Data Harmonization Process for Creating the National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention Atlas

Kim Elmore, PhD, MA[a]
Rob Nelson, MPH[b]
Zanetta Gant, PhD, MS[a]
Carla Jeffries, MPH[c]
Lance Broeker, MBA[d]
Massimo Mirabito, MBA[d]
Henry Roberts, PhD[e]

ABSTRACT

In 2009, the CDC National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) initiated the online, interactive NCHHSTP Atlas. The goal of the Atlas is to strengthen the capacity to monitor the diseases overseen by NCHHSTP and to illustrate demographic, spatial, and temporal variation in disease patterns. The Atlas includes HIV, AIDS, viral hepatitis, sexually transmitted disease, and tuberculosis surveillance data, and aims to provide a single point of access to meet the analytical and data dissemination needs of NCHHSTP. To accomplish this goal, an NCHHSTP-wide Data Harmonization Workgroup reviewed surveillance data collected by each division to harmonize the data across diseases, allowing one to query data and generate comparable maps and tables via the same user interface. Although we were not able to harmonize all data elements, data standardization is necessary and work continues toward that goal.

[a]Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Division of HIV/AIDS Prevention, Atlanta, GA

[b]Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Division of STD Prevention, Atlanta, GA

[c]Northrop Grumman Corporation, Falls Church, VA

[d]Agency for Toxic Substances and Disease Registry, Geospatial Research, Analysis, and Services Program, Atlanta, GA

[e]Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Division of Viral Hepatitis, Atlanta, GA

Address correspondence to: Kim Elmore, PhD, MA, Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, 1600 Clifton Rd. NE, MS-E48, Atlanta, GA 30333; tel. 404-639-8719; fax 404-639-8642; e-mail <kelmore@cdc.gov>.

An atlas is a specific and sophisticated mapping tool that displays spatial relationships, patterns, and trends. Maps, generated within an atlas application, provide a unique method for examining data. In fact, it can be argued that maps are the most powerful method for displaying statistical information (i.e., the utility of maps over tables appears to increase as the quantity of data increases).

In 2009, the Centers for Disease Control and Prevention (CDC) National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) began developing an online, interactive atlas. At the time, there was no integrated data store for human immunodeficiency virus (HIV), viral hepatitis, sexually transmitted disease (STD), and tuberculosis (TB) data that could be used to meet the analytical and data dissemination needs of NCHHSTP. While limited data were available to the public via CDC Wonder, the resource did not integrate with NCHHSTP data and lacked the ability to cross-link to NCHHSTP online resources.[1]

The overarching objective of the NCHHSTP Atlas project is to provide a single, unified access point to NCHHSTP data that meets the analytical and data dissemination needs of the agency and its national, state, and local partners, as well as the general public. The goal of this project is to develop an application that can be used to illustrate demographic, spatial, and temporal variation in disease patterns; identify gaps in health-care access and delivery; and understand geographic variation in services. The primary intended audience for these data is staff of state and local health departments, community-based organizations (CBOs), and other domestic partners that are involved in HIV, viral hepatitis, TB, and STD prevention. It is also useful to other federal agencies, hospitals, community health centers, and private health-care providers. This application adds value to the field of public health, in particular HIV and STD prevention, treatment, and care. The Atlas will use creative geographical displays of data to spark public interest in the spatial and temporal patterns of disease.

The initial data for the Atlas include state-level HIV, acquired immuniodefiency syndrome (AIDS), acute viral hepatitis (types A, B, and C), STD (chlamydia, gonorrhea, and primary and secondary syphilis), and TB case surveillance data collected by state and local health departments. Having these data centrally located makes the process easier for users to collect data across diseases, as the diseases within NCHHSTP share a number of commonalities. For example, they have similar at-risk populations: racial/ethnic minority groups, men who have sex with men, and injection drug users. These diseases also have significant health interactions and share similar social determinants of health. In 2010, NCHHSTP released a strategic plan outlining goals and objectives for the NCHHSTP.[2] The data harmonization effort contributes to the plan by promoting greater collaboration across the organization, as well as with an expanded array of external partners, and supporting a more comprehensive approach to prevention.

