wells than control wells, and (ii) at least double the number of SFCs in the experimental wells than in the control wells (Larsson et al., 1999; Subklewe et al., 1999; Alter et al., 2002). Similar criteria appear throughout the immunology literature (Lalvani et al., 1997; Kaul et al., 2000; Larsson et al., 2000; Loing et al., 2000; Dhodapkar et al., 2000). In what
follows, we refer to the criterion based on (i) and (ii) above as the empirical method. While the empirical method is reasonable, it is not clear how the false positive rate is being controlled. For example, it is not known whether the criterion is too liberal or conservative as the background frequencies from the control wells of the assay vary, or if the strictness of the criterion varies according to the assay format. Furthermore, the empirical criterion gives no formal guidance with regards to using results from multiple antigenic stimulations to assess an overall response. The goal of this study was to employ statistical methods in developing an alternative positive criterion that addresses these issues. Consideration is given to optimizing sensitivity while controlling the false positive rate of the assay and accommodating multiple comparisons. The effect of assay format (i.e., the number of replicate experimental and control wells) is also considered.

2. Materials and methods

2.1. Study population

Assays were performed using cryopreserved PBMC from 36 HIV-1 seropositive and 45 HIV-1 seronegative individuals. Of the seropositive population, there were 10 long-term non-progressors, 8 antiretroviral treated chronically infected patients, 12 untreated chronically infected patients, and 6 untreated primary infected patients (see Russell et al., 2003 for details on a majority of these patients). Of the seronegative population, 35 were placebo recipients in a recent HIV Vaccine Trials Network vaccine trial (HVTN 203), the results of which are forthcoming. These volunteers received multiple doses of a placebo combination which included both an aluminum hydroxide adjuvant (AIDSVAX placebo, VaxGen, Brisbane, CA) and a mixture of 10 mM Tris–HCl buffer pH 9.0, virus stabilizer, and freeze drying medium (PLACEBO-ALVAC, Aventis Pasteur, Lyon, France). The remaining 10 seronegative individuals were previously enrolled in an AIDS Vaccine Evaluation Group protocol (AVEG 201); PBMC from pretreatment time points were used for these experiments. All study participants provided written consent, and all aspects of the study were approved by the human subjects institutional review board.

2.2. ELISpot assay

An IFN-γ ELISpot assay was used to detect HIV-1-specific IFN-γ producing T cells from cryopreserved PBMC as previously described (Russell et al., 2003). Ninety-six-well hydrophobic polyvinylidene difluoride membrane-bottomed plates (Millipore) were coated with anti-IFN-γ mAb overnight at 4 °C. After washing, plates were blocked at room temperature for 1 h. Cryopreserved PBMC were thawed and incubated overnight at 37 °C with 5% CO2. T cell subsets were purified using anti-CD4+ Ab-coated immunomagnetic beads (Microbeads, Miltenyi Biotec) according to the manufacturer’s instructions. Cells (either whole PBMC, CD4+, or CD4-depleted) were then added to pre-coated plates at concentrations of 2 × 10^5 or 1 × 10^5 cells/well. Two hundred twelve HIV-1 Env (MN strain), 248 Pol (HXB2), and 49 Nef (BRU) 15-mer peptides overlapping by 11 amino acids were synthesized by SynPep (Dublin, California) and 122 HIV-1 Gag (HXB2) 15-mer peptides overlapping by 11 amino acids were provided by the National Institutes of Health AIDS Research and Reference Reagent Program (Bethesda, MD). The peptides were used for antigenic stimulation in pools of 25 or 50 peptides, each at a final concentration of 1–2 μg/ml. Negative control wells contained media alone (no peptide). Cells stimulated with phytohemagglutinin (PHA) (1 μg/ml) served as a positive control. After antigenic stimulation overnight (16-20 h), a secondary biotinylated anti-IFN-γ mAb was added, followed by avidin biotinylated enzyme complex (VectastainABC Elite Kit, PK-6100, Vector) and subsequently, AEC peroxidase substrate (Vectastain). Plates were developed for approximately 7 min and colored SFCs were counted using an automated ELISpot reader (Immunospot, CellularTechnology).

