(Mis)Use of Statistics in Science – Interview with Dr. Gerta Rücker

Dr. Gerta Rücker,[1] a mathematician by training, works as a biostatistician at the Medical Center – University of Freiburg, Germany. Her special area is meta-analysis, and she is associated with Cochrane Germany. She has written a large number of research papers on statistical methods, and co-authored a number of Cochrane reviews. Additionally, she is engaged in teaching meta-analysis methods and is one of the authors of a book 'Use R for meta-analysis'.

[1]email: ruecker@imbi.uni-freiburg.de

JUnQ: Dear Dr. Rücker, you are familiar with misconduct and errors in scientific publications and studies as well as with the publication bias. Furthermore, you are an expert in the statistical sector. Can you please explain the importance of good statistics? Or more importantly, how easy or difficult is it to look at statistics objectively?

Dr. Rücker: Two different questions. Good statistics is important because all empirical science results in data, sometimes "Big data", that is, numbers, lots of numbers. Whether big data or very small data sets (as often found in the medical science) – it is impossible to make inference without certain skills in how to analyze them. This is what statisticians are qualified for. To the second question: First, admittedly, statisticians, particularly in Germany, sometimes have failed to explain their methods sufficiently clearly to the public (it is a little different in Britain, with its much older and better developed statistical culture). Secondly, there is a broad range of statistical methods, and often more than one approach is appropriate for a research question. But finally, undoubtedly, statistics is a science with rigorous methods, and statistical education should play a greater role in the curricula of all sciences, especially the life sciences, and even the humanities.

JUnQ: What is, in your opinion, the biggest problem with interpretation from statistics in the life sciences?

Dr. Rücker: The dominance of the p-value! This recently has been spoken out clearly in a statement of the American Statistical Association (ASA), published in the American Statistician and accompanied by twenty invited commentaries very much worth reading.[1] There, leading statisticians from all over the world plea for abandoning the unfortunate big role of hypothesis testing, p-values and the arbitrary 5% threshold. Instead, statistical modeling, estimation with uncertainty and Bayesian methods should be preferred. Unfortunately, statisticians over decades have failed to convey this message when teaching statistics.

JUnQ: In the last few years several studies showed that up to 80% of preclinical studies are not reproducible, i.e., in spite of scientific publications stating positive results, the data is erroneous. How is such a big percentage possible?

Dr. Rücker: For a long time, clinical researchers have been told that they should produce as many "significant" results as possible. Consequently, they do lots of experiments, look at lots of hypotheses, look at lots of outcomes or models, until they find a significant p-value, which they finally decide to publish. If someone else repeats just this experiment, he must fail. There are many names for this kind of conduct – p-hacking, data dredging, publication bias.

JUnQ: Glenn Begley, a former employee of Amgen, recently cited a conversation he had with a scientist about a cancer study published in Cancer Cell.[2] He and his team tried to reproduce the published results, but failed a hundred times. The author of the study said: "Oh yes, we performed this experiment about a dozen times and achieved the published result only once. But in the end, we decided to publish exactly this." How can it be that experienced scientists deal with experimental data in that way? It is obviously incorrect and can even be very dangerous in the case of medical studies.

Dr. Rücker: Just as I said, this example is by no means an exception. It also demonstrates that the problem goes far beyond statistics. I cite Stephen Senn, who in his commentary to the ASA statement noted psychological reasons for these kinds of misconduct: "An impatience with the necessary nuances of expression that good statistical reporting requires; the (usual) prejudice of scientific journals in favor of 'positive' results; the common habit of transforming

[1]R. L. Wasserstein, N. A. Lazar, *Am. Stat.* **2016**, *70*, 129–133.

[2]http://www.deutschlandfunk.de/wissenschaftsmuell-wenn-forschung-nicht-haelt-was-sie.740.de.html?dram:article_id=330956 (last access on 24.05.2016).

shades of gray into either black or white; and the desire of individual scientists for recognition and reward."

JUnQ: During our research about the subject, we found two different extremes. First, there are huge studies with a great amount of data albeit with an unclear formulation of the question. And second, very small studies with too few participants. To deal with the first mentioned extreme of immense amount of data: It is obvious that one can find proof for nearly every hypothesis if the amount of data is large enough. In case of studies in which the initial problems are not clearly identified, this allows the researcher to find some positive results. Can you please explain why this is not a good scientific practice? Is it not often the case that you find the greatest results in science serendipitously?

Dr. Rücker: We distinguish two approaches: the exploratory or hypothesis-generating approach, and the confirmatory or hypothesis-proving approach. If you have a huge data set and no particular idea (this is often the case in genome-wide association studies) it is completely legitimate to look at millions of p-values first. But you have to be very clear that this is only the very first step, it is completely exploratory. Statisticians working in the area of high-dimensional data are currently working on refined methods of model selection for this huge challenge. Only after further steps of model selection in a statistically principled way, we can expect reproducible results.
I definitely agree that serendipity plays a great role in science. However, both before and after the great moment of a discovery, science is very hard unspectacular work, first in order to come up with the idea and to recognize its meaning, and afterwards to prove its usefulness.[3]

JUnQ: And for the other extreme of studies with too small a data set: Why is it so that a researcher is almost guaranteed to find the desired results if there are too few participants in the study? What is meant by "regression towards the mean"?

