

The need for national HIV databases

Scott P. Layne, Thomas G. Marr, Stirling A. Colgate, James M. Hyman and E. Ann Stanley

Researchers and public health officials involved in surveying and forecasting the course of the HIV epidemic require complete and unfiltered information from many sources. Governments should respond by establishing national HIV databases.

IN THE United States, the cumulative number of reported AIDS cases is growing as a cubic function of time, and it is estimated that between one and two million people are infected with the human immunodeficiency virus (HIV)^{1,2}. The mean period from HIV infection to AIDS diagnosis is estimated to be 8 years³ and the mean survival time after AIDS diagnosis is about 12 months⁴, which indicates that most of those who develop symptoms after being infected with HIV will die within the next 10 years unless medical breakthroughs are made.

The HIV epidemic calls for a rapid and coordinated scientific effort across a broad range of disciplines. To that end, there must be effective mechanisms for the sharing of raw data among researchers and their use to answer several crucial questions — How does HIV transmissibility vary with time after infection and with disease stage? What is responsible for the large variance in incubation time from infection to the onset of AIDS? Will current trends in the spread of HIV infection continue? Can we explain the past rate of growth of the epidemic? Can we predict the most effective ways of slowing the spread of infection? And how extensive must national surveillance programmes be reliably to monitor the prevalence of HIV infection?

Information on HIV is surrounded by sensitive social, legal and ethical issues⁵⁻⁷, so researchers attempting to answer the above questions do not have access to many important elements of raw data. Finding sources of such data and entering into collaborative relationships happens haphazardly, and formal agreements for sharing data and ensuring their confidentiality are difficult to achieve without a precedent or formal framework. The work of those such as epidemiologists, sociologists and mathematical modellers, who all use raw data from a wide range of fields, is severely restricted by these obstacles.

To reduce the obstacles, we propose the creation of national HIV databases to coordinate the sharing of raw data. At first such databases would address the information needs of researchers in their respective countries, but in time they could be linked internationally. Here we argue for a national HIV database for the

United States, but the same principles apply to other countries.

A large number of commercial, educational and government organizations have developed information services on HIV. These services are used by many HIV researchers, but there are still significant gaps for those who require detailed data on social and sexual behaviour and unprocessed epidemiological data on HIV.

Surveillance

The biggest database summarizing epidemiological surveillance for the United States is the *AIDS Public Information Data Set*. This database contains one record for each reported AIDS case and is updated quarterly by the Centers For Disease Control with the following information: age, sex, race, area of residence, month and year of diagnosis, month and year of report, date of the patient's death, primary risk group, additional risk factors, country of birth and opportunistic diseases at diagnosis⁸. To ensure confidentiality, case reports are assigned to six Standard Metropolitan Statistical Areas that have more than one million inhabitants. This large granularity greatly reduces the value of the data. It is not possible to identify the AIDS cases for a particular city nor is it possible to identify diagnosis dates of AIDS cases before 1982. As a result, researchers cannot use these data to analyse the progression of the epidemic for a particular city.

There is, too, a lack of useful raw data in journal articles, government reports and

conference proceedings. Most of the data contained in these publications are either highly processed or combined into large blocks of statistical information. As a result, the rapidly growing number of interactive databases and preprinted reports that review the literature on HIV do not help researchers identify useful sources of raw data. Similarly, there are a large number of continuing or completed cohort studies that examine the detailed social and sexual behaviour of groups such as homosexual and bisexual men, haemophiliacs, intravenous drug users, transfusion recipients, heterosexual couples and prostitutes⁹. These studies are the primary source of information on HIV serological status as a function of the rate of practice of specific high risk behaviours and on the amount of sexual contact (or mixing) within a particular risk group. But current information services have not systematically identified these cohort studies nor have they collected details on their individual questionnaires, operational procedures and main findings.

To develop mathematical models of HIV transmission and the progression to AIDS, researchers need several distinct kinds of data sets — they need detailed data from a well-defined cohort study for testing and refining of the model; data from a similar cohort study for validating the predictive value of the tuned model; and, for larger-scale applications of the validated model in epidemiological studies, raw data from a broad range of

Table 1 Example of a standard agreement between users and providers of data

- List the names and affiliations of all the users and providers of data on the agreement
- Obtain prior approval from the providers of data before adding new names to the agreement's user list
- Show no raw data to persons outside the agreement's provider and user list
- Protect the security of the raw data when it is stored outside the database system
- Do not use the names of individuals as identifiers in the raw data subset
- Do not disclose data that would allow identification of individuals by inference
- Submit for review to the providers of data all manuscripts prior to submission for publication
- Cite the providers of data as authors or give them acknowledgement on all manuscripts
- Dispose of all user copies of the raw data held outside the database after completion of research
- Agree to abide by any privacy laws or other special handling restrictions that may apply to the raw data

sources dictated by the particular set of equations in the model.

Regardless of the model's complexity, two general sets of parameters, biological and behavioural, are essential for understanding the HIV epidemic^{10,11}. Biological parameters are to do with pathogenesis of the disease and include factors such as variations in transmissibility during the course of the infection, length of carrier state and influence of cofactors (for example, sexually transmitted diseases and age). They are influenced by environment and by factors associated with immune stimulation¹². Behavioural parameters concern activities that may transmit infection between people and include factors such as level and type of sexual activity, and choice of sexual partners. Because of a scarcity of biological and behavioural information, few of these parameters (of which we give examples only) have been calculated to a precision that is required by the mathematical models of the spread of HIV currently proposed¹⁰. Nevertheless, it may be possible to obtain better estimates on several of them if more raw data from many different studies were available and were analysed together. For that, it will often be necessary to collate information from a large body of epidemiological and sociological studies.

