ÜBERSICHTSARBEIT

Big Data in Medical Science—a Biostatistical View

Part 21 of a Series on Evaluation of Scientific Publications

Harald Binder, Maria Blettner

SUMMARY

Background: Inexpensive techniques for measurement and data storage now enable medical researchers to acquire far more data than can conveniently be analyzed by traditional methods. The expression "big data" refers to quantities on the order of magnitude of a terabyte (10^{12} bytes); special techniques must be used to evaluate such huge quantities of data in a scientifically meaningful way. Whether data sets of this size are useful and important is an open question that currently confronts medical science.

Methods: In this article, we give illustrative examples of the use of analytical techniques for big data and discuss them in the light of a selective literature review. We point out some critical aspects that should be considered to avoid errors when large amounts of data are analyzed.

Results: Machine learning techniques enable the recognition of potentially relevant patterns. When such techniques are used, certain additional steps should be taken that are unnecessary in more traditional analyses; for example, patient characteristics should be differentially weighted. If this is not done as a preliminary step before similarity detection, which is a component of many data analysis operations, characteristics such as age or sex will be weighted no higher than any one out of 10 000 gene expression values. Experience from the analysis of conventional observational data sets can be called upon to draw conclusions about potential causal effects from big data sets.

Conclusion: Big data techniques can be used, for example, to evaluate observational data derived from the routine care of entire populations, with clustering methods used to analyze therapeutically relevant patient subgroups. Such analyses can provide complementary information to clinical trials of the classic type. As big data analyses become more popular, various statistical techniques for causality analysis in observational data are becoming more widely available. This is likely to be of benefit to medical science, but specific adaptations will have to be made according to the requirements of the applications.

► **Cite this as:**
Binder H, Blettner M: Big data in medical science—a biostatistical view. Part 21 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2015; 112: 137–42. DOI: 10.3238/arztebl.2015.0137

Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) at the University Medical Center of the Johannes Gutenberg University Mainz: Prof. Dr. oec. publ. Harald Binder; Prof. Dr. rer. nat. Maria Blettner

Big data" is a universal buzzword in business and science, referring to the retrieval and handling of ever-growing amounts of information. It can be assumed, for example, that a typical hospital generates hundreds of terabytes (1 TB = 10^{12} bytes) of data annually in the course of patient care (1). For instance, exome sequencing, which results in 5 gigabytes (1 GB = 10^9 bytes) of data per patient, is on the way to becoming routine (2). The analysis of such enormous volumes of information, i.e., organization and description of the data and the drawing of (scientifically valid) conclusions, can already hardly be accomplished with the traditional tools of computer science and statistics. For example, examination of the exomes of several hundred patients requires sophisticated analytical approaches and the selection of statistical methods that optimize computation time to avoid exceeding the available storage capacity.

This is a challenge for the discipline of statistics, which has traditionally analyzed data not only from clinical studies but also from observational studies. Inter alia, techniques have to cope with a number of characteristics per individual that greatly exceeds the number of individuals observed, e.g., in the acquisition of 5 million single-nucleotide polymorphisms from each of a cohort of 100 patients.

In the following description of scenarios, techniques, and problems we focus on medical science, i.e., on the question of where and how big data approaches to the processing of large volumes of information can contribute to the advancement of scientific knowledge in medicine. While the description of the corresponding data analysis techniques takes a predominantly scientific perspective, the three scenarios preceding the discussion of techniques are intended to guide the reader in how these approaches can be used in handling routine data.

Because clinical studies are our reference point, applications that have little in common with the structure of such studies, e.g., the prediction of disease spread from search engine data *(Box)*, will not be discussed. Furthermore, concepts for technical implementation, e.g., cloud computing (5), will not be presented.

