Title

Beyond the Primary Endpoint Paradigm: A Test of Intervention Effect in HIV Behavioral Intervention Trials with Numerous Correlated Outcomes

Permalink
https://escholarship.org/uc/item/3t112413

Journal
Prevention Science, 18(5)

ISSN
1389-4986

Authors
Harwood, Jessica M
Weiss, Robert E
Comulada, W Scott

Publication Date
2017-07-01

DOI
10.1007/s11121-017-0788-y

Peer reviewed

Published in final edited form as:

Prev Sci. 2017 July ; 18(5): 526–533. doi:10.1007/s11121-017-0788-y.

Beyond the primary endpoint paradigm: A test of intervention effect in HIV behavioral intervention trials with numerous correlated outcomes

Jessica M. Harwood[1], **Robert E. Weiss**[2], and **W. Scott Comulada**[3]

[1]Department of Medicine, Division of General Internal Medicine and Health Services Research, University of California Los Angeles, 911 Broxton Avenue, Los Angeles, CA 90024

[2]Professor, Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, 650 Charles E. Young Drive South, Los Angeles, CA 90095

[3]Department of Psychiatry and Biobehavioral Sciences, Semel Institute Global Center for Children and Families, University of California Los Angeles, 10920 Wilshire Blvd Suite 350, Los Angeles, CA 90024

Abstract

Behavioral interventions are increasingly based on holistic approaches to health with an understanding that health-related behaviors are linked. A motivating example is provided by the Philani study, an intervention trial conducted to improve the health of South African mothers and their children. Inter-related health problems around maternal alcohol use, malnutrition, and HIV were addressed; multiple endpoints were targeted. The traditional hypothesis testing paradigm that tests significance on one primary outcome did not suffice. Past multiple endpoint studies have utilized a sign test on the number of estimated differences between treatment and control that favor the intervention. However, in order to preserve type 1 error, one must account for correlations among the outcomes. We propose an alternative approach that counts the number of significant treatment-control differences. Monte Carlo simulation is used to adjust for correlation, providing updated critical values and p-values. Our method is implemented through an R package and applied to the Philani data to test the intervention's overall effect.

Keywords

HIV; randomized controlled trial; multiplicity; correlated outcomes; statistical methodology

Phone: 310-794-8518. Fax: 310-794-0732. jharwood@mednet.ucla.edu.

Disclosure of potential conflicts of interest: The authors declare that they have no conflict of interest.

Research involving Human Participants and/or Animals: The Philani study was approved by the Institutional Review Boards of the University of California, Los Angeles, Stellenbosch University, and Emory University and was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the Philani study.

INTRODUCTION

Primary endpoint study designs provide a natural basis for clinical trials where success can be quantified by a single measure, such as survival time. The need for a multiple endpoint paradigm has been acknowledged in a number of fields of study, such as chronic pain management. Treatment success is measured through multiple physical and emotional measures of pain (Turk et al., 2008). In HIV prevention trials, markers of success often entail multiple biological endpoints, including the presence of sexually transmitted infections, as well as HIV (Hartwell et al., 2013; Holtgrave, Leviton, Wagstaff, & Pinkerton, 1997; Fishbein & Pequegnat, 2000). HIV transmission behaviors are also key endpoints that include multiple measures of sexual behavior and substance use. The scope of interest is further widened in behavioral interventions where HIV prevention and treatment are intertwined with other markers of health (Gibson & Young, 1994; Aral & Peterman, 1996). For example, the Philani study (le Roux et al., 2013), a cluster randomized controlled trial conducted to improve the well-being of South African mothers and their children in the 6 months after childbirth, targeted 28 outcomes regarding HIV, nutrition, healthcare, mental health, and social well-being.

