# From Medical Geography to Computational Epidemiology – Dynamics of Tuberculosis Transmission in Enclosed Spaces

Joseph R. Oppong, Armin R. Mikler, Patrick Moonan, and Stephen Weis

## Introduction

Medical geographers study the geographic distribution of health and health-related phenomena such as diseases, and health care facilities. Seeking to understand *who* is getting *what* diseases or health services *where* and *why*, they examine spatial disparities in access to health care services, and the geographic distribution of health risks. Medical geographers apply tools of geographic enquiry such as disease mapping and geographical correlation studies to health-related issues (Elliot *et al.*, 2000; Pickle, 2002). Some have called this research endeavor spatial epidemiology (Cromley, 2003; Rushton, 2003a).

Disease mapping is an important tool for medical geographers. Such maps help to identify associations between disease and related factors such as environmental pollution. Inevitably, disease maps stimulate the formation of causal hypothesis. By enabling the simultaneous examination of multiple factors associated with disease linked by location, Geographic Information Systems (GIS) facilitate medical geography research. In fact, recent developments in GIS and proliferation of spatially referenced health data sets are spawning new ways to examine health-related issues. Projects such as the *Atlas of United States Mortality* (Pickle *et al.*, 1996) have prompted researchers to explore various measures of morbidity and mortality (Goldman and Brender, 2000; Pickle et al., 1999), their visual representation in geographic contexts (James *et al.*, 2004), and the application of spatial statistics to morbidity and mortality data (James *et al.*, 2004; Pickle, 2002; Rushton 2003).

GIS has revolutionized the way researchers explore the geography of health (Gatrell, 2002; Gatrell and Senior, 1999; Melnick, 2002; Ricketts, 2003), and their utility for the study of health issues is widely documented (de Lepper *et al.*, 1995; de Savigny and Wijeyaratne, 1995; Scholten and de Lepper, 1990). GIS and health research focuses on the quantitative analysis of health-related phenomena in spatial settings (Gatrell and Senior 1999:925) and, thus, isolates locations of health-related phenomena for analysis and interpretation.

While GIS has enabled Medical Geographers to address previously inconceivable complex health-related phenomena, their ability to deal with the dynamic processes of disease transmission among population groups, which usually requires complex interactions among numerous variables, is quite limited. High Performance Computing provides the requisite tools for breaking this barrier and is the focus of a new field of endeavor that we have called computational Epidemiology.

T. Böhme et al. (Eds.): IICS 2004, LNCS 3473, pp. 189–197, 2006.

## Epidemiology, Medical Geography and High Performance Computing

Although the role of Epidemiologists and Medical Geographers has become more pronounced in light of public health threats, computational tools that would enhance quality of information, facilitate prediction, and accelerate the generation of answers to specific questions are still lacking. In fact, at a time when global health threats make precise epidemiological information a critical necessity, epidemiologists continue to draw conclusions and make predictions using sparse, widely dispersed, incomplete or compromised data. Meanwhile, the complexity surrounding disease diffusion continues to escalate. Diverse populations traveling extremely long distances in unprecedented short times due to increased globalization mean that disease causing organisms circulate freely in a rapidly shrinking global village. An imperative response is to develop new tools that leverage today's cyber infrastructures for disease tracking, analysis, surveillance, and control.

The ability to predict how a disease might manifest in the general population is essential for disease monitoring and control strategies. Traditionally data collected during previous outbreaks are used. However, for newly emerging or re-emerging infectious diseases, such data is often unavailable or outdated. Changes in population composition and dynamics require the design of models that bring together knowledge of the specific infectious diseases with the demographics and geography of the region under investigation. New scientific methods that enhance understanding of the intricate interplay of disease and population are needed.

In a world of bioterrorism, where new and reemerging local disease outbreaks threaten all mankind, disease monitoring cannot continue to be fragmentary and inadequate focusing on small spatial domains (CDC 1994). As the recent outbreak of Severe Acute Respiratory Syndrome (SARS) showed, effective surveillance is critical to an effective defense against global disease threats, and requires consideration of huge volumes of data from other parts of the world. Developing tools that will accelerate epidemiological research, disease tracking and surveillance is thus, imperative. Computational models for the simulation of global disease dynamics are required to facilitate adequate what-if analyses.

