PAA 2004 Abstract Submission
Session 106, Sexually Transmitted Diseases, Sevgi Aral, Chair
Extended Abstract

Sociodemographic Correlates of STD Biomarker Outcomes in a National Sample of Young Adults Dawn M. Upchurch, Ph.D., William M. Mason, Ph.D., and Edward W. Hook III, M.D.

Introduction and Background

Adolescents and young adults have the highest rates of sexually transmitted diseases (STDs) and comprise the target population of national public health strategies aimed at reducing transmission and acquisition of STDs. Despite the clear need for it, there has been little research characterizing the distribution of STDs among youth and young adults. Clinical studies that use biomarker data to measure the presence of STDs typically demonstrate contrasts within selected groups and hence are not suitable for population inference. Population-based sample surveys are well suited to the characterization of the distribution of STDs, but until quite recently it has not been feasible to combine the sampling design advantage of surveys with biomarker measurement of STD presence. Instead, studies based on surveys have relied on indirect reports. An indirect report is one in which a respondent reports the results of a biomarker test, which typically was first reported to the respondent by a health care worker. The primary disadvantage of indirect reporting is that it leaves uncontrolled the possibility of random and especially nonrandom reporting error introduced by the respondent. This is in addition to measurement error inherent in biomarker testing. The goal of the present paper is to report results on the distribution of STDs using a population-based sample survey *and* per-individual STD biomarker test findings. The research is made possible by the inclusion of biomarker measurement in the third wave of the National Longitudinal Study of Adolescent Health ("Add Health").

The work to be presented in this paper is part of a larger ongoing project (*i*) to model STD risk as a function of key sexual and protective practices, (*ii*) to explore how the interplay of relationship factors affects these practices, and (*iii*) to examine how the social environment shapes related attitudes, beliefs, and behaviors. Our research integrates more established sociological theories of social structure and social relations with new conceptual developments in social epidemiology and recognizes the biological contributions to STD risk. We have substantial experience and expertise working with Waves I and II of the Add Health data, we have already obtained Wave III, and we have begun data cleaning and coding.

This paper will be the first in our work with Wave III. In it, we will investigate the effects of a comprehensive set of individual-level sociodemographic characteristics on STD outcomes. Age, gender, race and ethnicity, nativity, religion and religiosity, education and enrollment status, family background, marital status, and socioeconomic

status define individuals' placement within social strata, and these factors structure social and sexual relationships. We expect to find both additive and interactive effects. For example, we anticipate that the effects of gender on STD risk are contingent on race/ethnicity. Because of the large sample size of the Add Health study and the considerable detail with which race/ethnicity is recorded, it will be possible to investigate risk among more homogeneously defined ethnic groups than has been previously possible (e.g., Mexican, Puerto Rican, Cuban and to a lesser extent, Chinese, Korean, Japanese, and Philipino). It will also be possible to investigate the potential influence of nativity and other measures of acculturation. The use of detailed race/ethnicity categories in the Add Health study allows not only for greater precision relative to other national datasets and CDC surveillance data, but also for better understanding of differentials than is possible using the less detailed, and more commonly seen, categorizations. More generally, the Add Health study has more comprehensive as well as more detailed socioeconomic measurement than is found in probably the majority of STD studies. This is important, for example, in the study of race and ethnic differentials. Technically, there is a reduction in omitted variable bias. Substantively, we are better able to assess the extent to which STD risk differences between race/ethnic groups are due to *non*-socioeconomic factors.

In the paper we will argue that, in modeling STD presence at the individual level as a function of sociodemographic characteristics, we are estimating a reduced form model. A full model will include factors that are predetermined with respect to STD presence, but endogenous with respect to sociodemographic characteristics. Such factors include sexual and protective practices, and psychosocial characteristics of respondents.

Data

The data to be used for these analyses are from the National Longitudinal Study of Adolescent Health, which is well-suited to the investigation of STD risk in adolescents and young adults. No other nationally representative data set: (*i*) includes such a large and diverse sample of adolescents followed through young adulthood; (*ii*) contains detailed sexual histories and relationship-specific information; and (*iii*) assesses STD histories and statuses using indirect reports and biomarker test findings, per individual.

