Estimating Human Immunodeficiency Virus (HIV) Prevention Effects in Low-incidence Settings

Jacqueline E. Rudolph, Stephen R. Cole, Joseph J. Eron, Angela D. Kashuba, and Adaora A. Adimora

Background: Randomized controlled trials (RCTs) for determining efficacy of preexposure prophylaxis (PrEP) in preventing human immunodeficiency virus (HIV) infection have not been conducted among US women because their lower HIV incidence requires impractically large studies. Results from higher-incidence settings, like Sub-Saharan Africa, may not apply to US women owing to differences in age, sexual behavior, coinfections, and adherence.

Methods: We propose a novel strategy for evaluating PrEP efficacy in the United States using data from both settings to obtain four parameters: (1) intention-to-treat (ITT) and (2) per-protocol effects in the higher-incidence setting, (3) per-protocol effect generalized to the lower-incidence setting, and (4) back-calculated ITT effect using adherence data from the lower-incidence setting. To illustrate, we simulated two RCTs comparing PrEP against placebo: one in 4000 African women and another in 500 US women. We estimated all parameters using g-computation and report risk ratios averaged over 2000 simulations, alongside the 2.5th and 97.5th percentiles of the simulation results.

Results: Twelve months after randomization, the African ITT and per-protocol risk ratios were 0.65 (0.47, 0.88) and 0.20 (0.08, 0.34), respectively. The US ITT and per-protocol risk ratios were 0.42 (0.20, 0.62) and 0.17 (0.03, 0.38), respectively. These results matched well the simulated true effects.

Conclusions: Our simple demonstration informs the design of future studies seeking to estimate the effectiveness of a treatment (like PrEP) in lower-incidence settings where a traditional RCT would not be feasible. See video abstract at, http://links.lww.com/EDE/B506.

Keywords: Generalization; HIV; Intention-to-treat effect; Per-protocol effect; Preexposure prophylaxis; Simulation

(Epidemiology 2019;30: 358–364)

Submitted July 2, 2018; accepted January 2, 2019.

From the aDepartment of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC; bSchool of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC; and cSchool of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC.

This work was supported by grant T32ES007018 from the National Institute of Environmental Health Sciences and grants R01AI100654, U01AI103390, UM1 AI068619, and P30AI50410 from the National Institutes of Health.

The authors report no conflicts of interest.

Randomized controlled trials (RCTs) are the gold standard for determining the efficacy of a treatment, but variations in adherence can compromise the generalizability of RCT results to groups that differ from the study population in which the RCT was conducted. Adimora et al1 discussed an approach to estimate the intention-to-treat (ITT) effect in a target population whose outcome incidence is substantially lower than the incidence in the population where the treatment was formally tested using a RCT. The authors outlined how to use key data from both samples, such as biologically determined adherence and baseline effect measure modifiers, and the per-protocol effect (i.e., the effect of remaining on trial protocol, usually of remaining adherent to the trial drug) as an intermediary step to obtain the desired estimates in the target population. The approach discussed by Adimora et al. is important when the event of interest is uncommon enough in the target population to render a local RCT infeasible. In brief, the steps of this approach are (1) to estimate both the ITT and (2) per-protocol parameters in the higher-incidence setting, then (3) to generalize the per-protocol parameter estimate to the lower-incidence setting, and finally and innovatively (4) to estimate the ITT in the lower-incidence setting by combining the generalized per-protocol estimate with information on observed compliance in the target setting. Although we believe this approach has many potential applications, we describe here how it might be implemented for the example described by Adimora et al.1, namely, determining the effectiveness of pre-exposure prophylaxis (PrEP) as prevention for human immunodeficiency virus (HIV) transmission. We provide a summary of simulation experiments that detail how the proposed approach works in typical sample sizes, and we give examples of the types of variables one would need to measure and include when adapting our approach to real-world data.