BOX

The debate about a big data showpiece: Google Flu Trends

In the Google Flu Trends project (3), the frequency of Google searches for certain terms is used to predict the influenza activity at regional level in a large number of countries. The original publication (3) shows that this method enables precise prediction of data that have traditionally been acquired in much more cumbersome fashion, e.g., by the United States Centers for Disease Control and Prevention (CDC), and did not use to be available until some time later. The possibility of rapid reaction opened up by the Google approach is often cited as a successful application of big data. However, later investigations (4) showed serious systematic deviations from predicted values in the period covered by (3). These may have been caused by modification of the search engine algorithm for business reasons, i.e., to optimize the primary function, with resulting impairment of the secondary function of influenza prediction.

Instead, we focus on biostatistical aspects, such as the undistorted estimation of treatment effects, which represent a crucial precondition for progress in medical science (6).

Big data scenarios

Diagnosis on the basis of high-resolution measurements

The introduction of microarray methods enabled characterization of patients at several molecular levels simultaneously at the time of diagnosis, e.g., via single-nucleotide polymorphisms, DNA methylation, mRNAs, or microRNAs (7). These techniques yield several million items of information per patient. Statistical analysis of these data could potentially lead to identification of parameters that distinguish among various diseases or point to the most suitable treatment option.

New sequencing techniques (referred to as next-generation sequencing) offer higher resolution and increase the number of variables that can be examined (8). Where small numbers of patients are concerned, however, the volume of data remaining after preprocessing is no longer so large and no special approaches to data handling are required. For example, the gene expression data for 22 000 genes from 400 patients amounts to less than 1 GB and can be processed on a standard PC. The data on 5 million single-nucleotide polymorphisms in 400 patients have a volume of circa 100 GB and can be analyzed using the RAM of the kind of server typically available to a small scientific working group.

Genuine big-data challenges arise when, for instance, the raw data, or data at several molecular levels, from several thousand individuals have to be considered together. In such a case the data volume becomes an important factor in the choice of analytic strategy, because not all statistical procedures lend themselves equally well to large volumes of information. This applies not only to epidemiological cohorts but also to a diagnostic scenario in which a patient's data need to be compared with external sources. The Cancer Genome Atlas (TCGA), for example, offers data from several different molecular levels. Automated comparison is a challenge for computer science and statistics (10).

Continuous monitoring of healthy individuals

In the framework of the 100K project, healthy individuals first have their genome sequenced and are then examined several times a year for a number of years. On each occasion classic parameters of clinical chemistry, parts of the microbiome, and organ-specific proteins are determined and cardiac, respiratory, and sleep parameters are all recorded within a short period of time (11). A preliminary trial of this measurement program in 108 individuals, launched in 2014, has as one of its goals evaluation of the technical feasibility of the project and the potential applications of the data. The idea behind the 100K project is that relevant changes in the observed parameters may take place long before a disease is diagnosed, so that early initiation of continuous monitoring could permit timely corrective measures (12).

Thus measurement of a potentially large number of parameters is complicated by the dimension of time. In order to discern problematic developments in timely fashion, the data analysis must include an explicit search for temporal patterns in high-dimensional data.

The complexity of continuous monitoring is increased by interventions such as individual nutritional guidance. The complexity of the approaches required for estimation of the consequences of interventions is comparable to that of the tracking of the treatment process subsequent to diagnosis in a clinical context (13).