What is needed is a novel interdisciplinary research program that facilitates epidemiology and Medical Geography research through high performance computing (HPC). Specifically, we envision the collaboration of Biologists, Medical Geographers, Epidemiologists, Computer Scientists, Biostatisticians, and Environmental Scientists to develop and implement computational tools in support of epidemiological research. These tools include simulation, visualization, and analysis tools that require HPC infrastructure.

Researchers in computational biology and medical informatics have relied on the availability of HPC infrastructures consisting of parallel computing architectures, cluster computing, and high performance visualization. Computational biologists have concentrated primarily on computational models at the molecular level, addressing specific computational problems in genomics, proteomics (protein folding), drug design etc. Most notable is the use of HPC in the design of drugs to cure or prevent specific diseases such as HIV. The field of medical informatics, particularly in Europe, has utilized HPC to manage the vast volumes of patient data. Further, HPC and high performance visualization tools have been used to design medical devices and test their functions in a simulated environment. To the best of our knowledge, there exists no single comprehensive program that aims at utilizing HPC in the field of Epidemiology or Medical Geography to build and analyze computational models of how a given disease manifests in the general population. This may include models of Tuberculosis (TB) outbreaks in different environments (homeless shelters, factories etc.), a West Nile Virus outbreak in a specific geographic region, or the progression of infectious diseases such as measles in the United States.

For these models to yield adequately precise information, many different factors must be considered. These factors may include socio-economic status of geographic regions, travel behavior of people, or airflow in a factory building. This clearly requires the use of a computing infrastructure that is capable of yielding computational results in a reasonable amount of time. To interpret the data, visual metaphors or data visualization that permits the epidemiologist to interact with the data are needed. For example, we envision an investigator immersed into a simulated model of a factory in which a TB outbreak is being investigated. The scientist is thus able to interact with the model, change functional parameters and thus engage in a what-if-analysis that currently is not available.

One computational challenge is to combine the spatially and temporally disparate datasets. This necessitates a fundamental knowledge of database systems, data management, and data retrieval. Even if a comprehensive dataset, containing individual health data for a large section of the population could be constructed, the extraction of relationships among the data constitutes a second computational challenge. For example, the domain of Artificial Intelligence and Machine Learning has been successfully used in Bio-informatics and is likely to be a valuable tool for discovering relationships among epidemiological data. Geographical Information Systems (GIS) help visualize spatial relationships of epidemiological data.

Whereas the collection of epidemiological data is essential for research, the need for Homeland Security, Disease Tracking and Surveillance requires sharing data among different federal, state, and local agencies. Healthcare providers in hospitals and private practice may be required to provide information to respond quickly and decisively to possible community health threats. A corresponding communication infrastructure can dramatically improve the precision with which health threats are analyzed, predicted, or traced. Such a system requires the combined effort of epidemiologists and computer scientists, each with a detailed understanding of each other's domain. For instance, issues that are central to the analysis of health-related data may dominate the requirements for a network infrastructure to interconnect healthcare providers. Examples of such issues include the type of information to be shared, format of information, and possible privacy and security issues.

Recent breakthroughs in sensor technology and wireless communication have led to the concept of Sensor Networks. Sensors of different types (biological, chemical, physical) have been deployed to monitor conditions in a variety of diverse environments. Ecologists, meteorologists, soil scientists, and others rely on such sensors to collect information about the environment. Connected via wireless networks, sensors can cover extended geographic areas, generating information instantaneously. The National Science Foundation (NSF) has announced special interdisciplinary programs to advance the field of Sensor Networks at a fast pace. This technology facilitates the monitoring for biological and/or chemical agents, and is expected to play a major role in Homeland Security. Environmental surveillance is essential in the field of public health, as it leads to early detection of adverse conditions and hence the ability to alert the population. However, to optimize this technology, scientists need to understand the technical (computational) as well as the epidemiological domains. New algorithms that autonomously extract data from sensors and auto-correlate sensor events must be developed. This leads to the design of intelligent systems, capable of learning from sensor data, and being able to classify events expediently.