The sampling frame consists of all US high schools. Each participating school provided a student roster that constituted the student-level sampling frame. From that listing, a *baseline sample* was drawn consisting of a core sample and several special over-samples. The *core sample* is a probability sample of size 12,105 that is nationally representative of adolescents enrolled in grades 7 through 12 during the 1994-95 academic year. The *over-samples* include: (a) ethnic—African-Americans from well educated families, Chinese, Cubans, and Puerto Ricans; (b) saturated schools—16 schools with all enrolled students selected; (c) disabled individuals; and (d) genetically related individuals—siblings residing in the same household. The combined baseline sample size is 20,745. The Wave II sample consists of all adolescents interviewed at Wave I, except for (*i*) the deletion of twelfth graders at Wave I who were not part of the genetic sample, (*ii*) the deletion of the Wave I disabled over-sample, and (*iii*) the addition

of 65 adolescents from the genetic sample who were not interviewed in Wave I. The sample size at Wave II is 14,738. *The Wave III sample consists of all respondents in the Wave I baseline sample, regardless of age or grade.* Wave III was collected in 2001-2002; individuals in the original, Wave I sample, are now 18-26 years old, and are at the ages of highest STD risk. The sample size for Wave III is 15,197; the follow-up rate from Wave I was 73 percent. Because of the possibility of attrition bias, our work will include an analysis of the extent to which sociodemographic factors are associated with dropout between Waves I and III. If necessary, we will model dropout jointly with STD incidence.

The STI and HIV status of each respondent was measured by bioassay at the time of the Wave III interview.[1] The Ligase Chain Reaction (LCR) assay was used to detect *Chlamydia trachomatis* and *Neisseria gonorrhoeae* (Abbott LCx). The assay involves two tests (one for each pathogen) on the same aliquot of urine. The LCx urine assay demonstrates excellent sensitivity and specificity (approximately 90 percent and 99 percent, respectively) for both men and women and for both pathogens. For HIV detection, ELISA and Western Blot tests were used (Orasure kits). Trichomoniasis was assessed using the Polymerase Chain Reaction (PCR) assay developed at UNC. About eight percent of respondents refused to provide the necessary biomarker specimens for evaluation. As part of our research, we will analyze the probability of refusal, since refusal potentially introduces selection bias.

Analytic Strategy

The goal of our analysis is to simultaneously estimate gender, race/ethnicity, and socioeconomic differentials in risk, including interactions between factors, and to do so in considerable detail. As part of this work, we will provide some of the first national prevalence estimates for gonorrhea, chlamydia, and trichomoniasis, using the Wave III Add Health data.[2] Our primary purpose here will be to examine relative differentials in STD risk.

The clustered sampling design at Wave I *potentially* induces dependence between observations within clusters. Since the response variables of interest were measured at Wave III, the dependence may turn out to be negligible. For now, this remains an empirical question to be answered. The answer can be obtained through the use of generalized linear mixed models. Because the initial design was hierarchical with several layers of nesting, the error structure, even for a model involving only a single response variable (e.g., a single STD), may be complex. We are experienced with models of this kind, and routinely allow for layered random errors in our work with the Add Health data.

[1] Because so few HIV cases (16) were detected by the Wave III bioassay, we exclude HIV from our analyses. However, we include HIV here, in the discussion of biomarker testing, in order to provide a complete description of the types of bioassay used by Add Health.
[2] Handcock et al. (2003) presented initial results for gender crossed with African American vs. non-African American.

Initially, we will explore the impact of sociodemographic covariates separately for the risk of gonorrhea, chlamydia, and trichomoniasis. Let θ be the conditional expectation of suitably transformed values of *Y*, where Y is one of the tests, such that θ can be expressed as a linear combination of covariates. (For a binary response, the logit and probit transformations are "suitable.") We will estimate

$$\theta_i = \boldsymbol{x}_i \beta \,, \qquad (1)$$

where $\boldsymbol{x}_i$ is a vector of individual-level covariates that includes gender, age, race and ethnicity, nativity, marital status, religion, education and enrollment status, employment status and earnings, size and type of place of residence, and region; β is a vector of their effects on biomarker outcomes. Based on surveillance statistics and the findings of Handcock et al. (2003), we expect higher STD incidence among women, African Americans, among those who are single, younger, of lower SES, and among those who live in central cities and in the South. We also expect to find important interactions, such as that between gender and race/ethnicity. For example, we expect to find higher STD rates among black and Hispanic women and among black men, but lower rates among white and Hispanic men, when compared to white women. We have no priors on what will be found when all of the other relevant sociodemographic characteristics are simultaneously controlled. That is one of the reasons for exploring this specification and its specializations. No previous national study has assessed STD risk using biomarkers, and neither surveillance data nor data from impacted populations can provide these detailed comparisons.

Thus far, we have specified separate equations for gonorrhea, chlamydia, and trichomoniasis. It will be appropriate and desirable to examine the estimated regressions across the STD outcomes. Depending on the outcome of initial analyses, it may be helpful to estimate models that take account of within-individual clustering. We will return to this point in the following section.

