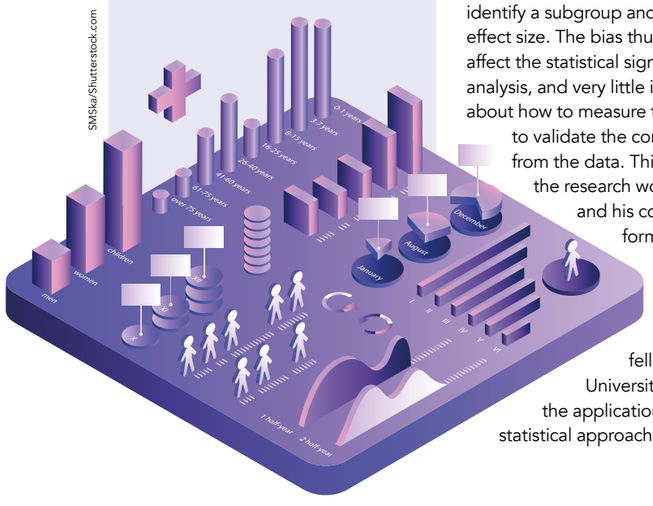
Understanding statistical significance of subgroups in the data science era

Statistical models are routinely used to derive inferences from large amounts of data. They directly impact several disciplines, including precision medicine and individualised learning, which rely on information concerning individuals and groups to make predictions on the expected effects of a specific intervention on population subgroups. Although it is relatively easy to find impressive-looking associations in big data, for instance through the use of data mining, these associations can be spurious. Professor Xuming He at the University of Michigan shows how a better understanding of subgroup selection in the big data era is necessary for providing valid statistical analysis to aid decision making under uncertainty.



SMSkar/Shutterstock.com

The ability to analyse and extract useful information from sets of structured or unstructured data that are too large or too complex to be dealt with using standard data-processing techniques is a crucial and very sought after goal in the Information Era, particularly after the diffusion of internet technologies. Big Data is nowadays a buzzword with profound ramifications in fields ranging from medicine, learning, social surveillance and e-commerce. One of the features that big data often exhibit, and that is currently only partially understood, is data heterogeneity, that is the fact that different subpopulations from which the data are collected often behave or react differently to specific interventions. This is particularly important in the case of precision medicine, in which different subgroups may respond differently to new drugs or experimental therapies. One of the problems, in this case, is the risk of overstating subgroup effects, intentionally or unintentionally. This can happen when the same data are used to identify a subgroup and to quantify the effect size. The bias thus introduced can affect the statistical significance of the analysis, and very little is currently known about how to measure this bias and how to validate the conclusions drawn from the data. This is the focus of the research work of Prof He and his collaborator and former doctorial student Dr Xinzhou Guo (currently a postdoctoral fellow at Harvard University), who, through the application of rigorous statistical approaches, are proposing methods to correct for the bias and draw statistically valid conclusions in cases where a "best subgroup" has been identified within a set of data.

QUANTIFYING STATISTICAL SIGNIFICANCE

A statistical test can be used to assess whether a given claim concerning a population ("null hypothesis") is credible. For instance, suppose a take-away restaurant claims that they can deliver to your house within 30 minutes on average, but you believe that this claim is incorrect. Your counter claim ("alternative hypothesis") is that it will take more than 30 minutes on average for the delivery. How can you test your hypothesis in a statistically meaningful way? An obvious approach is to sample randomly a sufficient number of deliveries and compute a probability value (*p*-value) to test the null hypothesis. If the *p*-value is small (typically lower than 0.05), there is strong evidence against the null hypothesis, and you will be more likely to reject the restaurant's claim. Although p-values are widely used, they may not always lead to statistically significant conclusions and they can lend themselves to misinterpretation and misuse.

DATA DREDGING

One of the common dangers of statistical analysis is the identification of statistically significant patterns in data by performing multiple statistical tests and only reporting those cases that yield significant results, thus dramatically increasing the risk of false positives. In essence, this amounts to performing a data-driven variable selection and using the resulting model to derive statistical inferences ("post-selection inferences"). For instance, the *Harkonen v. United States* Supreme Court trial of 2018 examines a situation in which the misuse of *p*-values in statistical significance testing led to spurious conclusions. It also highlights the difficulties (and legal ramifications) involved in the interpretation of *p*-values in statistical analysis in medicine and drug discovery.

WHYFRAME/Shutterstock.com

Professor He shows how a better understanding of subgroup selection is necessary for providing valid statistical analysis to aid decision making, for example in clinical trials.

A better understanding of subgroup selection in the big data era is necessary for providing valid statistical analysis to aid decision making.

In 2009, Scott Harkonen, the CEO of the drug company InterMune, was found guilty of wire fraud for reporting results on the activity of a new drug (*Actimmune*) developed by the company, which had been approved for clinical use (and successfully increased its sales). However, the Food and Drug Administration (FDA) did not approve the use of the drug in the treatment of a widespread, and lethal, lung disease, on the basis of insufficient statistical significance in drug effectiveness studies. Using post-data dredging, Harkonen was subsequently able to unearth a non-prespecified population subgroup in which he identified a nominally statistically significant survival benefit. Despite the FDA refusal to approve the drug based on this evidence, Harkonen issued a press release reporting on the drug's statistically significant survival benefits within the population subgroup that his company identified.

The Harkonen case is a compelling example that shows the ambiguity and complexity of statistical analysis involving subgroups, and it points to the need for more clear-cut and robust approaches to statistical analysis, particularly in drug discovery. In clinical trials, a new treatment might turn out to be only marginally effective with the overall study population, but it might be very promising for a subgroup of the population. For this reason, we do not want to abandon subgroup identification but need to have the right tools for analysing the data.

