“Big Data” Versus “Big Brother”: On the Appropriate Use of Large-scale Data Collections in Pediatrics

AUTHOR: Janet Currie, PhD

Center for Health and Wellbeing, Woodrow Wilson School, Princeton University, Princeton, New Jersey

KEY WORDS
consent, risk, benefit, administrative records, vital statistics, natality, common rule, HIPAA

ABBREVIATIONS
HIPAA—Health Insurance Portability and Accountability Act
IOM—Institute of Medicine
IRB—institutional review board
PHI—protected health information

www.pediatrics.org/cgi/doi/10.1542/peds.2013-0252c

doi:10.1542/peds.2013-0252c

Accepted for publication Jan 23, 2013

Address correspondence to Janet Currie, PhD, Woodrow Wilson School of Public and International Affairs, Princeton University, 316 Wallace Hall, Princeton, NJ 08544. E-mail: jcurrie@princeton.edu

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275).

FINANCIAL DISCLOSURE: *The author has indicated she has no financial relationships relevant to this article to disclose.*

FUNDING: No external funding.

abstract

Discussions of “big data” in medicine often revolve around gene sequencing and biosamples. It is perhaps less recognized that administrative data in the form of vital records, hospital discharge abstracts, insurance claims, and other routinely collected data also offer the potential for using information from hundreds of thousands, if not millions, of people to answer important questions. However, the increasing ease with which such data may be used and reused has increased concerns about privacy and informed consent. Addressing these concerns without creating insurmountable barriers to the use of such data for research is essential if we are to avoid a “missed opportunity” in pediatrics research. *Pediatrics* 2013;131:S127–S132

There have lately been many discussions of so-called big data and the potential for using data sets with hundreds of thousands, if not millions, of people to answer important questions. In medicine, much of this discussion revolves around gene sequencing and the use of biosamples. It is perhaps less recognized that big data also exist in pediatrics in the form of vital records, hospital records, insurance claims, disease registries, and other administrative records.

These types of data are can be of great help in elucidating questions that more detailed and focused but less expansive examinations fail to answer. For example, large data sets can be used to uncover the effects of exposures that may have small effects on individuals but large cumulative effects on populations. They may also be used to identify subgroups in which there are effects (compared with clinical studies in which it may be difficult to recruit enough members of different subgroups to be able to identify significant differences) or to study rare conditions. Third, administrative records can be used to track individuals over time, and so are ideally suited to measuring the long-term impacts of health conditions or interventions. Although these advantages are well recognized, the use of big data also raises some well-recognized ethical concerns. The most common of these concerns focuses on the privacy of research subjects. Big data, by their very nature, allow researchers to learn many things about many people and some of those things may be personal, sensitive, or secret. Addressing these concerns without creating unnecessary barriers for research is essential to avoid a "missed opportunity" in pediatrics research.

EXAMPLES OF PEDIATRICS RESEARCH USING ADMINISTRATIVE DATA

Researchers have used vital statistics natality data (ie, information from birth certificates) to track birth outcomes for many years. It is through these data that we know, for example, that the incidence of low birth weight in the United States has been increasing, that teen motherhood is decreasing, and that cesarean section rates have been rising. Given their key role in public health surveillance, these data were explicitly exempted from the Health Insurance Portability and Accountability Act (HIPAA) of 1996. Under the terms of the HIPAA, the keepers of protected health information (PHI) may disclose this information without individual authorization for the purposes of preparing vital records, such as birth and death certificates, or for the purposes of controlling disease, injury, or disability or for public health surveillance.[1]

A few state health departments take the view that, with proper safeguards, their public health mandate is liberal enough to permit academic researchers to access vital statistics data, even with identifying information. This access has allowed researchers to answer important questions about infant health that would be difficult to answer in any other way. For example, given information about the mother's name, birth date, birth place, and/or social security number, it is often possible to link siblings. Comparing siblings can help to control for many possible confounders. In a recent article, Ludwig and Currie[2] used this design and revealed that in a sample of >500 000 women and 1.1 million offspring, the odds of giving birth to an infant weighing >4000 g was 2.26 for women who gain >24 kg during pregnancy, relative to women who gain only 8 to 10 kg during pregnancy.

Such studies are just the tip of the big-data iceberg. Much more could be done if researchers were able to link health records to other administrative records. This sort of linkage is relatively routine in Scandinavian countries, although, to be sure, these countries also have significant safeguards in place. One important study linked data from birth records for all of the children born in Norway between 1967 and 1981 to information about their schooling attainment, labor market participation, and earnings.[3] For men, it was also possible to access IQ scores from military records. The study found that each 10% increase in birth weight was associated with a 1% increase in the probability of finishing high school and with slightly smaller positive effects on earnings and IQ scores. These results were true for the general population and were even true in comparisons of siblings, including twins.

