LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

LA-UR -88-125

MASTER

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE **A NATIONAL HIV DATABASE THAT FACILITATES DATA SHARING**

LA-UR--88-125

DE88 005384

AUTHOR(S) **Scott P. Layne, T-7**
Thomas G. Marr, T-10
E. Ann Stanley, T-7
James M. Hyman, T-7
Stirling A. Colgate, T-6

SUBMITTED TO **Proceedings of the National Academy of Science**

DISCLAIMER

1/12/88

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

A NATIONAL HIV DATABASE THAT FACILITATES DATA SHARING

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

January 12, 1988

The purpose of this communication is to stimulate discussion on a National HIV Database that facilitates and coordinates data sharing. We argue for the creation of a new database because significant gaps exists in the type of information that are available on HIV. Databases that extensively survey the published literature on HIV are widely available, however, databases that contain either raw data or that describe ongoing HIV research efforts are not widely available. For epidemiologists, sociologists and mathematical modelers, who need to draw on raw epidemiologic and behavioral data from a broad range of fields, the existing databases are inadequate. In this paper we emphasize the particular requirements of epidemiologists, sociologists and modelers, and suggest a plan to accommodate their database needs.

INTRODUCTION

Human Immunodeficiency Virus (HIV) causes the Acquired Immunodeficiency Syndrome (AIDS) by infecting and killing cells in the immune and central nervous system. The long asymptomatic, but infectious, carrier state of this lentivirus and its transmission by sexual contact indicates that HIV has a genuine potential to keep increasing its prevalence in the population. Because of the prolonged incubation period, this increase may occur silently and long before enough AIDS cases appear to signal an alert. In the United States, the cumulative number of reported AIDS cases is growing as a cubic function of time and currently, it is estimated that 0.5 to 1.5 million people are infected with HIV.

The threat of an expanding HIV epidemic necessitates that we have a rapidly acting and openly communicating scientific effort from a broad range of disciplines. A few of the specialists that are working on the HIV epidemic include: immunologists, research biologists, public health officers, vaccine and pharmaceutical developers, neurologists, hematologists, pathologists, internists, pediatricians, epidemiologists, sociologists, and mathematical modelers. However, due to the overload of social, legal and ethical issues related to the epidemic, some important elements of raw data on HIV are not straightforwardly available.

The sharing of raw data on HIV are difficult to arrange for several reasons: first, researchers have to find sources of data; second, researchers have to establish contact and develop working relationships; third, researchers have to enter into formal agreements to share data; and fourth, researchers have to ensure and protect the confidentiality of individuals associated with raw data. Often these steps are very time consuming, due to rapid expansion of information on HIV and sensitivity of raw epidemiological data, and as a result data sharing is often frustrated. For researchers that need to use raw data from a broad range of fields, such as epidemiologists, sociologists and mathematical modelers, this reduction in sharing is especially problematic.

In order to alleviate this serious obstacle, we propose the creation of a National HIV Database that facilitates and coordinates the sharing of raw data between researchers. There are a number of justifications for proposing and considerations for implementing this new database, and in the following sections we address them. First, we will discuss some of the parameters that epidemiologists, sociologists and mathematical modelers need to know to understand the growth of the HIV epidemic. Next we will review the current HIV databases that are available to all researchers and discuss the limitations of them. After this, we will point out some potential sources of data that are not readily available to researchers and present a list of questions regarding the security and appropriate use of this raw data. Finally, we will suggest one possible plan for implementing this new database.

PARAMETERS FOR PREDICTING THE FUTURE COURSE OF HIV PREVALENCE

Predicting the future course of the HIV epidemic not only requires input from the broad range of specialties mentioned above, it also requires a concerted inter disciplinary effort by epidemiologists, sociologists and mathematical modelers. Each of these three disciplines are addressing somewhat different yet closely related questions. For example, epidemiologists are examining how rapidly HIV is spreading within groups at high risk for infection, and are monitoring to detect whether HIV is spreading from these groups to the rest of the population. Sociologists are examining how groups at high risk for infection are mixing with the rest of the population, and are looking to see whether sexual behavior is changing as a result of the AIDS epidemic. Mathematical modelers are investigating how different types of models simulate the dynamics of the epidemic, and are asking whether their results agree with the past history of the AIDS epidemic.

