Top 5 Things in Ethical Challenge for Statistician

Shaopeng Liu

Being ethical in practice of statistical analysis, especially in biomedical researches, is a big challenge. On the one hand, ethical traps may show up implicitly other than explicitly, causing problems that we are unaware of (e.g. data sharing, transparency and data abuse); on the other hand, the duty of biostatistician may involve broader range than it is expected: statistician should work as collaborator who participate in all necessary phases rather than consultant to ensure effective and valid data construction. Even given the point that we already realize the ethical issue in empirical situation, however, making a right decision would not be easy due to human nature. It's astonishing to realize how easily our mind would fall into bias traps. Attentions are needed everywhere!

(1) Working as a collaborator

There are much more duties than we expect for statisticians. We should regard ourselves as "collaborator" rather than "consultant". During research, it's statisticians' duty to participate in all phases of study to make sure that 1) the study design is reasonable regarding adequate sample size, random sampling process as well as collection of interested variable; and 2) the data collection process is carried out as planned; and 3) data cleaning step is reasonable and recorded in manuscript; and 4) the analysis is reliable regarding potential confounding factor and outliers.

The reason why knowing all those duties is vitally important is the fact that the best analysis can't save biased or insufficient data. To ensure the efficacy of available data, as a statistician, we need to change the expectation of our roles from "consultant" to "collaborator". By actively participate in each phase of a research, we can contribute our expertise as much as possible, and therefore the data would be justified from the very beginning. Analysis is only

the last step, while the duty of statistician arises from the beginning of a research. It would be ethical for both researchers and statisticians to carrying out valid researches without wasting of resources.

To better fit our role in research, we could follow the criteria as a collaborator as Roger (2002) summarized: "1) conception and design, or analysis and interpretation of data; and 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published". By setting high and clear expectation, statisticians can follow those duties as the guideline to do their job in a research. Especially for the 3rd item, we need to hold the moral requirements for ourselves that once we publish our conclusion, that means we justify the data as well as the analysis in an ethical manner.

Besides, one additional duty is that we also need to construct public belief that statisticians are reliable by insisting on the ethical requirements we have. Credibility is the core of our job. Therefore, I want to emphasize the duty in public image, as Hurwitz and Gardenier (2012) mentioned: "the credibility depends not only on professional skills with years of hard work, but also on a public perception that those skills are used ethically."

(2) Think twice before making decision

This is the most impressive lesson I learned about myself. By strategically reviewing my routine thinking process, I come to realize how biased I was in several situations.

"Think twice" means every decision during our analysis should be made deliberately with the consideration of reasonable foundation and possible bias. We should always be aware that "in making decisions, your own mind may be your worst enemy" (Hommond et al 2009), and follow a strategic framework (e.g. SMART strategy) to organize our thinking process.

Although it sounds cliché, the importance of this lesson lies on the challenges in making our decision reasonable all the time. Like Don & George (2004) said, "contrary to the common

view both lay among people and social scientists that decisions are made by weighing costs and benefits, the reality is that many decisions are made based on impulse and intuition". By following pre-defined thinking framework, we can do our best to be objective and get rid of bias as well as impulsive decisions.

The first challenge in making ethical decision is that we need to consciously be aware of the ethical problem we encounter to make rational decision. During the way of thinking process, Hammond et al (2009) has shown to us how different bias such as "anchoring trap", "Status quo", and "sunk cost" can influence our decision unconsciously. What's worse, Ann & David (2004) has revealed that in some situation our cognition "deceits" ourselves by falsely convincing us that the ethics are upheld. Keeping in rational mind is never an easy task, the combat of automatic process and controlled process depend a lot on your mindset state. Those give rise to the second challenge: it's very hard to be objective and get rid of bias and impulse.

Through the understanding of thinking process, I came to realize that I am still easily biased by the cognition traps and it's also useful for myself to utilize similar strategy in daily life. The challenges and solutions open a new window to better understand myself. I used to label myself "rational person" because I think I always consider both sides of a situation. But it turns out that the pre-existing bias is always misleading the direction of my thoughts unconsciously, when I utilize the "manage emotion" and "test the assumption" strategy from "SMART" by simply writing down my thoughts and the facts. When I ask myself again by "are you still angry because of the fact A", I would realize that how stupid I was in previous situation. Especially when a quarrel or unpleasant debate happened with someone, although it's my advantage that I will always patiently listen to and abstract opponents' points, I was biased by the unconscious malevolent impression of the other.

Developing a reviewing and thinking model would be very helpful for us. By listing all the potential bias and choices, we can consciously evaluate the situation to achieve an ethical decision.

(3) Be alert of misconduct

Research misconduct has a broad definition. US federal has a common definition of misconduct in research as "fabrication, falsification, or plagiarism in proposing, performing, or reviewing research or in reporting research results. FFP is the bottom line that every researcher should hold due to the catastrophes consequences and moral requirement. Although there were few guys who tried to break the red line, finally it turned out to ruin their career. Raymond et al (2006) showed that FFP is a minor problem, but instead there are more problems in interpretation of the data and the rule of science. It's more common that there are minor flaws within a promising result, such as unexpected different center effect, poor data quality control, arbitrary remove of outliers and data management issue etc.