Although it is difficult to do this sort of study in the United States, some enterprising researchers have been able to gain permission to construct such linkages, to dramatic effect. For example, a team of researchers in Florida has linked birth records of a cohort of all of the children born in that state in 1996 and 1997 to assessments of school readiness and the need for special education services.[4] By using these data, they were able to show that the risk of developmental delay or disability was 36% higher among late preterm infants (those born between 34 and 36 weeks' gestation) than among those with gestations of between 37 and 41 weeks. This important result contributed to a change in medical thinking with regard to tocolosis for late preterm pregnancies and the advisability of scheduling early cesarean deliveries.

Similarly, a team of researchers in California has made important progress unraveling some of the mysteries that underlie increases in the number of reported cases of autism by linking state autism registries back to birth data. Given that birth data with identifiers allow siblings to be identified, it is possible to use these data to compare children with autism with their own siblings who do not have the disorder. The birth data also have a variety of demographic information, including maternal and paternal age at the time of the birth. With the use of data on almost 5 million births, including those of 18 731 children with diagnosed autism and their siblings, the team was able to compare the relative contribution of maternal and paternal age to autism diagnoses.[5] Unlike previous researchers, they found that in each birth cohort between 1992 and 2000, older maternal age has a stronger effect on autism than older paternal age. They showed that previous studies' practice of pooling birth cohorts together (because of the limited sample sizes) resulted in an exaggerated correlation between paternal age and autism, because advances in diagnosis of the condition occurred during a time in which paternal ages were rising.

Some studies have also linked birth records to features of specific locations. For instance, 1 study took advantage of the implementation of electronic toll collections (E-ZPass; E-ZPass Group, Wilmington, DE) on highways in New Jersey and Pennsylvania to assess the effect of pollution due to traffic congestion on infant health.[6] E-ZPass greatly reduced idling and emissions around toll plazas. The researchers compared all infants born to women near toll plazas before and after E-ZPass implementation with all women located along the same highways but farther away from toll plazas serving as the controls. They were able to show that the implementation of E-ZPass reduced the incidence of low birth weight and prematurity by 8% to 10% in the vicinity of the toll plazas, a result that also sheds light on the toll pollution takes in terms of infant health.

LIMITATIONS ON DATA ACCESS

Given the examples above, a reader might be tempted to conclude that current norms of data access for projects involving big data in pediatrics are sufficient and that the way forward is unobstructed. However, this view is too rosy for several reasons.

First, only a handful of states currently have procedures in place to allow access to individual vital statistics natality data with identifiers of any kind. Second, the number of such states is falling rather than rising. For example, the state of Texas, which until recently had a well-defined process for regulating and allowing access to its natality data, has recently ended access for most classes of academic researchers.

Third, although there are certainly many important questions that can be investigated regarding newborns under current norms, the real "low-hanging fruit" of big-data–driven research will not be harvested unless researchers can get access to datasets that are not available to them under our current system of research regulations. For example, suppose it were possible to link vital statistics natality data to hospital discharge data and data on emergency department visits to follow a child from birth onward. One could use such data to investigate questions that have been addressed with the use of birth cohort studies in other countries, such as the relationship between birth weight and childhood asthma, the relationship between assisted reproductive technology and child health, and the long-term effects of maternal smoking during pregnancy. Arguably, one can obtain a more definitive answer by using administrative data than those in a birth cohort study because the latter may be subject to bias due to attrition from the sample. One recent study benchmarked estimates from the Danish Birth Cohort Study against estimates obtained by using administrative data and found that the estimated relationship between maternal smoking during pregnancy and attention-deficit/hyperactivity disorder in the child was biased upward by 33% in the birth cohort study.[7]

New research outside of pediatrics indicates the sorts of questions that could be answered given the ability to link data across administrative data systems. For example, the state of Oregon recently expanded Medicaid coverage to uninsured adults via a lottery. The state has teamed with researchers from Harvard, the Massachusetts Institute of Technology, and other institutions to link those who "won" and "lost" the Medicaid lottery to hospital claims data as well as data on credit scores from Experian. By using this extraordinary data set, the researchers revealed that randomly gaining Medicaid coverage increased utilization of care and costs, but reduced the probability that winners had unpaid medical debt.[8] Another, remarkable, finding was that coverage had no significant effect on the usage of emergency rooms. The study has obvious and important implications for what we may expect from implementation of the Medicaid expansion portion of the Affordable Care Act.

