Position Papers of the Federated Conference on Computer Science and Information Systems pp. 99–102
DOI: 10.15439/2018F378
ISSN 2300-5963 ACSIS, Vol. 16

Challenges in Causal Inference from Personal Monitoring Devices

Tomasz Wiktorski
Department of Electrical and Computer Science
University of Stavanger
Stavanger, Norway
Email: tomasz.wiktorski@uis.no

***Abstract*—Personal Monitoring Devices (PMDs) collect immense amount of data about health and wellness of hundreds of millions of people. One of the obstacles of the prevailing data analytics approaches to PMDs' data is limited value of correlation-based conclusions in a health context. Causal inference seems a natural solution, but general causal inference methodologies are difficult to apply to PMDs data due to size and complexity of observational data. Some methods, such as randomized trials, are largely infeasible in PMDs' context due to lack of control over the investigated population. In this paper, we overview existing approaches to causal inference including recent works that attempt to take advantage of time series data to automatically derive causality using extended difference-in-deference or Granger methods. We then outline challenges and opportunities for causal inference in the health context. Finally, we propose a following challenge: can we establish a new standard of evidence and a study design process that: (1) allows for drawing causal conclusions from large observational datasets and (2) can suggest interventions to enforce causal links discovered in the data.**

I. Introduction

OFTEN repeated phrase "correlation is not causation" became with time an offhand apology for providing only correlation-based conclusions for research questions, rather than motivating efforts to develop improved methods for determining causation. Even though, it might require a considerable extra effort to arrive at causal relations, the utility of causal results is far greater than correlation results, even if causal link would be of limited precision.

Many researchers and practitioners settle on correlation only, but this trend appears to change recently. Highly publicized cases, such as flops of Google Flu Trends[1], show dangers of ignoring proper causal analysis. If we want to plan and evaluate interventions or make safe long-term predictions, a tool stronger than correlation is necessary.

Pearl, originally known for development of Bayesian Networks and probability-centered approach to AI, suggests in his new book [2] that current developments in machine learning and artificial intelligence focus too much on improved curve fitting, that is correlation and probability. To lead to the next breakthrough machines have to reason beyond probability. He brings an example of malaria and fever. It is important to understand that malaria causes fever and not only that they are correlated. Introducing causality language into existing data analyses could help us reason about possible results of interventions using observational data.

Correlation prevails as for now in big data publications, as documented by Ekibia et al. review [3]. Other authors, e.g., George et al. [4] also notice that causality does not inform the design of big data as a domain. Grimmer [5] points out that large amounts of data will not solve the problem of variable selection in causal inference. Scientists require more training to better understand causal inference. With time, problems, such as those with Google Flu Trends, uncovered issues with such approach. Lazer et al. in their Science paper [6] suggest changes to approaches to data analysis based on the problems with Google Flu prediction. They show that analysis based on like Google Correlate is difficult to reproduce. Even Google Flu cannot be reproduced using Correlate. Analyses that base on sources produced by users are susceptible to blue team and red team problem. Blue team problem is the unintended influence on algorithm results due to changes to data generation as a result of changing company policies and goals. Red team problem is the unintended influence on algorithm results due to attack on data generation mechanism (generating fake data based on knowledge of algorithm functioning).

More and more authors stress importance of causal analysis for big data and data science. A good example is an article by Provost and Fawcett [7]. They stress, in particular, the importance of careful analysis of assumptions on confounding factors, causal links are directly dependent on these assumptions. Confounding factor is characterized by associations both with outcome and evaluated cause variable, but does not lie in the cause path between them. Kitchin [8] describes his reservations about the shift from knowledge-driven science to data-driven science. At the same time he notices benefits of big data and suggests that they could lead to creation of more sophisticated models.

After Introduction, in Section II we explain basics of causal inference including most common models and methods, such as Neyman-Rubin Causal Model, Structural Equation Modeling, Structural Causal Model, and Difference in Differences. In Section III we present current trends in causal inference in big data, healthcare, and time series leading to automation. We conclude in Section IV where we propose a challenges.

II. Basics of Causal Inference

Gelman [9] organizes causal reasoning in two general types of questions one can ask: forward and reverse. These two

questions are sometimes formulated in a shorthand form as "effects of causes vs. causes of effect", which traces back to Mill [10]. Forward question asks about results of a certain intervention one might perform. Reverse question asks about causes of a certain observed effect. The consensus based on Gelman's overview is that effects of causes can be usually traced, so forward question can be answered. Typical methods used to answer forward questions include: Neyman-Rubin causal model (RCM), Structural Equation Modeling (SEM), and Structural Causal Model (SCM).

