Commentary: Protecting Human Subjects and Their Data in Multi-site Research

Harold S. Luft, PhD†*

Abstract: Comparative effectiveness research is sometimes best done using routinely collected data from multiple real world settings. This raises important issues with respect to the authorizations to use the data, dealing with the multiple IRBs that may have oversight authority, and protecting it from breaches. Careful delineation of what is quality improvement vs. research-driven interventions can help with IRB reviews, especially for low-risk projects. It is impossible to totally de-identify data without markedly reducing its value for research, but creative strategies can markedly reduce the risk of advertent disclosures of patient identities.

Key Words: IRB reviews, quality improvement research, HIPAA privacy concerns

(*Med Care* 2012;50: S74–S76)

Comparative effectiveness research (CER) is most valuable when drawn from multiple real-world settings, because this addresses issues of generalizability and rapid acquisition of enough cases. This, however, often means dealing with multiple sites [and their Institutional Review Board (IRB) reviews] and ensuring that the research is Health Insurance Portability and Accountability Act (HIPAA) compliant. However, for CER to be maximally useful, it must be accomplished in a timely manner; therefore, effective approaches to IRB reviews and HIPAA compliance are fundamental. The papers by Marsolo[1] and Kushida et al[2] offer useful summaries of these issues. I focus here on a subset of data governance, setting aside questions of the controls an organization may want over the use of its data and addressing instead the issues faced by willing collaborators.

The fundamental challenge is to design strategies that (1) protect human subjects and the privacy of their data while (2) facilitating research. IRBs are charged with the first task and accomplish it to varying degrees. There are concerns that federal regulations and IRB practices unduly burden certain types of research, especially those posing minimal risk to subjects, without adding significant protections. The Office of Human Research Protection suggested changes to the federal regulations in its Advance Notice of Proposed Rule Making[3] that garnered 1110 public comments. Some of the suggested changes would facilitate multi-site CER, but underlying challenges would remain, and it may be years before IRBs fully adopt any changes pursuant to new regulations.

To understand the issue at hand, it is useful to consider a simple 2×2 table (Table 1). If researchers have any influence over the use of an intervention, an IRB should review the risks and benefits of the intervention. However, if clinicians or others decide on their own to alter practice, then it comes under the rubric of quality improvement and does not require IRB review. If the data to be used by the researcher are fully deidentified so that the privacy of the patients involved is not at risk, there is no need for an IRB review of HIPAA authorization forms or a waiver of authorization. An important result of the Pronovost decision was that IRB review was not needed for the use of deidentified data used to assess a quality improvement intervention.[4] Thus, if a CER researcher is studying "naturally occurring clinical behavior," effective deidentification can eliminate many hurdles.

However, as Kushida and colleagues point out, deidentification of important aspects of electronic data is not a simple task. There are 2 issues here. The first is that certain combinations of variables, such as 5-digit zip code, birth date, and sex, can allow data to be matched with publicly available information to reidentify records.[5] The second is that text and images necessary for research may include direct identifiers, such as names or recognizable faces. These are quite different problems and require different solutions.

The first problem involves purposeful attempts to identify individuals within a data set by matching the values of the data with information from outside the data set. Avoiding the possibility of any record being reidentified would require eliminating so much information as to render the data useless for most work. The only viable solution is to forgo creating "public-use" data sets accessible by anyone wanting to do reidentification. Instead, data sets should be made available through data use agreements with researchers agreeing to not attempt reidentification. Additional protection can be achieved from institutional data security processes (firewalls), encrypting removable devices, limiting the variables on a specific analytic data set to just those

From the *Palo Alto Medical Foundation Research Institute, Palo Alto; and †Caldwell B. Esselstyn Health Economics and Health Policy, University of California, San Francisco, CA.

The author declares no conflict of interest.

Reprints: Harold S. Luft, PhD, Palo Alto Medical Foundation Research Institute, Ames Building, 795 El Camino Real, Palo Alto, CA 94301. E-mail: lufth@pamfri.org.

ISSN: 0025-7079/12/5007-0S74

TABLE 1. Issues Arising from Data and Human Subjects Involvement in Research

	Researchers Influence the Use of the Intervention	
HIPAA Privacy Concerns	**Yes**	**No**
Yes	Protocol and data issues	Data issues
No	Protocol issues	No issues?*

*Many journals require an IRB determination that the study is exempt; organizations may use the pathway to the IRB as a way to address other governance issues.

needed for the project, and more complicated multi-keyed links described by Kushida et al.[2]

A trusted data processing intermediary can be a useful adjunct in this regard. It can maintain under highly secure conditions large quantities of reidentifiable data and create on demand files with just the information needed. Research projects are often built around the "available data" that can be obtained in a single request. If the intermediary can quickly add variables, then the researcher can request far less, reducing the risk to privacy.