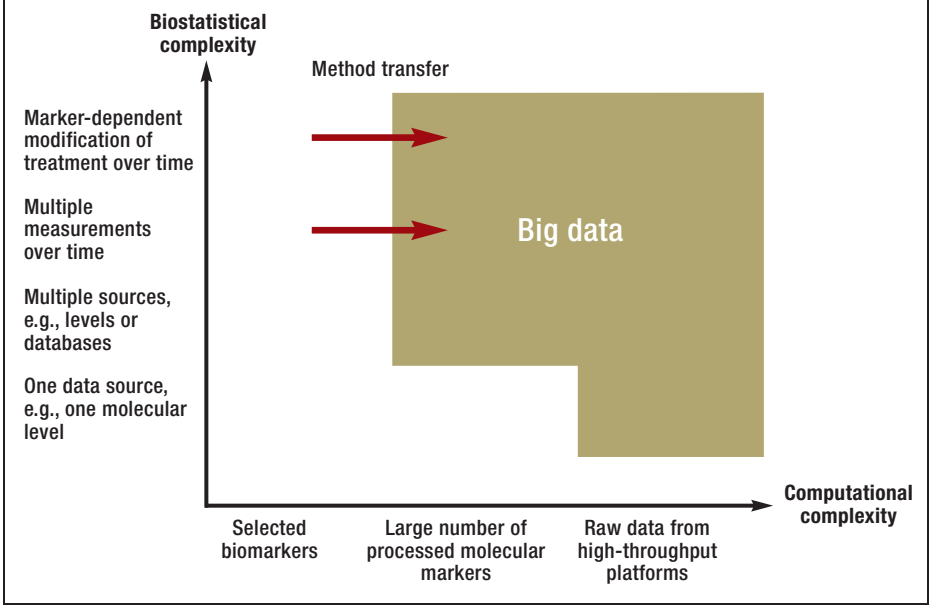
Prediction and treatment decisions

A third scenario is the monitoring of molecular characteristics in the course of treatment. For a certain number of biomarkers this is already routine practice in clinical registries. The registry for hepatocellular carcinoma in Mainz, Germany (14), for example, contains data from in some cases over a dozen time points for more than 1000 individuals. Thus, large amounts of information accumulate for each individual in the course of treatment.

Based on these data, measurements at a defined time point reveal the probability of future events, e.g., metastasis or death. It can be expected that in future clinical registries will include high-resolution molecular and/or imaging data, enabling cancer patients, for example, to be divided into groups with high and low risk of death, with implications for treatment (15). As in the diagnostic scenario, data from cohorts in the low hundreds of patients can, after preprocessing, be analyzed on standard PCs at one single time (16). The temporal dimension of the measurements considerably increases the volume and complexity of the data.

An additional challenge lies in the parallel treatment process. Treatment decisions are continuously made for

FIGURE



Computational and biostatistical complexity of various big data analyses/problems

each patient, based on the measured characteristics and influencing future measurements. Therefore, not only the temporal nature of the measurements but also the temporal pattern of treatment decisions have to be taken into account when, for instance, comparing patients with one another and determining the best treatment options. Precisely this combination forms the foundation of personalized medicine.

Figure 1 depicts the computational and biostatistical complexity of the various big data analyses in different scenarios.

Techniques

A characteristic feature of big data scenarios is that the accumulated data are hard to handle by means of conventional methods. The difficulty begins with the very first step of analysis, namely description of the data. With 10 potential biomarkers, for example, a table of mean values would typically be generated, but with 10 000 or more possible markers such a table is no longer useful. Particularly in big data applications, pattern recognition, i.e., the detection of relevant, potentially frequent patterns, has to be supported by machine learning techniques that automatically recognize patterns and can yield a dimension reduction or preselection (17).

So-called "unsupervised" machine learning techniques seek to detect, for instance, the frequent simultaneous occurrence of certain patient characteristics. One example is the "bump hunting" approach, which gradually refines the definition criteria of frequent groups of individuals (18). Alternatively, clustering approaches identify groups of similar patients (19). If, for example, identification of biomarkers that show similar patterns with reference to these patient groups is desired, biclustering procedures are available (20).

In contrast, "supervised" approaches have a particular target criterion, e.g., prediction of 1-year survival based on the gene expression profile of the tumor at the time of diagnosis. A crucial aspect of these procedures is the automated selection of a small number of patient characteristics or, for example, gene expression parameters, that are suitable for prediction. Another important distinction lies in whether and to what extent the respective approaches are based on a statistical model, i.e., a mathematically explicitly specified form of connection between the observed parameters. Model-based approaches stem from classical statistics (see [21] for extensions of regression models), while model-free approaches are often rooted in computer science (22). Prominent model-based approaches are regularized regression procedures (23) and logic regression (24). Well-known model-free approaches include random forests (22) and support vector machines (25).