WEIGHING BENEFITS AND RISKS

The use of administrative records for pediatric research raises 2 sets of ethical issues. The first has to do with weighing the benefits of the research against the risks to subjects. The preceding discussion has sought to illustrate the potential benefits by discussing some of the striking results that have been obtained using these kinds of data. Although these kinds of investigations are essentially descriptive, description is an important part of science. Establishing new facts, such as the deficits suffered by late preterm infants, is a first step toward improving health.

The benefits of this type of research generally accrue to society rather than to individual subjects, and it may be possible to reach subjects only through broader public dissemination of research results. Therefore, state institutional review boards (IRBs) generally and appropriately demand that researchers describe the way that results will be disseminated to the public.

What, then, are the risks? There are no physical risks to subjects, so the main risk is that sensitive medical information could be disclosed. One can think of many examples of sensitive data, including treatment of sexually transmitted diseases, abortion services, or chronic conditions, that individuals might not want to disclose to potential employers. States that allow access to administrative health data generally require their own IRB review in addition to the approval of the principal investigator's own IRB.

Many different approaches have been taken to minimize the risk of disclosure. Some states do any matching in their own offices and produce anonymized data for researchers. Others mask key variables or add small amounts of "noise" so that individuals cannot be identified. Researchers are generally required to suppress small cells in any published data to protect individuals with rare characteristics or conditions.

De-identified data are generally not subject to the common rule that has governed human subjects research for several decades, because data without identifiers are not considered to pertain to human subjects. However, in the summer of 2011, the US Department of Health and Human Services invited public comment on proposals to revise the rule in view of many concerns about the functioning of the IRB system, including the fact that de-identified data can potentially be re-identified.[9] This concern is particularly true of biosamples, because if genetic information can be extracted, then it is in theory possible to identify specific individuals in large data sets even if the researcher does not have access to conventional identifiers such as names and addresses. That is, technology has advanced to the point that biosamples are inherently protected health information (PHI).

The same is not true of administrative data. It is possible, in principle, to mask or suppress small cells so that no individual can be uniquely identified by any combination of the information that remains in a data set. One part of the Department of Health and Human Services proposal is that the HIPAA Privacy Rule be applied to define de-identified data. Whereas the HIPAA rule allows a qualified expert to make a judgment about whether data are sufficiently de-identified, in practice most HIPAA-covered entities have taken the law's second approach to de-identification by stripping data of 18 categories of information, which include geographic identifiers smaller than the state name as well as dates of service.

Many entities, including the Institute of Medicine (IOM), have concluded that the HIPAA's requirements often seriously impede research without adequately protecting patients' privacy.[10] The IOM recommends replacing HIPAA with an oversight regimen that focuses on desired outcomes rather than on prescriptive regulation, that mandates strict data protection standards, that involves legal penalties for re-identification of data, and that applies to all users of health data rather than only to HIPAA-covered entities.

As ethical researchers, we are required to weigh the benefits of proposed research against the risks. An important aspect of this comparison that is generally neglected has to do with the social cost of conducting different forms of research. As the IOM points out, good stewardship demands that we use health information efficiently to advance the public good. Much administrative health data are collected and maintained by government, and much health research is ultimately paid for by government. In cases in which it would be much more cost-effective to use existing administrative data to answer a research question than to undertake a de novo data collection, we need to ask whether it is an irresponsible use of public funds to do the latter rather than the former. Similarly, if access to existing data could answer the question at hand, it is perhaps immoral to needlessly subject a new set of individuals to risk.

THE QUESTION OF CONSENT

A second and even thornier set of ethical issues surrounds the question of consent. Most research that uses large administrative data sets would be impossible if it were necessary to "re-consent" each person. Even if it were possible to track each person down, often after many years, it would be prohibitively expensive to find millions of people. Research based only on samples of people who could be re-consented, would be likely to yield seriously biased results. IRBs often waive informed consent for this type of research on that very basis.

Yet, this may be the most controversial question facing researchers. A recent article in *Nature* highlighted concerns about the reuse of biosamples.[11] The issue is that under current rules, subjects who give consent for their samples to be used for 1 type of research may unwittingly have the same samples used for other research as long as the samples have been de-identified. Whereas such reuse is often valuable to science, patients are not given the opportunity to opt out of research they might object to. The article also discusses (and raises objections to) approaches such as asking subjects to give very broad consent for their data to be used in all types of research or using an opt-out (rather than what is de facto an opt-in) approach to consent.

On the other hand, the IOM report draws a clear distinction between "interventional research" and research that is "exclusively information based." It recommends that the latter type of research be allowed without individual consent as long as researchers have policies and procedures in place to protect data privacy, and as long as the research is done for clearly defined and approved purposes. The IOM envisages federal guidelines defining the policies and procedures that are adequate to protect data privacy, whereas local IRBs would presumably decide, as they do today, whether the benefits of the research justify a waiver of informed consent.

