Published in final edited form as:

Curr Opin Allergy Clin Immunol. 2014 June ; 14(3): 181–185. doi:10.1097/ACI.0000000000000051.

Refining Gold from Existing Data

Ted D. Wade^a

^aNational Jewish Health, Denver, Colorado, USA

Abstract

Purpose of review—The purpose is to review the arguments for increasing use of existing data in health research.

Recent Findings—The reuse of data in observational, exploratory and outcome studies, as well as in confirming other findings, is being justified on epistemological grounds as the major path to new knowledge and to the generalizing of findings to “real world” populations. It is also justified on grounds of cost, power and efficiency, even though data reuse comes with real informatics, scientific culture, societal and regulatory issues.

Summary—Data reuse is becoming more compelling. When contemplating new research for any purpose other than RCTs for efficacy, it is prudent to consider whether partnership with existing data holders should be part of the optimal research plan.

Keywords

observational research; big data; data mining; data sharing

Introduction

The last decade produced the widely-held opinion that we are entering an era of overwhelming importance of data to the advancement of health research and practice [1]. The opinion grows largely from acceleration of [2]:

- data-producing processes such as electronic medical records (EMRs), and both the variety and quantity of biomolecular (aka “omics”) studies,
- our ability to store, distribute and compute on massive amounts of data, and
- the belief that we simply cannot waste this treasure trove of potential knowledge.

Fundamentally this trend is about the *reuse of existing* data and information [2,3,4]. The rising cost of, and decreasing resources available for, vital *experimental* research, such as randomized clinical trials (RCTs) are viewed as rate-limiting, or even slowing, the translational process. The National Research Council [4] strongly criticized the prevalent methods of clinical research, relying on expensive data, gathered once and discarded. They said this approach was underpowered, under-general, high-cost, and closed to data reuse.

Correspondence to: Ted D. Wade, PhD, Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, CO 80206-2761, USA, +1 303-398-1877, wadet@njhealth.org.

No conflicts of interest to report.

They further noted that it segregated caregivers and researchers, lacked long term follow-up, and gave inadequate direct feedback to clinical care.

The rate of progression from existing data into new knowledge has significant impediments. One set is factors which impede access to data, such as needs for preservation of confidentiality and autonomy of patients [5], or proprietary interests in data. Another large set of factors is a wide variety of unsolved informatics issues around data comparability [2,6**,7]. The reuse of data is generally thought of as observational research, as opposed to experimental research. This review will broadly characterize the nexus of: the need for data reuse, the availability of what is being called “big data”, observational research, and exploratory data methods.

The case for observational research - empiricism versus experimentalism

It is ideal to guide behavior (e.g., policy, treatment, institutional operations, etc.) as much as possible by use of tested, evolved theories rather than just generalizing from observations. Some recent philosophers of science even believe that empirically predicting events – i.e., inductive reasoning -- is way more dangerous than useful [8,9], because empirical predictions can fail catastrophically. When the Institute of Medicine issued its recent [10] report on transforming clinical research to address the translational gap, only one paragraph out of 129 pages was spent on the role of observational research -- as a generator of hypotheses to test with experimental methods. While they acknowledged that “Many consider the RCT [randomized controlled trial] to be unsustainable as an approach to addressing the large number of research questions that need to be answered because of the time and expense involved” [10, page 8], they still concluded “... registries do not provide the conclusive evidence necessary to change clinical practice.” [10, page 8] This is a strong contrast to the National Research Council’s position on the inadequacy of prevalent research methods, as noted above.

Principles of study design help us make valid experimental tests of theories. But the principles of sampling statistics that we use to analyze the studies also tell us that statistical inference only applies to the population that was actually sampled. The populations in RCTs are necessarily narrowly defined to control various sources of error variance. The resulting restriction of the sampled population is what underlies the current assertions [11,12] that observational studies are better than RCTs in telling us about the “real world”.

There is extensive support (reviewed in [13]) for viewing post hoc, i.e., un-planned, analyses of study data with caution. However, reuse of existing data cannot just suggest new hypotheses, but -- through non-experimental methods such as case-control and retrospective cohort designs -- can also replicate previous findings and allow testing of new ideas. In a major clarification of the roles of experimental and observational research, Vandenbrouke [14, page e67] said, “When the validity of observational research is doubted, it is usually not because of fear of chance events, but because of potential bias and confounding”.

Techniques for controlling bias, such as the use of propensity scores to control for “confounding by indication” are well-established [12]. Vandenbrouke also notes that the ills

of post-hoc subgroup analysis can be cured by replication, perhaps by doing the same analysis over multiple existing datasets.