For clinical trials, the Consolidated Standards of Reporting Trials (CONSORT) Statement glossary warns that testing multiple study outcomes increases the chance of false findings of significance: “multiple statistical comparisons increase the probability of making a type 1 error, i.e. attributing a difference to an intervention when chance is the more likely explanation” (“Consort - Glossary,” n.d.). Thus, for trials that specify and test more than one primary outcome, best practice requires that analyses be adjusted for multiple comparisons (Tyler, Normand, & Horton, 2011). When testing multiple endpoints, several statistical methods are available to strongly control the familywise type 1 error rate, the probability of erroneously rejecting at least one true null hypothesis of no effect, regardless of which and how many of the null hypotheses are true. Examples of these methods include the Bonferroni, Holm, Hochberg, and Hommel procedures (Turk et al., 2008). Rather than controlling the familywise error rate, another option is to control the false discovery rate, the expected proportion of false findings of significance (Benjamini & Hochberg, 1995). However, lack of independence of outcome measures, when combined with testing a large number of primary outcomes like the 28 endpoints in the Philani study, can challenge the effectiveness of these methods. Other options for controlling type 1 error in the event of multiple hypothesis tests include composite outcomes, but composite outcomes can wash out effects on a few significant component measures (Turk et al., 2008; Holtgrave et al., 1997; Cordoba, Schwartz, Woloshin, Bae, & Gøtzsche, 2010; Lauer & Topol, 2003). Other considerations include methods for combining p-values, such as the Fisher, Lipak, Tippett, Sidak, and Simes combination tests (Westfall, PH, 2005), and the binomial method (Wilkinson, 1951; Brozek & Tiede, 1952; Sakoda, Cohen, & Beall, 1954). Again, this is only appropriate when p-values can be considered independent (Jones & Fiske, 1953).

In the context of comparing two treatments (intervention vs. control) on multiple outcomes, the sign test of group differences is a common application of the binomial method for combining multiple p-values from one study. This sign test of group differences is based on the nonparametric sign test (Daniel WW, 2009). However, rather than testing multiple observations of a single outcome, the interest lies in testing multiple primary outcomes in a single study by evaluating the number of estimated treatment-control differences that favor the intervention (Siegel S, 1988). It is assumed that the probability of success (the intervention being better than the control) is 0.5, the same as the probability failure. Using Monte Carlo simulation, Grønhaug et al. and Onwuegbuzie et al. show that, to preserve type 1 error, correlation between outcomes must be accounted for when performing a sign test of group differences, and provide tables of adjusted critical values to this end. The limitation is that when correlation between outcomes gets big enough, it is impossible to reject the null hypothesis of no intervention effect while preserving type 1 error < 0.05 (Grønhaug, Heide, & Gitlesen, 2000; Onwuegbuzie, Levin, & Ferron, 2011).

Thus, we propose an alternative approach that evaluates the number of significant estimated treatment-control differences that favor the intervention. We apply the binomial method but define “success” as a significant positive intervention effect, the result of a one-sided test. It is important to emphasize that the binomial method pertains to the count of successes and not the distributions of the study outcomes; the method is applicable for both continuous and discrete outcomes. Since a commonly-used type 1 error level for a 2-sided test is 0.05, we define the probability of success (1-sided test favoring the intervention) as half of 0.05, or 0.025. We account for the correlation between outcomes by using Monte Carlo simulation, providing updated critical values and p-values. The test is implemented via an R package. This method was applied in the Philani study. In this article, we discuss our test in more detail, lay out the general method that can be used across different study scenarios, and use a simulation study to generalize the results so that our method may be used in other HIV behavioral intervention trials.

METHODS

A test of intervention effect: multiple independent outcomes

To explain our methodology, we start with a simple case, assuming independence between outcome measures. Our test is based on the binomial distribution; the n Bernoulli trials are our n 1-sided univariate tests comparing intervention to control for n primary outcome measures. Under the null hypothesis of no intervention effect, we assume the n univariate tests are independent and identically distributed and that each univariate test may result in one of two mutually exclusive outcomes: (1) success, defined as a significant 1-sided univariate test favoring the intervention (“significant positive test”: 1-sided upper-tail p-value < 0.025), or (2) failure, defined as a non-significant 1-sided test (1-sided upper-tail p-value ≥ 0.025). The hypotheses test proportion p, with null hypothesis H_0: $p \leq p_0$ and alternative hypothesis H_A: $p > p_0$, where $p_0 = 0.025$. Our test statistic is x, defined as the count of successes (significant positive tests) out of the n univariate tests. In other words, in a study with n outcomes, x is the number of significant outcomes. We assume x follows a binomial distribution with n trials and a probability of success for each trial set to 0.025, $x \sim \text{Bin}(n, 0.025)$. For a one-sided binomial test, we set our type 1 error α equal to 0.05, and choose a critical value c such that c is the minimum value that satisfies the inequality

$$\sum_{i=c}^{n}\binom{n}{i}{p_0}^{i}(1-p_0)^{n-i}<0.05$$

We reject the null hypothesis if $x \geq c$. In practical terms, c gives us a cut off point for the number of significant results that are needed to declare the intervention is a success. For example, if we conduct separate regressions on each of n primary study outcomes, then we need to obtain at least c significant results to be confident that the intervention has a positive effect across outcomes.