CONCLUSIONS

Pediatrics researchers have become more aware of the unique risks and possibilities of big data, primarily through exposure to genetic studies. However, many secrets are currently locked up in more prosaic and conventional sources of big data such as vital statistics data, hospital discharge records, insurance records (including Medicaid and Medicare claims data), registries for specific diseases, and so on. The possible uses of such data are multiplied when they can be linked together, or linked to other outside data (such as the Experian credit records in the Oregon study or early-childhood assessment data in Florida). But such linkages require the use of personally identifiable information.

To be more concrete about the possibilities, here are 5 examples of interesting questions that could be answered by using such linked data:

1. using linked birth records, hospital discharge data, and emergency department visit records, it would be possible to ask whether children born with the aid of assisted reproductive technology are more likely than other children (or their own siblings) to have subsequent health problems;
2. using birth records linked to educational records, it would be possible to determine whether children whose mothers smoked during pregnancy were more likely to have attention-deficit/hyperactivity disorder than siblings born when the mothers did not smoke;
3. using hospital records linked with education records, it would be possible to examine the impacts of head injuries on educational outcomes;
4. using data on hospital and emergency department visits for asthma over time linked to data from air pollution monitors, it would be possible to see whether new cases of asthma were more likely to develop in high pollution areas, as well as how children with asthma responded to variations in pollution levels; and
5. using birth data linked to data from autism registries and special education records, it would be possible to see when children who eventually end up in autism registries enter the special education system and what sort of diagnoses they receive.

These are just some of the many possibilities.

Efforts to ensure patient privacy in the era of big data should strive not to throw out the baby with the bathwater. Stripping data sets of all identifiers as required by 1 version of the HIPAA Privacy Rule would render them useless for many purposes and make linkages impossible. Despite the similarities between administrative databases and biosample data banks, there is a critical difference: whereas biosamples are intrinsically PHI, individual identifiers can be successfully masked in administrative data bases while retaining enough information to conduct research into important unanswered questions.

The possible uses of administrative big data are only starting to be explored. It would be a tragic waste if this promising line of inquiry was nipped in the bud by policies that failed to recognize its possibilities. Policies such as those suggested by the IOM offer a way forward. Scientific exploration of existing data is in all of our interests.

REFERENCES

1. Summary of the HIPAA privacy rule. 45 CFR §164.512. Available at: www.ecfr.gov/cgi-bin/text-idx?c=ecfr&SID=41e688f8c23a34cf965d781dc88b18fe&rgn=div8&view=text&node=45:1.0.1.3.79.5.27.8&idno=45. Accessed February 18, 2013
2. Ludwig DS, Currie J. The association between pregnancy weight gain and birthweight: a within-family comparison. *Lancet*. 2010;376(9745):984–990
3. Black S, Devereux P, Salvanes K. From the cradle to the labor market: the effect of birth weight on adult outcomes. Institute for the Study of Labor, Discussion Paper No. 1864, November 2005. Available at: http://ftp.iza.org/dp1864.pdf. Accessed February 18, 2013
4. Morse SB, Zheng H, Tang Y, Roth J. Early school-age outcomes of late preterm infants. *Pediatrics*. 2009;123(4). Available at: www.pediatrics.org/cgi/content/full/123/4/e622
5. King M, Fountain C, Dakhallah D, Bearman PS. Advancing paternal and maternal age are both important for autism risk.

Am J Public Health. 2009;99(9):1673–1679

6. National Bureau of Economic Research (NBER) Working Paper 15413. Available at: www.nber.org/papers/w15413.pdf. Accessed February 18, 2013
7. Greene N, Greenland S, Olsen J, Nohr EA. Estimating bias from loss to follow-up in the Danish National Birth Cohort. *Epidemiology*. 2011;22(6):815–822
8. Baicker K, Finkelstein A. The effects of Medicaid coverage—learning from the Oregon experiment. *N Engl J Med*. 2011;365(8):683–685
9. Department of Health and Human Services, Office of the Secretary. Human subjects research protections: enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators. *Fed Regist*. 2011;76(143):44512–44532. Available at: www.gpo.gov/fdsys/pkg/FR-2011-07-26/html/2011-18792.htm. Accessed February 18, 2013
10. Nass SJ, Levit LA, Gostin LO, eds. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Washington, DC: Institute of Medicine, National Academies Press; 2009
11. Hayden EC. Informed consent: a broken contract. *Nature*. 2012;486(7403):312–314