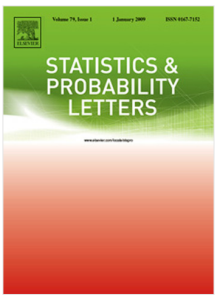
# Accepted Manuscript




The role of statistics in the era of big data: Electronic health records for healthcare research

Linda D. Sharples

PII: S0167-7152(18)30089-0
DOI: https://doi.org/10.1016/j.spl.2018.02.044
Reference: STAPRO 8166

To appear in: *Statistics and Probability Letters*

Please cite this article as: Sharples L.D., The role of statistics in the era of big data: Electronic health records for healthcare research. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.044

# The role of statistics in the era of big data: electronic health records for healthcare research

Linda D Sharples, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London WC1E 7HT; Linda.Sharples@lshtm.ac.uk

## Abstract

The transferring of medical records into huge electronic databases has opened up opportunities for research but requires attention to data quality, study design and issues of bias and confounding.

## Introduction

Traditionally, statistical analysis in healthcare was limited by the numbers of patients available and difficulty in combining and reconciling information from different sources. Huge investment in transferring medical notes into electronic health records (EHR) and provision of efficient software have largely overcome these limitations. New data sources offer opportunities to gain deeper insight into biological mechanisms, effectiveness of healthcare delivery and medical therapies, based on real-time availability and low-cost data. EHR can enhance quality and/or efficiency of evaluating quality of care, designing improvements, monitoring safety and effectiveness of therapies when applied to a population, and assessment of cost effectiveness. Studies of patient heterogeneity in treatment allocation and response may bring stratified medicine closer.

The reliability of statistical analysis of EHR depends on the type of database used and how it arose (e.g. administrative, disease registry). Many databases now have national coverage, are population-based, longitudinal and linkable to other databases. Others depend on some sampling process, detailed data collection is restricted to a short time interval, possibly surrounding an event or a health intervention, or both; these selection procedures must be taken into account in analysis and interpretation. Moreover, different data sources will have different levels of quality, consistency of recording and definitions, completeness, levels of detail and speed of accumulation. This paper discusses practical and statistical issues surrounding the use and analysis of routinely-collected EHR.

## Data access and quality

To protect patient confidentiality and data security, the body that collates, links and provides the dataset (data guardian) is usually separate from the researchers who analyse it. It is essential to work

with data guardians to understand the final dataset, assess risk of bias and ensure that it is produced on time and with minimum cost. Co-opting data guardian staff onto the project team gives them a stake in the work and fosters a culture of finding solutions to information governance issues.

Data quality is fundamental to any statistical analysis and is particularly important for EHR, in which selective sampling and imperfect recording may result in amplification of bias. The size and complexity of existing datasets means that post hoc identification of changes in data definitions can be difficult.

Some perceived data quality issues are highlighted in panel A; none are specific to EHR. However, with very large databases and the potential for multiple outcomes and (correlated) independent variables, commonly used methods for identifying and accommodating inaccurate data can be cumbersome. New methods for visualization to address this issue and guidance on automated quality assessment would be welcome. For databases that include patient-level longitudinal measurements, mixed-effects methods that iterate between fitting and outlier identification/removal have been investigated and applied to primary care data (Welch et al., 2014). The utility of these methods in other settings remains to be investigated, as does the transferability of methods developed for risk-based monitoring in multi-centre trials. (Hurley et al., 2016)

For data fields that are not recorded, it may be possible to impute missing data under an assumption of *missing at random*, conditional on observed fields.(Carpenter and Kenward, 2013) If data are not likely to be *missing at random,* it is important to work with data guardians to understand the sampling processes and investigate reasons for missing data, although with large datasets this may be computationally intensive.

In traditional cohort studies and clinical trials, the sampling process is usually known and a joint model for sampling and medical condition can be constructed. However, any dataset arising from EHR may be incomplete due to selective sampling, inability to fully link data sources or some other unknown systematic process. Whilst close collaboration with data guardians will provide insight into

linkage problems, other issues may remain unclear. The extent of bias that may be introduced due to incomplete sampling deserves further investigation.