Model-based approaches bear a greater resemblance to the statistical methods used in clinical studies. However, while clinical studies are designed to quantify the influence of a parameter—typically the effect of a treatment—precisely, i.e., unbiased and with low variability, analysis of a large number of potential parameters, e.g., candidate biomarkers, comes at a price. Namely, important markers are identified, but their effects can no longer be estimated without distortion (26).

In model-based approaches the available data are summarized in the shape of an estimated model, on the

TABLE

Different classes of machine learning procedures with typical applications and examples of approaches*

	Model-free	Model-based
Unsupervised	Description, pattern recognition, e.g., bump hunting (18)	Distribution of (unknown) groups, e.g., mixture models (32)
Supervised	Prediction, e.g., random forests (22)	Prediction, identification of predictors, e.g. regularized regression (23)

* "Unsupervised" means searching for patterns without a quantifiable target criterion (e.g., prediction performance in relation to the survival status known from the data), while "supervised" means the presence of a target criterion

basis of which, for instance, predictions for future patients can be formulated. For model-free approaches this aggregation takes another form. In the random forest procedure, for example, a large number of decision trees (typically 500 or more) are formed, each from a slightly modified version of the data (27). For new patients a prediction, e.g., the probability of a fatal outcome, is generated from each of these trees and the predicted values are combined (typically by averaging). However, it is difficult to assess the influence of individual patient characteristics on the prediction (28). Model-free approaches are therefore better suited for prediction than for increasing understanding of the underlying process (27).

An extreme form of model-free approach uses the data of all previously observed individuals directly, for example to make predictions for new patients. "Nearest neighbor" approaches, for instance, identify those individuals who are most similar to the new patients and predict clinical endpoints on the basis of the observations in these similar individuals (29). Based on this idea, "case-based reasoning" approaches (30) intuitively correspond to the way a physician might proceed on the basis of experience with previous patients. A further variant consists in developing prediction models for groups of similar individuals (31). The *Table* shows an overview of the various approaches with examples of the techniques and the typical applications.

Particularly with large amounts of data it is important to distinguish whether there is an aggregation (e.g., on the basis of a model) or whether the data of all individuals have to be employed, e.g., to make predictions for new cases. Permanent access to large volumes of patient data, possibly distributed between different sites, is also problematic from the perspective of data protection (33). From the technical point of view, further problems arise when the collection of patient data is continually growing and thus repeatedly requires updating, e.g., for purposes of prediction. For this kind of learning from data streams, either adaptations can be carried out at regular intervals, e.g., by reestimation of a regression model, or specially modified procedures can be used to adapt the prediction model individual by individual (34).

Distinctive features of medical science

The approaches described in the foregoing section were often not designed with the specific demands of medicine in mind. This is particularly true with regard to:

- The different types of patient characteristics
- The time structure
- The treatment information.

Without special modification, machine learning routines, i.e., procedures that recognize patterns in automated fashion and yield a dimension reduction or preselection, treat all measurements or patient characteristics in the same way. For example, similarity determination, a component of many procedures, assigns no greater weight to characteristics such as age or sex than to any one of 20 000 measured gene expression values. Even just for optimization of prediction accuracy, however, it is advantageous to distinguish between clinical features and other characteristics, e.g., high-dimensional molecular measurements (35).

In continuous monitoring of individuals and when measurements during the course of treatment have to be considered, the potentially high dimension of the measured values is joined by the time structure as an additional dimension that has to be taken into account in data analysis (36). For example, the time of diagnosis is an important reference point if machine learning procedures are to be used to compare subsequent molecular measurements among patients or to determine similarities. The situation is further complicated by the different follow-up periods for different individuals. This corresponds to the censoring problems that are tackled in clinical studies by using procedures such as Kaplan–Meier estimation or Cox regression for examination of the end point of interest. Particularly machine learning procedures have to be specially adapted for such time structures. Simplifying reduction, e.g., to a binary end point despite censoring, can lead to severely biased results (21). Even without censoring an irregular grid of measurement time points, often dictated by clinical routine, may result in bias (37).