Vandenbrouke noted that the prior odds of a hypothesis are something we are controlling in a planned experiment, where the null hypothesis and the experimental alternative are something like 50:50. In contrast, “fishing” for significance among, for example, a large number of SNP genotypes means that most findings may appear significant by chance. However, prior odds may also be high in a specific hypothesis being tested with existing data, because we know more from other sources, including previous post-hoc or exploratory work. Vandenbrouke also advocates for more exploratory studies because we need new discoveries, and because the loss function (cost of being wrong) in exploratory work is lower than in RCTs, where the wrong result can clearly lead to the harm of wrong treatment or policy.

Data Mining

The extreme of exploratory research is usually called data mining, and even, pejoratively, data dredging [14]. Data mining *in strictu sensu* is the hypothesis-free search for patterns in data, and it has its own array of quantitative methods and logic that are distinct from traditional statistical ones [15]. When research simply needs to predict future behavior, such as clinical resource usage, by induction from data on past behavior, data mining methods can guide one to the best predictions. These predictions may be helpful, e.g. in running a healthcare business, or they may fail badly because they don’t reflect any true understanding of what is being predicted [8].

Still, induction and reuse of existing data has its strong adherents today because there have been successes in solving otherwise intractable problems in intelligent computation, such as automated language translation or vehicle driving, by using data mining/induction on gigantic data sets. Proponents of so-called “big data” [16**] even claim that “correlation” of past events (i.e., induction) is rapidly becoming the most efficient and popular way to solve problems and make decisions, because of the availability of unprecedented volumes of data. And in so doing, they say, it is reducing the role of experiments, theories, hypothesis-testing and reliance on causal explanations. A much-publicized example in health – that of disease surveillance by using search engine query data [17] – probably means instead that some problems currently lack powerful theories, and so empirical methods can be highly useful.

The recent review by Yoo et al [15] found the only successful uses of strict data mining techniques in biomedicine were to make new or better empirical prediction or decision models, or to reduce the dimensionality of large datasets, such as gene expression data. They did not find dramatic breakthroughs, as one might hope for, in understanding phenomena or suggesting important hypotheses to test.

Nevertheless, approaches that mix data mining logic with observational research designs, or with methods for controlling data biases, are starting to make progress in finding potentially useful hypotheses in large data sets, and replicating previous findings. These can often be thought of as the elucidation of knowledge networks from large databases, so that very large amounts of data become more usable to others in a variety of ways. Some of these

applications of data mining are quite novel in concept, essentially offering new paradigms for data reuse.

At Mt Sinai a major investment is being made to mine clinical data for patterns that predict re-admission risk, to target interventions on individual patients while they are out of the hospital [18*]. The goal is to eventually use risk models, derived from clinical and biological data on all patients, to guide treatment of individual patients.

Some studies illustrate great ingenuity in mining existing -- sometimes public-- data in building knowledge networks. A study from Vanderbilt [19] used its biobank-linked clinical database, BioVU [20] to pioneer the concept of phenome-wide association scans (PheWAS). It used ICD9 codes to identify cases and controls populations for 776 diseases, then looked at their associations with 5 important SNPs. It was able to confirm several previous SNP-disease associations and find some new potential associations.

A Stanford study [21] used correlations of the adverse effects across drugs and indications to find those effects that were unlikely to be un-confounded. Similarities of “side effects” were also related to protein targets, which could lead to new targets for existing drugs. A large set of drug class interactions were found, including a new one between thiazides and SSRIs. The findings and two derived intermediate databases were public results of this study. An earlier iteration [22] of this methodology found a hitherto unknown adverse effect of pravastatin with paroxetine on hypercholesterolemia. The finding was discovered in an FDA adverse events database, then confirmed in its specificity with data from three separate hospital EMR systems. The same signal was subsequently detected in a focused search [23*] of massive logs from three internet search engines.

Another Stanford study [24] inferred a network of human disease similarities by correlating how their mRNA expression profiles showed their involvement across 4,620 functional protein network models. The resulting 138 significant associations between 53 different diseases were related to drugs and gene drug targets, yielding possible new uses for drugs, and pointing to a large complex of protein modules that might be involved in many diseases. Besides the basic individual-level Gene Expression Omnibus [25] data, the researchers used 14 other public databases in the study.