## Data linkage

Linking of two or more datasets is a key issue. The ability to construct the complete patient care pathway would provide great potential to evaluate care quality and treatment decisions on short and long-term outcomes. However, data security is a concern, and data guardians for different datasets may be different. This implies a need for very clear specification of the linkage and associated data processing protocols.

Cleaning of datasets prior to linking is crucial to maximise the likelihood of finding matches. High sensitivity and specificity of database linking avoids some biases, although in practice, the trade-off between sensitivity and specificity will depend on specific applications. For example, large population-based EHR databases have potential for study of rare diseases, for which sensitivity of matching is extremely important, as loss of even a small number of cases would have a serious impact on power. In this case, highly sensitive (probabilistic) methods are attractive.

The two main types of data linkage in common use, deterministic (rule-based) and probabilistic (score-based) linkage, each with a number of versions, were developed over 50 years ago and have been described in detail. (Winkler, 2006, Dusetzina et al., 2014, Harron et al., 2015) Reporting of linking methods in journal submissions is now expected and an extension of the STROBE statement for observational studies that use linked databases (RECORD) is established. (Benchimol et al., 2015) The simplicity and efficiency of deterministic linkage is attractive for large databases, but may result in poor sensitivity, particularly if data quality is poor. (Zhu et al., 2015) The poorer performance may be amplified if more than two datasets are to be linked. Probabilistic linkage methods are often reported to have better performance, especially if data quality is poor, important identifiers are not available and/or matching on alphanumeric fields is required, but they have the added problem of choosing matching thresholds, and computation time may be prohibitive when several datasets have to be matched. (Zhu et al., 2015) Methods for improving the match accuracy in deterministic linking and strategies for increasing the speed of probabilistic linkage, would increase use of these methods

and may reduce bias, particularly for large datasets. However, the threshold for data matching will depend on the data quality.

Assessment of the quality of linkage is important. There are a number of methods available to assess this quality, which are reviewed elsewhere. (Harron et al., 2015)

A range of sophisticated linkage algorithms exist. For example, EM algorithms using matching weights, (Belin and Rubin, 1995) distance-based algorithms, (Herranz et al., 2015) and prior-informed imputation, (Harron et al., 2014) have all been investigated. Others have explored machine learning methods, which have potential for very large linkage problems (Elfeky et al., 2003) and further guidance on the value of these methods is needed.

### Analysis using observational study design

Although randomised controlled trials remain the gold standard for establishing causal relationships between interventions and outcomes, EHR, along with modern causal inference methods, can provide information when randomisation is not practical or ethical. EHR databases are often population-based, conferring strong external validity. On the other hand, naïve EHR analyses may exacerbate bias and confounding, which are common in observational research. Requirements for sound analysis and statistical themes to be re-assessed in this context are discussed below.

### The need for protocols

The way in which the dataset has been constructed is crucial; it should be made explicit and should be reflected in the statistical design and analysis of the study. Statisticians are concerned about how the population of interest is defined and sampled, how an intervention or exposure is defined, the definition of the impact of exposures and crucially, the structure of the model that links these components. These issues remain fundamental to analysis and interpretation. Apart from good registry design, a key issue is consistency of data definition across data-entry sites and through time, which relies on strong central management and co-ordination. Although these concerns may be straightforward to manage across similar sites, say primary care centres, when combined with secondary care or other databases, consistent definitions may be difficult to achieve. For example, diagnosis of an infection may be based on a serum marker in primary care but a tissue sample in

secondary care. Research into methods for detection and accommodation of conflicts between data sources would improve confidence in statistical analyses.

