Published in final edited form as:

J Acquir Immune Defic Syndr. 2016 August 1; 72(4): 462–464. doi:10.1097/QAI.0000000000001005.

Re-thinking data collection for HIV prevention trials

Ofer Harel, PhD and Valerie Nazzaro, PhD

University of Connecticut

Abstract

There is a strong push to conduct large-scale randomized controlled study designs in HIV prevention studies. In these randomized controlled studies, the primary research objective is typically to determine the treatment effect based on some biological outcome (e.g. HIV infection). But many unutilized self-reported outcomes are also being collected. We illustrate the extent of this problem using the EXPLORE data as an example.

Keywords

HIV prevention trials; biological outcomes; self-report outcomes; data collection; study design

Introduction

There is a strong push to conduct large-scale randomized controlled study designs in HIV prevention studies. In these randomized controlled studies, the primary research objective is typically to determine the treatment effect based on some biological outcome (e.g. HIV infection). Millions of dollars are spent on these studies and lots of resources are spent towards collecting data on that biological outcome but also on many self-reported outcomes. However, many of these self-reported outcomes, while of great interest, are rarely used in subsequent analyses. Does the extent they are used justify the cost? Does it justify the researcher's and subject's burden? What are the alternatives?

Extensive data collection is a general problem and the EXPLORE study is just one example that can be used to compare the volume of data collected against what actually ends up getting used. The EXPLORE study was a multi-site randomized control trial to test the efficacy of a behavioral intervention to prevent acquisition of HIV among men who have sex with men (MSM). It was the first study of its kind to use HIV infection as the primary endpoint. The behavioral intervention consisted of 10 one-on-one counseling sessions followed by maintenance sessions every 3 months. The standard condition was twice-yearly Project RESPECT individual counseling. Participants under both conditions were tested twice each year for the HIV antibody.

Contact: Ofer Harel, Department of Statistics, University of Connecticut, 215 Glenbrook Rd., Storrs, CT, USA, 06269, Phone: (860) 486-6989, Fax: (860) 486-4113, ofer.harel@uconn.edu.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

At baseline, participants were asked to fill out a survey consisting of 229 questions regarding their risk behaviors. At each follow-up visit, participants were asked 219 questions regarding their risk behaviors. Over the course of the study, on average, participants in the treatment group spent 9 hours answering these risk assessment questions. But this is only a fraction of the data that was collected. Even more time was spent for participants to answer questions pertaining to counseling, STD's, use of post exposure prophylaxis, etc. By the end of the 4 years, approximately 5000 data points for each participant was collected.

When assessments are lengthy, there is risk that respondents may tire of a particular survey and not answer certain questions, they may change their responses in order to prompt fewer continuing questions, or may even drop out of the study entirely. Recruitment for these studies is already difficult, so keeping participants interested is an important concern. If all of the data being collected was used for analysis, then the lengths of these assessments may be justified. But in reality, it is not clear how much of the data are actually used.

Methods

Our review was performed using Google scholar. It enabled us to search across many sources (articles, theses, books, online repositories, universities, academic publishes, and professional societies). The study period ended in July, 2003. The article that presented the main findings of the EXPLORE study was used as our starting point. On November 26, 2014 all citations that met the following criteria were retrieved:

1. Referenced the main findings of the EXPLORE study.
2. Used the risk assessment data from the EXPLORE study for at least some portion of analysis
3. Not a meta-analysis

The search resulted in (n=13) citations that met the inclusion criteria. A reviewer assessed the studies and marked off what questions from the EXPLORE study had been used for analysis. In particular, the focus was to examine the risk assessment questions that were used, since those are the focus of most analyses of HIV prevention research.

Results

The results of the reviewing process are summarized in Table 1.

The primary analysis did not utilize any risk assessment behaviors (just HIV status and group assignment). In the secondary analyses, serodiscordant unprotected receptive anal intercourse, serodiscordant unprotected anal intercourse, and unprotected anal intercourse were examined across all time-points in the study. These analyses also utilized baseline risk assessment behaviors regarding injection and non-injection drug use, sex with HIV-positive males/females, unprotected receptive/insertive anal sex, alcohol use, and depression. Looking at the baseline survey alone, we estimate that data from between 33 to 66 of the 229 questions were utilized in these analyses. Looking collectively across all risk assessment surveys given over the 4 year study on risk behaviors, we estimate that data from 185 to 482 of the 1,981 questions were utilized in the secondary analyses.