The administration issued the "Memorandum on Transparency and Open Government" in January 2009, which defined three principles—transparency, participation, and collaboration.[3] As a direct result of this memorandum, the Data.gov initiative was launched in May 2009 to increase public access to data generated by federal agencies. In the spirit of the Data.gov initiative, the data harmonization process allows us to make NCHHSTP data more accessible and interpretable via a standard format.

To provide unified access to NCHHSTP data, an NCHHSTP-wide Data Harmonization Workgroup was established to develop a common format for aggregated surveillance data. Harmonizing data across disease conditions allows users to dynamically query NCHHSTP data and generate comparable maps, charts, tables, and other graphics for each disease. For the purposes of this project, data harmonization is defined as a process whereby an organization reviews separate but similar data and attempts to align the data elements in a way that can be combined and used for comparison or stratification. This alignment can be accomplished by translating data into a common format or by associating metadata that allow for the comparison to occur. Complete harmonization would be reached if all of the data elements in the Atlas were defined in the same manner, using the same categories; for example, if all diseases used five-year age groupings. However, given differences in program focus and disease epidemiology, data are collected, stored, and presented in different formats by each NCHHSTP division. In the following section, we describe our efforts to harmonize these data to maximize comparability. These efforts, championed by NCHHSTP's program integration leadership, resulted in unprecedented program collaboration for the purpose of integrating HIV, TB, STD, and viral hepatitis surveillance data.

METHODS

Each division within NCHHSTP receives de-identified case report data at the individual level. Although each division collects a distinct set of data, there is a common core collected by each, which includes year of diagnosis, geographical area of residence, age, race,

and sex.[4–7] The harmonization effort focused on aligning the datasets for a given disease and the values for each of the core elements in the most comparable manner (Figure 1).

Once the workgroup identified which core variables would be displayed in the Atlas (Phase 1), we determined that the data would be released in phases (Figure 1). The workgroup decided to incorporate and go live with HIV and STD surveillance data first (Phase 2); six months later, viral hepatitis and TB data were incorporated into the Atlas (Phase 3). At this point, all NCHHSTP data were harmonized into the Atlas and released to the public. Phase 4 (in progress) includes making the data available via Data.gov.

Data elements

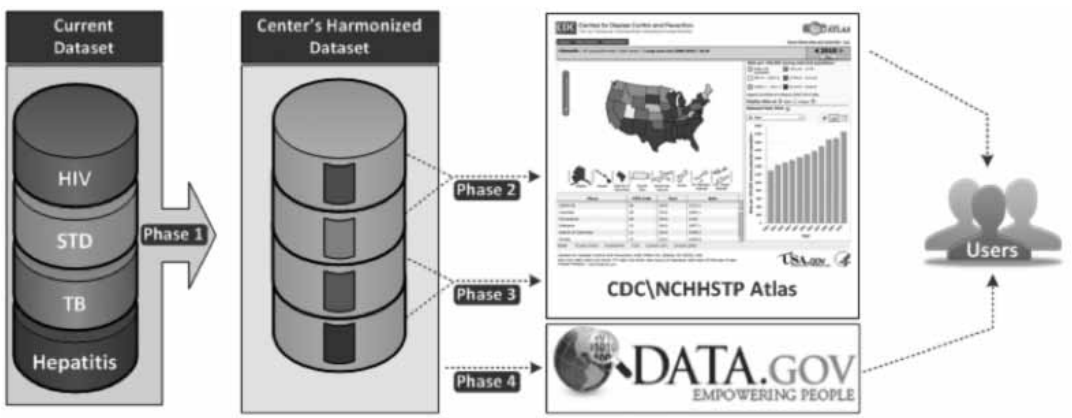
Although each division collects similar variable attributes, differences were found in how the divisions typically categorize these data for dissemination, such as annual surveillance reports. Therefore, issues arose during the harmonization process.