Problems of Data Sharing and Shareability

Analytic methodology is not the only issue in reuse of data. There are important problems in how to get data shared by others and in making it usable when combined with other data. An Institute of Medicine workshop on “sharing clinical research data” [6**] found a strong need and definite benefits to be gained from wider sharing of clinical study data. It also found legal, cultural and technical barriers. One participant likened journal articles to a mere synopsis of, or even advertisement for, the “immense quantities of data from which they are derived” [6**, page 4], nearly all of which remains inaccessible –sometimes, because of inadequate documentation and preservation, to even the original investigators. The workshop found sufficient examples of the value of pooling data from existing sources, but these efforts nearly always highlighted concomitant costs of the reuse of data. There were also cautions about a number of ways to make missteps and misinterpretations when using

data originally collected for other purposes. Standardization of data was said to be very helpful, but much more effective and efficient when applied pro-actively rather than retrospectively.

The workshop gave several examples of more or less successful, if expensive, projects based on the idea that data from control groups from different studies could be logically pooled to provide larger or reusable control samples. The most straightforward and successful was the ePlacebo project at Pfizer [26]. They used placebo control groups from 203 of their studies to obtain a well characterized set of about 20,000 patients that could be reused as controls in future studies. For example, the database can show the empirical variability of many different lab values in control populations. Internal comparability of the data was both helped by, and limited to, data from two corporate databases where consistency of data collection was sufficiently standardized. The need to make data comparable reduced the number of studies to 0.8% of the original large inventory.

The workshop mentioned another promising approach to building shared data sources, exemplified by the FDA's Mini-Sentinel system [27]. This is a data federation in which data are kept by their originating institution, but software allows investigators to query data from multiple institutions and return combined results from all those queries.

The IOM report [6**] concluded that success with data sharing would depend on better engagement of all the stakeholders:

- patients to appreciate the value of research and to allow and even encourage the reuse of their data;
- journals to encourage better documentation of the reproducibility of research, the provenance of data, and even its publication;
- investigators to balance career competitiveness with ways to incentivize collaboration,
- funders to be firmer in encouraging sharing, and
- regulators to make it easier to share data legally and safely.

Data sharing has been more successfully encouraged in omics-centered research, as exemplified by, among a number of others, the Gene Expression Omnibus [25] and dbGap [28] data collections. Nevertheless, an influential call for more sophisticated systems biological models [29] said that by far most data is unavailable, calling for a cultural change in sharing. Some of the authors of that paper founded a non-profit research organization, Sage Bionetworks, to promote both genomic data sharing and predictive modeling of disease phenotypes from the data [30*]. Sage has formed a number of partnerships to share data, tools and modeling. It has published a free software tool, Synapse, that was used in NIH's The Cancer Genome Atlas (TCGA) project. Synapse enabled 60 different research projects to share data from 1,930 data files on 6 different types of molecular data for 12 different cancer types [31*].

Conclusion

There is a growing realization that reuse of existing data is: philosophically sound, supported by advancing computational means, improving in research methodology, and economically necessary to the elimination of bottlenecks in the process of biomedical research at all translational levels.

Acknowledgments

This work was supported by National Jewish Health and the Colorado Clinical and Translational Science Institute, grant UL1RR025780 from NIH.

References

1. Milgrom H, Tran ZV. The rise of health information technology. *Curr Opin Allergy Clin Immunol*. 2010; 10:178–180. [PubMed: 20386434]
2. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? *Ann Intern Med*. 2009; 151(5):359–360. [PubMed: 19638404]
3. Safran C, Bloomrosen M, Hammond WE, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *J Amer Med Inform Assoc*. 2007; 14:1–9. [PubMed: 17077452]
4. National Research Council. Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. National Research Council (US). National Academies Press; 2011.
5. Desmond-Hellmann S. Toward Precision Medicine: A New Social Contract? *Sci Transl Med*. 2012; 4:127ed3.
- 6**. Olson, S.; Downey, AS. Sharing Clinical Research Data: Workshop Summary. National Academies Press; 2013. Is good because (a) it provides detailed rationales for data sharing, and (b) it presents model solutions for the significant barriers to sharing. A large number of academic, government and industry experts participated, but this advantage is weakened by it being only a workshop summary, without the force of a strong policy statement. Another flaw: it should have referenced other systematic studies and proposed solutions, such as the knowledge network in reference [4]
7. Scheuerman RH, Milgrom H. Personalized care, comparative effectiveness research and the electronic health record. *Curr Opin Allergy Clin Immunol*. 2010; 10:168–170. [PubMed: 20431366]
8. Deutsch, D. The Beginning of Infinity: Explanations That Transform the World. New York: Viking Adult; 2011.
9. Taleb, NN. The Black Swan: the Impact of the Highly Improbable. 2. New York: Random House; 2010.
10. English, R.; Lebovitz, Y.; Griffin, R. Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary. National Academies Press; 2010.
11. Dreyer NA, Garner S. Registries for robust evidence. *JAMA*. 2009; 302(7):790–791. [PubMed: 19690313]
12. Gliklich, RE.; Dreyer, NA. Registries for Evaluating Patient Outcomes: a User's Guide. 2. Rockville, MD: US Dept. of Health and Human Services, Agency for Healthcare Research and Quality; 2010.
13. Everett D, Milgrom H. Post-hoc data analysis: benefits and limitations. *Curr Opin Allergy Clin Immunol*. 2013; 13:223–224. [PubMed: 23571411]
14. Vandembrouke JP. Observational research, randomized trials, and two views of medical science. *PLoS Med*. 2008; 5:339–343.
15. Yoo I, Alafaireet P, Marinov M, et al. Data Mining in Healthcare and Biomedicine: a Survey of the Literature. *J Med Sys*. 2012; 36:2431–2448.