### Adjustment for confounding in observational studies

In recent years substantial contributions to the methodology for adjustment for bias due to confounding have been made, and include propensity scoring and similar methods, and instrumental variables approaches including Mendelian randomisation (Pearl, 2016, Stuart, 2010, Williamson et al., 2012, Smith and Ebrahim, 2003) Perhaps the most widely used methods in healthcare settings are propensity score methods. (Stuart, 2010) The conceptual simplicity of matching methods and their straightforward assumptions may explain their more frequent use in health research. However, the assumption that all important factors are included in the propensity score is often given scant attention. Whilst propensity score methods will be extremely useful if the datasets are largely complete and linkage error is low, they may be complicated if many covariates are required for the propensity score (potential for missing covariates) or more than two data sources are to be linked (potential for linkage error). There are strategies for high-dimensional propensity scores in a medical setting (Schneeweiss et al., 2009), and missing data methods are established. (Carpenter and Kenward, 2013) However, in EHR confounding and missing covariates must be tackled simultaneously, for which more methodological work is required. (Leyrat et al., 2017)

Other flexible, but more complicated, methods for dealing with confounding are available including G-computation, (Robins, 1986), inverse-probability-of-treatment (IPTW) estimators (Austin and Stuart, 2015), and doubly robust methods such as targeted maximum likelihood estimation (Targeted MLE). (Gruber and van der Laan, 2009) Targeted MLE is attractive in the context of EHR since it focuses on minimising bias for key parameters, treating all others as nuisance parameters. Such targeted inference, if it can be made accessible, may have a place in the analysis of EHR. More generally, a greater understanding of the utility, advantages and disadvantages of more complex causal methodologies is needed. Validation of sophisticated causal methods in this context is challenging. Where possible attention should focus on efficiency of implementation if the methods are to make a difference in practice.

## Multiple dimensions

The complexity of EHR opens up opportunities to investigate multiple dimensions of healthcare.

**Multi-level models:** Hierarchies (e.g. patients within hospitals within regions) within EHR introduce complicated correlation structures. Failure to recognise this may result in inaccuracy in both point estimates and measures of uncertainty, especially if statistical units vary in size. Statistical analysis of EHR requires exploratory methods to investigate clustering of statistical units, via hierarchical and cross-classified models, (Goldstein, 1987) and to identify latent constructs for further investigation.

**Multi-state models:** Patient monitoring through time for repeated events, possibly of different levels of severity is made possible by EHR. For example, multi-state models can be constructed to simultaneously predict survival, time to the next hospitalisation and total time spent in hospital, incorporating fixed and time varying covariates. (Ieva et al., 2017) Open questions pertain to the choice of sojourn time distribution (parametric or semi-parametric intensities), effects of linkage errors, the level and nature of missing records, assessment of model fit and computational issues. Although software packages for fitting interval censored (*msm*) and fully observed (*mstate, flexsurv*) data improve accessibility of methods, their utility for analysis of EHR and the accuracy of approximate methods need investigation. (Jackson, 2011, de Wreede et al., 2011, Jackson, 2016)

**Multivariate models:** The richness of EHR allows for multiple correlated variables to be investigated simultaneously. As the population ages, the number of people with multiple, possibly related, comorbidities will increase. These co-morbidities may be treated as vectors of response variables or correlated independent predictors, depending on the research question. EHR can provide important information on strategies for the joint management of multiple comorbidities. The size of EHR databases may also allow investigation of interactions between comorbidities. In this context it is important to understand the causal pathway. For example, when assessing the joint effects of smoking and COPD on lung cancer risk, one might wish to model smoking as an explanatory variable for COPD and both as explanatory variables for lung cancer. Care is needed in choosing the statistical model, including the covariance structure.