THE MOTIVATING EXAMPLE

PrEP has been shown in some populations to be effective at preventing transmission of HIV. For instance, among men who have sex with men (MSM), tenofovir/emtricitabine reduced incidence of HIV by 44% compared to placebo (hazard ratio [HR]: 0.56; 95% CI = 0.37, 0.85).2 One trial in serodiscordant, heterosexual couples in Kenya and Uganda found that tenofovir/emtricitabine reduced HIV incidence by 75% compared to placebo (HR: 0.25; 95% CI = 0.13, 0.45).3
However, we currently have a limited understanding of how well PrEP prevents HIV among women in the United States.

There are two potential reasons for the lack of information on the effectiveness of PrEP among US women. First, it would be infeasible to conduct a phase III RCT in this population because the incidence of HIV is too low. Even among US women at exceptionally high risk, whose HIV incidence rate was found by one prior study to be 320/100,000 person-years, one would need to enroll 10,000 women to have adequate statistical power to detect the same effectiveness observed in MSM. This would be infeasible not just because of the required sample size but also the difficulty of enrolling women who are at highest risk for HIV in the United States. Second, results from populations of women in Southern Africa, where incidence rates have been observed to be as high as 5700/100,000 person-years, are likely not directly generalizable to US women. There is concern regarding generalizability because distributions of key demographic, socioeconomic, clinical, and behavioral factors are likely to differ between US women and African women at risk for HIV. Generalization of results is made even more difficult because the trials conducted in African women (e.g., the FEM-PrEP [Preexposure Prophylaxis Trial for HIV Prevention among African Women] and VOICE [Vaginal and Oral Interventions to Control the Epidemic] trials) had to be stopped owing to the lack of demonstrated effectiveness of tenofovir/emtricitabine. In these trials, adherence to study protocol was only 30% to 40%, so the observed ITT analyses were essentially futile (i.e., undetectable regardless of actual efficacy). Additionnally (and more importantly for our purposes), the differences between African and US women in demographic, clinical, and behavioral factors make it likely that the adherence patterns observed in the African trials differ from what would be seen among US women.

We can overcome the above challenges by combining key data (particularly a valid measure of adherence) from both the higher-incidence African and lower-incidence US settings with modern analytic methods to obtain both the per-protocol and ITT effects in the target population under stated assumptions.

**METHODS**

**The Proposed Approach**

Our approach consists of four steps: (1) obtain the ITT and (2) per-protocol effects in the higher-incidence African setting, (3) generalize the per-protocol results to the lower-incidence US setting, and (4) estimate the ITT effect in the United States. Although per-protocol results have been generalized to a target population, to our knowledge the last step is novel.

Before conducting the analyses, we generated 2000 simulations of the two populations of interest. Our higher-incidence sample mimicked a placebo-controlled RCT of the effect of tenofovir/emtricitabine on prevention of HIV transmission in 4000 women in Southern Africa (specifically, the FEM-PrEP and VOICE trials). Our lower-incidence sample mimicked a RCT of 500 women in the United States (with variable distributions being informed by HPTN 064). In both settings, the populations were women at risk of acquiring HIV, with the expectation that some but not all women would be in serodiscordant partnerships. In each trial, we randomized half the participants to receive tenofovir/emtricitabine and half to placebo and, if they were assigned to PrEP, determined their adherence (treated simply as a binary variable). In Africa, half of the women randomized to PrEP were expected to take it at a level required by the study protocol, e.g., a stipulation that women take 80% to 90% of dispensed pills or have 10 ng of tenofovir per milliliter of plasma. In the United States, 75% of those randomized to PrEP were expected to follow protocol, using the same definition as in Africa. Whether women adhered was dependent on the woman’s age (defined as ≥21 years or not). We then simulated whether they acquired HIV 12 months after randomization, based on adherence, age, and the interaction between adherence and effect measure modifier lack of a sexually transmitted infection (STI) at baseline. Lack of an STI was selected as our example modifier because the distribution of STIs differs between the two settings and having an STI increases risk of acquiring HIV. Figure 1 illustrates the causal relationships between the example variables included in our simulation.