2.3. Univariate criteria

In this section we discuss several potential positive criteria. First we introduce some notation. Suppose that we have \( m_E \) replicate wells for cells stimulated with a particular peptide pool and \( m_C \) common replicate wells containing cells serving as a negative control, all at the same concentration (often 200,000 cells/well). Let \( Y_{1C}^E, \ldots, Y_{mC}^E \) be the number of SFCs/well for the \( m_C \) control wells, and \( Y_{1E}^E, \ldots, Y_{mE}^E \) be the number of SFCs/well for the \( m_E \) peptide pool wells.
Let the distribution governing the number of SFCs/control well be denoted by $F_C$ with mean $\mu_C$. Similarly define $F_E$ and $\mu_E$ for the peptide pool replicate wells. Let $\bar{Y}_E = \sum Y_{Ei}/m_E$ be the per-well average for a particular peptide pool and let $\bar{Y}_C$ be the corresponding average for the control wells.

The empirical criterion described above for a positive response is then given by

$$\bar{Y}_E - \bar{Y}_C \geq 10$$

That is, experiments satisfying Eq. (1) are assumed to provide sufficient evidence to conclude that the peptide pool wells have significantly more SFCs/well than the control wells. Criterion (1) is typically used when the experiment is done at 200,000 cells/well (Russell et al., 2003). Should a different concentration be employed, Eq. (1) would be modified accordingly. For example, at $10^6$ cells/well, we would require $\bar{Y}_E - \bar{Y}_C \geq 50$ and $\bar{Y}_E \geq 2\bar{Y}_C$.

Alternatively, statistically based positive criteria might be employed by testing the null hypothesis

$$H : \mu_C \geq \mu_E,$$

versus the alternative $H' : \mu_C < \mu_E$, or

$$H : F_C(y) \leq F_E(y) \text{ for all } y,$$

versus the alternative $H' : F_C(y) \geq F_E(y)$ for all $y$ with at least one strict inequality. To avoid statistically significant but biologically or clinically irrelevant results, one might choose more conservative null hypotheses. For example, we might require statistical evidence that the mean SFCs/well for stimulated wells is at least twice background, i.e., $H : 2\mu_C \geq \mu_E$. Below we list several statistically based positive criteria, some of which are currently employed in the ELISpot literature and others of which are new.

### 2.3.1. t and Wilcoxon ranksum

Two statistically based criteria previously employed in the analysis of ELISpot data are the $t$ and Wilcoxon ranksum tests. The two-sample (or unpaired) $t$-test has been applied to determine if the difference in the mean number of SFCs in the experimental wells is significantly greater than in the control wells (Herr et al., 1996, 1997, 2000). In particular, a one-sided two-sample $t$-test can be employed to test the null hypothesis (Eq. (3)) under the assumption of equal variance or the more general null (Eq. (2)) using a Satterthwaite approximation (Rosner, 1995). The $t$-test assumes that the underlying distribution is approximately normal, a potentially dubious assumption for count data. Thus, a log transformation might be considered prior to performing such a test. Alternatively, the Wilcoxon ranksum (or Mann–Whitney) test, which is the non-parametric analogue of the two sample $t$-test, has been employed in a similar fashion (McCutcheon et al., 1997).

#### 2.3.2. Binomial

An alternative approach is to employ standard categorical data methods by considering whether the proportion of total spots attributable to the experimental wells is significantly greater than to be expected given the assay format (Flanagan et al., 1999; Self et al., 2001). Since the ELISpot assay produces count data, one might assume that $Y_{1E}, \ldots , Y_{mE} \sim \text{Poisson}(\mu_E)$ and $Y_{1C}, \ldots , Y_{mC} \sim \text{Poisson}(\mu_C)$. Letting $Y_E = \sum_i Y_{Ei}$, $Y_C = \sum_i Y_{Ci}$, and $N = Y^E + Y^C$ it is easy to show that given $N$, $Y^E$ is distributed Binomial$(N, p)$ where $p = m_E/(m_E + m_C)$. Note that $p = 1/2$ for equal number of replicate wells for the experiment and the control. Because the conditional distribution of $Y^E$ is binomial, an exact one-sided test for the null hypothesis given by Eq. (3) can be employed. Alternatively, a normal approximation can be applied wherein the quantity

$$Z_{\text{bin}} = \frac{Y^E - pN}{\sqrt{p(1-p)N}},$$

is treated as a standard normal deviate.