A large part of this work requires the estimation of two general classes of parameters that are listed below. Most of these parameters are common elements for constructing any framework

that explains the spread of HIV infection and predicts the number of future AIDS cases. These parameters may also be used for developing more specialized models that predict the outcome of various public health measures such as educational programs, therapeutic drug programs and vaccination programs.

Biological parameters:

- initial distribution of infected people
- transmissibility of HIV as a function of time after infection
- duration from infection to stages of disease (ARC, AIDS and dementia)
- cofactors that modify HIV transmissibility,morbidity and mortality
- effects of therapeutic intervention on transmissibility, morbidity and mortality

Behavioral parameters:

- distribution of population according to number of sexual partners
- number of contacts with each sexual partner
- membership in a risk group
- mixing within the risk group
- Mixing between risk groups and rest of the population
- behavioral changes secondary to AIDS epidemic
- Behavior as a function of population density

Due to a scarcity of biological and behavioral information, a number of the above parameters have not been calculated with adequate precision. Nevertheless, it may be possible to get better estimates on several of them if more raw data were readily available. For most of these parameters, it will be necessary to collate information from a large body of basic science, epidemiological and cohort studies.

A LIST OF CURRENT HIV DATABASES

Over the past several months we have been compiling a list of data sources and their contents. Some of these sources are specifically focused on HIV and AIDS while others are broadly based in scope. The majority of sources consist of references to journal articles, government reports, and conferences and proceedings. A few sources contain epidemiological surveys and occasionally some fraction of their associated raw data. None of the sources that we have identified contain an exhaustive set of information and in general, these sources do not reference each other. Each of these databases have unique modes of access, some have service charges, and all have different formats of data presentation. The following is a list of these sources by general category:

COMPUTERIZED DATABASES

Massachusetts Medical Society
1440 Main Waltham, MA 02154
617-893-4610
AIDS Knowledge Base is published on a monthly basis in both CD ROM and online formats from San Francisco. Intended for clinicians, researchers, nurses, public health workers, hospital administrators, and educators. Provides basic science information on clinical, pathogenic, diagnostic, and public health topics. Several major journals participate in generating information for the database.

General Video Text
Cambridge, MA 02139
800-544-4005 and 213-464-7400
CAIN (Computerized AIDS Information Network) is a computerized database that focuses on AIDS news, general awareness and some epidemiology. Aimed primarily for service providers; it is accessible on the DELPHI network.

Centers for Disease Control
Statistics and Data Management Branch
AIDS Program, CID Building 6, Room 270
1600 Clifton Road
Atlanta, GA 30333
303-639-3775
Quarterly AIDS Incidence Report is a floppy disk that contains information abstracted from AIDS case reports received by the Centers for Disease Control. These data have been reported voluntarily to CDC by State and local health departments, and are protected under an Assurance of Confidentiality of the Public Service Act which prevents disclosure of any information that could be used to either directly or indirectly identify individual patients or establishments.

DIALOG Information Service
3460 Hillview Avenue
Palo Alto, CA 94304
800-334-2564
An online service that provides access to number separate databases. Users read these databases by using various intra- and inter-database searching tools. Relevant databases may include: *BIOBUSINESS, BIOCOMMERCE ABSTRACTS, BIOSIS PREVIEWS, CANCERLIT CLINICAL ABSTRACTS COMPUTER DATABASE, CONFERENCE PAPERS INDEX DISSERTATION ABSTRACTS ONLINE, EMBASE, FEDERAL RESEARCH IN PROGRESS, FOUNDATION DIRECTORY, FOUNDATION GRANTS INDEX, HEALTH PLANNING & ADMINISTRATION, INTERNATIONAL PHARMACEUTICAL*

ABSTRACTS, LIFE SCIENCES COLLECTION, MEDLINE, NTIS, ONTAP BIOSIS PREVIEWS, ONTAP MEDLINE, PAIS INTERNATIONAL, PHARMACEUTICAL NEWS INDEX, PSYCHINFO, SOCIAL SCISEARCH, SOCIOLOGICAL ABSTRACTS, ZOOLOGICAL RECORD.