Another example of convergence of epidemiology and computer science is the modeling and simulation of infectious disease outbreaks. Such an endeavor requires modeling demographics of the geographic domain within which a simulated outbreak is to take place. It further requires a high-fidelity representation of the disease pathology. Although very small models may be executed on a single computer, the simulation of a geographic region of moderate size will require computational resources beyond those of a single workstation. This leads to the use of a high performance computing infrastructure or computing clusters with tens or even hundreds of processors. Similar to computational biology and bio-informatics, computational epidemiology can utilize modern communication and computation infrastructures to solve computationally complex problems. The next generation of national (and international) cyber-infrastructure to provide access to high-bandwidth networks and high performance computing is about to be developed. The field of Epidemiology must develop tools that will enable scientists to effectively use such an infrastructure. To illustrate the need for computational epidemiology, two case studies are presented below.

## Tuberculosis Transmission in Enclosed Spaces – A Homeless Shelter and a Factory

The dynamics of localized TB transmission within enclosed facilities such as homeless shelters or factories is little understood. Traditional Medical Geography, involving disease diffusion mapping precludes detailed analysis of the dynamics of TB transmission in enclosed spaces. For example, the spatial patterns of individual movement, pathogen characteristics, airflow and other specifics of the facility that trigger transmission are not easily modeled in a GIS.

The case studies cover tuberculosis transmission in a factory and a homeless shelter. Results of initial analysis suggest that proximity of workspace to infected person is a major determinant of infection. After showing the shortcomings of traditional medical geography and disease mapping for modeling dynamics of disease transmission, preliminary results of a simulation model using advanced computational tools, are presented. The new tool of computational epidemiology allows the spatial distribution of risk to be defined not in terms of large regions but in micro-space, literally feet and inches. The potential of such computational tools for disease transmission in enclosed spaces is demonstrated.

The number of tuberculosis cases in the United States is at its lowest point in history, with 15,075 cases reported in 2002 and a TB case rate of 5.2 per 100,000 [17]. Consequently recent research suggests that molecular based studies focusing on dynamics of TB transmission in specific locations, such as homeless shelters, and social settings such as bars [15, 18, 19] is a much-needed final push to TB control. For example, the homeless and those living in marginal housing and overcrowded areas [22] constitute reservoirs of TB infection. Recent research conducted in Los Angeles and Houston suggest that locations at which the homeless congregate are hot spots of tuberculosis transmission, and measures that reduce tuberculosis transmission should be based on locations rather than on personal contacts [23, 24]. Yet, little research exists on the dynamics of localized TB transmission within a homeless shelter or other enclosed facility.

While much emphasis has been placed on homeless shelters, little attention has been paid to other enclosed facilities such as factories, warehouses and classrooms where long-term exposure usually in close contact situations that may facilitate transmission of pathogens, is usually the norm. For example, tuberculosis transmission is a recognized risk to patients and workers in health-care facilities (CDC 1994). Factories and warehouses, where people usually work in close proximity for long periods, may also be areas of concern.

## Methodology

Data for these case studies were based on data collected prospectively on all persons newly diagnosed with culture positive tuberculosis at the Tarrant County Health Department (TCHD) between January 1, 1993 and December 31, 2000. Each eligible patient was prospectively enrolled and participated in a structured interview as part of their routine initial medical evaluation. As part of an on-going Center for Disease Control and Prevention (CDC) study of the molecular epidemiology of tuberculosis, all positive isolates obtained from persons residing in Tarrant County are sent to the Texas Department of Health (TDH) for DNA fingerprinting. All patients are interviewed at the time of the initial evaluation, using a data collection instrument designed to obtain demographic information and medical history. The results of the DNA fingerprinting were incorporated into the database using patient identification numbers. Epidemiological factors included in this study were age, country of birth, date of entry, race/ethnicity, onset of symptoms, date of diagnosis and physical address. Any patient who did not have both PCR-based spoligotyping and RFLP-based IS6110 analysis performed on their corresponding MTB isolate, and/or did not live within Tarrant County at the time of collection were excluded from the geographical analysis.

*M. tuberculosis* culture isolation, identification, and drug susceptibility was conducted at the Texas Department of Health Bureau of Laboratories. Clinical isolate IS6110-based RFLP and PCR-based spoligotyping methods were utilized to identify patients infected with the same *M. tuberculosis* strain using published methods (van Embden JD, et al., 1993; Kamerbeek, et al., 1997). Since the discriminatory power of the IS6110 probe is poor for strain differentiation among specimens with five or fewer of the insertion elements, additional genotyping using the PCR-based spoligotyping was utilized (Kamerbeek, et al., 1997). We consider isolates representing a cluster when two or more patients had identical number of band copies, IS6110-RFLP, and spoligotyping patterns.