#### 2.3.3. Severini

While the exact binomial test discussed above is optimal under the Poisson distribution (Lehmann, 1959), it relies on the simple but potentially unrealistic assumption that the number of SFCs/well is governed by a Poisson distribution. It is not uncommon, however, for count data to fail to satisfy the assumption of the Poisson distribution that the variance equals the mean. In general, the existence of variability in the data that exceeds the nominal variance expected under the assumed model is called overdispersion. For count data, heterogeneity in the average event rates across the population under study
is often a cause of overdispersion. In the presence of overdispersion, an alternative positive criterion is needed since the Poisson assumption of the statistical positive criterion does not hold, which in turn can result in drawing incorrect inferences. Recently, Severini (1999) proposed an extension of the binomial test to be applied in case of overdispersion. Severini’s test entails a slight modification to the normal approximation to the binomial test wherein $Z_{ser} = Z_{bin}/\sqrt{1 - T}$ is treated as a standard normal deviate with $\gamma$ equal to the ratio of the sample variance to the sample mean of combined counts $Y_1^C, \ldots, Y_m^C, Y_1^E, \ldots, Y_m^E$.

2.3.4. Resampling

Bootstrap or permutation-based methods (Efron and Tibshirani, 1993) offer two alternative positive criteria for comparing the distributions of SFCs/well without making the parametric assumptions of the two-sample t-test or the exact binomial test. The permutation test corresponds to testing hypothesis (3) whereas the bootstrap approach may test either hypothesis (2) or (3) depending on the algorithm employed (Efron and Tibshirani, 1993). The permutation and Wilcoxon ranksum tests are closely related, with the main difference being that the permutation test permutes the actual counts, whereas the ranksum test permutes the ranks of the counts.

2.4. Multiple comparisons

Often in application of the ELISpot assay several peptides or pools of peptides are used in an effort to determine the breadth of response. Therefore, when determining whether a subject has a positive response, a multiple comparison adjustment may be appropriate to control the false positive rate (Huang et al., 2000; Nagorsen et al., 2000). Below we give a brief outline of the main issues of multiple comparisons germane to the analysis of data arising from the ELISpot assay in a clinical trials setting.

In general, the approach employed for a multiple comparisons problem depends on the scientific goals of the study or trial. In clinical trials it is often important to control the false positive rate at the participant level. For example, consider a Phase II HIV-1 vaccine trial where the primary endpoint is the proportion of participants who have a positive response by ELISpot to at least one HIV-1 peptide pool (e.g., HIV Vaccine Trials Network protocol 203 (NIAID Office of Communications and Public Liaison, 2002; Cohen, 2002)). In this case, it is desirable to control the family-wise error rate (FWE), i.e., the probability of having any false positives among the multiple hypotheses being tested (Costigan, 1998). On the other hand, if the setting is of an exploratory nature (e.g., epitope mapping), it may be more appropriate to control the false discovery rate (FDR), i.e., the expected proportion of rejected hypotheses that are true (Benjamini and Hochberg, 1995). Methods that control the FDR can be more powerful than those controlling the FWE, but are only valid when the overall conclusion of a set of hypothesis tests is not erroneous, even if some of the individual null hypotheses are falsely rejected. Motivated by the use of the ELISpot assay as an endpoint in HIV-1 vaccine trials, the focus of this paper concerns controlling the FWE. For further discussion regarding different types of FWE (e.g., strong or weak), see Westfall and Young (1993).

Among multiple comparison procedures (MCPs) that control the FWE, one seeks to maximize sensitivity or power, i.e., the probability of correctly rejecting false hypotheses. In the multiple comparisons setting, there are several definitions of power, two of which we will consider. Complete power is the probability of rejecting all false hypotheses; minimal power is the probability of rejecting at least one false hypothesis. For the clinical trials setting motivating this work, minimal power is important if the primary endpoint is defined as having a positive response to at least one peptide pool. On the other hand, if the focus is on the breadth of response (i.e., the number of positive responses across all peptide pools), then complete power might be of greater interest.