The second problem is quite different, involving the unwanted presence of directly identifiable information. Free-text fields (eg, notes in medical records) are often the only source of critical data. Patient and other names may be embedded in such text. Kushida and colleagues describe various programs to remove names from text. None are perfect, and attempts to deidentify images are even more fraught. A combination solution is probably necessary. Because the privacy risk is inherent in accessing the data, access should only be granted to researchers bound by strong ethical principles and policies. (Misuse of inadvertently identified information is comparable with unauthorized access to specific medical records by a health care worker.) Access to data that could include identifiers should be restricted to the minimum number of staff; they can then share the extracted data without any identifiers with the rest of the team. Risks may be further reduced by sequentially using 2 deidentification programs with complementary weaknesses. Generic programs will usually be less effective than those leveraging a deep understanding of how the data were generated. Instead of substituting names with numeric codes, which makes any name stand out, it may be better to replace real names with randomly chosen sex-matching names—"hiding in plain sight" those happening to be missed by the deidentifying programs.[6]

If it will be impossible to create true "public-use" data files for CER (ie, few projects will be in the lower right box of the table), IRB reviews will be needed, along with data use agreements. If the research goes beyond passive observation of normal or independently generated quality improvement-focused activities, the IRB will want to examine protocols describing the interventions to assess any risks to patients, who would now also be serving as consented or unconsented subjects. (When the intervention is considered clinically appropriate and part of a quality improvement effort, consent might not be necessary. Involvement of researchers, however, may bring the question to the IRB.) To increase sample size and generalizability, one would like to study the intervention in multiple settings, which then involves multiple IRBs.

Marsolo[1] describes various strategies that can facilitate multi-site studies by reducing the need for complete protocols to be approved by each IRB with the risk that a late-reviewing IRB may require changes unacceptable to an early approver. Most of these strategies were developed for clinical trials that often involve significant risks to the subjects. (These appear on the left-hand side of the table.) Having a central or lead IRB with the expertise to carefully review and require protocol modifications by the principal investigator can be a major benefit to other (local site) IRBs that can then decide whether to accept the project or not.

This model works well for large collaborative clinical trials in which individual sites are primarily responsible for enrolling and managing a few patients each. The situation is far more complex if the intervention involves substantial effort by both the local investigator and the organization. Applying this model to a CER study of process changes in hospitals, would be a challenge, because the "intervention protocol" is likely to be far less specific than that of a clinical trial. Moreover, if randomization is even part of the study, it is likely to be by hospital unit, or month, rather than by patient. Asking for individual subject consents/authorizations does not make sense in such cases. Researchers will have to work with their IRBs to carefully tread the line between quality improvement and research. One option would be to have the organization view the stepped implementation as a quality improvement effort and then the IRB would review the collection of data as research.

Collaborative clinical studies may have protocols with more subject risk, but usually have the ability to obtain direct patient consent and HIPAA authorization eliminating privacy issues. Even with prospective methodologies, CER studies are more likely to rely on HIPAA waivers, introducing the concerns about data privacy discussed above. When a set of researchers and organizations are frequently working together, as is the case in the HMO Research Network, they can establish the processes for sharing data, the standardized data use agreements, the ability for one IRB to accept the review of another, and the necessary trust to make this all work well. Such arrangements would also facilitate the creation of a trusted third party data processing intermediary to handle linkage and deidentification challenges.

CER raises IRB and HIPAA concerns relatively uncommon in classic prospective research or in research with public-use data files. The ability of clinical data to answer the critical questions CER address, however, makes the extra effort worthwhile.

REFERENCES

1. Marsolo K. Approaches to facilitate Institutional Review Board approval of multi-center research studies. *Med Care.* 2012; 50(suppl 1):S77–S81.
2. Kushida C, Nichols D, Jadrnicek R, et al. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care.* 2012; 50(suppl 1):S82–S101.

3. HHS 2011. Human subjects research protections: enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators. Available at: http://www.gpo.gov/fdsys/pkg/FR-2011-07-26/pdf/2011-18792.pdf. Accessed April 23, 2012.
4. HHS 2008. *Letter Correspondence to Peter Pronovost*. Available at: http://www.hhs.gov/ohrp/policy/Correspondence/pronovost20080730letter.pdf. Accessed April 23, 2012.
5. Sweeney L. *Computational Disclosure Control: A Primer on Data Privacy Protection*. Cambridge: Massachusetts Institute of Technology; 2001.
6. Hirschman H, Aberdeen J. Measuring risk and information preservation: toward new metrics for de-identification of clinical texts. 2010. Available at: http://aclweb.org/anthology/W/W10/W10-1111.pdf. Accessed April 23, 2012.