The first case study describes the dynamics of tuberculosis transmission within a homeless shelter with 800 beds providing both long and short-term occupancy for homeless people in Tarrant County, Texas. We seek to understand how location within a homeless shelter influences risk of tuberculosis infection. The data set comprises screening records for each case including age, race, date tested, status of tuberculosis, location in the facility, length of time spent in the facility, and other variables. Within the Shelter the 800 beds can be assigned to major areas – Men's Mats, Men's Beds, Men Over 50, Female Mats and Females Over 50 Beds. Each of these areas varied in bed density, floor space and occupants (Figure 1). The Mats area, (both male and female), is occupied by transients with no regular source of food, shelter, or shower. The Beds and Over 50 areas (male and female) are less dense overall, with more permanent residents. Results of initial analysis suggest that TB risk is not uniformly distributed but depends on the location of the sleeping bed and duration and frequency of stay at the night shelter. For example, 12 of the 17 active cases (63.2%) had been visiting the shelter for more than 5 years.

**Table 1.** Who Sleeps Where and TB

| Category | Total People | Active TB |
|---|---|---|
| Men's Mats | 1220 | 1.0% |
| Men's Beds | 51 | 3.9% |
| Men's Over 50 | 87 | 3.4% |
| Female Mats | 265 | 0.0% |
| Female Over 50 | 63 | 0.0 |

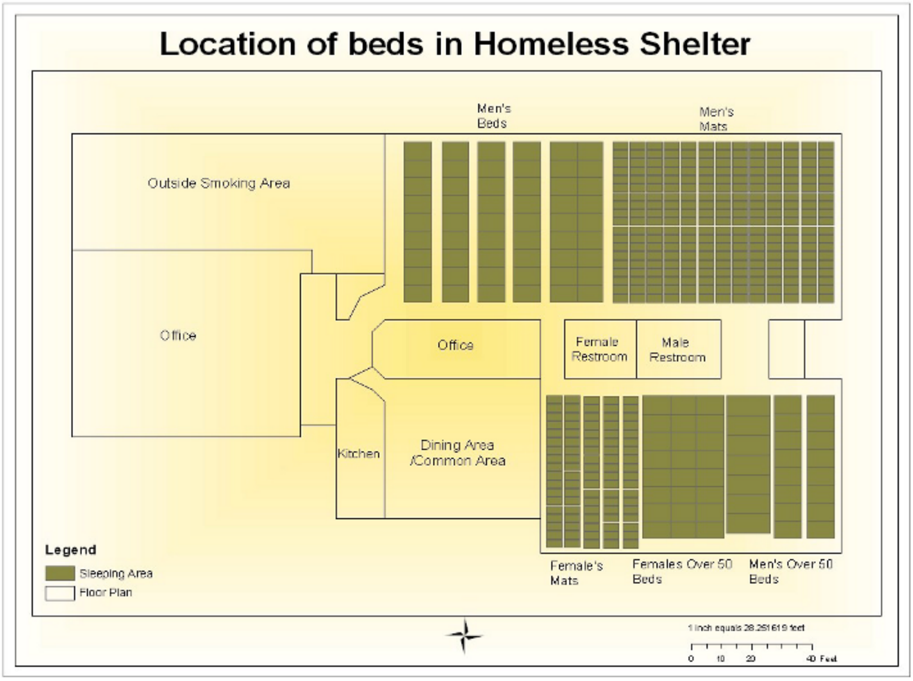

**Fig. 1.**

We need to examine certain dynamics of the Homeless Shelter that promotes or inhibit TB transmission such as the air circulation system. For example, while the men's and women's areas had different airflow systems, it appeared that air flow in the women's section was more effective than in the men's section. Is this a factor in TB transmission? The movement patterns of residents also need to be addressed. How much time do they spend in common areas such as the dining area, the smoking area, the TV area and rest rooms? Another factor to consider is the lighting in different parts of the Homeless Shelter. Does the heat put off by the lighting affect dispersion of the bacilli?