Given that a multiple comparisons adjustment is needed to maintain control of the FWE, there are several methods to consider (Hochberg and Tamhane, 1987; Westfall and Young, 1993; Hsu, 1996; D’Agostino and Russell, 1998). By comparing different sets of experimental wells for each distinct peptide pool to the same set of control wells, the ELISpot assay is an excellent example of the many-to-one problem (Dunnett, 1955; Hochberg and Tamhane, 1987; Westfall and Young, 1993; Hsu, 1996; Hochberg and Westfall, 2000). The challenge in choosing the appropriate MCP in the many-to-one problem arises from the
dependence among the multiple test statistics induced by using the same control for each comparison. Bonferroni-type (Holm, 1979), normal-based (Dunnett, 1955; Hsu, 1996), and resampling (Westfall and Young, 1993; Westfall et al., 1999) MCPs have been advocated for either general situations entailing dependent test statistics or the many-to-one problem in particular.

2.5. Simulation study

A set of simulation studies were employed to assess the performance of the different univariate positive criteria and multiple comparison procedures discussed in the previous sections. The goal was to find at least one MCP that maintains the FWE for various

(i) numbers of distinct peptide pools ($K$),
(ii) assay formats (i.e., $m_C$ and $m_E$),
(iii) levels of background (i.e., mean number of SFCs/ negative control well, $\mu_C$), and
(iv) underlying distributions governing the number of SFCs/well.

To address (i), we considered values of $K = 1$ (i.e., no multiple comparisons), 10, and 25. To address (ii), several combinations of $m_C$ and $m_E$ were considered. Duplicate ($m_C = 2$, $m_E = 2$) and triplicate ($m_C = 3$, $m_E = 3$) formats are common combinations found in the ELISpot literature; however, statistical theory suggests that choosing $m_C > m_E$ may be more optimal for many-to-one problems such that ($m_C = 6$, $m_E = 2$), ($m_C = 10$, $m_E = 2$), and ($m_C = 6$, $m_E = 3$) formats are also considered. To address (iii), the mean number of SFCs/ control well was varied over four values: $\mu_C = 3, 10, 30, 100$. To address (iv), count data were generated by assuming that the counts follow a mixture model in which the mean of the Poisson distribution is itself a random variable with a gamma distribution having mean $\mu$ and variance $(\phi - 1)\mu$. The resulting counts are then negative binomial random variables with mean

![Fig. 1. Histogram of spot forming cells (SFCs) per well from simulated control wells under each distributional assumption. The mean number of SFCs is given by $\mu$. The ratio of the variance to the mean is denoted by $\phi$.](image-url)
and variance $\phi\mu$. For $\phi = 1$ the mean of the Poisson distribution is fixed at $\mu$ such that the resulting counts are simply Poisson with mean and variance $\mu$. Simulations were done for a variety of levels of overdispersion: none ($\phi = 1$), mild ($\phi = 2$), moderate ($\phi = 10$), and severe ($\phi = 25$). Illustrative histograms of simulated control wells are given in Fig. 1. Looking across the columns of Fig. 1 corresponds to an increase in the mean SFCs/well ($\mu_c$), while looking down the rows corresponds to an increase in overdispersion ($\phi$). For each combination of (i)–(iv), 10,000 simulated data sets were computed; for the resampling-based methods, bootstrap and permutation samples of size 500 were drawn for each simulation.

For each set of simulations, both control of the FWE and power (i.e., sensitivity) were considered. Alternative hypotheses were simulated by assuming that the