Finally, the treatment information and the time points of treatment decision and treatment change play an essential part in the search for patterns in potentially large volumes of data. In routine clinical practice the treatment decision is influenced by measured values, but in turn it will influence (future) measurements. For instance, if in such a constellation the effect of a treatment on survival is to be determined and is, for the sake of comparability among patients, viewed conditionally on a repeatedly measured laboratory parameter, typically via adjustment in a regression model, this adjustment can mask a part of the treatment effect, which, however, in turn, affects the laboratory parameter. This problem, which can lead to bias of estimated treatment effects in any direction, is generally termed "time-dependent confounding" (38).

For classical biostatistical analyses of observational data, approaches that can cope with censored data have been developed, along with procedures for joint consideration of continually measured data and a potentially censored clinical end point (39). There are also various approaches to dealing with the time-dependent confounding problem. While these approaches have so far rarely been combined with machine learning procedures, in principle there is no reason not to do so. Thus, for example, the sequential Cox approach for time-dependent confounding is based on transformed data, i.e. machine learning procedures can also be employed (40).

Discussion

The term "big data" embraces a wide variety of disciplines and applications and many different statistical and computational approaches.

Medical science applications have to take into account different types of patient characteristics, time structure, and treatment information. Although some machine learning approaches already satisfy these requirements and could be used for big data applications in this area, there remains considerable potential for the development of appropriate approaches for automated pattern recognition. Many of these yet to be developed approaches will probably also prove useful for applications whose position on the continuum of complexity is such that they are not yet viewed as big data problems. This may facilitate the utilization of all observational data, above all routine data. Although it is extremely difficult to obtain bias-free results by means of big data approaches, the latter promise at least to provide valuable complementary information for the benefit of medical science.

KEY MESSAGES

- The field of medical science is also confronted with big data problems, particularly where molecular measurements on multiple levels or routine data with continual monitoring are concerned.
- Automated pattern recognition, e.g., via clustering, can provide descriptive analysis for large volumes of data, the traditional first step of statistical analysis.
- It is a special feature of medical data that procedures for data analysis have to consider the weighting of individual patient characteristics, e.g., age and sex in relation to thousands of gene expression values.
- Analysis of causality from jointly acquired treatment data and molecular markers must take account of the temporal sequence, e.g., by adaptation of existing procedures for observational data.
- The increasing popularity of big data approaches is leading to wider availability of the corresponding data analysis techniques, with beneficial consequences for medical science.

Conflict of interest statement
The authors declare that no conflict of interest exists.

Manuscript received on 8 May 2014, revised version accepted on 18 November 2014.

Translated from the original German by David Roseveare.

REFERENCES

1. Sejdic E: Adapt current tools for handling big data (Correspondence). Nature 2014; 507: 306.
2. Tripathy D, Harnden K, Blackwell K, Robson M: Next generation sequencing and tumor mutation profiling: Are we ready for routine use in the oncology clinic? BMC Med 2014; 12: 140.
3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L: Detecting influenza epidemics using search engine query data. Nature 2009; 457: 1012–4.
4. Lazer D, Kennedy R, King G, Vespignani A: The parable of google flu: Traps in big data analysis. Science 2014; 343: 1203–5.
5. Marx V: The big challenges of big data. Nature 2013; 498: 255–60.
6. Chiolero A: Big data in epidemiology. Epidemiology 2013; 26: 938–9.
7. Cho YJJ, Tsherniak A, Tamayo P, et al.: Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J Clin Oncol 2011; 29: 1424–30.
8. Marioni J, Mason C, Mane S, Stephens M, Gilad Y: RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genom Res 2008; 18: 1509–17.
9. Huerta M, Munyi M, Expósito D, Querol E, Cedano J: MGDB: crossing the marker genes of a user microarray with a database of public-microarrays marker genes. Bioinformatics 2014; 30: 1780–1.
10. Robbins DE, Grüneberg A, Deus HF, Tanik MM, Almeida JS: A self-updating road map of the cancer genome atlas. Bioinformatics 2013; 29: 1333–40.
11. Hood L, Price ND: Demystifying disease, democratizing health care. Sci Transl Med 2014; 5: 225.
12. Hood L, Friend SH: Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011; 8: 184–7.
13. Gibbs WW: Medicine gets up close and personal. Nature 2014; 506: 144.
14. Weinmann A, Koch S, Niederle IM, Schulze-Bergkamen H, et al.: Trends in epidemiology, treatment and survival of hepatocellular carcinoma patients between 1998 and 2009: an analysis of 1066 cases of a German HCC registry. J Clin Gastroenterol 2014; 48: 279–89.
15. Simon R: Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005; 23: 7332–41.
16. Horn JDV, Toga AW: Human neuroimaging as a big data science. Brain Imaging Behav 2013; 2: 323–31.
17. James G, Witten D, Hastie T, Tibshirani R: An introduction to statistical learning. New York: Springer 2013.
18. Friedman JH, Fisher NI: Bump hunting in high-dimensional data. Stat Comput 1999; 9: 123–43.
19. Andreopoulos B, An A, Wang X, Schroeder M: A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform 2009; 10: 297–314.
20. Eren K, Deveci M, Küçüktunç O, Çatalyürek ÜVV: A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform 2013; 14: 279–92.
21. Binder H, Porzelius C, Schumacher M: An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. Biom J 2011; 53: 170–89.
22. Breiman L: Random Forests. Mach Learn 2001; 45: 5–32.
23. Witten DM, Tibshirani R: Survival analysis with high-dimensional covariates. Stat Methods Med Res 2010; 19: 29–51.

