JMRO

Journal of Medical and Radiation Oncology

Journal homepage: www.jmedradonc.org

In Pursuit of Meaningfulness in the Biomedical Literature – Notes from a Scrap Booklet

Monica-Emilia Chirilă[1], Søren M. Bentzen[2,3]

[1] *Clinical Development Department, MVision AI, Helsinki, Finland*
[2] *Department of Radiation Oncology, University of Maryland School of Medicine, Baltimore, USA*
[3] *Department of Epidemiology & Public Health, University of Maryland School of Medicine, Baltimore*

Corresponding author: Monica Emilia Chirilă; **e-mail:** monica.emilia.chirila@gmail.com

About our guest

Professor Søren M. Bentzen is the Director of the Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, USA, and he is also an Adjunct Professor of Radiobiology and Medical Physics, University of Copenhagen, Denmark.

Dr. Bentzen received a Master of Sciences in physics and mathematics, a PhD in medical image analysis, and he is a doctor of medical science in quantitative clinical radiobiology. He was one of the leaders of the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) initiative, and serves as a member of the Pediatric Normal Tissue Effects in the Clinic (PENTEC) steering committee. He was also a member of the American Society for Radiation Oncology (ASTRO) Task Forces to develop evidence-based guidelines on the appropriate fractionation for whole breast irradiation and for localized prostate cancer. He has served as chair or member of 17 (seven currently active) Independent Data and Safety Monitoring Committees or Trial Steering Committees and as co-chair for Translational Research for two Radiation Therapy Oncology Group (RTOG) phase III trials.

He has published more than 500 original papers and book chapters, has presented >370 invited lectures, and he currently serves on 10 international cancer journal editorial boards. His main research interests include bioeffect modeling; biomathematics; applied biostatistics; clinical trial design; evidence-based medicine; late effects of radiotherapy; clinical radiobiology; integration of data from genomics, proteomics, and molecular imaging into novel therapeutic

Interviews

https://doi.org/10.53011/JMRO.2023.02.11

strategies. He received more than 30 awards and honors by now, including the ASTRO Gold Medal and European Society for Radiation Oncology (ESTRO) Breuer Gold Medal (1).

This interview is meant to present in an interactive form some general elements that could help young clinicians to develop a critical approach when reading a medical research manuscript.

Keywords: *interview, radiation oncology, clinical trials, statistics, critical thinking, medical literature*

MC: Dear Professor Bentzen, you are, among many other roles, the Course Director of the ESTRO teaching course on Quantitative Methods in Radiation Oncology: Models, Trials and Clinical Outcomes. One of the topics presented during the course was "Critical Appraisal". I believe the ideas and concepts you mentioned during the lecture would be useful for young clinicians and researchers. What would be in your view the first rule for critically assessing a paper?

SMB: Read the paper – the whole paper. Sometimes, totally different statements are found in the title, in the abstract or in the body of the paper, on the same topic. We are overwhelmed by the quantity of published papers. In 2020, there were more than 30 million records in Pubmed, with 800 000 new records added each year. The abstract is the part of the paper which has the biggest chance of being read. An interesting analysis was done by Pitkin *et al*, assessing random samples of 44 articles published during one year in each of five major medical journals at that time: *Annals of Internal Medicine*, *British Medical Journal*, *Journal of American Medical Association, Lancet* and *New England Journal of Medicine*, and 44 consecutive papers in the *Canadian Medical Association Journal*. Abstracts were considered deficient if they contained data that were inconsistent with data in the article or not found in the body of the paper at all. The proportion of the deficient abstracts varied from 18% to 68% (2). It is not evident whether the situation has improved since then; so, read the whole paper not just the abstract.

MC: This makes citation based on abstracts a risky bet, isn't it? What do you think about citing a paper based on its citation in another paper?

SMB: Never use secondary citations. This is another trap. Just an example to illustrate this is a paper by Evans *et al*, in which they evaluated fifty randomly selected references from a single months' issue of three journals: the *American Journal of Surgery*; *Surgery, Gynecology and Obstetrics*; and *Surgery.* A major error of citation was assigned if the referenced article failed to substantiate, was unrelated to or even contradicted the assertion made by the quoting authors. Of the fifty randomly selected citations, thirteen major and 41 minor citation errors were found (3).

MC: We agree that we have to read the paper, but how to read it? What should we take into account?