Preliminary Results

Table 1 presents unweighted, *univariate* percentaged distributions for the Wave III chlamydia, gonorrhea, and trichomoniasis biomarker test results. Because each participating individual contributed one urine sample, which was then subsampled for each test, the number of refusals in each distribution is necessarily the same (7.8 percent). As noted above, in the paper we will model refusals. In particular, it is of considerable interest to determine whether the probability of refusal varies with the same socioeconomic covariates that will be used to model presence/absence of STDs. If there are important factors in determining refusal/cooperation that are not socioeconomic in nature, and if these factors are measured in the Add Health data, treatment of the selection problem will be relatively straightforward.

Table 1. Unweighted, univariate percentage distributions of biomarker tests for chlamydia, gonorrhea, and trichomoniasis

Outcomes	Chlamydia	Gonorrhea	Trichomoniasis
No results	4.60	10.63	5.28
Positive	4.24	0.38	2.24
Negative	83.36	81.19	84.68
Refused	7.80	7.80	7.80
Total	100.00	100.00	100.00

Source: Wave III, Add Health Study. $N = 15{,}197$ for each distribution.

Table 1 also shows that there were non-negligible numbers of tests that failed to produce either a positive or a negative result. Test failures were most likely in the gonorrhea test, due to technical decisions in the testing process. It is uncertain whether re-tests of the "no result" gonorrhea tests will be carried out. Although we would prefer a lower "no result" rate for gonorrhea, we plan to work with the data as they have already been made available. In particular, we will model whether a test is "no result" vs. "result," conditional on the test having been conducted. We expect to find no association with any covariates available in the Wave III Add Health data.

We next consider whether there is within-individual association in test outcomes. Table 2 is a three-way table of frequencies for chlamydia (*C*) results by those for gonorrhea (*G*) by those for trichomoniasis (*T*). The table is limited to individuals for whom all three test outcomes were decisive, and it permits the computation of odds ratios between the three tests. The *CT* odds ratio is 3.01, that for *CG* is 35.96, and the odds ratio for *TG* is 5.42. Clearly the bioassay test results are within-individual associated.

Table 2. Unweighted test outcome cell frequencies for chlamydia (*C*) by gonorrhea (*G*) by trichomoniasis (*T*)

	T-		*T*+	
	G-	*G*+	*G*-	*G*+
C+	524	29	33	7
C-	11,410	21	275	0

Source: Wave III, Add Health Study. Table $N = 12{,}299$.

This simple demonstration of within-individual association between test outcomes suggests that, if the covariate coefficients of eq. (1) are approximately proportional across

tests, it will be straightforward—and indeed reasonable—to treat the problem of correlated outcomes either from a multilevel modeling or a generalized estimating equation (GEE) perspective. There is a plausible substantive rationale for advancing this kind of formulation: The entire set of STDs is communicated through sexual contact, and exposure to risk can be greatly reduced by condom use.

Returning to the notation introduced earlier, now let i=1,2,3 denote a test, and j denote an individual. Tests are nested within individuals; for brevity's sake we omit details concerning higher levels of clustering. Then the model of eq. (1) might be extended to an item-response formulation as

$$\begin{aligned} \theta_{ij} &= \pi_{0j} + \pi_1 C_j + \pi_2 G_j + \pi_3 T_j \\ \pi_{0j} &= x_j \Gamma + \alpha_{0j} \end{aligned} \qquad (2)$$

where C, G and T are dummies for positive results on the chlamydia, gonorrhea and trichomoniasis tests. Eqs. (2) is a simple two-level item-response model of STD propensity. The model incorporates disease-specific propensities. We expect to estimate models of this kind as part of the paper. Allowing gender to interact with STD will almost certainly be necessary. In addition, it is straightforward to allow for clustering above the level of the individual, and we plan to do so.

Models like that of eqs. (2) are not the only kind we will estimate, but if the data are consistent with this framework, we gain the advantage of descriptive economy. Moreover, the locus of discussion shifts somewhat from the unique characteristics of specific sexually transmitted diseases or infections to the factors that affect sexual behavior more generally.

Reference

Handcock, Mark S., Martina Morris, William C. Miller, Carol A. Ford, John L. Schmitz, Marcia M. Hobbs, Myron S. Cohen, Kathleen Mullan Harris, J. Richard Udry. 2003. "HIV and STD Prevalence in Wave III." Paper presented at the Add Health Users Workshop, July 28-29, 2003, National Institutes of Health, Bethesda, Maryland.