**Table 1**

<table>
<thead>
<tr>
<th>Name</th>
<th>Multiple comparison procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial-Holm</td>
<td>Normal approximation to Binomial with Holm (1979) multiplicity adjustment</td>
</tr>
<tr>
<td>Severini-Holm</td>
<td>Severini’s (1999) Binomial adjustment with Holm (1979) multiplicity adjustment</td>
</tr>
<tr>
<td>$t$-test-Holm</td>
<td>Two-sample $t$-test with Holm (1979) multiplicity adjustment</td>
</tr>
<tr>
<td>Empirical</td>
<td>Empirical criterion (Eq. (1))</td>
</tr>
<tr>
<td>Bootstrap-WY</td>
<td>Westfall and Young (1993) step-down bootstrap</td>
</tr>
<tr>
<td>Permutation-WY</td>
<td>Westfall and Young (1993) step-down permutation</td>
</tr>
<tr>
<td>Dunnett</td>
<td>Dunnett (1955)</td>
</tr>
</tbody>
</table>
number of SFCs in experimental wells can be decomposed into the sum of background SFCs and “true” SFCs corresponding to IFN-γ emitting T cells. The background SFCs were assumed to have the same distribution as SFCs in the negative control wells (with mean \( \mu_C = 3, 10, 30, \) or 100). The antigen-recognizing cell populations were assumed to be governed by the same distribution, but with a mean of 20 SFCs/well. Under the Poisson background, this implies that the SFCs in the experimental wells have a Poisson distribution with mean \( \mu_E = 23, 30, 50, \) or 120. For \( K = 10 \) and 25, both minimal and complete power were considered.

2.5.1. Implementation

Each of univariate positive criteria considered is simple to implement in today's statistical computing environment. On the other hand, many of the multiple comparison procedures discussed here are relatively recent developments. SAS (Cary, NC, USA) offers two procedures, PROC GLM and PROC MULTTEST, which allow implementation of several of the tests described above. More recently, the Bioconductor project (www.bioconductor.org) has developed a multtest package for use in R (Ihaka and Gentleman, 1996) which implements many of the same multiple comparisons procedures available in SAS. Other statistical packages offer limited multiple comparison capabilities. All simulations and analyses in this paper were implemented in SAS (Westfall et al., 1999). Sample code for implementing these procedures in SAS is available upon request. Code for the permutation criterion is given in Appendix A.

Fig. 3. Minimal power for various multiple comparison procedures (as described in Section 2.4 and Table 1) based on 10,000 simulated data sets with \( K = 25 \) peptide pools and two false null hypotheses for assay format \((m_C, m_E)=(6, 2)\). The shaded bars correspond to settings where the observed FWE was less than or equal to the nominal \( \alpha = 0.05 \) level. The mean number of SFCs for the negative control wells is given by \( \mu_C \) (muc). The ratio of the variance to the mean is denoted by \( \phi \) (phi).
3. Results

3.1. Simulation study

3.1.1. Univariate results

False positive rates for the various univariate positive criteria (i.e., for \( K = 1 \) peptide pool) are displayed by the trellis plot (Venables and Ripley, 1999) given in Fig. 2. The figure is composed of 16 panels corresponding to each combination of background level (\( \mu_C \)) and dispersion (\( \phi \)). Within each panel, the FWE is given for each positive criterion. Results are shown for the \((m_C, m_E)=(6, 2)\) format only; similar trends were observed for other formats. Comparisons with the dotted vertical line depicting the nominal FWE of 0.05 show that only the Permutation test consistently controls the FWE. The Binomial and Empirical criteria can give especially high FWEs in the presence of overdispersion (e.g., \( \phi = 25 \)). Simulations under the alternative hypothesis (results not shown) demonstrate (i) superior power of the Binomial test for the Poisson simulations (which is consistent with statistical theory (Lehmann, 1959), and (ii) a general decrease in power with an increase in either overdispersion or background for all criteria. The Permutation and Wilcoxon ranksum tests gave nearly identical results (not shown) with regard to the FWE and power.

3.1.2. Multiple comparison results

Given the considerations of the section above, the univariate simulation results, and implementation complexity, only the MCPs given in Table 1 were considered in the simulation study. Fig. 3 shows the minimal power (i.e., the probability that at least one false null
hypothesis is rejected) when two of \( K = 25 \) null hypotheses are false for the \((m_C, m_E) = (6, 2)\) format. The shaded bars indicate scenarios where the observed FWE was less than 5%. Generally, most MCPs demonstrate excellent control of the FWE under the Poisson distribution. However, as in the univariate simulations, only the Permutation-WY MCP consistently controls the FWE under all scenarios considered. Additionally, the Permutation-WY approach performs well with respect to power compared to other MCPs which control the FWE. For example, consider the results in Fig. 3 for \( \mu_c = 30 \) and \( \phi = 2 \) where only the Permutation-WY, Severini-Holm, Dunnett, and Empirical criteria have FWE less than or equal to 5%; of these four criteria in this scenario, the Permutation-WY is the most powerful. In general, the Empirical criterion lacks sensitivity in the presence of high background (i.e., \( \mu_c = 30, 100 \)) and lacks specificity in the presence of overdispersion (i.e., \( \phi = 10, 25 \)). Similar trends were found for complete power, other assay formats, and \( K = 10 \) peptide pools (results not shown).