Beginning with year of diagnosis, all cases of disease are described by the attribute "years." Next, we reviewed the range of years available and determined that 2000 was the earliest year for which data were available for all diseases at the most detailed level. The divisions' respective raw data are available by a wide range of geographic areas, including region, state, metropolitan statistical area, county, ZIP code, and census tract. State (or state's equivalent) was selected as the common geographical area for which data are available for all diseases.

Sex (at birth) is collected in the same manner by each division. Age at time of diagnosis is collected as ungrouped, single year of age by each division. However, each division categorizes age differently to highlight different populations for its individual disease. The Division of HIV/AIDS Prevention (DHAP) defines adults and adolescents as people ≥13 years of age, but all other divisions define adults and adolescents as people ≥15 years of age. The Division of STD Prevention (DSTDP) typically uses five-year age groups when releasing data; the other three divisions use wider groupings (Figure 2). Case counts of STDs are larger than case counts of other NCHHSTP diseases; thus, STD data can be presented in smaller age groups without compromising confidentiality. In addition, it is often more important to STD prevention efforts to identify and target specific ages. Therefore, we decided divisions would display data in their preferred format.

There is a notable difference among divisions in how race/ethnicity data are collected and displayed. In 1977, the Office of Management and Budget (OMB) implemented reporting standards for race consisting of the following categories: American Indian or Alaska Native, Asian or Pacific Islander, black, and white.[8] At the same time, reporting Hispanic ethnicity was mandated (i.e., Hispanic origin and not of Hispanic origin). These standards were revised in 1997 to modify available race categories and allow reporting of more than one race category. The new minimum standard values for race are American Indian or Alaska Native, Asian, black or African American, Native Hawaiian or other Pacific Islander, and white.[9] DHAP and the Division of TB Elimination (DTBE) have adopted these revised race categories. DSTDP and the Division of

Figure 1. Data harmonization process for NCHHSTP Atlas



NCHHSTP = National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention

HIV = human immunodeficiency virus

STD = sexually transmitted disease

TB = tuberculosis

Figure 2. NCHHSTP Atlas variables

Variable	*HIV*	*Sexually transmitted disease*	*Tuberculosis*	*Viral hepatitis*
Disease	• HIV • AIDS	• Chlamydia • Gonorrhea • Primary and secondary syphilis	Tuberculosis	• Acute viral hepatitis A • Acute viral hepatitis B • Acute viral hepatitis C
Measure	• Prevalence • Incidence • Deaths	Reported cases	Reported cases	Reported cases
Disease and measure	• AIDS diagnoses, deaths, and prevalence • HIV diagnoses, deaths, and prevalence	• Chlamydia • Gonorrhea • Primary and secondary syphilis	TB incidence	• Acute viral hepatitis A • Acute viral hepatitis B • Acute viral hepatitis C
Year of diagnosis	• 2007–2010 (HIV) • 2000–2010 (AIDS)	2000–2010	2000–2010	2000–2009
Sex	• Male • Female	• Male • Female • Unknown	• Male • Female • Unknown	• Male • Female • Unknown
Age (in years) at time of diagnosis	13–24 25–34 35–44 45–54 ≥55	10–14 15–19 20–24 25–29 30–34 35–39 40–44 45–54 55–64 ≥65	0–4 5–14 15–24 25–34 35–44 45–54 55–64 ≥65	Hepatitis A and C: <5 5–14 15–39 40–59 ≥60 Hepatitis B: <15 15–39 40–59 ≥60
Race/ethnicity	• American Indian/ Alaska Native • Asian • Black/African American • Hispanic/Latino • NHOPI • White • Multiple races • Unknown	• American Indian/ Alaska Native • African American • Hispanic • White • Unknown • Asian/Pacific Islander • Other (OMB-compliant race not yet available from all areas)	• American Indian/ Alaska Native • Asian • Black/African American • Hispanic/Latino • NHOPI • White • Multiple races • Unknown	• American Indian/ Alaska Native • African American • Hispanic • White • Unknown • Asian/Pacific Islander • Other
Geographic areas	50 states, DC, American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands	50 states, DC, Guam, Puerto Rico, and the U.S. Virgin Islands	50 states and DC	50 states and DC
Transmission category (HIV, AIDS only)	• MSM • IDU • MSM/IDU • Heterosexual contact • Other			