While the variables utilized in each of the analyses were always clearly stated, it was less obvious what particular questions from the survey were used in data management to define each of the variables. In particular, information regarding serodiscordant unprotected receptive anal sex was mentioned (in some form) across 24 questions on the baseline survey and there was (at times) uncertainty from the reviewers whether all 24 instances were used in any given analysis. To account for this uncertainty, we give a range of our estimates and anticipate that the true value of questions used lies somewhere within the range provided.

At baseline, we estimate that between 75 (38%) and 99 (43%) of the 229 variables collected on risk assessment were used in at least one of the papers that met the inclusion criteria. Furthermore, we estimate that between 435 (22%) and 747 (38%) of the 1,981 variables collected across the 4-year study on risk assessment were used in at least one of the included papers.

Discussion

One reason we focused on risk assessment questions was to highlight that even the most prominently used questionnaire of the study is not being fully utilized. The EXPLORE team collected dozens of other surveys with thousands of other data points that are even more sparsely used than the risk assessment questions.

We realize that data is collected for a variety of reasons – at times it is collected with a specific analysis in mind, and at times it is collected for the purpose of discovering new relationships to be confirmed in future studies. Although it seems harmless to collect as much data as possible, the extraneous data requires extra work to extract and manage. These other variables could cause oversight on the more critical variables of interest. While the cost of an additional question may seem marginal during the survey design phase, the true cost comes from the impact on data quality. As data collection procedures move into more electronic formats, decline in attention span and its impact on data quality becomes even more problematic. This stresses the need to make data collection as efficient as possible.

It is common today to collect biological (blood) specimen, use part of it and freeze the rest for future use. Some would argue the same reasoning for the self-report questionnaires. However, the HIV epidemic is changing drastically all the time and most of the information from these studies will not be applicable for more than several years. There is definitely an expiration date on self-report and biological outcomes data. For example, while the EXPLORE data was collected, HIV drugs were very different than what is available now, treatment as prevention was not considered and community viral load was not an issue.

When surveys are constructed collaboratively, it is possible that researchers all come in with their own favorite set of questions. Perhaps as a way of appeasing all collaborators, all questions become part of the survey. However, at some point, only a specific set of variables are used for analysis, making inclusion of all questions superfluous.

Even if individuals agree to participate, researchers must keep in mind the value of a participant's time. Since HIV studies are at times conducted on marginalized and

socioeconomically disadvantaged populations, time spent on these surveys could be critical time away from earning money, seeking shelter, or caring for children.

From a statistician's perspective, there seems to be room for improvement in terms of data collection and study design. In particular, it is important that the whole research group will attempt to reduce patient burden and cost by using more appropriate and innovative study designs and procedures. It is possible to introduce "designed missingness", where each participant is randomly assigned a subset of questions or assigned to only participate on a subset of time points. With this approach, there will still be flexibility to explore new relationships, but will hopefully be less-burdensome for participants. There is evidence that suggests that this can be achieved without sacrificing the integrity of the inferences to be made. These ideas are used by the education research community. It may be time to start rethinking how these large-scale studies are designed to adequately justify how resources are spent.

Acknowledgments

This project was partially supported by Award Number K01MH087219 from the National Institute of Mental Health.

References

1. Koblin BA, Team ES. Effects of a behavioural intervention to reduce acquisition of HIV infection among men who have sex with men: the EXPLORE randomised controlled study. *The Lancet*. 2004; 364(9428):41–50.
2. Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*. 1993; 57(3):289–304.
3. Wakeman SE, Zaller ND, Flanigan TP, et al. HIV among marginalized populations in Rhode Island. *Medicine and health*. 2009; 92(7)
4. Strauss WJ, Ryan L, Morara M, et al. Improving cost-effectiveness of epidemiological studies via designed missingness strategies. *Statistics in medicine*. 2010; 29(13):1377–1387. [PubMed: 20527011]
5. Adams RJ, Lietz P, Berezner A. On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assessments in Education*. 2013; 1(1):1–27.
6. Koblin BA, Team ES. Effects of a behavioural intervention to reduce acquisition of HIV infection among men who have sex with men: the EXPLORE randomised controlled study. *The Lancet*. 2004; 364(9428):41–50.
7. Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*. 1993; 57(3):289–304.
8. Strauss WJ, Ryan L, Morara M, et al. Improving cost-effectiveness of epidemiological studies via designed missingness strategies. *Statistics in medicine*. 2010; 29(13):1377–1387. [PubMed: 20527011]
9. Adams RJ, Lietz P, Berezner A. On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assessments in Education*. 2013; 1(1):1–27.
10. Koblin BA, Husnik MJ, Colfax G, et al. Risk factors for HIV infection among men who have sex with men. *Aids*. 2006; 20(5):731–739. [PubMed: 16514304]
11. Colfax G, Coates TJ, Husnik MMJ, et al. Longitudinal patterns of methamphetamine, popper (amyl nitrite), and cocaine use and high-risk sexual behavior among a cohort of San Francisco men who have sex with men. *Journal of Urban Health*. 2005; 82(1):i62–i70. [PubMed: 15738319]