Dr. Gerald Myers
T-10, MS-K710
Los Alamos National Laboratory
Los Alamos, NM 87545
505-665-0480
Database that contains DNA, RNA and protein sequences from all AIDS virus isolates and animal viruses that are related to the AIDS virus. Contains prepublication information and unpublished public information. Funded by NIAID interagency agreement with the Department of Energy.

National AIDS Clearinghouse
Aspen Systems, Inc.
P.O. Box 6003
Rockville, MD 20850
301-25.-5000
On September 1987 the CDC awarded a grant to Aspen Systems to develop two categories of information service. This information will be both electronic and hard copy, and will be maintained on a dial-in computer.
1. Database of Organizations that Provide Resources on AIDS. This database is intended to identify state, community and local organizations that provide services, testing, counseling, and medical care for AIDS patients. it is expected that this database will provide a comprehensive listing of all services provided for AIDS in the United States.
2. The Publicatior and distribution of Educational Material on AIDS. These materials are intended for wide ranges of audience including the general public, members of high risk groups, health care professionals, and etc.

COMPUTERIZED REVIEWS OF PUBLISHED LITERATURE

London School of Hygiene and Tropical Medicine
Bureau of Hygiene and Tropical Disease
Kepple Street
London WC1E 7HT
ENGLAND
011-44-01-636-8636
1. *Current AIDS Literature* is a monthly bibliographic database with coverage from 1984 to

present. Focuses on public health publications with broad coverage of European and developing country literature. Available online from BRS in the United States. Survey from 1984 to 1986 focuses on most important papers with abstracts and comments. Survey from 1986 to present is exhaustive with annotation and critical abstracts.
2. *AIDS Newsletter* is unpublished 17 times a year this newsletter reviews news and media in the UK and abroad, social and occupational issues, and science and medicine.

National Library of Medicine
National Institutes of Health
Bethesda, MD 20894
303-496-5116
Monthly bibliography of AIDS publications obtained by computer searches at NLM.

Peter O. Way, Ph.D.
Chief, Africa and Latin America Branch
Center for International Research
U. S. Department of Commerce
Bureau of Census
Washington, D.C. 20233
301-763-4086
Compiling epidemiological data on AIDS in third world countries. Information is based on extensive literature searches and stored on a PC network.

PRINTED REVIEWS OF PUBLISHED LITERATURE

Abt Associates, Inc.
55 Wheeler Street
Cambridge, MA 02138
617-492-7100
1. *An Annotated Bibliography of Scientific Articles on AIDS for Policymakers*. July 1987. Conducted under PHS Task Order Contract # 282-85-0064.
2. *An Inventory of Research Studies Regarding AIDS or HTLV-III/LAV Infection*. January 1987. Conducted under PHS Task Order Contract # 282-85-0064.

AmFAR (American Foundation for AIDS Research)
40 West 57th Street, Suite 406
New York, NY 10019
212-333-3118
1. *AmFar directory of Experimental Treatments for AIDS & ARC*. June 1987. List experimental drugs and ongoing experimental treatment programs for AIDS.
2. *AIDS Targeted Information Newsletter*. Reviews and abstracts AIDS related literature on a

monthly basis. Published by Williams & Wilkins, 428 East Preston Street, Baltimore, MD 21202.