NCHHSTP = National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention

HIV = human immunodeficiency virus

AIDS = acquired immunodeficiency syndrome

TB = tuberculosis

NHOPI = Native Hawaiian/other Pacific Islander

OMB = Office of Management and Budget

DC = District of Columbia

MSM = men who have sex with men

IDU = injection drug user

Viral Hepatitis (DVH) are in the process of implementing these standards, as the most recent case report formats are OMB-compliant as of 2008. Almost all jurisdictions submit compliant race/ethnicity data for STDs and viral hepatitis; once the remaining areas are compliant, all divisions will be able to report fully compliant data at the national level. Until that point, and because divisions use particular standards for race categories, the harmonization of race/ethnicity is not completely possible.

Denominator data

The Atlas presents both cases and rates of infection (per 100,000 population). Rates are calculated using several population data sources. DHAP uses the U.S. Census Bureau's annual population estimates that match their multiracial category morbidity data. DTBE uses the Current Population Survey for U.S.-born and foreign-born population denominators and the American Community Survey for geographic population estimates. DSTDP and DVH use CDC's National Center for Health Statistics bridged population data file, which is based on Census data but features race categories comparable to the STD and hepatitis source data.[10]

In addition, DHAP uses the most recent vintage of the Census Bureau's estimates, while the other divisions prefer to use "frozen" estimates. For instance, DSTDP uses the 2007 vintage estimates to generate rates for 2007, focusing on being able to replicate rates; DHAP uses the 2009 vintage estimates to generate 2007 rates. Again, because the divisions use slightly different population files, the harmonization of denominator data is not completely possible.

Stratifications and suppression rules

Each division's surveillance data are maintained in separate, secure data stores. The layout and formatting of these data are unique to each division and reflect the development of the surveillance data stream during the past 20 or more years. A separate database with combined data from all four divisions is maintained by the Atlas development team. Before being transferred from the divisions to the development team, data were pre-summarized using SAS®,[11] as well as pre-suppressed by removing all small cells following the data re-release guidelines (for HIV, data were additionally limited to two-way stratifications of demographic/transmission category variables), and placed into a uniform structure that preserved rows containing null values. Pre-summarization is aggregation of data into a cube-like data structure that contains rows for every combination of demographic variables by year, geography, and disease. The most detailed level for a given disease and year is by state, race/ethnicity, age group, and sex (and, for HIV, transmission category). An example database row at this level is: Chlamydia, 2010, Alabama, Asian, 15- to 19-year-olds, male. At the other end of the spectrum, the most generalized summarization level is simply the total case count and rate for a given disease and year. An example of a most general summarized level is: Chlamydia, 2010, all states, all races, all age groups, and all sexes.

Each division is bound by agreements with the states, territories, and the Council of State and Territorial Epidemiologists (CSTE) to release only data that have been reshaped in a way that protects confidentiality. DHAP and DTBE maintain individual agreements with each state and territory, while DSTDP and DVH use the 2005 data release document from CSTE.[12] However, DSTDP plans to update data release guidelines during 2014. Therefore, data confidentiality concerns influenced the overall data management. To prevent Atlas users from identifying specific individuals by stratifying on multiple variables in areas with low denominators, records were suppressed using division-specific suppression criteria. Suppression is a method of disclosure limitation used to protect individuals' confidentiality by not showing the cell values when the value does not meet a certain minimum threshold (primary suppression). Primary suppression may be augmented by complementary or secondary suppression of other cells to avoid inadvertent disclosure through back-calculation.[13] For a full discussion of the suppression rules used in the Atlas, please see the "about these data and footnotes" section on the Atlas main page.[14]

OUTCOMES

In 2012, NCHHSTP released the first version of an online, interactive Atlas containing HIV, STD, viral hepatitis, and TB data (Phase 4). The work completed by the Data Harmonization Workgroup allowed for the development and release of the Atlas. NCHHSTP data alignment is in agreement with the NCHHSTP's Strategic Plan and the three primary goals of the White House's National HIV/AIDS Strategy: reducing HIV incidence, increasing access to care and optimizing health outcomes, and reducing HIV-related health disparities.[2,15]

The Center-wide data harmonization project was a first step in tackling the complex issues on data integration, reporting, and dissemination. The following is a brief summary of the key findings and lessons learned:

1. Organizations that are considering data harmonization should not underestimate the complexities and potential for unexpected delays.