Starting with the higher-incidence African setting, our first parameter of interest was the ITT effect of randomization to PrEP versus placebo on HIV incidence. We estimated the ITT risk ratio (RR) in two ways: (1) using a log binomial generalized linear regression model and (2) using the Snowden adaptation of Robin’s generalized computational algorithm formula (g-formula). It was not necessary to include any covariates in the models to estimate the ITT.

To account for noncompliance in the PrEP arm, we then obtained an estimate of the effect of remaining adherent to trial protocol on HIV seroconversion in the African trials, i.e., the per-protocol parameter. This method corrects for adherence by analytically censoring participants when they no longer adhere to trial protocol, following a predetermined rule.
such as described above. Because this process involves analysis of adherence to protocol, which is by definition a post-randomization variable (and is not guaranteed to be balanced across the trial arms), we must account for (but not stratify on) those variables that confound the relationship between protocol adherence and the outcome.\(^7,8,19\)

The first method we used to estimate the per-protocol effect was a log binomial model among women who adhered to the protocol, weighted to represent the entire trial using stabilized inverse probability of censoring weights.\(^8\) Variables included in the censoring weight models were those that affect adherence and the outcome, so our weight models included our example confounder age. The second method was g-computation. The key to using the g-formula to estimate the per-protocol effect is in specifying the counterfactuals being included our example confounder age. The second method was g-computation. The key to using the g-formula to estimate the per-protocol effect is in specifying the counterfactuals being compared: (1) set all participants to receive PrEP and adhere to protocol versus (2) set all participants to receive placebo.\(^20\)

When modeling the outcome, the model must include confounders of compliance and the outcome. Using the parametric g-formula (g-computation), we estimated risk of the outcome under counterfactual 1 and under counterfactual 2 and then compared those risks to obtain the RR. After estimation of the per-protocol effect in the higher-incidence African setting, one can standardize the results to the target population using the known distributions of effect measure modifiers that differ between the populations.\(^21\) In our scenarios, lack of an STI was our example baseline effect measure modifier. We again estimated this parameter in two ways. First, we estimated stabilized inverse odds of selection weights in a combined data set of the African and US trials, where selection was defined as inclusion in the African trial. The only variable that needed to be included in the weight models was lack of an STI.\(^22\) Those weights were multiplied by the censoring weights, and we ran inverse-probability-weighted log binomial models in the African data to obtain the RR for the per-protocol estimate generalized to the lower-incidence US setting. Second, we adapted our g-computation approach for the per-protocol estimate to include an initial step where we took a weighted sample from the African women, where participants were selected with replacement and with a weight determined by our previously estimated inverse odds of selection weights. This ensured that, although the distribution of all other variables remained the same as in African, the distribution of any baseline modifiers would be (on average) the same as in the United States. The per-protocol effect was then estimated in the same way as above but now within this weighted sample.

In the final step, we obtained the ITT effect generalized to the United States, using the reweighted African sample. We did this by combining information on adherence in the target, lower-incidence US setting with the g-computation estimators of the generalized per-protocol parameter. This involved adding a step in which we used the US data to model adherence among those randomized to PrEP. The coefficients from this US adherence model were then used when generating the outcomes, instead of using the coefficients from the African adherence model. Finally, we obtained RRs by comparing the risks under the scenarios where (1) all participants were set to receive PrEP but adherence was allowed to be what it would have been in the United States and (2) all participants were set to receive placebo.

