Data Aggregation and Individual Privacy: 
A Critical Evaluation of the Policies of the CDC
Kathy Goodson (Captain), Eleanor Birrell, Katie Fifer, Amelia Sagoff

The CDC:

Founded in 1946 to help control Malaria\(^1\), the Center for Disease Control and Prevention (CDC) is now a government agency responsible for collecting and publishing data concerning all nationally notable diseases.

The CDC occupies an important place in this nation’s efforts to protect and defend its citizens. In accordance with its stated mission (to remain at the forefront of public health efforts to prevent and control infectious and chronic diseases, injuries, workplace hazards, disabilities, and environmental health threats\(^2\)), the CDC has assumed two basic roles: that of research, and that of education.

The research side of the CDC takes two forms. The first type of research consists of looking for solutions to current health problems in a laboratory setting; this research is conducted at the CDC’s fifteen research centers located across the country. The second type of research consists of looking for current health problems and patterns in infectious diseases. This research involves gathering, aggregating, and analyzing large quantities of data from individual medical records and other sources. This research has been greatly facilitated by the implementation of modern data aggregation techniques, which allow the CDC to process more data more quickly and more reliably.

The education side of the CDC focuses primarily on providing current, accurate statistics on a variety of current public health issues. It also provides the public with accurate information about infectious diseases, including their causes and treatments.

Throughout its many activities, the CDC tries to improve public health in the United States in accordance with its stated mission.

The Privacy of Data Records:

In order to fulfill its mission, the CDC collects, aggregates, and retains a lot of private health information about individuals. They maintain records of people involved in CDC

studies, people who have been reported as exposed to toxic substances, people whose medical records have been randomly selected for analysis, and many other groups that are believed to contribute valuable information to the study of public health. The ability to aggregate and analyze these data allows the CDC to better understand health trends and threats, and thus better protect the public from disease.

However, although the CDC aggregates large quantities of data, it maintains the privacy of each individual’s information through complete compliance with the Privacy Act and HIPPA, as well as through additional security measures.

There is no general policy for who can have access to a given database. However, most databases will only be released to research groups (if it is determined that this data is vital to the organization’s research and if such a release does not violate the conditions under which the information was gathered), the private group that contracted the research (if applicable), health departments and cooperating medical authorities, or the judicial department. If additional expertise is needed from outside the CDC, data may be released to collaborators. However, such releases are always contingent on the collaborator agreeing to appropriate confidentiality protections and the guidelines of the Privacy Act. In all cases, data is not released if such a release would violate either individual privacy or the confidentiality agreement.

Additionally, the CDC maintains separate databases of information for different purposes. For example, the Health Care Statistics database is maintained separately from the Records of Subjects in Health Promotion and Education Studies. Thus, the CDC can treat information differently according to its purpose, and ensure that only relevant information is shared with outside contractors, research organizations, or other government offices.

Finally, the CDC follows guidelines to ensure the security of the data. The CDC keeps all hard copy data not in use under lock. Most CDC sites require ID badges and cardkeys for access, and maintain daily logs. The CDC stores electronic data on tapes CDs, or in electronic files. Access to electronic data requires a password from an

---

authorized user. Files completely overwritten when the data is required to be erased, thus ensuring that all records are completely destroyed. Through these precautions, the CDC tries to ensure that unauthorized groups and individuals do not gain access to the data aggregated by the CDC for public health purposes.

Although individual data is usually only considered in the initial editing phase, when the data are checked for accuracy, all information is linked to the person’s name and social security number in order to ensure accurate statistics. In order to ensure that data is valid, anyone can provide the CDC with their name and social security number, and the CDC will report back all the information linked to that person. Individuals have the right to review information about themselves, and to contest its accuracy and request a change.

Therefore, although the CDC collects and aggregates large quantities of data about individual Americans, it tries to ensure that these records will not invade the privacy of individual citizens.

**Policy Evaluation:**

The CDC plays a critical role in our national and global health. Without access to information about current and past relevant data, the CDC would be rendered ineffective. Alternatively, the more data related to disease spread and control that the CDC is able to aggregate and analyze, the better it will be able to fulfill its vital role. Therefore, the CDC is absolutely justified in collecting both data on individuals and more generally on population trends whenever such data may aid in the fight of infectious disease.

Although there is no good definition for when data “may aid public health,” it seems the CDC has not been abusing its position as a data aggregator. All records maintained by the CDC include both a list of the information associated with the individual and the purpose for which that data was collected or aggregated⁷. This level of transparency helps ensure that the CDC remains true to its goal of improving public health and does not abuse its power to collect and aggregate data. Moreover, in any questionable case, the CDC is justified in erring on the side of collecting and attempting

---

⁷ Ibid. April 6, 2007.
to analyze more data rather than less, as the effective tracking and control of epidemics requires huge amounts of data.

While the CDC’s data aggregation and analysis efforts are certainly warranted, the fact that it does store such a wealth of personal, often very sensitive, information mandates that it take strict measures to ensure privacy protection. It appears that the CDC does make a serious effort to protect individual’s information, through compliance with HIPPA, sharing only ‘relevant’ information with contractors, protecting databases, and allowing people to find out what information of theirs is stored with the CDC. The attempts to make data anonymous when possible are also very valuable.

In evaluating any privacy challenge, it is vital to weigh the costs and benefits – here it seems the cost to individuals is minimized, by the various privacy protections the CDC implements. Meanwhile, it seems that the benefits of the CDC’s data aggregation far outweigh any costs. With this data, the CDC is given a much greater ability to operate as it should - monitoring and fighting the spread of disease, while studying past outbreaks. As long as the CDC continues to collect and aggregate only relevant data and as long as it continues the current access controls, its data aggregation will continue to respect a reasonable level of personal privacy.

Current standards notwithstanding, we advise the CDC to continue to evaluate its privacy protections and improve on them whenever possible. Any time that large amounts of data are stored somewhere, there are risks to individuals’ privacy; the CDC should continue to focus on minimizing those risks. Additionally, the CDC should make more publicly and easily accessible, its policies on the data it collects. While we do not advocate publishing exactly what data is collected by the CDC, we do advocate making a clear privacy policy covering treatment of this data central to the CDC website. Publishing a clear, accessible policy will reassure concerned citizens that their privacy is being protected, and it will not impede the CDC’s work.

Overall, we firmly support the CDC’s work in fighting disease, and, given its current privacy protections, we support its use of data aggregation in meeting this important end.