SCANDINAVIAN JOURNAL OF UROLOGY
2022, VOL. 56, NO. 1, 12–13
https://doi.org/10.1080/21681805.2021.2019305

EDITORIAL COMMENT

No self-evident interpretation of a randomized study

In this issue of the journal, Schmidt-Andersen et al. address a relevant problem: how does prospective and retrospective data collection of postoperative complications compare [1]? The paper raises several interesting themes, some of which will be discussed here: Was randomization the best option for a study of this problem? Is this a study 'only' of prospective versus retrospective data collection? Should we focus on the statistical significance or on quantitative estimates?

Was randomization helpful?

Patients were randomized 'to either standard collection of complication rates retrospectively through medical journals or prospectively through questionnaires and interviews'. The study 'was conducted at a hospital where a strict policy for reporting complications after urological surgery already was incorporated in the clinic'. Those randomized to prospective data collection also had the same standardized information in the digital medical records as those in the retrospective data-collection arm.

The concept behind randomization is to approach the counterfactual definition for drawing causal inference from a clinical study: can we design a study so that two groups are exchangeable – that is, identical with the exception from the exposure of interest [2]? The ideal would be to observe the same individuals twice during the same circumstances and ideally even the same calendar time, once with and once without the exposure. This cannot be done when the exposure is a clinical intervention.

The Schmidt-Andersen et al. study however is not a study of a clinical intervention, but about the correctness of prospective or retrospective data-collection in patients. A threat to a study of this kind is that a care-givers' knowledge that a study of this nature is ongoing, improve their reporting in the medical records to a level that could not be upheld under routine care and may give a false impression of the accuracy that medical records have in standard care. In the particular setting of the Schmidt-Andersen et al. study, it is unclear if randomization helps, because reporting to the medical records would be better for all, since the caregivers were blinded to who was in which group. Furthermore, in their setting with an institutional strict policy for reporting, such a temporary improvement in data collection most probably is marginal.

So, an alternative here would have been to collect prospective information for a larger number of patients and compare back where all are their own controls. This design is even closer to the counterfactual ideal than the randomized design chosen by the authors. It had also allowed for comparison of item by item, not only comparing an average level of complications.

The strength with the randomization in this study is said to be to 'decrease the risk of residual confounding which was successful in the present study'. The authors are right that we by randomizing *hope* to minimize what is called residual confounding, that is, to make the groups comparable not only in terms of known risk factors but also with respect to unknown disturbing factors that may distort the result. If we in a given study are successful or not, we cannot observe and measure, we can only hope that this is the case. The larger the study, the higher the probability that the group assignment is successful in terms of making the groups exchangeable in terms of known and unknown disturbing factors. This is a small study and there is no guarantee that residual confounding was avoided.

Even a large study, randomized with state-of-art methods where selection bias has been avoided can be destroyed [3,4]. One example is the information bias regarding the outcome parameters that can be introduced in non-blinded studies. Hence the use of the double-blind study concept in pharma trials. In Schmidt-Andersen et al's study, an information bias could have been introduced despite the blinding to the study arm if, for example, an experienced urologist had reviewed the medical records and an untrained research assistant had summarized the questionnaires and interviews. The authors state that a degree of interpretation had to be done in coding the prospective patient data. Unfortunately, we have no information about who the data collectors were. This may be a more important point than the potential differential misclassification the authors discuss.

Only a study of prospective versus retrospective information?

The strict protocol for reporting complications generating the retrospective data probably to most readers lies very close to an actual prospective data collection; the paper can be read as a comparison between two prospective protocols, one built on standardized digital medical records, one on patient-reported outcomes. A question arises if the differences in how the protocols perform pertain mostly to whether the data collection is pro- or retrospective, or to how they differ in the primary method of data collection. For example, do the caregivers and patients differ in which complications they value as important and worthwhile to report? Patients could have interpreted the questionaries as a positive sign that the institution cares about their views, which could influence the rating of complications.

