Published in final edited form as:

Epidemiology. 2009 March ; 20(2): 169–171. doi:10.1097/EDE.0b013e318196ad0f.

Constraints on Data Sharing:

Experience From the Nurses' Health Study

Graham A. Colditz

From the Alvin J. Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO

The visibility of the Nurses' Health Study grew through the 1980s and 1990s¹ and a broad array of lifestyle and biomarker exposure data were accumulated.² With this visibility grew an interest among colleagues to combine Nurses' Health Study data with other studies, as exemplified by the Oxford Collaborative group. By 1996, interest in data sharing was made explicit during peer review for continued funding of the cohort. In response to the input from peer reviewers, we established an external advisory committee and formalized a data enclave to facilitate data sharing. The external advisory committee was charged with review of external requests for data access, which at that time were limited largely to questionnaire data. Guidelines for data access were approved by the Advisory Committee and made available to potential outside users of the cohort data. Subsequently the guidelines were posted to the Nurses' Health Study³ web site as guidelines for external collaborators. With this approach an ad hoc system was formalized that had already led to sharing of data with the Oxford Collaborative Group on hormonal factors in breast cancer.⁴

One early task of the external advisory committee was to balance the requests from outside users against the research plans of investigators within the team, which included future renewals of the cohort and its substudies. Such specific aims may be stronger with additional cases accrued through further follow-up before embarking on the evaluation of a hypothesis. In addition to the plans of faculty members, another concern was the training needs of doctoral students at the Harvard School of Public Health, many of whom were funded through National Institutes of Health (NIH) training grants (such as T32 CA09001). Over 70 doctoral theses have drawn on the data resources of the Nurses' Health Study; many have included a methodologic paper from the cohort exposure assessment within their thesis. As noted by Dr. Samet,⁵ this balancing of the needs and priorities of original investigators and data analysts remains a challenge for the long-term viability and scientific integrity of cohort studies, and more generally for epidemiology.

By 2003, the NIH had implemented a policy requiring data sharing requirements for grants over \$500,000 direct costs per year.⁶ As a data enclave was 1 of the options that met this NIH requirement, the NHS Advisory Committee endorsed this approach at its meeting in 2003. By 2005, over 50 requests had been received from outside users. Each was reviewed for feasibility given existing data resources. Over time, access to questionnaire data was supplemented by use of biomarkers in pooling studies,⁷ in addition to the creation of subsets of cohort data for teaching purposes. Investigators have included National Cancer Institute (NCI)-funded initiatives such as CISNET (Cancer Intervention and Surveillance Modeling Network)⁸ and others⁹ including some who have submitted ROIs or completed sabbatical stays in Boston.¹⁰ Dietary data have also been pooled from many prospective cohort studies through ongoing data sharing.¹¹

Among the long-term challenges of implementing this data sharing plan was the expectation by NIH that the cohort study would provide the infrastructure necessary to support the data sharing system through ongoing funding. The 2005 renewal of the Nurses' Health Study

cohort funding for continued follow-up and confirmation of incident cancers, added new challenges. The NCI mandated a broader data sharing plan before funding approval. This raised the potential for conflict between the NCI mandate of greater access than could be achieved through an enclave, and existing implementation of NIH policies by other Harvard-based cohorts (that were approved under the same Partners Health Care, Institutional Review Board). Dr. O'Rourke (Director of Human Research Affairs, Partners Health Care System) was unable to resolve NCI versus NIH policy demands, though discussions with NCI were ultimately successful in retaining the data enclave as a suitable approach to data sharing.

Although the NIH policy states that "privacy must be protected at all times" NCI expressed interest in release of limited data sets, perhaps in a manner similar to the sharing of data sets implemented by the National Heart, Lung, and Blood Institute and exemplified by the Framingham data.¹² Data sharing must be free from identifiers that could lead to identification of individual research participants and deductive disclosure of the identity of individual subjects. Profession-based cohorts raise additional concerns regarding this deductive disclosure in that they are based on a limited universe of participants. Cohorts based on race, such as the Black Women's Health Study,¹³ may also have unique considerations regarding public access.

To evaluate how easily a participant might be identified with widespread access to questionnaire data, we considered the 2-year follow-up interval during which individual cases of breast cancer are diagnosed, age in 5 years, smoking history (yes/no), BMI (3 categories, <25 kg/m², 25–29.9, 30+), and history of childbirth (yes/no). Cross-classifying these variables produced 6 cells with only 1 case. Thus, inclusion of approximate date of diagnosis (within a 2-year interval) in any public access data file looked to increase greatly the likelihood of deductive disclosure. Furthermore, breast cancer was the most frequent cancer diagnosis. How would one set a lower limit on the number of cases diagnosed in a time period to allow easy public access? How would analysis completed on the public access data set without date of diagnosis approximate analysis done by primary investigators who used time varying covariates and time to outcome?