### Generalisation of clinical trial results

A major benefit of EHR is the ability to assess how interventions and exposures impact patients in a general health setting. Participants in RCTs are selected to a greater or lesser extent and may differ from the target population to be treated, so that expected treatment effects may differ. Using EHR and propensity matching methods, the treatment effect in the target population can be evaluated, so that prescribing practices can be optimised. (Stuart, 2011)

The regression discontinuity design (RDD) has been used where treatments are applied according to a fixed rule and results from RCTs are not consistently replicated in routine healthcare. (Geneletti et al., 2015) For example, statins are recommended when risk of a cardiovascular event exceeds a particular threshold. Patients who have a risk prediction close to the threshold are considered approximately comparable and those just above the threshold are compared with those below. (Geneletti et al., 2015) These designs applied to data from EHR can identify the extent of compliance with clinical guidance, reasons for non-compliance, and the treatment effect in the target population. Practical guidance is needed on application of these methods, the window within which patients are considered comparable and the estimation approach. In particular, for patients who are likely to be under-represented in trials, (e.g. older people, rare diseases), these methods provide an indication of the likely value of treatment.

### General evidence synthesis

General evidence synthesis describes the combination of data sources in a statistical model and has applications in, for example, health economics and epidemic modelling. In the UK, healthcare decisions are assessed by the National Institute of Health and Care Excellence, based on decision models that typically combine RCTs or meta-analyses of RCTs (robust treatment effects), observational studies (non-trial outcomes), disease registries (longer-term disease-related outcomes, stratification for baseline risk), hospital records (resource use) and population statistics (death from non-related causes) (NICE, 2013). The improvement in availability and quality of EHR has conferred an increase in quality of these analyses. However, results, and therefore policy decisions, are sensitive to the structure of the decision model, and it is difficult to give general advice on this structural uncertainty. (Jackson et al., 2011) Moreover, information on a parameter in the model can come from

more than one database, which sometimes results in conflicts and difficulty in model fitting. Statistical and computational methods that identify and accommodate conflicts are required.

### Randomised trials

Despite the benefits of EHR they are not likely to obviate the need for tightly controlled randomised studies. Of course RCTs have limitations: obtaining funding and regulatory approvals is challenging and time consuming; they often represent selected populations and unrealistic practice (limited external validity); difficulties in patient recruitment and retention are well documented; follow-up is short-term; data collection is limited; most compare a single experimental treatment with a control. (Olsen et al., 2013, Young, 2010). However, well designed and conducted RCTs provide evidence of *causal* effects of interventions. EHR may improve efficiency of design and analysis of RCTs.

In the UK, the National Institute for Health Research (NIHR) have issued calls for efficient trial designs that both replicate conditions in the UK National Health Service and maintain randomised treatment allocation. In this context, EHR can be used in design of trials, providing objective information on population characteristics, inclusion and exclusion criteria, recruitment rates and possibly attrition rates. Interrogation of existing databases would allow realistic estimation of numbers of available patients and of clinicians/treatment centres that might contribute cases. National mortality databases have long been used to define or verify survival outcomes for trials; EHR, especially disease registries, include a range of routinely-recorded outcomes (e.g. hospital admissions, myocardial infarction, cancer diagnosis). EHR may be used to assess the potential effect of an intervention, providing a more evidence-based sample size calculation. If a sufficiently mature and high quality database exists, trial outcomes may be wholly or partially available, thereby reducing resources required for patient follow-up. Further, EHR can enable health economists to capture all calls on health resources in the trial arms.

Several groups have advocated innovative trial designs embedding randomised trials within an existing large cohort. (Relton et al., 2010). For example, in the *cohort multiple RCT* a cohort is constructed, perhaps around an initial RCT, and patients are approached and consented to data follow-up. As new treatments emerge a sample of enrolled patients are selected using a random process and

offered the new treatment. Routinely collected outcome data are compared between selected and non-selected patients. This design retains random allocation, via random selection, thereby providing unbiased treatment effect estimates, minimises selection bias and retains the characteristics of the target population. It is possible to achieve high recruitment rates at lower cost than conventional RCTS, as demonstrated in the TASTE trial. (Frobert et al., 2010) Although the scope and success of such designs is yet to be established they are worth pursuing for a number of reasons. Importantly, they promote the implementation of successful new treatments since their utility in practice will have been established. Open research questions surround treatment of missing data, maintaining data quality, the reaction of patients in the control group to not being told they are in the trial and acceptability of the design to patients, clinicians, sponsors, funders and regulators.