A test of intervention effect: multiple correlated outcomes

However, if outcome measures are correlated, then our univariate tests are not independent. This would mean that the result of one univariate test would be influenced by the result of another univariate test, and the assumptions of a binomial test would be violated. Correlation among outcomes does not affect the expected number of significant positive univariate tests, but does affect the variance of the number of significant positive tests. We use Monte Carlo simulation, performed in R (version 2.11.1), to evaluate the effect of correlation between outcome measures on the binomial test's type 1 error behavior. To study the effects of global positive correlation among all outcomes on the number of significant positive tests assuming no intervention effect, we assume that each of the n univariate tests is a normal z-test. Z-statistics were assumed to come from an equi-correlated multivariate normal distribution. For each level of correlation ρ, from ρ=0 to 0.9 in steps of 0.1, we simulate 1,000,000 sets of n tests. For each set, we count x, the number of significant positive tests (defined as z >1.96) out of the total n tests. Next, we calculated the probability of observing x or more significant positive tests out of the 1,000,000 trials, for x running from 0 to n. For each level of correlation, for various n, we identify the critical value c such that $P(x \geq c) < 0.05$. We present these adjusted critical values of the number of significant positive tests needed to reject the null hypothesis for n=5 to n=50 outcomes in Table II.

Motivating Example

Our motivating example for this methodology is the Philani study, a cluster randomized controlled trial conducted to improve the health of South African mothers and their children in the 6 months after childbirth. Study details and results have been reported elsewhere (le Roux et al., 2013). Briefly, the Philani study evaluated the effect of home visits by Community Health Workers (CHW) on maternal and infant well-being for a sample of Cape Town, South Africa township women. Data on 28 inter-related health outcomes regarding HIV, nutrition, maternal alcohol use, healthcare, mental health, and social well-being were collected for participants in the intervention (PIP: 12 neighborhoods, 644 mothers) and control (SC: 12 neighborhoods, 594 total mothers) conditions (see Table I). Participants were assessed during pregnancy and reassessed at one week and six months post-birth. We found seven positive intervention effects and one negative intervention effect. Aware of the variety of primary outcomes of interest, our objective was to evaluate the overall effect of the Philani intervention on maternal and child well-being and to control for multiple comparisons.

RESULTS

Analysis of simulation results

The adjusted critical values are presented in Table II for n=5 to n=50 outcomes, at levels of correlation from 0 to 0.9 in steps of 0.1. Overall, correlation between outcomes does affect the type 1 error behavior of a binomial test. Generally, higher correlations lead to inflated type 1 error above the nominal level of 0.05. However, for moderate correlations of ρ=0.1 or ρ=0.2, only one or two more significant positive tests are needed to reject the null compared to when ρ=0. Furthermore, for a small number of outcomes, correlation does not change the critical value at all: for n=5 to n=9 outcomes, the adjusted critical value c is equal to two for all correlations. However, at n=10 and after, an increase in correlation leads to a gradual increase in the critical value up to a certain point, after which (often at about ρ=0.7 or ρ=0.8) the critical value decreases slightly. Furthermore, as the number of outcomes increases, the critical values increase more rapidly as correlation increases. For example, for n=25 outcomes, the difference in critical values (the number of significant results needed to declare a positive overall intervention effect) between ρ=0 and ρ=0.6 is 2 (ρ=0: c=3; ρ=0.6: c=5). However, at n=50 outcomes, the difference is an extra 4 significant results needed to reject the null (ρ=0: c=4; ρ=0.6: c=8).

For a given number of outcomes being tested, our advantage over the sign test is that, for outcomes n=5 to n=50, it is possible to reject the null of no intervention effect at all levels of inter-measure correlation, while still preserving a type 1 error of less than 0.05. When using the sign test of group differences under correlation, probability of success is 0.5. Thus, the test saturates fairly quickly, meaning that success for all outcomes appears with high probability even under the null. Therefore, for a given number of outcomes, if the correlation between outcomes gets big enough, statistical significance with type 1 error below 0.05 is not possible even in the case of success for all outcomes (Grønhaug et al., 2000; Onwuegbuzie et al., 2011). Alternatively, we have a probability of success of 0.025 under the null hypothesis, so our test does not saturate so quickly. For outcomes n=5 to n=50, it is possible to reject the null at all levels of correlation while preserving type 1 error < 0.05.