12. Mimiaga MJ, Noonan E, Donnell D, et al. Childhood sexual abuse is highly associated with HIV risk-taking behavior and infection among MSM in the EXPLORE study. *Journal of acquired immune deficiency syndromes (1999)*. 2009; 51(3):340. [PubMed: 19367173]
13. Chin-Hong PV, Husnik M, Cranston RD, et al. Anal human papillomavirus infection is associated with HIV acquisition in men who have sex with men. *Aids*. 2009; 23(9):1135–1142. [PubMed: 19390418]
14. Brown E, Wald A, Hughes J, et al. High risk of human immunodeficiency virus in men who have sex with men with herpes simplex virus type 2 in the EXPLORE study. *American journal of epidemiology*. 2006; 164(8):733–741. [PubMed: 16896053]
15. Salomon EA, Mimiaga MJ, Husnik MJ, et al. Depressive symptoms, utilization of mental health care, substance use and sexual risk among young men who have sex with men in EXPLORE: implications for age-specific interventions. *AIDS and Behavior*. 2009; 13(4):811–821. [PubMed: 18709453]
16. Menza TW, Hughes JP, Celum CL, Golden MR. Prediction of HIV acquisition among men who have sex with men. *Sexually transmitted diseases*. 2009; 36(9):547. [PubMed: 19707108]
17. Philip SS, Yu X, Donnell D, Vittinghoff E, Buchbinder S. Serosorting is associated with a decreased risk of HIV seroconversion in the EXPLORE Study Cohort. *PLoS One*. 2010; 5(9):e12662. [PubMed: 20844744]
18. Donnell D, Mimiaga MJ, Mayer K, Chesney M, Koblin B, Coates T. Use of non-occupational post-exposure prophylaxis does not lead to an increase in high risk sex behaviors in men who have sex with men participating in the EXPLORE trial. *AIDS and Behavior*. 2010; 14(5):1182–1189. [PubMed: 20490908]
19. Bedoya CA, Mimiaga MJ, Beauchamp G, Donnell D, Mayer KH, Safren SA. Predictors of HIV transmission risk behavior and seroconversion among Latino men who have sex with men in Project EXPLORE. *AIDS and Behavior*. 2012; 16(3):608–617. [PubMed: 21390540]
20. Barresi P, Husnik M, Camacho M, et al. Recruitment of men who have sex with men for large HIV intervention trials: analysis of the EXPLORE Study recruitment effort. *AIDS education and prevention: official publication of the International Society for AIDS Education*. 2010; 22(1):28. [PubMed: 20166785]
21. Eaton LA, Kalichman SC, Kenny DA, Harel O. A reanalysis of a behavioral intervention to prevent incident HIV infections: Including indirect effects in modeling outcomes of Project EXPLORE. *AIDS care*. 2013; 25(7):805–811. [PubMed: 23245226]
22. Tieu H-V, Li X, Donnell D, et al. Anal sex role segregation and versatility among men who have sex with men: EXPLORE Study. *Journal of acquired immune deficiency syndromes (1999)*. 2013; 64(1):121–125. [PubMed: 23945255]

Table 1

Number of risk assessment questions utilized by paper

Paper	Range of estimated number of baseline risk assessment questions used in analysis	Range of estimated number of risk assessment questions used over all time points
Koblin (2004)	33–66	185–482
Koblin et al. (2006)	43–64	387–576
Colfax et al. (2005)	21–27	189–243
Mimiaga et al. (2009)	63–99	207–531
Chin-Hong et al. (2009)	0	329–518
Brown et al. (2006)	27–36	243–324
Salomon et al. (2009)	30	30
Menza et al. (2009)	27–45	27–45
Philip et al. (2010)	33	185
Donnell et al. (2010)	0	224–368
Bedoya et al. (2012)	61–69	205–213
Barresi et al. (2010)	19–37	19–37
Eaton et al. (2013)	18	252
Tieu et al. (2013)	32–53	288–477
Total number of unique risk assessment questions used in analysis across all studies	75–99	435–747
Total number of risk assessment questions	229	1,981