2. The data harmonization process must happen prior to technical implementation.
3. Disease domain experts should be brought together to determine how to harmonize data to ensure that variables can be directly compared with one another.
4. Senior leadership should be involved to assist in defining the vision and pushing forward the integration effort.
5. Everyone should be prepared to negotiate on data sharing to achieve common ground.
6. It is important to spend time modeling the outcomes to ensure that the application does not encourage users to draw false conclusions.
7. Granularity, data resolution, and confidentiality rules can become challenging when multiple disease are combined. However, the outcomes could be that states and local jurisdictions standardize collection and reporting parameters.

This Atlas is noteworthy for five reasons. First, the Atlas is the first application to use harmonized NCHHSTP data to query, analyze, and graphically display the disease data collected by NCHHSTP. The Atlas is a user-friendly application that can be accessed by a variety of users. Internal and external partners now have a single tool to obtain NCHHSTP surveillance data. Second, prior to the Atlas, the general public would have to contact each division individually, submit a detailed data request, and wait for the data to be available. This process could take days or weeks depending on the request, time of year, and/or disease. Now, partners and the public can submit their query of interest and obtain user-friendly results (including data in a Microsoft® Excel spreadsheet and a map in a PowerPoint slide) in a matter of seconds. Third, making mapped data easily available is very useful for the end user, as maps are visually striking and, at times, easier to view than tabulated data. Tufte indicates that "no other method for the display of statistical information is so powerful"[16] in allowing users to more easily identify trends and patterns in the data. Fourth, the ability to share data across programs is essential to identifying areas with co-occurring high rates of infectious disease, as surveillance data are typically collected and analyzed independently by separate disease-specific programs within many health departments across the U.S. And fifth, on the other side of the data request process, the Atlas has been beneficial to the surveillance divisions that provide the data, as data requests within each division have significantly decreased. Overall, there are a significant number of benefits to this data harmonization process.

LESSONS LEARNED

Although the Atlas has provided many benefits in terms of data accessibility and visualization, our data harmonization efforts were only partly successful. All divisions defined year of diagnosis similarly and with the same range of years, sex at birth is defined the same for all divisions, and all divisions re-release and display data for the 50 states and the District of Columbia (although the divisions differ regarding from which U.S. territories they collect and re-release data). This process made harmonizing the data fairly easy. As discussed, the two variables that were the most difficult to harmonize were race/ethnicity and age groupings. In terms of race/ethnicity, because all areas are not yet compliant with the most recent OMB standards on race[8] and because each division collects these race/ethnicity data differently, data for the Atlas depict both new and old race categories. Once this issue is resolved, the Atlas will be able to display data solely using the new race categories. At this time, however, divisions only allow for their specified standard to be used for race/ethnicity. Finally, although all divisions collect age in the same manner, they have different rules for releasing data by age, based on the age ranges that are epidemiologically and programmatically important for each division and confidentiality agreements with the areas.

Limitations

While the Atlas has proven to be a beneficial tool for data dissemination and visualization, there are some limitations. Currently, it does not include comorbidity data; for example, data cannot be integrated to show that an individual has HIV and hepatitis C, or chlamydia and TB. Therefore, while side-by-side data comparisons can be made, full integration of NCHHSTP data is not possible because individual-level information is unavailable. Preexisting confidentiality agreements between specific divisions and their state and local partners prevent the display and stratification of certain data, and pre-suppression of the data altered the functionality of the Atlas application, which limits the user from obtaining certain aspects of data. Some states still collect race/ethnicity data using a non-current definition of race. Additionally, there are disease-specific limitations. For example, confirmed hepatitis cases require patient follow-up, and a viral hepatitis surveillance system does not exist. These limitations greatly impact the completeness of acute and chronic viral hepatitis reporting. Complete harmonization of all data elements is not possible, as it is important for the divisions to disseminate and display data in a manner that is in line with current data re-release agreements with the states and territories. However, all

four divisions and NCHHSTP leadership realize that data standardization is necessary and work continues toward that goal.