Applying the test to the motivating example

First, for the n=28 Philani primary outcomes, we estimated the correlation ρ via calculating the average of the absolute Pearson correlations between the outcomes. Absolute values of Pearson correlations are used because we expect positive correlations between outcomes in most instances, but wanted to allow for the possibility of negative correlations in the Philani data. For example, the Pearson correlation between “Asked partner to test for HIV” and “Discussed HIV status with sexual partner” was 0.41, while the Pearson correlation between “Asked partner to test for HIV” and “One feeding method” was −0.005; we took the absolute value of each, using 0.41 and 0.005. There was a total of 378 pair-wise Pearson correlations for our 28 outcomes; we took the absolute value of each one, then averaged the absolute values. Because variables included “true dichotomies” (e.g. “Asked partner to test for HIV”) and indicators created by dichotomizing continuous outcomes (e.g. “Weight-for-age z-score ≥ −2”), we also calculated the average of the absolute tetrachoric correlations, planning to use whichever method (Pearson vs. tetrachoric) produced a higher average absolute correlation. The average absolute correlation between the 28 measures using the Pearson and the tetrachoric correlations was 0.1 and 0.2, respectively. Next, using the critical values from Table II, we see that for n=28 and ρ=0.2, 4 outcomes are needed to reject the null. As explained elsewhere (le Roux et al., 2013), we then tested PIP's effect on 28 binary measures of maternal and infant well-being at a one-sided upper-tail alpha=0.025 using random effects logistic regression models to account for neighborhood clustering. Outcome variables were arranged so that higher positive z-statistics are good for the intervention, negative results are better for the control. As shown in Table I, PIP out-performed SC on 7 of 28 outcomes. Thus, we can declare PIP to have resulted in significantly better overall maternal and infant well-being over the first 6 months post-birth compared to SC. Using our R package we obtain a p-value of 0.005.

DISCUSSION

In sum, our test of the number of significant estimated treatment-control differences provides an overall index of the intervention's benefits. Importantly, we accounted for correlation between outcomes using Monte Carlo simulation, and provided tables of adjusted critical values. We then provided an example to illustrate how the test would be applied to study results. The test is implemented in the R-package BINOMCORR.TEST (R Development Core Team, 2006).

This test is a novel alternative to the sign test of group differences. Both provide an overall test of an intervention's effect and therefore a solution to the multiple testing problem. However, the binomial test is a useful alternative because it has the advantage of being able to reject the null of no intervention effect at all levels of inter-measure correlation while still preserving a type 1 error below 0.05.

We highlight that our test may be run on various types of primary outcomes. While our motivating example, the Philani study, chose binary variables as its primary outcomes, our method also works for trials testing other outcome types (e.g., continuous variables, count variables). Furthermore, our test may be run on a study that includes various types of analyses of primary outcomes. For example, each primary outcome could be analyzed in a different way, making it possible to determine the effect of an intervention that includes logistic regression (as in the Philani study), survival analysis, repeated measures analysis, and other analysis types. As explained in the Methods section, our test counts significant test statistics (a test showing the intervention performed significantly better than the control). Our simulation uses z-statistics (coming from a multivariate normal distribution), which are the test statistics used in the Philani analyses for logistic random effects regressions adjusting for neighborhood clustering. Z-statistics are also the test statistics for many other types of analyses used in clinical trials and research studies.

When using our test, an important distinction should be made between all variables for which data is collected in a study versus measures chosen as primary study outcomes. Our test is designed to be run on primary study outcomes only. It is not a substitute for established data combination and reduction practices; if for substantive and/or theoretical reasons several variables cannot each be considered important enough to be primary study outcomes and are better used as indicators of the same underlying process, then they should be combined into one primary outcome. For example, the Philani study collected multiple measures of risky drinking (from the Alcohol Use Disorders Identification Test-C [AUDIT-C]) and depression (from the Edinburgh Postnatal Depression Scale [EPDS]), combining each set of risky drinking and depression measures into a primary outcome (outcomes number 20 and 25 in Table I, respectively).