Given the results above, we also explored the effect of assay format on sensitivity using the Permutation-WY criterion (Fig. 4). Not surprisingly, power generally increases with the total number of wells \((m_C + K m_E)\). For example, in Fig. 4 the highest sensitivity is achieved under the \((m_C, m_E) = (6, 3)\) format, which requires 81 total wells when there are \( K = 25 \) peptide pools; in contrast, the duplicate format has the lowest sensitivity and requires only 52 total wells. Note that reasonable gains in power can be achieved over the duplicate format by simply increasing the number of control wells.

Fig. 5. Left panel: Scatterplot and least-squares regression line of mean SFCs/well for each peptide pool versus negative control in seronegative individuals. Right panel: Scatterplot of between replicate variances versus mean for all peptide pool and negative control wells in seronegative and seropositive individuals. The three lines denote observed overdispersion (i.e., variance to mean ratio) of \( \phi = 1, 10, \) and 100.
wells. For example, in the absence of overdispersion (i.e., upper left panel of Fig. 4) and a mean background of \( \mu_C = 40 \), minimal power increases from 0.61 for \((m_C, m_E)=(2, 2)\) to 0.75 for \((m_C, m_E)=(6, 2)\) and to 0.82 for \((m_C, m_E)=(10, 2)\). Given operational constraints of performing each experimental peptide pool in triplicate, increasing the number of control wells provides a simple method for increasing assay sensitivity.

### 3.2. Application

To illustrate application of the proposed statistical criteria, we considered ELISpot assay results from 45 HIV-1-seronegative and 36 HIV-1-seropositive volunteers. Several of the seronegative volunteers were tested at more than one time point, but for purposes of this analysis, these results are treated as independent observations. For each volunteer, up to three different cell populations (PBMC, CD4+ T cells, or CD8+ T cells) may have been assayed. For nearly all assays, experimental wells were tested in duplicate \((m_E = 2)\) while six wells were used for each negative control well \((m_C = 6)\). The number of HIV-1 peptide pools tested ranged from \(K = 3\) to \(K = 26\). Two displays of the data are given in Fig. 5. Using data from seronegative individuals only, the left panel shows a strong linear association between the level of background and the mean SFCs/well for HIV-1 peptide pools. Since no response to these pools is expected from seronegative samples, this plot indicates that the negative control wells are predictive of the background in the experimental wells and thus supports the assumption that a positive criterion should entail contrasts between experimental and negative controls wells. The right panel of Fig. 5 indicates moderate levels of overdispersion in these data.

Response rates, shown in Table 2, were calculated separately for each cell population. All statistically based positive criteria were employed using one-sided tests performed at 0.10, 0.05, and 0.01 significance levels. For assays from the HIV-1 seronegative volunteers, the Permutation-WY criterion yields results closest to the nominal false positive rate. In general, for the other criteria, the observed false positive rate substantially exceeds the nominal significance level for at least one of the cell populations. For example, the Empirical criterion yields reasonable false positive rates for CD4+ and CD4− cells, but incorrectly categorizes PBMC from 21% (5/24) of seronegative individuals as positive. These results are generally consistent with the simulation study from the previous section when there is extra-Poisson variability. While the Permutation-WY criterion gives rise to lower observed sensitivity in the HIV-1 seropositive individ-