CONCLUSION

The NCHHSTP Atlas has received positive feedback, and both internal and external partners continue to use the Atlas on a routine basis. Version 1, which contains state-level data, was released in January 2012.[14] The next step in this project is to display HIV and STD data at the county level, which will provide a more detailed perspective. Future plans include incorporating an advanced query component, a dashboard component, additional NCHHSTP data not limited to surveillance data (e.g., HIV testing data), and additional data from the U.S. Census and other sources.

The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

This study was considered exempt from institutional review board review.

REFERENCES

1. Centers for Disease Control and Prevention (US). CDC Wonder [cited 2012 Jun 26]. Available from: URL: http://wonder.cdc.gov
2. Centers for Disease Control and Prevention (US), National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. NCHHSTP strategic plan 2010–2015. February 2010 [cited 2013 Jul 27]. Available from: URL: http://www.cdc.gov/nchhstp/docs/10_NCHHSTP-strategicPlanBookSemi-final508.pdf
3. The White House (US). Transparency and open government: memorandum for the heads of executive departments and agencies [cited 2012 Jun 26]. Available from: URL: http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment
4. Centers for Disease Control and Prevention (US). HIV/AIDS statistics and surveillance: statistics center [cited 2012 Jun 26]. Available from: URL: http://www.cdc.gov/hiv/topics/surveillance/index.htm
5. Centers for Disease Control and Prevention (US). Sexually transmitted diseases (STDs): data and statistics [cited 2012 Jun 26]. Available from: URL: http://www.cdc.gov/std/stats/default.htm
6. Centers for Disease Control and Prevention (US). Tuberculosis (TB): data and statistics [cited 2012 Jun 26]. Available from: URL: http://www.cdc.gov/tb/statistics
7. Centers for Disease Control and Prevention (US). Viral hepatitis: statistics and surveillance [cited 2012 Jun 26]. Available from: URL: http://www.cdc.gov/hepatitis/Statistics/index.htm
8. The White House (US), Office of Management and Budget. Standards for the classification of federal data on race and ethnicity [cited 2012 Jun 26]. Available from: URL: http://www.whitehouse.gov/omb/fedreg_race-ethnicity
9. The White House (US), Office of Management and Budget. Revisions to the standards for the classification of federal data on race and ethnicity [cited 2012 Jun 26]. Available from: URL: http://www.whitehouse.gov/omb/fedreg_1997standards
10. Ingram DD, Parker JD, Schenker N, Weed JA, Hamilton B, Arias E, et al. United States Census 2000 population with bridged race categories: data evaluation and methods research. Vital Health Stat 2 2003(135).
11. SAS Institute, Inc. SAS®: Version 9.3. Cary (NC): SAS Institute, Inc.; 2011.
12. Centers for Disease Control and Prevention (US), Agency for Toxic Substances Disease Registry, and Council of State and Territorial Epidemiologists. CDC-CSTE intergovernmental Data Release Guidelines Working Group (DRGWG) report: CDC-ATSDR data release guidelines and procedures for re-release of state-provided data [cited 2012 Jun 26]. Available from: URL: http://www.cdc.gov/od/foia/policies/drgwg.pdf
13. Zayatz L. Disclosure avoidance practices and research at the U.S. Census Bureau: an update. Washington: Census Bureau (US), Statistical Research Division; 2007.
14. Centers for Disease Control and Prevention (US). NCHHSTP Atlas [cited 2012 Jun 26]. Available from: URL: http://www.cdc.gov/NCHHSTP/atlas
15. The White House (US), Office of National AIDS Policy. National HIV/AIDS strategy for the United States. Washington: Office of National AIDS Policy; 2010.
16. Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire (CT): Graphics Press; 2001.