Causes of effects are more difficult to deduce, especially if we are facing complex processes involving, e.g., politics, such as causes of war. Nevertheless, some reverse questions, even involving social matters, might still be answerable. It typically requires expert reasoning to pre-identify meaningful factors for actual analysis. The analysis would then account for these factors, which should lead to disappearance of differences in outcomes. This confirms causes of effects under assumption that factors are meaningful. This assumption naturally weakens such causal argument.

A. Model vs. Data

There are several general viewpoints on what is necessary to derive causal relation. These views are summarized and organized by Gelman. The main question that Gelman asks is how permissive a particular view is in allowing correlation and observational data as a basis for causal relations. Most strict approaches are based exclusively on strict models and randomized studies. Some scientists, such as Pearl, hold a view that observational data under strong models may lead to causal conclusions. In some domains it is also common to derive causality from covariance matrices, but it is unclear how widely it can be applied. The most permissive view covers automatic derivation of causality by computer from observational data. Gelman's organization can be seen as a scale that balances between model and data. All methods require data for confirmation of the actual relation, but few see data as a sole basis for such relation.

RCM is the method that often unconsciously guides people's thinking about causal inference. It introduced concept that are now taken for granted, such as: counterfactual, treatment, and control. Main methods that base on structure are SEM and SCM. They use both graphical and purely computational techniques to a different extent, particularly there are many different methods under SEM. Main method which uses minimal structure is DD (Difference in Differences), it also heavily focuses on use of data. It can be used as a basic calculation technique for some other methods that employ structural approach. An interesting example of a method with minimal expert structure is Google's CasualImpacts [11], which we mention in greater detail later in this paper. We will now shortly describe the four main methods: Neyman-Rubin causal model (RCM), structural equation modeling (SEM), structural causal model (SCM), and Difference in Differences (DD).

B. Neyman-Rubin Causal Model

The basis for Neyman-Rubin causal model was first formulated by Neyman [12] for randomized (alternatively called controlled) experiments and extended by Rubin [13] and others for both observational and experimental studies. This model introduced basic tools considered now a standard in any causal analysis. The model defines counterfactuals used to specify potential outcomes and then comparing outcomes of alternative exposures. In practice, it is usually impossible to check more than one intervention on the same subject. This problem is often named the fundamental problem of causal inference. The goal of randomized experiments is to solve this problem by creating two comparable groups that collectively can form a basis for causal reasoning.

C. Structural Equation Modeling

SEM is, as Kline [14] describes, a grouping of several methods used to verify a proposed causal model. Methods include path analysis, confirmatory factor analysis, structural regression models, latent growth models, covariance and correlation structure models. Sometimes path analysis is called causal modeling, which is considered an anachronism. Results of SEM cannot be assumed to be causal per se. Causality is dependent on the design of experiment which SEM methods then verify.

Confirmatory Factor Analysis is a form of Factor Analysis. It is used to compare researchers understanding of factors and their relation against the measurement. CFA is sometimes called measurement model in SEM and it does not specify structure. Factor analysis (Exploratory Factor Analysis, EFA) uses unobserved variables, called factors, to reduce amount of variables necessary to described variability in observed data. Principal Component Analysis is a simple version of EFA

Path analysis, is a method to specify directed dependencies between variables in the model. It only provides structure, no measurement. Latent growth modeling adds a time component to explanation of dependent variables, this way it can describe longitudinal change (change over time). Most applications limit themselves to slope and starting point only, but higher order methods are also used.

D. Structural Causal Model

SCM proposed by Pearl [15] attempts to integrate other existing approaches to causal inference into a unified model, in particular it subsumes SEM and various graphical models. It also claims to subsume RCM, but Aliprantis [16] indicates that an important difference in modeling approaches exist between these two methods. Even though, it is possible to find an SCM generating any RCM contingency tables, there is no unique SCM that generates a particular solution. It may result in critical discrepancies between each approach. However, this limitation does not necessarily take away all usability of SCM. It is still important because the integrative approach is helpful to clarify formal meaning of important concepts, such as: d-separation, interpretation of counterfactuals, and confounding.

E. Difference in Differences

Difference in differences method uses observational data to imitate experimental design. It calculates normal difference in

the outcome, using control group for normalization. Measurements need to be done at least once before and once after the intervention. The control group by assumption should: (1) show similar trends over time, (2) not be exposed to intervention enacted on or experienced by treatment group, (3) nothing other than treatment changes in only one group. These assumption can be collectively called a parallel trend assumption.