### Table 2

Response rates by different positive criteria

<table>
<thead>
<tr>
<th>Study population</th>
<th>Seropositive</th>
<th>Seronegative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effector cells</td>
<td>CD4+ (n=36)</td>
<td>CD4− (n=36)</td>
</tr>
<tr>
<td>PBMC (n=24)</td>
<td>CD4+ (n=74)</td>
<td>CD4− (n=74)</td>
</tr>
<tr>
<td><strong>Empirical</strong></td>
<td>16 (53%)</td>
<td>5 (21%)</td>
</tr>
<tr>
<td><strong>α = 0.10</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permutation-WY</td>
<td>22 (61%)</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>Bootstrap-WY</td>
<td>24 (67%)</td>
<td>4 (17%)</td>
</tr>
<tr>
<td>t-test-Holm</td>
<td>23 (64%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>Dunnett</td>
<td>27 (75%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>Severini-Holm</td>
<td>25 (69%)</td>
<td>4 (17%)</td>
</tr>
<tr>
<td>Binomial-Holm</td>
<td>25 (69%)</td>
<td>12 (50%)</td>
</tr>
<tr>
<td><strong>α = 0.05</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permutation-WY</td>
<td>20 (56%)</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>Bootstrap-WY</td>
<td>23 (64%)</td>
<td>4 (17%)</td>
</tr>
<tr>
<td>t-test-Holm</td>
<td>23 (64%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>Dunnett</td>
<td>26 (72%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>Severini-Holm</td>
<td>23 (64%)</td>
<td>3 (13%)</td>
</tr>
<tr>
<td>Binomial-Holm</td>
<td>25 (69%)</td>
<td>11 (46%)</td>
</tr>
<tr>
<td><strong>α = 0.01</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Permutation-WY</td>
<td>8 (22%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Bootstrap-WY</td>
<td>19 (53%)</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>t-test-Holm</td>
<td>16 (44%)</td>
<td>5 (7%)</td>
</tr>
<tr>
<td>Dunnett</td>
<td>21 (58%)</td>
<td>8 (11%)</td>
</tr>
<tr>
<td>Severini-Holm</td>
<td>18 (50%)</td>
<td>13 (18%)</td>
</tr>
<tr>
<td>Binomial-Holm</td>
<td>23 (64%)</td>
<td>10 (42%)</td>
</tr>
</tbody>
</table>

*Number and percent of positive responses for 36 HIV-seropositive and 45 HIV-seronegative volunteers. A volunteer’s response was categorized as positive if at least one peptide pool stimulated a positive IFN-γ secreting response.*
uals (especially at $z = 0.01$), the higher rates given by the other criteria likely include a substantial portion of false positives. Moreover, based on the results in Fig. 4, one would expect a greater observed sensitivity with the Permutation-WY criterion had the assay employed triplicate experimental wells (i.e., $m_E = 3$).

4. Discussion

Establishing a unified positive criterion for the IFN-γ ELISpot assay is critical to enabling valid comparisons across labs, studies, and populations. Our collective experience at the HIV Vaccine Trials Network suggests that the distribution of background SFCs/well can vary between labs and within labs, either over time or across protocols, and in individual responses over multiple time points. Until the sources of this variation are identified and controlled, it is important that a positive criterion be robust to such variability. To this end, we recommend adopting the permutation-based criterion using a resampling adjustment for multiple comparisons. Our simulation results suggest that in the multiple comparisons setting, this method consistently preserves the nominal false positive rate while remaining competitive with other methods in terms of sensitivity. These results also suggest that assay format is an important consideration given the significant between well heterogeneity that can occur with the ELISpot assay.

It is important to note that the statistical criteria presented here will not control the “biological” false positive rate. In the context of a vaccine trial, such positive responses are not induced by vaccination but may be due to cross-reactivity with other pre-existing immune effector cells. It is easy to imagine how the rate of such responses may vary considerably across populations, especially in the developing world where the breadth of the infectious disease burden is considerable. Use of concurrent placebo control groups and/or baseline, pre-immunization assays remains an important component of trial design to clearly separate the vaccine-induced positive responses, the “biological” false positive responses and the false positive responses.