Another consideration is that there will likely be certain subsets of primary outcome measures that will be strongly intercorrelated, and correlated only weakly with others. To account for any variation in inter-measure correlation, a conservative approach would utilize the correlation associated with the highest critical value in Table II. In other words, if the set of observed inter-measure correlations includes a correlation that would result in a higher critical value than would the average correlation, then this “correlation of highest critical value” could be used when running the test. For example, as seen in Table II and noted in the Results section, often correlations of 0.7 result in the highest critical values.

A further consideration regarding correlations is that in practice, for the Philani study, we estimated the correlation ρ from the outcomes (using the average absolute Pearson or tetrachoric correlation), while in the Monte Carlo simulation ρ is defined as the correlation among the test statistics. Because our test counts significant test statistics, ideally a researcher would be able to estimate the correlation ρ directly from the actual test statistics used in the study. Future research should explore this as well as the trade-offs associated with approximating ρ from outcomes. Considerations include the relationship between outcomes and test statistics, missing data, and study design. Longitudinal studies and clustered data come with additional considerations, such as correlations of the particular covariance model or random effects, residual variance, and residual correlation.

While our proposed method is designed specifically to address type 1 error (in other words, statistical significance), this does not preclude the need to distinguish between statistical significance and clinical significance, including discussion of effect size and individual outcomes’ importance. We argue that issues surrounding clinical importance should be addressed at the outset of a study when planning study outcomes, and again at the end of a study, when discussing trial results. Any statistical results from our method should be discussed in light of the data and outcomes upon which the method is performed, and in light of the effect sizes shown for the individual tests and for our method overall.

Our method is designed to test the intervention as a whole. Thus, a statistically significant result means that the intervention has significantly more benefits relative to the control, and one can declare the intervention overall to be better than the control. Clinically, it will be important to discuss the percentage of outcomes with significant intervention effects in order to put the test results into context; the higher the percentage of outcomes with a significant intervention benefit, the stronger the clinical evidence that the intervention is meaningful overall. For example, in the Philani trial, our method resulted in the intervention having significantly better overall maternal and infant well-being using 7 significant outcomes out of a total of 28 outcomes, or 25%. Clearly, the trial could have been judged as more successful if a higher percentage of outcomes were significant. On the other hand, this is more successful than a scenario of 4 significant outcomes (14%), the minimum number for a statistically significant result using our method.

This minimum percent of outcomes needed to be significant varies depending on the number of outcomes tested and the correlation level. Looking at results from Table II, across the 5–50 outcomes and 0-0.9 correlation levels, an average of 18% of outcomes must show a significant intervention effect in order to declare an overall intervention benefit. As described in the Results section, while the number of significant results increases as the number of study outcomes increases, the percent of outcomes that need to be significant decreases. For example, at zero correlation, 40% of outcomes must be significant for a study with 5 outcomes (2/5), as opposed to 8% in a study with 50 outcomes (4/50). However, this differential in percent of outcomes needed to be significant shrinks for higher levels of correlation, e.g., at a correlation level of 0.5, the 5-outcome study still needs 40% of outcomes to significantly favor the intervention, but the 50-outcome study now needs 16% (8/50). Thus, as noted earlier, as the number of outcomes increases, the critical values increase more rapidly as correlation increases.

Putting these percentages into context, our method uses a stringent definition of “success” for each outcome measure (1-sided test favoring the intervention must have a p-value < 0.025). Under our definition, if there were no differences between intervention conditions, we would expect 2.5% of outcomes to have significant tests. Thus, for Philani, when testing 28 outcomes we would expect 28*0.025=0.7 significant tests on average (i.e. less than 1). As mentioned earlier, this is part of the test’s flexibility to accommodate correlation and makes our test more useful than the sign test of group differences, particularly for studies with small numbers of outcomes and/or large inter-outcome correlations. For example, in a study with 5 outcomes, the sign test of group differences would need success for all 5 outcomes, unless the correlation was larger than .05, in which case it would no longer be possible to reject the null hypothesis of no intervention effect (Onwuegbuzie et al., 2011). In a study with 50 outcomes, it is not possible to find an intervention effect if the correlation is more than 0.6 (Grønhaug et al., 2000).