REMOVING BIASES IN STATISTICAL INFERENCE

The recent work of Prof He and Dr Guo has been addressing from a rigorous mathematical perspective the identification and use of population

J Main/Shutterstock.com

The Harkonen case is a compelling example that shows the ambiguity and complexity of statistical analysis involving subgroups.

Professor He takes a rigorous mathematical perspective to the identification and use of population subgroups to derive statistically valid inferences from results of clinical trials.

metamorworks/Shutterstock.com

subgroups to derive statistically valid inferences from results of clinical trials. The aim of their work is to provide a means to evaluate the effects of subgroup choice, taking into account the data-dependent search used to find the subgroup, by addressing the question of the statistical validity of post-hoc subgroup analysis. This has important implications for managerial decisions and regulatory deliberations on clinical trials, as clearly shown during the Harkonen trial.

In a recent publication, the team re-analysed a clinical trial of the effectiveness of an experimental treatment for patients affected with advanced nonsquamous nonsmall-cell lung cancer. An initial study (MONET1) seemed to indicate that East Asian patients were more responsive to the drug. However, a subsequent study (AMG-706) failed to confirm this claim.

The first problem in these studies is how to identify the best selected subgroup, that is the subgroup for which the drug is most beneficial, and various algorithms are available to carry out this task, based e.g. on machine learning or model-based methods. Once the best selected group has been identified, one needs to assess how good the subgroup choice is and whether it warrants further clinical trial.

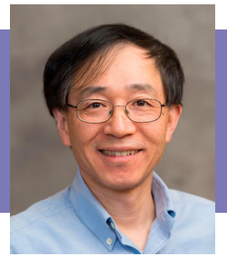
The Harkonen case is a compelling example that shows the complexity of statistical analysis involving subgroup-effect quantifications.

aodaodaodaod/Shutterstock.com

SUBGROUP SELECTION BIAS

Unfortunately, inference on the best selected subgroup identified from the same data suffers from over-optimism and is likely to lead to spurious correlations, a phenomenon that Prof He and Dr Guo label "subgroup selection bias". They propose a resampling-based method to address this problem, which is model-free, easy to implement and provides asymptotically sharp inference, regardless of whether the subgroups are pre-defined or identified post hoc from the data.

The application of this procedure to the MONET1 study shows that the subgroup selection bias can be correctly accounted for and that, depending on how many candidate subgroups had been considered, the initial trial may not exhibit statistical significance in the East Asian subgroup, in agreement with the subsequent AMG-706 study. Although a larger bias adjustment has been found to be required as the number of candidate subgroups increases, the adjustment has also been shown to level off quickly after a certain threshold. This makes the method proposed by He and Guo practically useful even in the case in which all potential subgroups are explicitly taken into account.



Behind the Research

Professor Xuming He

E: xmhe@umich.edu **W:** https://lsa.umich.edu/stats/people/faculty/xuhe.html
W: http://www.xuminghe.com/

Research Objectives

Professor Xuming He's research interests include broad areas of robust statistics, including quantile regression, post-selection inference, and semiparametric methods. His interdisciplinary research aims to promote the better use of statistics in biosciences, climate studies, concussion research, and social-economic studies.

Detail

Xuming He
1085 S. University, Department of Statistics, University of Michigan, Ann Arbor, MI 48103, USA

Bio

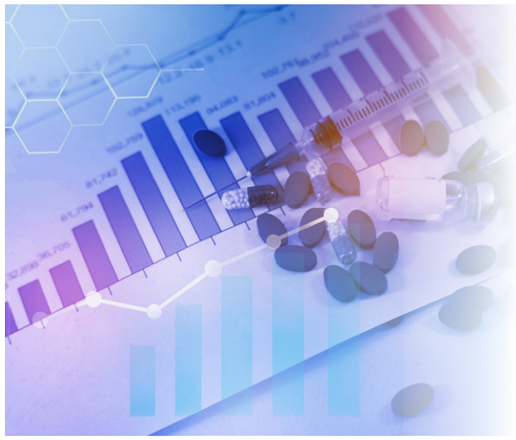
Professor Xuming He obtained his PhD in Statistics from the University of Illinois at Urbana-Champaign in 1989. He joined the University of Michigan as H. C. Carver Collegiate Professor in 2011. His prior appointments include faculty positions at National University of Singapore and University of Illinois at Urbana-Champaign. He is an elected Fellow of the American Association for the Advancement of Science.

Funding
National Science Foundation, USA

Collaborators

Xinzhou Guo, Postdoc fellow at Harvard University

janews/Shutterstock.com



References

Guo, X; He, X (2020). Inference on Selected Subgroups in Clinical Trials. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2020.1740096

Mayo, D (2020). P-Values on Trial: Selective Reporting of (Best Practice Guides Against) Selective Reporting. *Harvard Data Science Review*, 2.1, 1-20. https://doi.org/10.1162/99608f92.e2473f6a

Personal Response

The statistical methods you have developed to remove biases in the interpretation of large and heterogeneous sets of data have proved very successfully for clinical trial data. What other big data fields will likely benefit from your approach and the tools you have developed, and what are the remaining challenges that need to be addressed in those situations for which optimal data subgroups can be identified?

"Our recent work aimed at data from randomised experiments as commonly used in clinical trials, but the statistical methodology we have proposed can be further developed to subgroup analysis with observational studies.

Subgroup identification and quantification of subgroup effects are attractive options in the big data era, and their applications can be found in policy studies, personalised learning, marketing, and public health. With observational studies, one must account for multiple sourses of bias, not just the subgroup selection bias. Dr Guo is currently working with Professor Jingshen Wang at University of California, Berkeley, to extend our work to observational studies."

LSA COLLEGE OF LITERATURE, SCIENCE, AND THE ARTS
UNIVERSITY OF MICHIGAN