Human Genetics (2020) 139:1–3
https://doi.org/10.1007/s00439-019-02074-w

EDITORIAL

Special issue on 'Genetic epidemiology of complex diseases: impact of population history and modelling assumptions'

Amke Caliebe[1,2] · Michael Nothnagel[3,4]

Published online: 29 October 2019

The conduct of genetic epidemiological research on complex diseases has massively changed during the past decades, shifting from candidate-gene studies via array-based genome-wide association studies (GWAS) to sequencing approaches. In parallel, sample sizes have been increasing from small family-based studies to ever-larger datasets comprising up to several hundreds of thousands samples in worldwide consortia, with sample sizes exceeding a million being no longer an illusion. Part of this data explosion is technology driven by enabling access to genetic variation that was either out of reach before or only at prohibitive expenses, but it is also to a substantial extent due to power issues induced by decreasing effect sizes, complex etiological processes, the multitude of tested hypotheses and a widening array of potential confounding factors. A recurrent theme in biomedical research is that hopes run high with the introduction of a new technology, combined with expectations that we will now be able to obtain the complete picture and answer all questions, because we will be soon given a previously unimaginable amount of data. This happens sometimes even before the new technology kicks in. Also recurring is the inevitable disappointment after some low-hanging fruits have been harvested. This has been true for GWAS and, without much doubt, will be the case for the new wave of omics and single-cell studies, given their much higher complexity and variation.

Analytical methods' development and applicability often do not keep pace with new technologies and the data deluge. It usually takes several years of experience for the community to develop widely recognized and adapted quality and analysis standards for new study types, implying that early high-impact publications would be unacceptable by later standards. An emerging trend in analytical methods that somewhat mirrors the previously mentioned data explosion is the increasing application of machine learning approaches, often and misleadingly called artificial intelligence. 'Hypotheses-neutral research' lets 'the data speak for themselves', leaving the impression that such methods are supposed to pose research questions autonomously, a successful strategy for many grant applications and in the public.

In contrast, we argue that simply amassing data and computing power in itself will not be enough. Structural information underlying the data and data heterogeneity prevent success by such simple means. Examples are population structure, evolutionary processes, pleiotropy, complicated interaction relations, reduced penetrance and cell or tissue dependency. It takes what might be called 'organic intelligence' to arrive at meaningful, interpretable results. Just like in the "good old days" (or not?) of small sample sizes, we deem it important to follow the established routine of scientific methodology and to advance statistical genetics to be able to handle new types of data and study designs. Not surprisingly, this involves precisely formulated hypotheses before the study, a fitting design conducted with adequate analysis methods as well as a qualified interpretation of the results to arrive at knowledge. Most importantly, this also includes the laborious tackling of the presence of a potentially large number of confounders and inhibiting structures in the data. As was, for example, noted in 2013, "better conducted and better reported genetic association research may lead to less inflated results" (Aljasir et al. 2013). To keep up with the progress in technology, a continual development and improvement of statistical methods in genetic

✉ Amke Caliebe
caliebe@medinfo.uni-kiel.de

✉ Michael Nothnagel
michael.nothnagel@uni-koeln.de

[1] Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany

[2] University Medical Centre Schleswig-Holstein, Kiel, Germany

[3] Cologne Center for Genomics, University of Cologne, Cologne, Germany

[4] University Hospital Cologne, Cologne, Germany

epidemiology of complex diseases is mandatory and, to this end, may require novel ideas. To pick just one example, personalized risk prediction by polygenic risk scores and other means faces the danger of overfitting of those models to small and limited groups and populations, such as Europeans. Such issues have to be addressed. One step in the right direction might be the reduction of complexity in statistical models by improved incorporation of prior knowledge on biological processes by enhanced communication between statisticians, biologists and clinicians.