Several assumptions are required to identify the effects of interest in our approach. For all four steps, one must assume that the treatment is well defined (i.e., there is no interference and any variations in treatment are irrelevant),\(^23,24\) there is no measurement error, and all models are correctly specified. Perhaps more importantly for the design of a study seeking to use our approach, one must make the assumption of conditional exchangeability for internal validity\(^25\) and its external validity counterpart.\(^26\) For the ITT effect, randomization of treatment grants exchangeability in expectation. When estimating the per-protocol effects in steps 2 and 3, one has to assume that all confounders of adherence to protocol and the outcome have been accounted for, to ensure that those who adhered to protocol in the treatment arm are conditionally exchangeable with those who adhered to protocol in the comparator arm. Analogously, when generalizing to the target setting in steps 3 and 4, one has to assume that all baseline effect measure modifiers which also affect selection into one setting versus another have been accounted for, to ensure conditional exchangeability between those selected into the source population (higher-incidence setting) and those not selected (lower-incidence setting). Lastly, one has to assume there is a positive probability of being observed in all strata defined by the treatment (or, for external validity, selection) and those variables necessary to achieve conditional exchangeability.\(^26–28\)

Below we provide the results summarized across 2000 simulations. Key population variables are summarized using the median and interquartile range (IQR) of the percentiles in each simulation. We then report the RR at each step averaged across the simulations, alongside the 2.5th and 97.5th percentiles (hereafter referred to as the 95% central mass) of the 2000 RRs. We additionally give Monte Carlo simulation errors for each parameter, estimated by the standard error of the simulation RRs.

All analyses were conducted in SAS version 9.4 (SAS Institute, Cary, NC). Technical details on how each step was carried out using g-computation and SAS code are given in the eAppendices (http://links.lww.com/EDE/B458).

**RESULTS**

As shown in Table 1, across the 2000 simulations, 50.0% (IQR: 49.5%–50.5%) of the 4000 women enrolled in the African trial were randomly assigned to PrEP, but only 24.9% (IQR: 24.4%–25.4%) of the 4000 women actually received PrEP. One third (33.3%; IQR: 32.9%–33.9%) were >21 years of age, and 79.0% (IQR: 78.6%–79.4%) tested negative for...
TABLE 1. Distributions of Key Variables Across 2000 Simulations in the Higher HIV Incidence African Setting and Lower-Incidence US Setting

<table>
<thead>
<tr>
<th>Variable</th>
<th>African Trials (n = 4,000) Median (IQR) (%)</th>
<th>US Trials (n = 500) Median (IQR) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization to PrEP</td>
<td>50.0 (49.5–50.5)</td>
<td>50.0 (48.6–51.4)</td>
</tr>
<tr>
<td>Receipt of PrEP</td>
<td>24.9 (24.4–25.4)</td>
<td>37.0 (35.4–38.4)</td>
</tr>
<tr>
<td>Incident HIV</td>
<td>4.2 (4.0–4.4)</td>
<td>0.4 (0.2–0.8)</td>
</tr>
<tr>
<td>Age &gt;21 years</td>
<td>33.3 (32.9–33.9)</td>
<td>66.6 (65.4–68.2)</td>
</tr>
<tr>
<td>No STI at baseline</td>
<td>79.0 (78.6–79.4)</td>
<td>89.0 (88.2–90.0)</td>
</tr>
</tbody>
</table>

HIV indicates human immunodeficiency virus; IQR, interquartile range; PrEP, preexposure prophylaxis; STI, sexually transmitted infection (other than herpes simplex virus).

aAfrican age distribution based on Refs. 6 and 10. US age distribution based on Refs. 4 and 11.

bAfrican distribution based on Ref. 10, and US distribution based on Ref. 12.

TABLE 2. Intention-to-Treat and Per-Protocol Effects in the Higher HIV Incidence African Setting and in the Lower-Incidence US Setting as Estimated Using G-computation to Generalize from the African Trial and Directly Estimated in an Impractically Large Trial

<table>
<thead>
<tr>
<th>Parameter</th>
<th>African Trials (n = 4,000) RR (2.5th–97.5th)</th>
<th>US/African Composite Trials (n = 500/4,000) RR (2.5th–97.5th)</th>
<th>US Trial (n = 10,000) RR (2.5th–97.5th)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intention to treat</td>
<td>0.65 (0.47–0.88)</td>
<td>0.42 (0.20–0.62)</td>
<td>0.41 (0.19–0.68)</td>
</tr>
<tr>
<td>Per protocol</td>
<td>0.20 (0.08–0.34)</td>
<td>0.17 (0.03–0.38)</td>
<td>0.13 (0.00–0.31)</td>
</tr>
</tbody>
</table>

2.5th represents 2.5th percentile of the 2000 simulation RRs or, for the US trial, of 500 bootstrap resamples. 97.5th represents 97.5th percentile of the 2000 simulation RRs or, for the US trial, of 500 bootstrap resamples.

