Medicine in Data Intoxication

Big data makes everything new. Of course also medical diagnosis and therapy. These are the big promises. But nobody talks about the risks. A correction.

By Gerd Antes

Big data, digitization, machine learning and artificial intelligence are currently dominating the headlines when it comes to the future. The Chancellor thus connects the next industrial revolution. The auto industry sells autonomous cars, giving us time to control our home cooking and cleaning programs while driving the Internet of Things. Politicians, scientists, the media and the IT industry are turning to each other to bring us the benefits of the digital future. In huge amounts of data, the apologists of big data are seeing a golden future for us that comes with a paradigm shift in science. With such a profound intervention, common sense would expect a careful - and actually common - technology assessment, but one searches in vain. Warning voices indicate risks and possible damages. However, they go under in big data hype.

Data is the oil of the 21st century. What is more obvious than to use this treasure for our health? In this area, too, there is an enthusiasm that makes the usual benefit-risk-cost estimates superfluous. The most recent evidence of this is a statement by the German Ethics Council on "Big Data and Health", in which the Song of Songs on Big Data is uncritically voiced. However, you can only sell oil - or you have to refine it into usable products in complex processes. What is the image of the knowledge-based refinery that translates data into actionable knowledge and applicable procedures?

The model developed for this by Big Data is quickly described: The enormous technical possibilities in the data world make the tedious search for useful effects superfluous. We have left the era of causality and are already in the Age of Correlation. The old paradigm that theory and data lead to the generation of hypotheses, which in turn must be empirically confirmed by studies, is no longer valid. Correlations no longer lead only to hypotheses, but to confirmed knowledge due to arbitrarily expandable data sets. For its implementation, unlimited access to all data is necessary, for which the rights to the data must be transferred from the owner to the user. That would not be a problem as the users will handle the data responsibly. In addition, big data should provide snapshots of a changing world every second, so that the data basis for individual findings is not reproducible.

It is obvious that the big-data characteristics mean a frontal collision with the established scientific cognitive process. The restriction to correlations from observation data stands against the targeted generation of data by prospective planned experimental studies. For the comparison of these fundamentally different approaches, the consideration of the inevitable mistakes is indispensable. But that's exactly what Big Data dispenses with. Errors practically do not occur, they do not need either, as they can be corrected in any case by more data. However, that methods would have to be used to detect such errors is generously ignored.

Dealing with errors is of central importance for the scientific evaluation of medical research and its application in prevention, diagnostics, therapy and prognosis. The

inevitability of errors requires a clean statistical approach to data in order to arrive at valid statements that can never exceed the status of probability statements. That this is inseparable from a degree of uncertainty, dislikes doctors, patients and politicians, but is inevitable. Highest demands on quality in all phases of the knowledge acquisition are prerequisites for a faultless proof of causality. However, these requirements, which are standard in the science process today, do not play any role in big data processes. By redefining causality, all the associated requirements seem superfluous.

The incantations are unmistakable. Most impressive in the phrase "The end of the theory: The flood of data makes the scientific method superfluous". Under no circumstances may the authors of such lines engage in a discussion of possible errors. How this works can be read in the book "Big Data" by Viktor Mayer-Schönberger, and Kenneth Cukier (F.A.Z. from 24 October 2013). In interviews, Viktor Mayer-Schönberger says that big data will surely help save lives in the future. However, the interviewer missed the question of how many lives are threatened by big data, another example of avoiding any risk discussion.

Summoning the success of Big Data cannot hide the fact that there is little positive evidence for such promises. Negative evidence, however, there are very well. The abdominal landing of Google's "Flu Trends" for the prediction of influenza epidemics belongs to the author as well as IBM's Dr. med. Watson, who was recently quit by one of the world's largest cancer clinics, the MD Anderson Cancer Center in Houston, following $ 62 million of hospital investment. In both cases, the work of man proved to be better.

The damage-risk discussion is limited to the use of data and thus on the ethical and legal side. A serious discussion about the risk of big data is overdue, because the central rationale for the big data promises is wrong: more data does not automatically mean more knowledge. On the contrary, this intuition-resisting fact means that even adding more correct data can worsen the cognitive situation. Thus, the construction, the increase in knowledge and thus the basis for action on growing amounts of data is more than fragile. As the disturbing noise increases with increasing data, real effects are less easy to find and the proportion of misidentified, spurious effects increases. These false positives are one of the central problems of empirical research and can be brought to full bloom through the big data approach through the correlations. Under the slogan spurious correlations there is a wealth of examples in which two phenomena are highly correlated, but a causal relationship derived from this would be nonsense. Like small data, big data requires high-quality evaluation methods, so the proposed departure from theory would be a step in the wrong direction. It is more likely that the false positives will enable cherry-picking on an industrial scale. The shift to correlations as the main source of insight leads directly into the trap. Being exposed on a large scale would mean following either wrong tracks or spending more time getting the wrong signals out of the way than paying the right attention to the right one. The world is promised to enter a new era with big data. In fact, she's been hostage to GAFAM for years (Google, Apple, Facebook, Amazon, Microsoft) and accomplices. They also bow to those who were supposed to protect us, such as ministries, major research institutes and universities. The most recent proof of this is the statement by the German Ethics Council in the area of Big Data mentioned at the beginning.

What needs to be done to bring the old and the new world together and to get the most out of patients and healthy people? It takes a return to the patient, who often appears on the horizon in the discussion about big data, as the beneficiary of a development that is more of a vision than an imminent reality. It requires a complete discussion of benefits, risks and costs, and includes assessing the growing alienation of people from digital and commercialization medicine.

GERD ANTES
Gerd Antes is Scientific Director of the Cochrane Germany Foundation and of the Institute of Evidence in Medicine at the University of Freiburg.