The first step we should do is to seek help from other professional colleagues, experienced peers and maybe IRB. In research, we care about not only significance but also effect size. Sometimes it's useful to discuss with other professionals to understand what it means in clinic. Besides, there may also be empirical judgment in making the decision. So when we find potential minor misconduct, it's better to seek help first to understand the problem. In addition, it's possible that we may feel embarrassing and anxious of being ignorant or pointing out the problem. It's okay for the greater good of the team, so be calm down and be positive.

Then is to recognize ethical rules to deal with those challenges to check if there are anything else we can to correct or improve the situation, especially in data management and quality control steps. Researchers are apt to neglect the standard rules in those fields; we may

need to remind them or propose our suggestion with better practice. This part is a special requirement for we statisticians, because we should follow the rules to store, modify, and analysis data to make results reasonable and reproducible.

Next step is to anticipate the results based on the problem and the rules. Maybe we can present our findings and suggestions to the research managers, along with the potential consequences if we break the rule. But as we are not in control of the project (especially in the industry field), what we can do is to make a disclaimer in the manuscript describing the minor problem with data.

(4) Pay attention to data management and reproducibility

Mae Gordon's presentation showed us a framework to improve reproducibility in clinical measurement. But right here I'd like to describe a set of rules that should be followed for data management from study design to data analysis to make the data process ethical, reliable and adequate for sharing. Those guidelines can ensure the data is reasonable all the way down to analysis, and traceable for further adjustment or review.

In the phase of study design, other than clarity, anonymity and sufficient marker, we need also be careful about the consistency of data format, for example we don't know 02/02/10 means MMDDYY in US, or DDMMYY in Europe, or YYMMDD in Asia. It's better to use a standard design tool (like Redcap in WashU) for all the researchers. Otherwise it would be miserable in analysis, let along the data sharing. This why in the first point I stress that statistician should participate in all the phases of a research.

Then in the record step, we should follow the code to record data, so that everything can be traced. Every modification should be marked with date, reason and corresponding note. The same goes for exclusion/inclusion criteria, every modification should be traceable, reasonable and be recorded.

When it comes to the analysis step. Reproducibility is easier for statisticians because given the same seed, code and packages we use, the same software is expected to generate same results. Although Mike (2015) wrote that "reluctance to share data has been the norm rather than the exception ... (because) sponsors bear all the cost of data sharing". For statisticians, it's our jobs to justify the data as well as the analysis. We should include the details of our analysis to make it reproducible, including model choice, package, code, parameters and even the software edition.

(5) Be aware of potential risks in the era of "big data"

Thanks to the great advance of computer technology, the storage and computing ability has significantly increased, making the precision medicine possible based on the whole genome exon/epigenome/RNA sequencing etc. Personally, I would describe big data as "high-dimensional, comprehensive data" from which you can explore much more than you expect. Those challenges would make the ethical practice much more difficult, and may lead to evolution in our current data management manner.

First challenge: the data itself are much more than we need. A straightforward problem is "who you are", the encrypted information we don't want to reveal. Given the whole genome sequencing data, we can predict participants' height, color, eye, blood type, allergic agents, and even some potential disease. Those are much more than we need in a research, it's much easier to locate "certain patient" and I believe nobody would like to reveal those details. At least for insurance company it's a potential source of discrimination.

Second challenge: the hidden information that can be digged out. The "Target figured out a teen girl was pregnant before her father knew" story was well known to us. The clustering model is not difficult to design given the crucial data we have. Based on sequencing data and other electronic medical data, it's possible to predict other diseases of the participants. And

further digging can even unmask the name(id), personal feature, habit etc. Not only there are no privacy at all, but the leakage of information may also be dangerous.

Much more risks might be revealed along the implement of big data. Maybe advanced supervision in data usage and computing would be needed, and more strict rules need to be applied regarding safety issue. We need to be updating with the technology: knowing it is the prerequisite to handling it.

Conclusion

To maintain ethical practice as a statistician is never an easy task. In technical perspective, analysis is only the last step to run. We need to contribute our expertise from the very beginning of a research to ensure unbiased and sufficient data we have. To achieve this goal, a statistician should regard him/herself as a collaborator rather than the consultant. Man errs as long as he strives, the misconduct and ethical traps would be unavoidable in empirical situation. Intentionally train ourselves to think reasonably would be critical in decision making. Though it would be challenging and exhausting, some thinking frame work (e.g. SMART) would be helpful to deal with practical cases. Get used to and develop habits on them will grant you a desirable result. As a statistician, it's also our duty to focus on the data management, which doesn't draw enough attention from researchers. A valid analysis should include justified data as well as suitable model. And for the goal of reproducibility, we need to be transparent on our code, model as well as parameters. While we are benefiting from the big data era, potential risks are also on the way. Although the standard guidance would always take years of compromise, we'd better understand what's the challenge is and be prepared for them.