24. Ruczinski I, Kooperberg C, LeBlanc M: Logic Regression. J Comput Graph Stat 2003; 12: 475–511.
25. Evers L, Messow CM: Sparse kernel methods for high-dimensional survival data. Bioinformatics 2008; 24: 1632–8.
26. Porzelius C, Schumacher M, Binder H: Sparse regression techniques in low-dimensional survival settings. Stat Comput 2010; 20: 151–63.
27. Breiman L: Statistical modeling: The two cultures. Stat Sci 2001; 16: 199–231.
28. Boulesteix ALL, Janitza S, Kruppa J, König IR: Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2012; 2: 493–507.
29. Kruppa J, Liu Y, Biau G, et al.: Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. Biom J 2014; 56: 534–63.
30. Glez-Peña D, Díaz F, Hernández JM, Corchado JM, Fdez-Riverola F: geneCBR: a translational tool for multiple-microarray analysis and integrative information retrieval for aiding diagnosis in cancer research. BMC Bioinformatics 2009; 10: 187.
31. Binder H, Müller T, Schwender H, et al.: Cluster-localized sparse logistic regression for SNP data. Statl Appl Genet Mol 2012; 11: 4.
32. Reich BJ, Bondell HD: A spatial dirichlet process mixture model for clustering population genetics data. Biometrics 2010; 67: 381–90.
33. Toh S, Platt R: Is size the next big thing in epidemiology? Epidemiology 2013; 24: 349–51.
34. Gaber MM, Zaslavsky A, Krishnaswamy S: Mining data streams: a review. ACM Sigmod Record 2005; 34: 18–26.
35. Binder H, Schumacher M: Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics 2008; 9: 14.
36. Aalen Røysland O, Gran JM, Ledergerber B: Causality, mediation and time: a dynamic viewpoint. J R Stat Soc A 2012; 175: 831–61.
37. Andersen PK, Liest K: Attenuation caused by infrequently updated covariates in survival analysis. Biostatistics 2003; 4: 633–49.
38. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JAC: Methods for dealing with time-dependent confounding. Stat Med 2012; 32: 1584–618.
39. Ibrahim JG, Chu H, Chen LM: Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol 2010; 28: 2796–801.
40. Gran JM, Røysland K, Wolbers M, et al.: A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. Stat Med 2010; 29: 2757–68.

Corresponding author
Prof. Dr. oec. pub. Harald Binder
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI) der Universitätsmedizin der Johannes-Gutenberg-Universität Mainz
Obere Zahlbacher Straße 69, 55101 Mainz, Germany
binderh@uni-mainz.de