The authors aimed to make the classification of complications similar for the two groups by standardizing the coding of the patient-reported data. Seen from a standpoint of wanting to increase the comparability of the two datasets, this is relevant. Seen from an intent to understand how the two protocols function in how they report important clinical information about complications, this may not be the ideal thing to do.

The investigation of these two data-collection protocols is relevant. Both theoretically have pros and cons and several hypotheses about how they might differ are possible. The authors do well in forwarding a hypothesis to be challenged; a clear hypothesis helps structure the analysis and the interpretation. The authors anticipate that more minor complications would be reported in the patient-reported outcomes. It would have been enlightening if the authors had discussed why they thought so, which could have raised relevant concern about their refutation of the hypothesis, given the data.

Should we care about quantitative estimates or *p*-values?

The reporting of the results in this paper relies heavily on statistical significance. This practice is regrettably common in the medical literature, and in many ways problematic. For an enlightening discussion about this, see Greenland et al. [5]. This is a small study with low statistical precision: the description of the dimensioning of the study is difficult to understand, but maybe the target was to detect a difference as large as between 50 and 85% complications. We are not sure.

In a small study, we run the risk of false positives, but we also have a very large risk of missing a clinically relevant difference [5,6]. If we look at the quantitative estimates, the level of minor complications for the prospective versus the retrospective study arm was 58 versus 46% at 14 days and 32 versus 18% at 90 days. If those were true, would that not be a relevant difference? The reader had been helped in her/his interpretation if the confidence intervals for these estimates had been included [5].

A discussion of the possibility of missing a clinically relevant difference in a small study should have been included in the interpretation of the results, especially since these data are at the heart of the hypothesis the authors forwarded for the study. The authors are convinced that their hypothesis is refuted, but the low statistical precision and the quantitative estimates raise a question mark for that conclusion. In the previous issue of this journal, Kaisa et al. [7] reported that Clavien-Dindo registration in medical records was less consistent for the milder complications after kidney cancer surgery. This lends further thought to that the caregivers may pick up fewer minor complications than the patients.

What did we learn?

Given the considerations above, it is uncertain how much we learned about pro- or retrospective data collection in general or for cystectomy. However, as a study in scientific design and interpretation, we learned several things. Although randomization, in this case, was not wrong, it was probably not the most effective study design. Letting the study subjects undergo both protocols for data collection would have been closer to the counterfactual study design. And a randomized study design is *per se* not enough to provide valid study results. In interpretation, a thorough characterization of the exposure is as important as defining the outcome to understand possible causal pathways. A focus on statistical significance for scientific inference should be discouraged. For interpretation, we need to look at the totality of the hypothesis, the study setting, data quality, statistical precision, quantitative estimation with information about uncertainty, and if the study results show consistent patterns, not only at the influence of random errors.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Schmidt Andersen C, Thoft Jensen B, Nielsen Holck E, et al. Prospective versus retrospective recordings of comorbidities and complications in bladder cancer patients undergoing radical cystectomy – a randomized controlled trial. Scand J Urol. 2021:1–6.
[2] Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22:189–212.
[3] Rothman KJ. Six persistent research misconceptions. J Gen Intern Med. 2014;29(7):1060–1064.
[4] Hernán MA, Hernández-Díaz S, Robins JM. Randomized trials analyzed as observational studies. Ann Intern Med. 2013;159(8):560–562.
[5] Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–350.
[6] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485.
[7] Kaisa E, Veitonmäki T, Ettala O, et al. Does every Clavien-Dindo complication matter? A national multi-center study in kidney cancer surgery. Scand J Urol. 2021;55(6):441–447.

Lars Holmberg
Professor emeritus
Translational Oncology & Urology Research (TOUR), School of Cancer and Pharmaceutical Sciences, King's College London, London, United Kingdom. Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

Lars.holmberg@kcl.ac.uk

Received 6 December 2021; Accepted 7 December 2021