In the context of DD one might talk about natural experiment, which is non-randomized experiment. DD allows to estimate counterfactual from observational data from natural experiments. There preferably should be two groups with similar experience, but differently affected by an experiment or intervention. Intervention should be well planned, it should be exogenous that is not a reaction to a specific behavior, because it could cause people to change their behavior unpredictably to *game the system*. Resulting in inferior or even useless results.

III. Current Trends

A. Big Data

Emerging sources of data, often coming from sensors, are characterized by large volume (especially accumulative), sometimes high velocity, and wide variety of actual data sources and related data formats. These properties are commonly refereed as 3 Vs of big data. Moreover, such data that reflect various natural phenomena are often observational. By natural phenomena in this context we mean ones that actually occur due to normal human or other activity, in contrast to, e.g., simulations. While it is in theory possible to attempt randomized experimentation in such setting, it is unfortunately usually infeasible. Reasons are multiple, but major ones include: cost, ethical considerations, and general feasibility of recruitment to control groups.

Despite these obvious problems, new data sources provide important opportunities due to their granular and longitudinal character. Progress in sensor and storage technologies provides the possibility to capture data over long periods of time, large areas, and with ample frequency. These properties are not yet well explored in the existing research, with several exceptions which we mention here.

The most notable example might be work of Brodersen et al. [11] on CausalImpacts. The motivation of this approach was to infer causal impact for marketing campaigns, but it is not necessarily limited just to this one domain. In principle, it bases on difference in differences (DD) approach, but uses a state-space model of time series to predict the counterfactual. This is to compensate for many of the limitations of basic DD. It allows considering synthetic control, inclusion of Bayesian priors for parameters, and measuring evolution of impact of an intervention over time.

Three types of sources are used to construct the synthetic control. The first are properties of outcome time series before intervention. The second are properties of time series that could be used to predict outcome time series before intervention. The third source, if available, are parameter values from older related studies used as Bayesian priors.

Zigler and Dominici [17] propose a new Bayesian method to select Propensity Score variable, which is supposed to help causal inference in i.a. big data scenario. Beyond these examples, we have not found other explicit attempts to address data scale in causal analysis.

B. Time Series and Automation

Gelman in his aforementioned review suggests that time can be considered from two different non-exclusive perspectives in causal inference. First perspective considers contamination of results due to exposure to more than one treatment over time. Such contamination requires special attention in observational data, where exposure to more than one treatment is difficult to avoid and not always direct. For instance mere knowledge about alternative treatment might impact the results. Placebo method, which corrects for this issue, is difficult to deploy outside controlled randomized experiment setting.

Second perspective includes time as system variable to express changing effect of one treatment over time. Gelman only considers it for limited amount of future time points. Such limited treatment presumably results from combination of data available in traditional causal research and related methods. Some most resent developments address this limitation.

Granger causality test [18] has been more and more commonly used in the recent years. It provides well defined tool to determine if one time series is a predictor of another. In many cases this might be considered close to causality, but in principle it is a correlation with time precedence. The method relies on time series and allows for automation of the causality inference process. It defines determinate formulas eliminating (to some extent) necessity for human involvement, what makes it easier to scale with growing data, as long as we consider some form of windowing or sampling.

C. Health and Welfare Monitoring

Krumholz in his widely cited paper on "Big Data and New Knowledge in Medicine" [19] observes that improvements in any service (be it either health or movie rental) are possible despite lack of conceptual models, hypotheses testing, or randomized trials, but can come directly from observational data. Historically empirical insights derived from existing data were considered inferior to insights based on theory and experiments. However, many research questions can be answered based on observational data without understanding the underlining mechanism. In some case, these data might contribute then to understanding of the mechanism. Author provides an example of aspirin that was successfully used without physicians at the time knowing why it produced the results.

The paper opens for development of causal inference from non experimental studies. It is particularly important when experimentation might be unethical. Influence of smoking on development of cancer is widely accepted despite relying largely on observational data. Author identifies development of widely accepted criteria for evaluating causal conclusions from large observational datasets as crucial for further work, but does not attempt that in the paper.

Visvanathan et al. in the Research Statement of American Society of Clinical Oncology [20] notice that observational

studies can be complimentary to Randomized Controlled Trials (RCTs) by generating new hypotheses, revealing patterns and answering questions that cannot be answered by RCTs. To achieve this it is necessary to use a rigorous methodology and transparent reporting, in addition to ensuring data quality, interoperability and privacy.