BIOSIS (Biological Information Service)
2100 Arch Street
Philadelphia, PA 19103-1399
215-587-4800
1. *Collected Papers on AIDS Research.* Coverage of the international journal literature published between 1976 - 1986. Single volume contains 4500 items with 34 % entries abstracted.
2. *AIDS Research Today.* A monthly update of Collected Papers on AIDS Research. 1987 - present.

Gay Men's Health Crisis
Department of Medical Information
132 West 24th Street
New York, NY 10011
212-627-7737
Treatment Issues is a monthly newsletter on alternative therapies for AIDS.

Medical Data Exchange
445 S. San Antonio Road.
Los Altos Hills, CA 94022
415-941-3600
Online AIDS Update is a Monthly report on AIDS that is generated using MEDLINE and organized by six specific topics: 1) Epidemiology and Incidence, 2) Virology and Immunology, 3) Transmission and Prevention, 4) Clinical Management, 5) Social and Psychological Issues, and 6) Public, Economic and Legal Issues.

PRINTED SUMMARY OF EPIDEMIOLOGIC SURVEILLANCE

World Health Organization
1211 Geneva 27
Switzerland
011-41-22-912644
1. Dr. J. Chin. Surveillance, Forecasting and Impact Assessment. Special Programme on AIDS. WHO database of epidemiological data pertaining to AIDS.
2. *Weekly Epidemiological Record.* Every other edition contains an update of AIDS cases by country, date of first reported case, date of last reported case, total by country, and total by continent.

COMPUTER AND VOICE NETWORKS

Windom Health Enterprises
2926 Benvenne
Berkeley, CA 94705
415-848-6980
AIDSnet is an electronic mail network that promotes communication between AIDS researchers.

AIDS Teleforum
9150 Royal Lane, Suite 130
Irving, Texas 75063
800-654-2008 and 212-963-8193
AIDS Teleforum is a voice mail network with about 220 subscribers. The aim of such a voice mail system is to promote communication between AIDS researchers.

WHAT CRUCIAL FORMS OF DATA ARE NOT AVAILABLE

There are a number of examples of raw data that are either not readily available or completely unavailable to epidemiologists, sociologists, and mathematical modelers. These data have been collected by government agencies, state agencies, and university researchers. Some examples of these data are listed below.

Cohort studies in progress that are conducted in conjunction with the NIH and CDC for the following groups:

- Homosexual and Bisexual Men
- Hemophiliacs
- I.V. Drug Abusers
- Transfusion Recipients
- Heterosexual Couples Prostitutes

CDC seroprevalence surveillance data on the following groups:

- Defense Department military recruits
- Job Corps recruits
- Voluntary blood donors monitoring program
- Sentinel hospital monitoring program
- Drug treatment centers
- STD clinic patients
- Prostitutes
- Family planning clinics
- Child bearing women
- Federal prisoners

Immigrants
Hemophiliacs

CDC quarterly AIDS incidence data (listed in the previous section) with finer granularity prior to the beginning of 1982 and with finer granularity on a city by city basis.

AIDS incidence and HIV seroprevalence data on the state and community level collected by state, city, and local health departments. The more important cities include New York City and San Francisco.

QUESTIONS CONCERNING THE SECURITY AND APPROPRIATE USE OF RAW DATA

There are several important questions concerning the use of raw epidemiological data that should be addressed and answered before opening a National HIV Database for epidemiologists, sociologists, and mathematical modelers. These questions include:

How to protect the privacy of persons who are referenced in the database?

How to insure the security of the database?

How to grant access to the database?

How to insure that scientists are using the data accurately and under appropriate guidance of the researchers who collected the raw data?

How to insure that appropriate credit is given to researchers who collected and provided the raw data?

SPECIFIC DESIGN OF A NATIONAL HIV DATABASE

To promote coordination among diverse scientific efforts we propose a National HIV Database. In its simplest form, this database would be a compilation and description of the existing sources of raw and synthesized data. It would have sufficient manpower to keep abreast of new information and search for older information that is relevant to HIV research. This database would have to be experimental at first and with time would mature into a more refined information management system. It would incorporate the latest techniques in relational database technology and distribution.