## Concluding remarks

Investment in EHR has resulted in opportunities for insight into healthcare provision, epidemiology and evaluation of treatments. This paper gives a flavour of the issues that arise in EHR but is not a comprehensive review; areas such as data mining, supervised and unsupervised learning and sophisticated prediction methods, and a wide-ranging discussion of causal inference methods, have been left to others. The complexity of data access, linkage and database quality and design suggests that statistical analysis and study design, will be challenging.

**Panel A Data quality issues**

(i) **Inaccurate data records**. Since EHR are rarely constructed for research it is crucial to understand how the data fields have been defined and recorded. Data regularly and routinely change over time. Extreme records are easy to spot but other inaccuracies are not. Comparisons of summary statistics between different groups, examination of correlation structures and within-patient changes can highlight poor data quality. Consideration should be given to the life time (use-by date) of the database.

(ii) **Complete sampling but some data fields not recorded**. For some data fields, age or sex say, missing records are obvious. For other fields, say asthma attack, an empty field may mean that

no asthma attack took place, or it was simply not recorded. Bias and imprecision in estimates of asthma prevalence and treatment effects may result, and it is difficult to detect without external information.

(iii) **Incomplete sampling**. If a patient only seeks medical help when sick or if the data linking processes are inaccurate, the sampling process and the clinical situation under investigation will not be independent of each other. (Gruger et al., 1991, Zeng et al., 2015)

## References

AUSTIN, P. C. & STUART, E. A. 2015. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med,* 34**,** 3661-79.

BELIN, T. R. & RUBIN, D. B. 1995. A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association,* 90**,** 694-70.

BENCHIMOL, E. I., SMEETH, L., GUTTMANN, A., HARRON, K., MOHER, D., PETERSEN, I., SORENSEN, H. T., VON ELM, E., LANGAN, S. M. & COMMITTEE, R. W. 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med,* 12**,** e1001885.

CARPENTER, J. R. & KENWARD, M. G. 2013. *Multiple Imputation and Its Application (Statistics in Practice),* Chicester, Wiley.

DE WREEDE, L. C., FIOCCO, M. & PUTTER, H. 2011. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software,* 38.

DUSETZINA, S. B., TYREE, S., MEYER, A. M., MEYER, A., GREEN, L. & CARPENTER, W. R. 2014. Linking Data for Health Services Research: A Framework and Instructional Guide. *In:* QUALITY, A. F. H. R. A. (ed.). Chapel Hill, NC: The University of North Carolina at Chapel Hill.

ELFEKY, M. G., VERYKIOS, V. S., ELMAGARMID, A. K., GHANEM, T. M. & HUWAIT, A. R. 2003. Record Linkage: A Machine Learning Approach, A Toolbox, and a Digital Government Web Service *Purdue e-Pubs Computer Science Technical Reports.* USA: Purdue University.

FROBERT, O., LAGERQVIST, B., GUDNASON, T., THUESEN, L., SVENSSON, R., OLIVECRONA, G. K. & JAMES, S. K. 2010. Thrombus Aspiration in ST-Elevation myocardial infarction in Scandinavia (TASTE trial). A multicenter, prospective, randomized, controlled clinical registry trial based on the Swedish angiography and angioplasty registry (SCAAR) platform. Study design and rationale. *Am Heart J,* 160**,** 1042-8.

GENELETTI, S., O'KEEFFE, A. G., SHARPLES, L. D., RICHARDSON, S. & BAIO, G. 2015. Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine,* 34**,** 19.

GOLDSTEIN, H. 1987. *Multilevel Statistical Models*, Wiley

GRUBER, G. & VAN DER LAAN, M. J. 2009. Targeted Maximum Likelihood Estimation: A Gentle Introduction. *U.C. Berkeley Division of Biostatistics Working Paper Series*

Berkeley: University of California.