We have the good fortune that this special issue is published at a time of a huge expansion of knowledge, technology and data availability in genetics and genomics of complex diseases. This scenario offers a wide and rich field for new method development to provide unique insight into many decisive open questions. The special issue presents ten manuscripts: reviews, original research papers and one opinion paper. They can be loosely assembled into three groups of contributions dealing with evolutionary aspects, statistical methods and interpretation of genetic studies. In the first part, Uricchio (2019) gives an important evolutionary perspective on GWAS. He stresses that statistical inference in the genetic epidemiology of polygenic diseases is influenced by evolutionary processes, such as selection, possibly hampering the transferability of GWAS results across populations. Including evolutionary models into future statistical approaches might therefore be advisable. Population differences as one of the major confounders in genetic studies are also addressed in the review by Lawson et al. (2019). Crucially in the era of large meta-analyses, consortia and biobanks, even subtle population stratification can bias results, especially for prediction and causal inference. Commonly used approaches for adjustments like principal components might then not be adequate and the authors recommend applying and developing new methods, such as chromosome painting. This manuscript is complemented by a detailed look at the population substructure in West Africa by Chaichoompu et al. (2019). The authors introduce a novel method for fine-structure detection that relies on principal components and applies clustering as well as iterative pruning. The second part of the special issue comprises four statistical methods contributions. It includes the introduction of OpenMendel as a complete overhaul of the long established Mendel software by Zhou et al. (2019). This open source software, apart from offering a large variety of methods in statistical genetics, now embraces the challenges posed by modern large-scale datasets. It enables big data analytics, parallel and distributed computing and also allows cloud computing. With these innovations, OpenMendel promises to be an extremely useful, comprehensive and easily applicable tool for the genetic epidemiology community in the near future. Statistical learning has become more and more significant in statistical genetics during the last years, especially in the area of high-dimensional omics data. Boulesteix et al. (2019) give an overview on statistical learning approaches in genetic epidemiology and compare regression analysis to machine learning methods. They comment on the correct application of these procedures for high-dimensional data with a special focus on training and validation and the choice of the number of parameters. The review of Smeland et al. (2019) deals with the interesting subject of pleiotropy. It suggests the conditional false-discovery rate (condFDR) method to exploit cross-trait effects for identifying associated genetic variants. This is a Bayesian model-free approach based upon GWAS summary data. A comparison of condFDR to other cross-trait methods and examples about applications of condFDR is presented. Liu and Montgomery (2019) in their review give an overview on the timely topic of using cell type information as a follow-up to GWAS studies. The functional mechanisms of most variants detected by GWAS still remain in the dark and specialized cell types may help to discover causal mechanisms. Recent advances, best practices and remaining challenges in this area are discussed in detail. Three more contributions in the interpretation part take a look at the outcome of genome-wide studies from different angles. Genin (2019) revisits the mysteries of the missing heritability. She discusses the role of several potential causes such as an omnigenic model, rare or structural variants and interactions. As a consequence of her considerations, she warns about the abuse of the term heritability. She concludes that missing heritability is an ill-posed problem, because heritability is defined through an oversimplified model that does not fit the underlying complex biological structure. Weiss (2019) presents his personal conclusions on omics research in an opinion piece after working for decades in the field. He investigates the assumptions underlying meaningful studies in this area and also refers to the field of big data and personalized medicine. Finally, Sheehan and Didelez (2019) turn their eye on Mendelian randomization and explore its potential and pitfalls for inferring causality, drawing the circle of the scientific method to a close. In a very well-structured presentation, they explain and discuss thoroughly the core assumptions that need to be satisfied in a Mendelian randomization study. Additionally, they stress the implications of using, or not using, parametric modelling assumptions.

We are hugely grateful to all authors for their efforts and hard work which, combined with inspiration, have given rise to this excellent selection of contributions to the field of statistical genetics for complex diseases. We are confident that these papers will support discoveries in this area in the future and will help us not getting drowned in the ocean of data that is ahead of us all.

Acknowledgements The authors thank France Gagnon and Keerthana Govindarajan for their invaluable support in the preparation of this Special Issue.

References

Aljasir B, Ioannidis JP, Yurkiewich A, Moher D, Higgins JP, Arora P, Little J (2013) Assessment of systematic effects of methodological characteristics on candidate genetic associations. Hum Genet 132:167–178. https://doi.org/10.1007/s00439-012-1237-4

Boulesteix AL, Wright MN, Hoffmann S, Konig IR (2019) Statistical learning approaches in the genetic epidemiology of complex diseases. Hum Genet. https://doi.org/10.1007/s00439-019-01996-9

Chaichoompu K, Abegaz F, Cavadas B, Müller-Myhsok B, Pereira L, Steen KV (2019) A different view on fine-scale population structure in Western African populations. Hum Genet **(in press)**

Genin E (2019) Missing heritability of complex diseases: case solved? Hum Genet. https://doi.org/10.1007/s00439-019-02034-4

Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, Timpson NJ (2019) Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? Hum Genet. https://doi.org/10.1007/s00439-019-02014-8

Liu B, Montgomery SB (2019) Identifying causal variants and genes using functional genomics in specialized cell types and contexts. Hum Genet. https://doi.org/10.1007/s00439-019-02044-2

Sheehan NA, Didelez V (2019) Epidemiology, genetic epidemiology and Mendelian randomisation: more need than ever to attend to detail. Hum Genet. https://doi.org/10.1007/s00439-019-02027-3

Smeland OB, Frei O, Shadrin A, O'Connell K, Fan CC, Bahrami S, Andreassen OA (2019) Discovery of shared genomic loci using the conditional false discovery rate approach. Hum Genet. https://doi.org/10.1007/s00439-019-02060-2

Uricchio LH (2019) Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. Hum Genet. https://doi.org/10.1007/s00439-019-02040-6

Weiss KM (2019) The Four Horsemen of the 'Omicsalypse': ontology, replicability, probability and epistemology. Hum Genet. https://doi.org/10.1007/s00439-019-02007-7

Zhou H, Sinsheimer JS, Bates DM, Chu BB, German CA, Ji SS, Lange K (2019) OPENMENDEL: a cooperative programming project for statistical genetics. Hum Genet. https://doi.org/10.1007/s00439-019-02001-z

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.