- 16*. Mayer-Schönberger V, Cukier K. Big Data: A Revolution that Will Transform how We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt. 2013 Popular (Amazon sales rank 1,658), accessible book about big data by real experts. Gains credibility by clever explication of big data concepts and issues. But could do harm by too much emphasis on the demise of theory in favor of prediction by induction and correlation.
17. Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. *Nature*. 2008; 457(7232):1012–1014. [PubMed: 19020500]
- 18*. Humphries, C. [Accessed Oct 2 2013] A hospital takes its own big-data medicine. MIT Technology Review. Sep 26. 2013 <http://www.technologyreview.com/news/518916/a-hospital-takes-its-own-big-data-medicine/> Tells the technical public that a major AHC is investing hugely in mining of big clinical data for individualized care. Article avoids the obvious cost/benefit question, and it muddles the importance of the same institution's bet on genomics and multiscale biology research
19. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*. 2010; 26 (9):1205–1210. [PubMed: 20335276]
20. Pulley J, Clayton E, Bernard GR, et al. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci*. 2010; 3(1):42–48. [PubMed: 20443953]
21. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012; 4:125ra31.
22. Tatonetti NP, Denny JC, Murphy SN, et al. Detecting drug interactions from adverse-event pairs: interaction between Paroxetine and Pravastatin increases blood glucose levels. *Clin Pharmacol Ther*. 2011; 90(1):133–142. [PubMed: 21613990]
- 23*. White RW, Tatonetti NP, Shar NH, et al. Web-scale pharmacovigilance: listening to signals from the crowd. *J Am Med Inform Assoc*. 2013; 0:1–5. Shows another use (pharmacovigilance) of internet search logs analysis besides disease surveillance. It adds support for the Big Data (see ref [16]) hypothesis that data's original purposes do not predict their other future uses.
24. Suthram S, Dudley JT, Chiang AP, et al. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Comput Biol*. 2010; 6(2):e1000662. [PubMed: 20140234]
25. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res*. 2007; 35:D760–765. [PubMed: 17099226]
26. Desai JR, Bowen EA, Danielson MM, et al. Creation and implementation of a historical controls database from randomized clinical trials. *J Am Med Inform Assoc*. 2013; 20:e162–e168. [PubMed: 23449762]
27. Behrman RE, Benner JS, Brown JS, et al. Developing the Sentinel System—a national resource for evidence development. *New England Journal of Medicine*. 2011; 364(6):498–499. [PubMed: 21226658]
28. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007 Oct; 39(10):1181–1186. [PubMed: 17898773]
29. Butte AJ, Califano A, Friend S, et al. Integrative network-based association studies: leveraging cell regulatory models in the post-GWAS era. *Nature Precedings*. 2011 10.1038/npre.2011.5732.1
- 30*. Sage Bionetworks. [Accessed 9/30/2013] <http://www.sagebase.org/> Perusal of this site gives examples of the power of collaboration and openness, as promoted by charismatic champions
- 31*. Omberg L, Ellrott K, Yuan Y, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Analysis. *Nature Genet*. 2013; 45(10): 1121–1126. Deserves careful study to understand what made this massively successful collaboration. E.g., was it government sponsorship of the massive Cancer Genome Atlas project, or the data federating technology, or what? [PubMed: 24071850]

Key Points

- Calls to reuse existing data for health research are increasingly justified as the major path to new knowledge and as a vitally efficient means to confirm hypotheses.
- Hypothesis-driven use of existing data is justified epistemologically and is supported by sophisticated tactics to control bias.
- Exploratory mining of “big data”, popular outside of medicine because of successful empirical models and predictions, is also finding health-related uses.
- Sophisticated mixes of data mining and research methodology have produced shareable knowledge networks of both novel and confirmatory biomedical knowledge.
- There are persistent disincentives to share data and do the curatorial work to make it reusable.