Results from our method should also be interpreted in the context of the individual component outcomes and the data behind such outcomes. Clinical significance of a study depends on clear reporting of all component outcomes and individual test results, as is done in the Philani trial. Since those details have already been published, we do not repeat them in this paper, but urge any investigator using our method to report the results in a similar manner. This leads to a level of transparency necessary to provide a nuanced discussion of the clinical importance of the intervention, based on the effect sizes and p-values of significant (and insignificant) component measures. For example, as reported in the Philani study, looking at the individual outcomes tested, significant outcomes had a range of effect sizes. Significant estimated odds ratios ranged from 1.5 (outcome 3: Used a condom 10 of the last 10 times had intercourse at 6 months) to 3.6 (outcome 18: Exclusive breastfeeding first 6 months). Comparing the observed percentages between intervention conditions provides further intuitive context: PIP was three times more likely to follow exclusive breastfeeding for 6 months than the control (10% vs. 3%) and 1.3 times more likely to use a condom 10 of the last 10 times (44% vs. 34%). Significant outcomes had observed intervention-control percentage point differences between 5% and 26%, compared to 4% or less for insignificant outcomes. Based on these results, the study authors concluded that the Philani intervention produced modest effects (le Roux et al., 2013). Furthermore, statistically significant outcomes were those related to HIV prevention and child health/nutrition; thus, clinicians interested in improving these domains may judge Philani as more successful than investigators seeking to improve healthcare/monitoring, mental health, or social support.

Our method evaluates the overall effectiveness of an intervention and is designed specifically to control type 1 error in a study comparing intervention vs. control on multiple outcomes. As defined in the Philani study, our main analysis was the overall test of the intervention’s effect, and we defined our secondary analyses to be the tests of the intervention’s impact on individual outcomes. An important statistical note is that we considered our secondary analyses to be exploratory, and thus reported model p-values in lieu of a multiple testing adjustment when reporting individual outcomes’ results (le Roux et al., 2013).

While we compare our test directly to the sign test of group differences, future research should focus on comparisons to other solutions to the multiple testing problem. For example, the Bonferroni method assumes independence and is thus conservative relative to other procedures, while the false discovery rate and certain p-value combination tests are valid when data exhibit specific types of dependency (Benjamini & Yekutieli, 2001; Yekutieli, 2008; Westfall, PH, 2005). Further research should also focus on how to formally incorporate significant negative effects of the intervention into the test’s framework. For example, rather than only testing the number of significant intervention benefits (measures with a 1-sided upper-tail $p<0.025$), one could test the total number of significant positive and significant negative effects of the intervention (measures with a 2-sided $p<0.05$). Thus, rather than a test of intervention benefits, this would be a test of any overall intervention effect. Non-statistical judgment would weigh the individual positive and negative effects and decide if the intervention is better. This test can be implemented by changing the probability of success and the alternative hypothesis in the R package. Another option would be to perform two binomial tests, one for significant positive effects and a second for significant negative effects, and then adjust the two separate tests’ p-values using a multiple testing adjustment. Investigators could then compare the outcomes of the two tests to see if there is stronger evidence for intervention benefits or for negative intervention effects. Lastly, development of a multinomial test of significant positive results, significant negative results, and non-significant results could formally combine all possible results into one test.

Acknowledgments

Funding: This research was supported by the National Institutes of Health Grant numbers R01AA017104, P30MH058107, P30AI028697, UL1TR000124, K01MH089270, and R01DA032619. We would like to thank Dr. Mary Jane Rotheram for permission to use the Philani Project data.

References

Aral SO, Peterman TA. Measuring outcomes of behavioural interventions for STD/HIV prevention. International Journal of STD & AIDS. 1996; 7(Suppl 2):30–38. [PubMed: 8799792]

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995; 57(1):289–300.

Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics. 2001; 29(4):1165–1188.

Brozek J, Tiede K. Reliable and questionable significance in a series of statistical tests. Psychological Bulletin. 1952; 49(4:1):339–341. [PubMed: 12983453]

Consort - Glossary. (n.d.). Retrieved September 1, 2016, from http://www.consort-statement.org/resources/glossary#M

Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ (Clinical Research Ed). 2010; 341:c3920.

Daniel, WW. Biostatistics: A Foundation for Analysis in the Health Sciences. Hoboken, NJ: Wiley; 2009.

Fishbein M, Pequegnat W. Evaluating AIDS prevention interventions using behavioral and biological outcome measures. Sexually Transmitted Diseases. 2000; 27(2):101–110. [PubMed: 10676977]

Gibson DR, Young M. Importance of appropriate outcome measures in HIV behavioral research. Journal of Acquired Immune Deficiency Syndromes. 1994; 7(6):631–632.

Grønhaug K, Heide M, Gitlesen JP. Sign Tests when Observations are Correlated: A simulation study. Scandinavian Journal of Educational Research. 2000; 44(3):325–334. http://doi.org/10.1080/713696672.

