MEETING SUMMARY

November 2009

Distributed Data Networks for Active Medical Product Surveillance

Supported by the Food and Drug Administration

Background: The Sentinel Initiative

The FDA Amendments Act of 2007 mandated that, by 2012, FDA develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources, with the goal of including at least 100 million Americans.[1] In response, FDA began developing the Sentinel Initiative—a national, integrated, electronic system for active surveillance of medical product safety. In the 2008 announcement of the Sentinel Initiative, FDA proposed using electronic health care databases in a distributed network to accomplish these objectives. Unlike a centralized network where data are sent to a central location for storage and analysis, a distributed network allows all data sets to physically remain with the data holders behind their security and privacy firewalls: analysis is conducted via a remote system where each data owner processes the queries and returns the results to a coordinating center.

The initial phase of the Sentinel Initiative will likely focus on specific medical product-adverse event (AE) pairs for which there is a potential association based on previous clinical and/or epidemiologic information. As methods and data sources are developed, active surveillance efforts may be expanded to include detection of unanticipated medical product safety signals using data mining methodologies. FDA has indicated that the system will be capable of querying multiple existing distributed data environments of electronic health records and administrative claims that would continue to be owned and maintained by the original data holders.[2]

Building the Data Infrastructure for Active Surveillance

Building a distributed network for active medical product surveillance requires addressing a range of important technical and policy issues. These include: identifying and recruiting appropriate data environments, determining the data structure and methods for standardizing data from different sources, implementing a process for distributing queries and returning results, addressing security and privacy issues, and defining the responsibilities of the coordinating center and the data holders in executing queries. In the case of the Sentinel System, the initial network must be scalable over time. These data and infrastructure needs were a primary focus of the Brookings "think tank" on distributed data networks convened on November 23, 2009.

The specific objectives of this meeting were to: 1) identify key features of existing data networks conducting medical product safety surveillance; 2) discuss data infrastructure needs for a future

national active surveillance system; and 3) discuss barriers and explore solutions to private data environment participation in active surveillance. Think tank participants included approximately 50 individuals with expertise in safety surveillance and/or health care data networks, including a number of FDA staff. Data networks represented at the meeting included the HMO Research Network (HMORN), the Vaccine Safety Datalink (VSD), the Distributed Ambulatory Research in Therapeutics Network (DARTNet), Exploring and Understanding Adverse Drug Reactions (EU-ADR), and the Observational Medical Outcomes Partnership (OMOP). In addition, perspectives on participation in safety surveillance networks were shared by representatives from a range of for-profit and non-profit organizations including i3 Drug Safety, HealthCore, Kaiser Permanente, Hospital Corporation of America, and the American College of Cardiology.

Feasibility of Analysis Using a Distributed Network

Several large-scale safety surveillance efforts have employed distributed data networks. In the case of the VSD, a distributed approach is used for near real-time active surveillance of vaccine safety as well as traditional epidemiological studies. Before the VSD implemented its current processes for analyzing data and submitting results, participating data holders annually sent datasets directly to the Centers for Disease Control and Prevention (CDC) for storage, data quality assessments, and analysis. Because of several issues including privacy concerns, this process was re-structured into a distributed data model. Since the system was restructured, the VSD has been able to capitalize on the new structure to create additional data files in which new data become available for analysis on a weekly basis. Users are instructed to run only approved queries, which may generate summarized results, summarized datasets, or individual analytic datasets in which to conduct approved analyses. Queries are additionally used to ensure standardization, conduct data quality assessments, and conduct study feasibility assessments. Trust is an essential element of the relationship between the CDC and data holders participating in the VSD. However the system contains certain built-in restraints and/or monitoring mechanisms, which vary by participating sites to ensure compliance.

The VSD's automated structure allows for a large number of previously approved queries to be submitted on a weekly basis. This large number of queries is necessary to conduct several active surveillance projects simultaneously and allow weekly/monthly data quality assessment as well. The VSD is in the process of completing or has completed over ten active surveillance projects for newly licensed vaccines. Workshop attendees noted the importance of allowing data holders to maintain control of their data because data holders have the most comprehensive understanding of the limitations of their own data and are usually best equipped to analyze it.

Solutions were suggested for ensuring that data remains with the sites participating in a distributed network. HMORN, for instance, has incorporated manual steps in processes for query execution and data submission. Data holders in HMORN always control the execution of queries of their data. Queries are sent via e-mail to data holders as SAS programs, and data holders must explicitly choose to run the analysis. This manual opt-in process has made it easier for health plans to participate in HMORN. In addition, the lack of automation minimizes the need for extensive database expertise and ongoing maintenance of a complex data infrastructure.

On the other hand, expertise is also required to manually submit programs. The VSD has found that for a weekly active surveillance project, it would be cumbersome to use a non-automated process, since the approved programs do not change from week to week. For example, the VSD has a single analyst/programmer submitting the programs weekly to collect the data for several active surveillance projects. If a non-automated process were in place, the VSD would need several of these programmers/analysts to submit the programs weekly, rather than just one.