HIV indicates human immunodeficiency virus; RR, risk ratio.
of adherence were different. That adherence is a postrandom-
ization variable on the pathway between randomization and
the outcome means that it must be carefully handled during
the analysis. Estimation of the per-protocol effect is one way
to validly account for adherence, after which generalization
can be carried out in the usual way. The generalized per-
protocol effect can then be tweaked using measured data on
adherence in the lower-incidence target population to back-
calculate the ITT effect.

There were of course several limitations of this simula-
tion. Primarily, our design was simplified, with only one time
point and an unrealistically small variable set. Future applica-
tions of our approach will likely have many time points
at which they assess adherence and HIV incidence, and use
of such data will require adequately accounting for time-
varying adherence and time-varying confounders. However,
although we do not show it here, the methods used at each
step can be extended to handling the time-varying case, and
we plan in future work to adapt our approach to these more
complex settings. Moreover, assessment of variables in real
trials will generally not be as clean cut as shown here. For
instance, adherence can be difficult to measure, being highly
prone to measurement error, and may not always be a yes/no
binary variable. We also constructed our example such that
all women in both samples remained in the study until the
outcome was assessed and that there were no missing data;
no actual trial is likely to be so perfect. As above, though, our
approach could be extended to account for censoring or for
missing data. Additionally, we did not demonstrate here how
to obtain confidence intervals for the estimated RRs. When
adapting our approach, one way to obtain valid confidence
intervals would be to use bootstrap.

Above we discussed the assumptions necessary for our
approach. Whether these assumptions are reasonable will be
context dependent and will in part rely on careful study de-
design and data collection. For instance, here we only included
the minimal number of variables necessary to demonstrate the
approach. When using our approach, a researcher will need to
carefully consider the context and question of interest to de-
terminethe variable set sufficient to meet the exchangeability
assumption. For instance, although we only examined base-
line STI, a researcher using our approach might also attempt
to capture whether a woman was in a sero-discordant part-
nership or frequency of condom use, so these could also be
used as effect-measure modifiers. Our goal, though, as proof
of concept was simply to give examples of the types of vari-
ables one would need.

Furthermore, a particular implication of the no measure-
ment error assumption worth highlighting is the requirement
of using a valid measure of adherence in both populations
when estimating the per-protocol effect. Although we simply
used an arbitrary, dichotomized indicator of adherence to pro-
tocol, an investigator would ideally have an accurate measure
of drug concentration, by which they could ascertain whether

**FIGURE 2.** Boxplots of the simulation risk ratios with indi-
vidual simulation estimates for all four parameters estimated
using g-computation. The filled diamonds are the true risk
ratios, and the unfilled diamonds are the estimated risk ratios.
The boxes represent the interquartile range, with the me-
dian risk ratio being the line within the box. Each of the light
gray circles represents the result from a single simulation. The
breadth of the whiskers is 1.5 times the interquartile range.
Extreme outliers were removed from the picture (for the US
intention-to-treat [ITT], 15 risk ratios <1 × 10^{-4} and, for the US
per-protocol, 15 risk ratios <1 × 10^{-10}).