The simplest way to start a National HIV Database is to identify all relevant cohort, surveillance, behavioral, and laboratory studies. This process could be initiated during a national conference that invites HIV researchers from federal, state and local governments, universities, industry,

WHO, and other appropriate organizations.

Epidemiologists, sociologists and mathematical modelers could be used to point out what types of raw data are necessary for their work as well as providing data from their own fields. Other specialists, such as those listed in the second paragraph of this paper, could be used to identify comprehensive sources of raw data.

After this, the database could proceed by contacting researchers or institutions that are associated with these studies and asking them to participate in the National HIV Database on a voluntary basis. If they agreed to participate, the database would ask them to supply the information that is listed below:

general description of study
purpose of study
when study started
when study ended
number of persons investigated by study
criteria for selection of participants
basic demographics and descriptors (number males, females, etc.)
location and affiliation of investigators
names of investigators involved in data collection
list of publications generated by study
copy of study protocol
copy of all questionnaires used in study
description of how raw data is stored and organized by the study
funding source of study
researcher to contact for further information
comments on willingness to share data and collaborate on particular subjects
information on federal, state or local laws that govern use of data

After receiving answers from research groups that agreed to participate, the information could be stored in an online relational database. This type of database could be updated continuously and made available to all interested researchers.

In order to facilitate collaboration between researchers, the National HIV Database could furnish a standard contract that governs all agreements to share raw data. The purpose of this contract is to save time and energy for collaborations that proceed through the database. The contract also establishes a uniform code of conduct among researchers regarding security and privacy of the data. This contract could address the following points between users and providers:

agree to list names and affiliations of all the users
agree to obtain approval before introducing new users to the list

agree not to show raw data to persons outside the user list
agree to protect the security of the raw data
agree to remove and not use the names of persons as identifiers in the raw data set
agree to let the providers review all manuscripts prior to submission for publication
agree to include the names of providers on all manuscripts
agree to pay the cost of supplying the data
agree to return or dispose of the raw data after completion of research

The National HIV Database could serve the research community through a number of different scenarios. We list three possible ones below and mention some tradeoffs for each:

1. The database acts solely as a directory of sources for raw data. It supplies the descriptive information listed above and assists users in identifying providers. For more information, each user would have to contact a provider individually. The database does not engage in establishing formal agreements between users and providers, and consequently the standard contract is merely a guideline for collaboration. In this instance, the burden of supply falls on the provider.

2. The database acts both as a directory of sources for raw data and as a go-between for establishing formal agreements. It supplies the descriptive information listed above and assists users in contacting providers for more information. The database engages in establishing formal agreements between users and providers, and consequently the standard contract is an instrument for collaboration. In this instance, the database keeps abreast of collaborations that it fostered and helps the provider with the burden of supply.

3. The database acts as a storehouse of raw data. It supplies the descriptive information listed above, supplies the raw data, and supplies the standard contract for use of the raw data. According to prearrangements, the database may or may not be obligated to contact a provider before releasing their data to a user. In this instance, the burden of supply is minimized for the original provider. However, it requires a larger support effort for the database and it might reduce contact between researchers.

CONCLUSIONS

1. Currently there are a large number of electronic and hard copy databases on HIV and AIDS. These databases mainly contain summaries of published literature and synthesized information with very little raw data.

2. We believe a national HIV database that organizes raw information for epidemiologists, sociologists, and mathematical modelers is required. This information should be contained in an electronic database and would assist researchers in obtaining the raw data they need to verify

models and make predictions.

3. The HIV database should be designed to protect the integrity and security of the raw data. It should also to promote collaboration between the researchers who provide and use data.

4. The HIV database should incorporate the latest techniques in relational database technology and distribution.

5. This HIV database should be initiated expeditiously.

REFERENCES