We note that to overcome this type of concern, the National Health and Examination Survey (NHANES) has a hierarchical approach to data access such that a data enclave is implemented for sensitive data.¹⁴ Requested SAS runs are submitted by e-mail and only printouts of tables are sent to registered email addresses. No SAS proc list, proc print etc., are allowed. This implementation clearly has substantial administrative costs.

Other issues considered in response to the initial request from NCI for greater access to data included discussions of our Institutional Review Board with study investigators on the possibility of reconsenting all participants for data sharing and disclosure. Management of such a reconsenting process, with the expectation of less-than-complete response, was a substantial concern. It would lead to 2 different data sets; a smaller reconsented cohort for publicly accessible data and a full initial cohort for internal analysis by primary cohort investigators. Accordingly, without funds for such an undertaking, we did not attempt to reconsent the whole cohort. Subsequently, reconsenting has been implemented for data sharing associated with genome-wide association studies, but this is far more manageable in the context of the nested case-control subsets identified from the cohort.

Balancing the tensions among priorities for investigators, for trainees, and for broader access to data remains a challenge for epidemiology. Additional concerns have been raised regarding data quality control on sharing and open access. As part of the management of the Nurses' Health Study cohort, extensive quality control systems are in place to check all

programming for consistency before manuscript submission. Checks of numbers in text and tables against analysis printouts are also routinely mandated before manuscript submission. Many of these quality checks and broader systems for quality control are not described in the methods of research papers but rather in grant applications. Perhaps Journals will need to modify the methods section to call greater attention to data quality control systems that are implemented, should broader access to secondary analysis of epidemiologic data become standard.

As Dr. Samet notes, the field of epidemiology, practicing epidemiologists, and science administrators need to come to a common understanding of the many issues raised by data sharing. We need to develop a culture of data sharing that can maintain the integrity of studies and the high quality of data exemplified by the ongoing cohorts.

References

1. Colditz GA, Winn DM. Criteria for the evaluation of large cohort studies: an application to the nurses' health study. *J Natl Cancer Inst.* 2008; 100:918–925. [PubMed: 18577745]
2. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer.* 2005; 5:388–396. [PubMed: 15864280]
3. Available at: <http://www.channing.harvard.edu/nhs/questionnaires/index.shtml>
4. Collaborative group on hormonal factors in breast cancer. Breast cancer and hormone replacement therapy. Combined reanalysis of data from 51 epidemiological studies involving 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet.* 1997; 350:1047–1059. [PubMed: 10213546]
5. Samet J. Data: to Share or Not to Share? *Epidemiology.* 2009; 20:172–174. [PubMed: 19234412]
6. Office of Extramural Research; US Department of Health & Human Services National Institutes of Health, editor. NIH Data Sharing Policy and Implementation Guidance. 2003; NOT-OD-03-032
7. The Endogenous Hormones and Breast Cancer Collaborative Group. Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. *J Natl Cancer Inst.* 2002; 94:606–616. [PubMed: 11959894]
8. Meza R, Hazelton WD, Colditz GA, Moolgavkar SH. Analysis of lung cancer incidence in the Nurses' Health and the health professionals' follow-up studies using a multistage carcinogenesis model. *Cancer Causes Control.* 2008; 19:317–328. [PubMed: 18058248]
9. Ritz B, Ascherio A, Checkoway H, et al. Pooled analysis of tobacco use and risk of Parkinson disease. *Arch Neurol.* 2007; 64:990–997. [PubMed: 17620489]
10. Bain C, Feskanich D, Speizer FE, et al. Lung cancer rates in men and women with comparable histories of smoking. *J Natl Cancer Inst.* 2004; 96:826–834. [PubMed: 15173266]
11. Available at: <http://www.hsph.harvard.edu/poolingproject/about.html>
12. Available at: http://www.nhlbi.nih.gov/resources/deca/policy_new.htm
13. Rosenberg L, Adams-Campbell L, Palmer JR. The Black Women's Health Study: a follow-up study for causes and preventions of illness. *J Am Med Womens Assoc.* 1995; 50:56–58. [PubMed: 7722208]
14. Lochner K, Hummer RA, Bartee S, Wheatcroft G, Cox C. The public-use National Health Interview Survey linked mortality files: methods of reidentification risk avoidance and comparative analysis. *Am J Epidemiol.* 2008; 168:336–344. [PubMed: 18503037]

Biography

GRAHAM COLDITZ is the Niess-Gain Professor, Department of Surgery, Washington University School of Medicine. He was an investigator on the Nurses' Health Study from 1982 serving as project director from 1986 and co-principal investigator and principal investigator from 1996–2006.