The often employed empirical criteria offer an alternative strategy for determining the critical value for positivity (Jung et al., 1999; Lewis et al., 2000; Dhodapkar et al., 2000). This approach relies on determining the specificity of the assay based on response rates in seronegative, unvaccinated individuals. As a result, these empirical methods combine both statistical and biological false positive rates. Employing such an approach either requires population-specific positivity criteria based on large control samples from each population or the development of a common, conservative cross-population criterion that will necessarily lose sensitivity in those populations with low biological false positive rates. Defining population-specific criteria may not be operationally or economically feasible, whereas a conservative omnibus criterion will be a clear disadvantage when assessing low frequency vaccine-induced responses. Ultimately, either approach will complicate interpretation of results because one cannot clearly separate the two types of false (i.e., not vaccine-induced) positive responses.

In summary, we advocate employing a statistically-based positive criterion such as the Permutation approach especially combined with study designs that can distinguish positive responses that are vaccine-induced from those that are not. This approach has the potential to more accurately assess T cell responses in diverse geographic populations in the clinical trial setting. Its utility is likely to extend to clinical evaluation of immunogens and immunotherapeutics for other pathogens of public health importance.

Acknowledgements

The authors are grateful for the financial support from NIH U01-AI46703, R201 AI-29168, AI-47086, AI-46725, AI-48017.

Appendix A

In this section we provide a sketch of the algorithm for computing permutation adjusted $p$-values. Sample SAS code is also provided. For further details, see Westfall and Young, 1993 and Westfall et al., 1999.

The single-step permutation algorithm is given by:

1. For each of the $K$ distinct peptide pools, compute an unadjusted $p$-value comparing the observed SFCs/
well from the peptide pool replicate wells to the negative control replicate wells. These unadjusted $p$-values can be computed using any univariate statistically-based positive criterion, e.g., a two-sample $t$-test.

2. Permute the observed SFCs/well across all peptide pool and negative control wells.

3. Compute unadjusted $p$-values as in Step 1 using the permuted data set from Step 2. Take the minimum of these $K$ unadjusted $p$-values.

4. Repeat Steps 2 and 3 $B$ times. $B \geq 10,000$ is recommended.

5. For each peptide pool, the permutation adjusted $p$-value is then computed as the proportion of times that the unadjusted $p$-value is greater than or equal to the $B$ minimum permutation $p$-values.

A slightly more complicated (and powerful) step-down permutation algorithm is given in Algorithm 2.8 of Westfall and Young (1993). The step-down permutation algorithm, denoted Permutation-WY, was employed in computing the results for this paper.

SAS code to compute single-step and step-down permutation $p$-values is given below:

* tailored version of PROC MULTTEST;
%macro mymult (data_in=k=);
 proc multtest data=&data_in perm n=20000 order=data steperm;
 class well; test mean(count/upper);
 %do peps=1 %to &k;
 contrast "peptide&peps" -1 %vect(len=&k,place=&peps);
 %end;
 %mend mymult;

* macro that returns a vector of length len w/ zeros except a 1 at place;
%macro vect (len=, place=);
 %do nn=1 %to &len;
 %if (&nn=&place) %then 0;
 %else %then 1;
 %end;
 %mend vect;

*simple data set done mostly in triplicate w/ k=10 peptide pools;
data elispot;
 length well $9;
 input well $ count @a;
 cards;
 controlwe 10 controlwe 9 controlwe 9 peptide1 10 peptide1 11 peptide1 13 peptide2 52 peptide2 43 peptide2 53 peptide3 43 peptide3 41 peptide3 33 peptide4 44 peptide4 49 peptide4 53 peptide5 35 peptide5 41 peptide5 43 peptide6 36 peptide6 41 peptide6 33 peptide7 26 peptide7 21 peptide7 13 peptide8 27 peptide8 16 peptide8 13 peptide9 18 peptide9 11 peptide10 3 peptide10 4 peptide10 3 ;

* call macro;
%mymult(data_in=elispot,k=10);
References


Matloubian, M., Concepcion, R.J., Ahmed, R., 1994. CD4+ T cells are required to sustain CD8+ cytotoxic T-cell responses during chronic viral infection. Journal of Virology 68, 8056.


NIAID Office of Communications and Public Liaison, 2002. NIAID Phase III HIV vaccine trial to determine correlate of protection will not proceed. NIAID News.


Westfall, P.H., Young, S.S., 1993. Resampling-based Multiple Tes-