Addressing all of these is clearly beyond the capacity of any GIS or simple disease mapping. HPC is required to simulate the movement pattern of residents, TB bacilli, lighting, the air circulation system and other variables.

## Case Study 2 – Factory TB Outbreak

The second case study covers TB outbreak in a factory that produces airplane bridges. Out of a total of 64 workers, 32 were ultimately infected with the same strain of TB presumably from one Index case (Table 2).

**Table 2.** Workspace and Tuberculosis Risk

| TB Test Result | Assembly | Maintenance | Office | Painting | Welding |
|---|---|---|---|---|---|
| Positive | 8 | 0 | 2 | 10 | 12 |
| Negative | 18 | 2 | 2 | 2 | 2 |
| Total | 26 | 2 | 4 | 12 | 14 |

In the factory, Figure 2, in addition to basic screening records as collected for the homeless shelter, other available data include measures of duration and proximity to infected person such as hours per week in the factory, hours per week in the same workspace, hours per week within 3 feet of infected person, and usual work area. Results of initial analysis suggest that proximity of workspace to infected person was a major determinant of infection. In fact almost 100% of those who worked directly in the same space with one infected person were infected with the same strain of TB.

Hours spent each week in the factory was not a statistically significant determinant of TB risk. Rather, hours spent in the same work space and hours spent within 3 feet of Index case were the significant determinants of risk. In short active TB risk in the factory depends not so much on time spent in the factory but on time spent where in the factory. Simple mapping of TB occurrence in the factory does not go far enough. We need to simulate actual transmission considering factors such as the dynamic movements of individuals in the factory, shared common areas and amount of time spent there, the air circulation system, and related variables. Clearly more sophisticated tools are required to handle multiple variables in a dynamic system.

This rich data set provides the opportunity to implement a model to calibrate the dynamics of TB transmission in enclosed facilities using computational epidemiology.

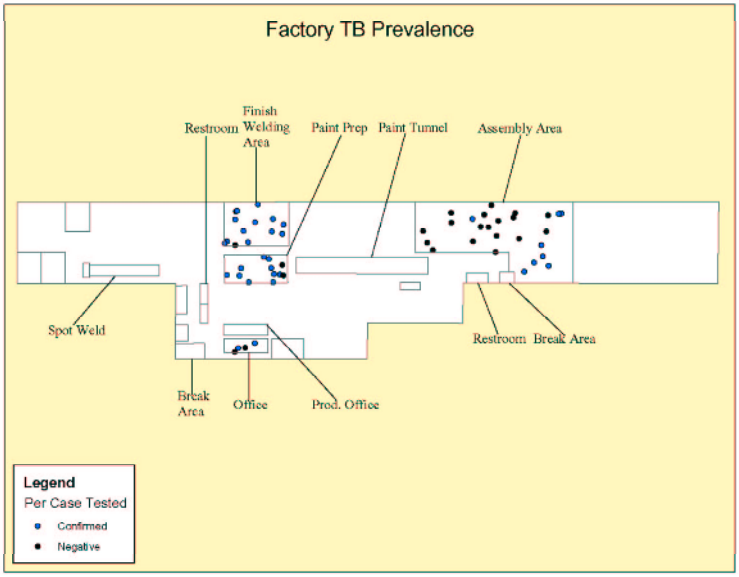

**Fig. 2.**

## Conclusion

This study examined tuberculosis transmission in a homeless shelter and a factory with ongoing TB transmission. To understand the dynamics and determinants of transmission in enclosed spaces, traditional medical geographic approaches such as GIS are not suitable for modeling dynamics of disease transmission. Computational epidemiology allows the spatial distribution of risk to be defined not in terms of large regions but in micro-space, literally feet and inches. The potential of such computational tools for simulating and explaining disease transmission in enclosed spaces is high.

## References

Acevedo-Garcia D. 2001. Zip-code level risk factors for tuberculosis: neighborhood environment and residential segregation in New Jersey, 1985 – 1992. *American Journal of Public Health*; 91(5): 735 – 741.

Barnes PF, El-Hajj H, Preseton-Martin S, et al. 1996. Transmission of tuberculosis among the urban homeless. *Journal of American Medical Association* 275:305 – 307.