While meeting participants emphasized that data should remain with data holders, they also noted the importance of linking administrative and clinical data to other data sources in certain circumstances. Links to death registries, for instance, can discern whether or not an absence of claims indicates that a particular patient has died. It is often feasible for data environments to link to external data sources while keeping their own information behind firewalls. This enables distributed network analysis on a richer dataset without transporting personally identifiable information. One-way hash function is a technical approach that is particularly useful for comparing population-level data sets without unnecessarily exposing patient data. Using probabilistic matching techniques, hash keys could be matched across distributed data sources to establish linkages.

Support for a Common Data Model

Workshop participants supported the use of common data models for conducting postmarket surveillance in distributed networks. A common data model is used to standardize particular data elements across data environments. Data holders that use a common data model can run distributed protocols and programming codes, ensuring more comparable results than when custom code is written for each disparate data environment.[3] Common data models may also reduce effort required by data holders to implement a study.

A practical approach to the development of a common data model in a distributed network requires maintaining the expectation that the model will need to be able to evolve to incorporate new data types, such as genetic data. While a common data model will likely change over time, it should also be designed to minimize the amount of information loss associated with transforming the data to fit the model.

The common data model development process should leverage the expertise of participating data holders and take into account the amount of effort required by sites to implement the model. A robust coordinating center can help ensure that data holders correctly implement the model so that data are transformed and analyzed in a manner that is consistent across sites. Additional checks may also be needed to validate source data. HMORN, for instance, checks data quality and completeness via distributed programs. The VSD also checks data quality and completeness via distributed data programs on weekly, monthly, and yearly bases.

Factors Affecting Participation of Private Sector Data Holders

In order to be included in safety surveillance initiatives, a data environment should meet at least three criteria. First, the data environment must capture relevant product exposures, outcomes, person-time data, and essential covariates in an acceptably complete manner. Potential data

environments include a number of public and private organizations such as insurance plans, public payers, hospital networks, integrated delivery systems, and registries. Second, the data environments should offer the possibility of confirming and augmenting claims data via full-text medical record review. And third, data environments should have the analytic infrastructure and staff to perform queries and return results to a coordinating center. While many private-sector data holders meet these criteria, a number of issues have the potential to limit their participation in the Sentinel System.

First, it is important to recognize that many potential private data environments are for-profit entities with proprietary tools and technology that—along with their data—are used to provide consulting or research services. Whether these organizations ultimately participate in a national surveillance network will depend on a number of factors, including its potential impact on their business activities and the value received in return for that effort. Private data environments will be more likely to participate if the organization's role involves utilizing their local expertise and analytic capacity to engage in the network's activities rather than just provide access to their data. It should further be noted that participation in the Sentinel System will require expenditure of resources and some organizations may require compensation for the development and/or maintenance of this capacity.

Second, as the experience of other distributed networks has demonstrated, private-sector data holders prefer an "opt-in" approach to participation in specific surveillance activities. The preferences and criteria for participation in any given activity may vary across environments.

Third, potential data environments need guidance on a range of potential legal and regulatory concerns. For example, they are interested in understanding whether or not their data could be "discoverable" in legal proceedings and/or requested under the Freedom of Information Act (FOIA) and therefore released into the public domain. Data holders' concerns about data privacy regulations could be addressed by clarifying how distributed data networks must be operated to comply with federal and state laws. Finally, potential data environments need to understand tort liability in the event that a signal is discovered in their data.

Summary and Next Steps

Some existing data networks in the U.S. already offer the capacity to evaluate safety signals, and their continued evolution to larger scale and more real-time distributed analyses should be a priority. Near-term technical objectives for network development include agreement on an initial common data/information model and terminologies/ontologies for defining variables of interest, the ability to securely enrich administrative data with other information, the capacity to distribute and operate queries in a more automated fashion, and standards for validating source data and query results.

A second objective is to expand the capacity of networks beyond signal strengthening, to include real-time signal identification (data mining) and rapid confirmatory studies. These activities will require advancements in safety science methods and development of common algorithms for analyzing data that has been transformed to fit the common data model.

A key enabler of expanded capacity will be policy support, including guidance in the areas of privacy, tort liability, data security, and the conduct of public health practice vs. research. Training for safety scientists, data environments, and the public on these issues will facilitate implementation.

Potential data environments must be developed over time in a way that preserves their control over when and how their data are used. The "business case" for developing and maintaining local infrastructure and analytic capacity must be developed, particularly for organizations that may be providing related research and consulting services already.

Finally, there is a need for continued guidance from FDA with respect to the overall goals and objectives of the Sentinel System, including desired capabilities and a vision for how the Sentinel System fits into the broader development of national health information architecture.

---

[1] Full text of the law is available at http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ085.110.

[2] The Sentinel Initiative: National Strategy for Monitoring Medical Product Safety. May 2008. Available at http://www.fda.gov/oc/initiatives/advance/reports/report0508.pdf

[3] Janet Woodcock, "Data and Infrastructure for Medical Product Surveillance," presentation at the Brookings Institution, Washington, DC, 02 Dec 2009.