In its simplest form, a national HIV database would be an annotated directory of raw data. In its most complete form, it would be a storehouse of raw data on HIV infection and the AIDS epidemic. At first, the database would be constructed to allow for iterative design of the data structures; later, it would mature into a high-quality information-management system incorporating the latest techniques in database technology.

Standard agreement

In order to facilitate collaboration between researchers, the national HIV database would furnish a standard agreement governing procedures on sharing raw data (see Table 1). This agreement would be signed by both the users and providers of data, and would apply to a particular set of raw data. The standard agreement would establish a code of conduct among researchers regarding the data's security and privacy, and would save time and energy in establishing collaborations that were based on information supplied by the database. In conjunction with this agreement, we have devised four options for the database that could serve the research community (see the box "Design options for a national HIV database").

Options 1-3 are enumerated according to increasing size of the database and scope of the services that they offer, whereas option 4 offers a decentralized repository managed by a decentralized staff. These four options are not mutually

Design options for a national HIV database

1. An annotated directory of sources for raw data, supplying the descriptive information listed in Table 2 and assisting users in identifying providers. Staff would not assist in establishing formal collaborations between users and providers, so the standard agreement is merely a guideline for sharing data. For more detailed information, each user would have to contact a provider individually; for each request by a user the burden of supplying the data would fall entirely on the provider.

2. Both a directory of sources for raw data and a go-between for the users and providers of data. Staff would arrange formal collaboration between users and providers, and so the standard agreement is an instrument for sharing data. The database service supplies the descriptive information (Table 2) and also assists users in finding data by collecting more detailed information from potential providers. Because the database collects detailed information from providers over time, providers are not repeatedly contacted with preliminary questions from various would-be users of their data, as is the case with option 1.

3. A storehouse of raw data, supplying descriptive information (Table 2), the raw data and the standard agreement for use of that data. According to prior formal arrangements, the database may or may not be obliged to contact a provider before releasing their data to a user. For the provider, the burden of supplying the data is reduced to a single shipment with periodic updates. This database service would require a larger support effort than the above two options, but reduces the need for preliminary discussions between the users and providers of data.

4. A decentralized repository of raw data. Raw data resides at the local, or institutional, level but the resulting information is presented at the national level. Access to the local resource is 'transparent' to the user of the national service. This design gives the local organization a greater degree of autonomous control over the data and how it is structured, but requires standardization of transaction types. Use of database staff and the standard agreement would be handled as in option 3.

exclusive; rather they offer a range of relationships that the users and providers of data could have within a single database service. Questions to be considered by the HIV research community in choosing a structure for the database service are — Which option best serves the needs of the

researchers? Which government agency should oversee development? What is the time scale for construction? Who is responsible for its operation? And what are the costs of construction and operation?

The establishment of a national HIV database will require an extraordinary level of commitment on the part of the research community. Individual researchers and institutions will have to share and protect large quantities of confidential data on the intimate behaviour of individuals. They will also have to share data that could otherwise be hoarded to build their own careers. But such a database is needed — and it is needed soon. □

Table 2 Example of a database questionnaire for requesting data

- General description of study
- Purpose of study
- Study start and end dates (some studies may be in progress)
- Number of persons investigated
- Criteria for selection of participants
- Basic demographics and descriptions (number of males and females, participant's economic status and so on)
- Location and affiliation of investigators
- Names of investigators involved in data collection
- List of publications generated
- Copy of study protocol
- Copy of all questionnaires used
- Description of raw data formats and content
- Funding source
- Person to contact for further information
- Restrictions that may apply to the use of the data (is it available for a fee, are there copyright protections and so on)
- Information on federal, state or local laws that govern use of data (Privacy Act, for example)
- A copy of the raw data

1. Centers for Disease Control. *MMWR* **36**(S-6), 1-48 (1987).
2. Morgan, W.M. & Curran, J.W. *Public Health Reports* **101**, 459-465 (1986).
3. Curran, J.W. *et al. Science* **239**, 610-616 (1988).
4. Rothenberg, R. *et al. New Engl. J. Med.* **317**, 1297-1302 (1987).
5. Lewis, H.E. *J. Am. Med. Ass.* **258**, 2410-2414 (1987).
6. Dickens, B.M. *Science* **239**, 580-586 (1988).
7. Walters, L. *Science* **239**, 597-603 (1988).
8. *AIDS Public Information Data Set* (Centers for Disease Control, Jan. 4 1988).
9. *An Inventory of Research Studies Regarding AIDS or HTLV-III/LAV Infection* Section A (Apt Associates, Cambridge, Massachusetts, Jan. 1987).
10. Hyman, J.M. & Stanley, E.A. *Math. Biosci.* (in the press).
11. May, R.M. & Anderson, R.M. *Nature* **326**, 137-142 (1987).
12. Haverkos, H.W. *J. infect. Dis.* **156**, 251-257 (1987).

Scott P. Layne, Thomas G. Marr, Stirling A. Colgate, James M. Hyman and E. Ann Stanley are in the Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.