Centers for Disease Control and Prevention 1994. Guidelines for Preventing the Transmission of Mycobacterium tuberculosis in Health-Care Facilities, 1994. *MMWR*; October 28, 1994 / 43(RR13);1-13200.

Centers for Disease Control and Prevention. 2003. *Reported Tuberculosis in the United States, 2002*. Atlanta, GA: CDC.

Cromley, E.K. 2003. GIS and Disease. *Annual Review of Public Health*. 24:7-24.

de Lepper, M.J.C., H.J. Scholten, and R.M. Stern (eds). 1995. *The Added Value of Geographical Information Systems in Public and Environmental Health*. Boston: Kluwer Academic Publishers.

de Savigny, D., and P. Wijeyaratne. 1995. *GIS for Health and the Environment*. Ottawa: International Development Research Center.

Diwan, V. K and A Thorson. 1999. Sex, Gender and Tuberculosis. *Lancet* 3/20/99, Vol. 353 Issue 9157, 1000-1001.

Gatrell, A. 2002. *Geographies of Health: An Introduction*. Oxford, Blackwell.

Gatrell, A., and M. Senior. 1999. Health and Health Care Applications. In: *Geographical Information Systems: Management Issues and* Applications - Volume 2, ed. P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, 925-938. New York: John Wiley & Sons, Inc.

Goldman, D.A., and J.D. Brender. 2000. Are Stanardized Mortality Ratios Valid for Public Health Data Analysis. *Statistics in Medicine* 19:1081-1088.

Hathcock, A.L., Greenberg, R.A., and A.W. Dakan, 1982. An Analysis of Lung Cancer on a Microgeographical Level. *Social Science in Medicine* 16:1235-1238.

James, W.L., R.E. Cossman, J.S. Cossman, C. Campbell, and T. Blanchard. 2004. A Brief Visual Primer for the Mapping of Mortality Trend Data. *International Journal of Health Geographics* 3:7.

Klovdahl AS, Graviss EA, Yaganehdoost A, et al. 2001. Networks and tuberculosis: an undetected community outbreak involving public places. *Soc Sci Med*; 52(5):681-94.

Leonhardt KK, Gentile F, Gilbert BP, Aiken M. 1994. A cluster of tuberculosis among crack house contacts in San Mateo County, California. *Am J Public Health*; 84(11): 1834 – 1836.

Melnick, A.L. 2002. *Introduction to Geographic Information Systems in Public Health*. Gaithersburg, Maryland: Aspen Publications.

Melnick, A.L. 2002. *Introduction to Geographic Information Systems in Public Health*. Gaithersburg, Maryland: Aspen Publications.

Pickle L.W. 2002. Spatial Analysis of Disease. In: *Biostatistical Applications in Cancer Research*. ed. C. Beam, 113-150. Boston: Klewer Academic Publishers.

Pickle, L.W., M. Mungiole, G.K. Jones, and A.A. White. 1996. *Atlas of United States Mortality*. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics.

Pickle, L.W., M. Mungiole, G.K. Jones, and A.A. White. 1996. *Atlas of United States Mortality*. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics.

Pickle, L.W., M. Mungiole, G.K. Jones, and A.A. White. 1999. Exploring Spatial Patterns of Mortality: The New *Atlas of United States Mortality*. *Statistics in Medicine* 18:3211-3220.

Ricketts, T.C. 2003. Geographic Information Systems and Public Health. *Annual Review of Public Health* 24:1-6.

Ricketts, T.C. 2003. Geographic Information Systems and Public Health. *Annual Review of Public Health* 24:1-6.

Rushton, G. 2003a. Epidemiology and Biostatistics: Public Health, GIS, and Spatial Analytic Tools. *Annual Review of Public Health* 24:43-56.

Rushton, G. 2003b. Public Health, GIS, and Spatial Analytical Tools. *Annual Review of Public Health* 24:43-56.

Scholten, H.J., and M.J.C. de Lepper. 1990. The Benefits of the Application of Geographical Information Systems in Public and Environmental Health. *World Health Statistics Quarterly* 44:160-171.

Yaganehdoost A, Graviss EA, Ross MW, et al. (1999). Complex transmission dynamics of clonally related virulent Mycobacterium tuberculosis associated with barhopping by predominantly human immunodeficiency virus-positive gay men. *Journal of Infectious Diseases* 180(4):1245-51.