GRUGER, J., KAY, R. & SCHUMACHER, M. 1991. The validity of inferences based on incomplete observations in disease state models. *Biometrics,* 47**,** 595-605.

HARRON, K., GOLDSTEIN, H. & DIBBEN, C. 2015. *Methodological Developments in Data Linkage*, Wiley.

HERRANZ, J., NIN, J., RODRÍGUEZ, P. & TASSA, T. 2015. Revisiting distance-based record linkage for privacy-preserving release of statistical datasets. *Data and Knowledge Engineering,* 100.

HURLEY, C., SHIELY, F., POWER, J., CLARKE, M., EUSTACE, J. A., FLANAGAN, E. & KEARNEY, P. M. 2016. Risk based monitoring (RBM) tools for clinical trials: A systematic review. *Contemp Clin Trials,* 51**,** 15-27.

IEVA, F., JACKSON, C. H. & SHARPLES, L. D. 2017. Multi-state modelling of repeated hospitalisation and death in patients with heart failure: The use of large administrative databases in clinical epidemiology. *Stat Methods Med Res,* 26**,** 1350-1372.

JACKSON, C. H. 2011. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software* 38.

JACKSON, C. H. 2016. flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software,* 70.

JACKSON, C. H., BOJKE, L., THOMPSON, S. G., CLAXTON, K. & SHARPLES, L. D. 2011. A framework for addressing structural uncertainty in decision models. *Med Decis Making,* 31**,** 662-74.

LEYRAT, C., SEAMAN, S. R., WHITE, I. R., DOUGLAS, I., SMEETH, L., KIM, J., RESCHE-RIGON, M., CARPENTER, J. R. & WILLIAMSON, E. J. 2017. Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Stat Methods Med Res***,** 962280217713032.

NICE 2013. Guide to the methods of technologyGuide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence.

OLSEN, R., BELL, S., ORR, L. & STUART, E. A. 2013. External Validity in Policy Evaluations that Choose Sites Purposively. *Journal of Policy Analysis and Management,* 32**,** 15.

PEARL, J., GLYMOUR, M., JEWELL, N. P. 2016. *Causal Inference in Statistics: A Primer* Wiley.

RELTON, C., TORGERSON, D., O'CATHAIN, A. & NICHOLL, J. 2010. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. *BMJ,* 340**,** c1066.

ROBINS, G. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7**,** 1393-1512.

SCHNEEWEISS, S., RASSEN, J. A., GLYNN, R. J., AVORN, J., MOGUN, H. & BROOKHART, M. A. 2009. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology,* 20**,** 512-22.

SMITH, G. D. & EBRAHIM, S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol,* 32**,** 1-22.

STUART, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science,* 25**,** 21.

STUART, E. A., COLE, S.R., BRADSHAW, C.P., AND LEAF, P.J. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A.* , 174**,** 18.

WELCH, C. A., PETERSEN, I., BARTLETT, J. W., WHITE, I. R., MARSTON, L., MORRIS, R. W., NAZARETH, I., WALTERS, K. & CARPENTER, J. 2014. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Stat Med,* 33**,** 3725-37.

WILLIAMSON, E., MORLEY, R., LUCAS, A. & CARPENTER, J. 2012. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res,* 21**,** 273-93.

WINKLER, W. E. 2006. Overview of Record Linkage and Current Research Directions. *In:* STATISTICAL RESEARCH DIVISION, U. S. C. B. (ed.) *RESEARCH REPORT SERIES.* Washington, DC: Statistical Research Division, U.S. Census Bureau.

YOUNG, R. C. 2010. Cancer clinical trials--a chronic but curable crisis. *N Engl J Med,* 363**,** 306-9.

ZENG, L., COOK, R. J., WEN, L. & BORUVKA, A. 2015. Bias in progression-free survival analysis due to intermittent assessment of progression. *Stat Med,* 34**,** 3181-93.

ZHU, Y., MATSUYAMA, Y., OHASHI, Y. & SETOGUCHI, S. 2015. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *J Biomed Inform,* 56**,** 80-6.