**DISCUSSION**

Here, we proposed an approach to generalize ITT and
per-protocol effects from a higher-incidence setting to obtain
estimates of those same effects in a lower-incidence setting.
In our example, HIV incidence in the United States was not
high enough to permit estimation of these effects without
enrolling an impractically large sample, but, using key data
from a trial of African women alongside modern statistical
methods, we were able to estimate both the generalized per-
protocol and ITT effects. Our results were consistent across
different estimators, i.e., the g-computation results matched
results obtained using inverse-probability-weighting methods.
Furthermore, our per-protocol results were comparable to the
“true RR” estimated from the individual counterfactual out-
comes, and our generalization results matched those we would
have seen if we enrolled a trial of 10,000 US women.

An approach like the one described here is necessary
if the goal is to estimate the ITT effect of a treatment in
the lower-incidence setting. This is because the ITT is a
function of the level of adherence within a population.
Therefore, if one generalized ITT from the higher-
icidence setting to the lower-incidence setting (which
requires measuring in both settings and analytically
controlling for the necessary effect measure modifiers), the
result would be biased if the patterns

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Africa ITT</th>
<th>Africa Per-Protocol</th>
<th>US ITT</th>
<th>US Per-Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Ratio</td>
<td>1.20</td>
<td>1.10</td>
<td>0.80</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.70</td>
<td>0.40</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.30</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.10</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
a woman received an effective dose of the drug. This information would then guide the formulation of the trial’s protocol regarding adherence. If an effective concentration was not known, the investigator could examine different cut points for drug concentration and compare the results of our approach for each cut point.

Some practical implications of our approach should also be noted. In our motivating PrEP example, the high-incidence trials from which we would generalize to the United States already exist. Thus, it would save a great deal of time and money to enroll the small US trial and apply our approach to estimate the desired parameters, in comparison to enrolling the 10,000 woman trial and estimate the parameters more directly. However, if our approach were to be used in scenarios where a high incidence study does not exist, this choice would not be so clean cut. The investigators would need to weigh the costs of designing and conducting studies in both settings sequentially or concurrently (as well as the “cost” of making the necessary assumptions for our statistical approach) against the cost of running one larger study in the low-incidence setting. This decision will likely be context-dependent, hinging on factors such as how large the study in the low-incidence setting would need to be (which would depend on the outcome incidence) and how well the investigators believe they could measure all the necessary variables to carry out our approach.

Despite its limitations, the proposed approach has the potential to be applied to a number of important public health questions. For instance, to return to our motivating example, understanding how well PrEP can prevent HIV incidence among at-risk women in the United States is an understudied but critical area of research. Assessment of PrEP efficacy in this population would require an impractically large trial because incidence is as low as 320 infections per 100,000 person-years (compared to 5700/100,000 person-years in Sub-Saharan Africa). However, our approach would allow researchers to estimate PrEP effectiveness while enrolling a much smaller trial of US women partly because it does not rely on measuring the outcome in the United States.

Moreover, our approach need not be limited to placebo-controlled RCTs in both settings. The strategy could be adapted for studies using an active comparator; such situations would require measuring adherence to the active comparator and accounting for this adherence in the per-protocol effect estimation. Observational studies could also serve as the data source. In an observational study, the ITT effect might compare women who were prescribed PrEP by a physician to a similar group of women who were not prescribed PrEP. Estimation of the per-protocol effect would, as in our example, require an accurate measure of adherence to PrEP. The use of observational data, however, would mean PrEP was no longer randomized, and exchangeability would no longer be expected when estimating the ITT. A researcher would thus need to additionally control for the confounders of being prescribed PrEP and HIV incidence. Accurate measures of all required variables (adherence, confounders, and effect measure modifiers) might also be more difficult to attain than in an RCT.

Thus, the simulations described here, while limited, inform the design of future studies that seek to examine the effectiveness and efficacy of a treatment (like PrEP) not just in higher-incidence settings but also in lower-incidence settings where a traditional RCT would not be feasible.

**ACKNOWLEDGMENTS**

We acknowledge our funding sources for support in completing this work.

**REFERENCES**

9. Lu H. Generalizing the adjusted per-protocol treatment effect using inverse probability weights. *Atlantic Causal Inference Conference, Chapel Hill, NC.*