Hartwell TD, Pequegnat W, Moore JL, Parker CB, Strader LC, Green AM, NIMH Collaborative HIV/STD Prevention Trial Group. The utility of a composite biological endpoint in HIV/STI prevention trials. AIDS and Behavior. 2013; 17(9):2893–2901. http://doi.org/10.1007/s10461-013-0501-5. [PubMed: 23748863]

Holtgrave DR, Leviton LC, Wagstaff DA, Pinkerton SD. Cumulative Probability of HIV Infection: A Summary Risk Measure for HIV Prevention Intervention Studies. AIDS and Behavior. 1997; 1(3): 169–172. http://doi.org/10.1023/B:AIBE.0000002977.08417.5e.

Jia J, Weiss RE. Common predictor effects for multivariate longitudinal data. Statistics in Medicine. 2009; 28(13):1793–1804. http://doi.org/10.1002/sim.3589. [PubMed: 19360840]

Jones LV, Fiske DW. Models for testing the significance of combined results. Psychological Bulletin. 1953; 50(5):375–382. [PubMed: 13100529]

Lauer MS, Topol EJ. Clinical trials–multiple treatments, multiple end points, and multiple lessons. JAMA. 2003; 289(19):2575–2577. http://doi.org/10.1001/jama.289.19.2575. [PubMed: 12759328]

le Roux IM, Tomlinson M, Harwood JM, O'Connor MJ, Worthman CM, Mbewu N, Rotheram-Borus MJ. Outcomes of home visits for pregnant mothers and their infants: a cluster randomized controlled trial. AIDS (London, England). 2013; 27(9):1461–1471. http://doi.org/10.1097/QAD.0b013e3283601b53.

Lin X, Ryan L, Sammel M, Zhang D, Padungtod C, Xu X. A scaled linear mixed model for multiple outcomes. Biometrics. 2000; 56(2):593–601. [PubMed: 10877322]

Onwuegbuzie AJ, Levin JR, Ferron JM. A Binomial Test of Group Differences with Correlated Outcome Measures. Journal of Experimental Education. 2011; 79(2):127–142.

R Development Core Team. BINOMCORR.TEST. Vienna, Austria: R Foundation for Statistical Computing; 2006. R: A Language and Environment for Statistical ComputingRetrieved from http://www.R-project.org

Sakoda JM, Cohen BH, Beall G. Test of significance for a series of statistical tests. Psychological Bulletin. 1954; 51(2:1):172–175. [PubMed: 13155709]

Siegel, S. Nonparametric statistics for the behavioral sciences. 2nd. New York: McGraw-Hill; 1988.

Turk DC, Dworkin RH, McDermott MP, Bellamy N, Burke LB, Chandler JM, Witter J. Analyzing multiple endpoints in clinical trials of pain treatments: IMMPACT recommendations. Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials. Pain. 2008; 139(3):485–493. http://doi.org/10.1016/j.pain.2008.06.025. [PubMed: 18706763]

Tyler KM, Normand SLT, Horton NJ. The Use and Abuse of Multiple Outcomes in Randomized Controlled Depression Trials. Contemporary Clinical Trials. 2011; 32(2):299–304. http://doi.org/10.1016/j.cct.2010.12.007. [PubMed: 21185405]

Weiss RE, Jia J, Suchard MA. A Bayesian model for the common effects of multiple predictors on mixed outcomes. Interface Focus. 2011; 1(6):886–894. http://doi.org/10.1098/rsfs.2011.0041. [PubMed: 22419987]

Westfall, PH. Encyclopedia of Biostatistics. New York: Wiley; 2005. Combining P values.

Wilkinson B. A statistical consideration in psychological research. Psychological Bulletin. 1951; 48(3):156–158. [PubMed: 14834286]

Yekutieli D. False discovery rate control for non-positively regression dependent test statistics. Journal of Statistical Planning and Inference. 2008; 138(2):405–415. http://doi.org/10.1016/j.jspi.2007.06.006.