Dr. Rücker: It is by no means guaranteed that a researcher finds the desired result in a small study. On the contrary, in most cases the researcher does not find the desired result and therefore switches to another outcome, a different analysis, deletes some observations, or doesn't publish the study at all. At the end of the day, we mainly see "desired results".
"Regression to the mean" is a different issue. If you have a strong headache and see the doctor (or not so), you will feel better the next day, the most probable reason being that you felt so very bad the day before. Likewise, if a study investigator selects only patients in the worst health, these patients are likely to benefit more from the treatment than patients with mild disease, whose health would improve anyway, with or without treatment.

JUnQ: "Statistical power" is a term for the necessary number of cases in a study to achieve significant results. Can you please explain how it is defined and why many studies lack statistical power even though most medical scientists should be aware of the problem?

Dr. Rücker: Statistical power is not the number of cases. Rather, it is a probability. The power of a statistical hypothesis test, designed to compare the outcome between two treatments, is the probability that a pre-specified difference between the two treatments can be demonstrated by the test. It depends, among other things, on the number of cases: the larger the sample size, the greater the power.
Why do many studies lack statistical power? There is a simple explanation, at least for the medical science: clinical trials, particularly if conducted thoroughly, are extremely expensive. Most doctors want to do their MD degree, they cannot invest much time or money, and simply collect some existing data for a simple analysis.

JUnQ: Even though this terminology – exploratory and confirmatory study – should be clear to researchers with statistical background, why are many exploratory studies published in a way so that they seem confirmatory? What is the biggest danger behind such an approach?

Dr. Rücker: Many clinical researchers are trained physicians. Physicians are used to making decisions. If there is a patient in want of a diagnosis or in need of a treatment just before you, you must make a decision. Even to do nothing and to send the patient home to bed is a decision. My experience is that physicians are not used to accepting uncertainty. They even don't accept continuous variables, they always want to classify and to dichotomize. It is all black and white. This is one reason why the p-value is so attractive to them, and why they only reluctantly accept that a result might be exploratory. They always want certainty.

JUnQ: Apart from these clear errors in basic statistical procedure, many studies are executed without proper systematic literature studies. The British journal Lancet published an article series "Research: increasing value, reducing waste" in 2014 about the problematic situation of results in medical studies, i.e., poor study design, inaccessible research, and selective reporting.[4] Accordingly, a lot of studies only repeat previous work leading to waste of time and money. Or they are predestined to fail. Are the scientists just too lazy to design their study properly or is there a different problem in dealing with published results?

Dr. Rücker: I already mentioned some of the reasons: lack of time, money, patience, education. Also, often there are soft factors with very negative impact like career needs, reputation of institutions or other dependencies in the academic world.

[3] J. Kimmelman, J. S. Mogil, U. Dirnagl, *PLoS Biol.* **2014**, *12*, e1001863.
[4] http://www.thelancet.com/series/research (last access on 24.05.2016).

JUnQ: Due to the publication bias, i.e., the predisposition to publish mostly positive results, many medical studies remain unpublished. Can you give us an estimate of the percentage of unpublished studies?

Dr. Rücker: 50%. It is a very stable figure that has been reproduced in many studies around the world.[5]

JUnQ: Why does this count towards being a huge problem?

Dr. Rücker: The answer to a research question provided by a study often helps only future patients. The patients who volunteered to participate in a clinical trial may not benefit from their participation. Their motivation for participating is the desire to help others. Having this in mind, we have a strong ethical obligation to make all research results available to the public by publishing it. Secondly, being published or not is strongly dependent on the results of a study. So-called "negative" results are much more likely to remain unpublished, resulting in a strongly biased evidence base for all following steps in the knowledge process. The result is strong over-optimism in the assessment of diagnostic tests and therapies.

JUnQ: What leads to such gross misbehavior? Is the education of junior scientists flawed or might there be a place for incorrect motivation set by the current publication system?

Dr. Rücker: Both. I don't think that "the publication system" is the biggest problem. A well conducted study with sufficient power for answering a relevant question will almost surely be published. But, undoubtedly, there are some deficiencies in the education of young scientists, at least in Germany.

JUnQ: Since this is no new problem, most scientists should be aware of the issue. However, the present procedure changes only very slowly. Can you think of an easy to implement viable solution?

Dr. Rücker: Some instruments already have been implemented, at least in clinical medicine, such as trial registration and reporting guidelines. The role of statistics is more acknowledged and better positioned in teaching and clinic than twenty or thirty years ago, think of the well developed methodology of systematic reviews and Cochrane. However, in basic research, animal experiments, and medical devices, we are far from the standards that would be desirable. There is not much hope that a viable solution will quickly lead to substantial improvements. It is rather an ongoing struggle to keep and improve quality, and a major and probably effective step would be to include these issues in the education of students and later in the training of scientists.

JUnQ: Thank you very much for the interview!

—Andreas Neidlinger and Soham Roy

[5] A. Blümle, J. J. Meerpohl, M. Schumacher, E. von Elm, *PLoS One* **2014**, *9*, e87184.