What emerges through this Research Statement is a need for a new standard of evidence to take advantage of observational data and compensate for the lack of randomness and controlled groups. It remains an open question whether elements of quasi-randomness could be effectively introduced to observational studies on large populations.

Janke et al. [21] discuss the potential of primitive analytics and big data in emergency car. The potential that they consider unexplored. They notice that observational data could improve patient care but weak causal inference is a limiting factor at the moment. One of the approaches they suggest is that systems could prompt for additional information after discovering patterns of interest. It could be done both on population and individual level.

They believe that observational data might be used to derive and validate new models, it cannot be used to evaluate the effects of implementing these models. Example of influence of smoking on cancer development shows that the mentioned limitation should not preclude wider use of observational approaches. Moreover, possible advancements in causal reasoning could at least partially reduce that limitation.

IV. Conclusions - The Challenge

In order to take advantage of large observational dataset coming, in the context of underdevelopment methodologies of causal inference from such datasets, propose the following challenge: can we establish a new standard of evidence and a study design process that: (1) allows for drawing causal conclusions from large observational datasets and (2) can suggest interventions to enforce causal links in these data.

We propose to focus on Personal Monitoring Devices, at least in the beginning. Large portion of population is already in possession of such device, data are already collected (though in proprietary systems) and causal analysis has potential to lead to meaningful health and well-being recommendations both on individual and societal level.

Based on the existing literature we suggest that the starting point should be causal inference based on time series properties as preliminary explored in, e.g., aforementioned Google's library CausalImpacts, possibly in combination with some formal causality formulation method, such as Pearl's Do-Calculus [22].

These developments could later also be applied wider to smart technologies and observational data, but we think that proposed initial limitation to would in fact stimulate faster development of an accepted methodology.

References

[1] "Google's flu project shows the failings of big data | time," http://time.com/23782/google-flu-trends-big-data-problems/, (Visited on 12/29/2015).

[2] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Penguin UK, 2018.

[3] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. R. Suri, A. Tsou, S. Weingart, and C. R. Sugimoto, "Big data, bigger dilemmas: A critical review," *Journal of the Association for Information Science and Technology*, 2015.

[4] G. George, M. R. Haas, and A. Pentland, "Big data and management," *Academy of Management Journal*, vol. 57, no. 2, pp. 321–326, 2014.

[5] J. Grimmer, "We are all social scientists now: How big data, machine learning, and causal inference work together," *PS: Political Science & Politics*, vol. 48, no. 01, pp. 80–83, 2015.

[6] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[7] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[8] R. Kitchin, "Big data, new epistemologies and paradigm shifts," *Big Data & Society*, vol. 1, no. 1, p. 2053951714528481, 2014.

[9] A. Gelman, "Causality and statistical learning," *American Journal of Sociology*, vol. 117, no. 3, pp. 955–966, 2011.

[10] J. S. Mill, *A System of Logic Ratiocinative and Inductive: Boeing a Connected View of the Principales of Evidence and the Methods of Scientific Investigation*. Bombay, 1906.

[11] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, S. L. Scott *et al.*, "Inferring causal impact using bayesian structural time-series models," *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247–274, 2015.

[12] J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, vol. 10, pp. 1–51, 1923.

[13] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[14] K. A. Markus, "Principles and practice of structural equation modeling by rex b. kline," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 19, no. 3, pp. 509–512, 2012.

[15] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[16] D. Aliprantis, "A distinction between causal effects in structural and rubin causal models," 2015.

[17] C. M. Zigler and F. Dominici, "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 95–107, 2014.

[18] C. W. Granger, "Causality, cointegration, and control," *Journal of Economic Dynamics and Control*, vol. 12, no. 2-3, pp. 551–559, 1988.

[19] H. M. Krumholz, "Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system," *Health Affairs*, vol. 33, no. 7, pp. 1163–1170, 2014.

[20] K. Visvanathan, L. A. Levit, D. Raghavan, C. A. Hudis, S. Wong, A. Dueck, and G. H. Lyman, "Untapped potential of observational research to inform clinical decision making: American society of clinical oncology research statement," *Journal of Clinical Oncology*, vol. 35, no. 16, pp. 1845–1854, 2017.

[21] A. T. Janke, D. L. Overbeek, K. E. Kocher, and P. D. Levy, "Exploring the potential of predictive analytics and big data in emergency care," *Annals of emergency medicine*, vol. 67, no. 2, pp. 227–236, 2016.

[22] J. Pearl and E. Bareinboim, "External validity: From do-calculus to transportability across populations," *Statistical Science*, pp. 579–595, 2014.