Table I

Philani study outcomes tested for differences between intervention conditions (le Roux et al., 2013)

HIV-related preventive acts	
Among mothers with a current sexual partner	
1. Asked sexual partner to test for HIV at 6 months	
2. Discussed HIV status with sexual partner at 6 months	
3. Used a condom 10 of the last 10 times had intercourse at 6 months	*
Among HIV+ mothers	
4. Mother knows last CD4 count at 6 months	*x*
PMTCT	
5. Mother took AZT prior to labour, or full-ARVs	
6. Mother took AZT during labour, or full-ARVs	
7. Mother took NVP tablet at onset of labour, or full-ARVs	
8. Infant given NVP syrup within 24 hours of birth	*
9. AZT dispensed for infant and medicated as prescribed	*
10. Took infant to 6-week HIV PCR test and fetched results	
11. One feeding method first 6 months: formula or breastfeeding	*
Child health and nutrition	
12. Birth weight ≥ 2500 grams	
13. Weight-for-age z-score ≥ −2 at 6 months	
14. Height-for-age z-score ≥ −2 at 6 months	*
15. Weight-for-height z-score ≥ −2 at 6 months	
16. Head-circumference-for-age z-score ≥ −2 at 6 months	
17. Number of months breastfed exclusively > median of 3	*
18. Exclusive breastfeeding first 6 months	*
19. Drank no alcohol the month prior to giving birth	
20. No risky drinking at 6 months (AUDIT-C score ≤ 2)	
Healthcare and monitoring	
21. 4 or more antenatal clinic visits (4 is standard practice)	
22. Mother free of post-birth complications through 6 months	
23. Mother tested for TB at 6 months	
24. Number of 6-month RTHC immunisations > median of 11 (16 total)	

Mental health

25. Not depressed at 6 months (EPDS ≤ 13)

Social support

26. 6-month number of close friends or relatives × frequency of contact > median of 16

27. Father acknowledged infant to family at 6 months

28. Receiving child support grant at 6 months

[*] Philani Intervention Program significantly better than Standard Care.

[x] Philani Intervention Program significantly worse.

Table II

Critical values of the number of significant positive results needed to reject the null hypothesis at $\alpha < 0.05$ for different numbers of outcomes (*n*), at each level of correlation between outcomes.

n	**Correlation between outcomes**									
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
5	2	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2
8	2	2	2	2	2	2	2	2	2	2
9	2	2	2	2	2	2	2	2	2	2
10	2	2	2	2	3	3	3	3	2	2
11	2	2	2	3	3	3	3	3	3	2
12	2	2	3	3	3	3	3	3	3	2
13	2	3	3	3	3	3	3	3	3	2
14	2	3	3	3	3	3	3	3	3	2
15	3	3	3	3	3	3	3	3	3	3
16	3	3	3	3	3	3	4	3	3	3
17	3	3	3	3	3	4	4	4	3	3
18	3	3	3	3	4	4	4	4	4	3
19	3	3	3	3	4	4	4	4	4	3
20	3	3	3	4	4	4	4	4	4	3
21	3	3	3	4	4	4	4	4	4	3
22	3	3	3	4	4	4	4	4	4	3
23	3	3	4	4	4	4	4	4	4	3
24	3	3	4	4	4	4	5	5	4	4
25	3	3	4	4	4	5	5	5	5	4
26	3	4	4	4	4	5	5	5	5	4
27	3	4	4	4	5	5	5	5	5	4
28	3	4	4	4	5	5	5	5	5	4
29	3	4	4	5	5	5	5	5	5	4
30	3	4	4	5	5	5	5	5	5	4

n	**Correlation between outcomes**									
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
31	3	4	4	5	5	5	6	6	5	4
32	3	4	4	5	5	6	6	6	6	4
33	3	4	4	5	5	6	6	6	6	4
34	4	4	5	5	5	6	6	6	6	5
35	4	4	5	5	6	6	6	6	6	5
36	4	4	5	5	6	6	6	6	6	5
37	4	4	5	5	6	6	6	7	6	5
38	4	4	5	5	6	6	7	7	6	5
39	4	4	5	6	6	6	7	7	6	5
40	4	4	5	6	6	7	7	7	7	5
41	4	5	5	6	6	7	7	7	7	5
42	4	5	5	6	6	7	7	7	7	5
43	4	5	5	6	6	7	7	7	7	6
44	4	5	5	6	7	7	7	8	7	6
45	4	5	6	6	7	7	8	8	7	6
46	4	5	6	6	7	7	8	8	8	6
47	4	5	6	6	7	7	8	8	8	6
48	4	5	6	6	7	8	8	8	8	6
49	4	5	6	7	7	8	8	8	8	6
50	4	5	6	7	7	8	8	8	8	6