SMB: It would be impossible to give a miraculous recipe or to cover all the details in a limited amount of text. A simple checklist could look like this:

- **Design – Is it appropriate for the question?**
- **Patients – Are they representative? Is there a selection bias?**
- **Interventions – Are they well defined and feasible? What was the compliance?**
- **Outcome measures (endpoints) – Are they relevant and clearly defined? Are any missing? Many oncology papers do not give quantitative data on adverse events (4).**
- **Results – How are they analyzed and presented?**

MC: Speaking about patients included in clinical trials, I remember discussions regarding the comparison of patients from the real world and patients from clinical trials. What is your opinion on this topic?

SMB: The cohorts would need both **internal and external validity**. The internal validity represents the extent to which the observed difference between groups can be correctly attributed to the intervention and it is threatened by bias and sampling variability. This is why we give the highest weight to data from

randomized controlled trials. The external validity is the extent to which study findings can be generalized to a population. Trial populations are often selected in order to limit the variability among cases within the trial – but this may make the trial population less representative of unselected cases seen in the clinic.

Another important element to evaluate is the **Intention-to-treat principle**. Are the patients analyzed as members of the treatment group to which they were randomized, regardless of actual treatment? Randomization only controls bias at the time of randomization. This should be the primary analysis of the outcome of any trial.

MC: I understand that the more changes are made to the initial plan, the more errors and bias can occur. Sometimes we read about results from sub-group analysis. What are your thoughts about their relevance?

SMB: That depends if the sub-group analysis was planned from the beginning or not. The primary clinical endpoint would have to be selected before the trial starts and investigators should be blinded to accumulating data. **We make our bet before the race begins**.

A retrospective cohort study on data from protocols and publications of 62 clinical trials showed that information on sample size calculations and statistical methods were often not specified or had unacknowledged discrepancies between what was in the protocol and what was included in the publication of the results (5).

MC: It seems that it would be necessary that a clinician have solid statistics knowledge, in order to spot the flaws. What do you think?

SMB: I don't think you need a degree in statistics to read a paper in a medical journal. However, there are multiple statistical methods used in medical papers and having at least a basic understanding of these would be useful. Also, there are a number of pitfalls in the analysis and interpretation of clinical data and it is extremely helpful also for biologists and clinicians to be aware of many of these. An example is the analysis of Marsh and Hawkins, dating from almost three decades ago, when they analyzed 44 reports on multicenter randomized controlled trials (RCTs). If someone would have wanted to understand at least half of them, it would have had to be familiar with 11 statistics techniques, including t-tests, power estimation, life tables, survival statistics, the Cox regression model and analysis of variance (6). Anyway, since then a variety of statistical methods became even more widely used in medical research, in part as a result of the availability of inexpensive statistical software packages. However, you still need input from biostatisticians in the context of the increasing amount of available data. According to the US Bureau of Labor Statistics, data scientists and statisticians are among the top 10 fastest growing occupations estimated for 2021-2031 (7).

MC: That's good news for the clinicians and researchers – when starting a clinical trial we will have a better chance to design it well if we will have a statistician within the team.

SMB: Sir Ronald A. Fisher - a British mathematician, statistician, biologist and geneticist, considered as one of the founders of modern statistical science – stated that: "To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination – he may be able to say what the experiment died of." Talk to a statistician, already from the design stage,

MC: Continuing the parallel from this inspiring quote, what are the "signs and symptoms" that we would have to assess when examining a "body of evidence"?

SMB: It would depend on the degree of complexity of the "case". We can mention some simple aspects that should not be missed.

Ideally, a RCT should be designed to answer a specific question. Two types of errors can occur – concluding that the hypothesis is false, when it is not, or the opposite. The size of the trial strongly influences its power and the statistical significance of an observed effect. However, this also means that the *P* value itself can give a false impression of the importance of trial findings. C**onfidence intervals around the estimated effect size** are so much more informative that just the *P* value (8).

However, there are observational, or registry studies in which a large amount of data is available. Here, we can fall into another trap. When the sample size is large enough, a very small difference becomes

statistically significant. Large databases like the US National Cancer Data Base (NCDB), which includes patient and treatment data from more than 1500 cancer programs and more than 34 million unique cases, is such an example of a "tall data" source. In the "tall data", the number of cases is high and the number of tested features is low. A proof of concept was done testing the hypothesis that the incidence of nodal metastasis as a biomarker in two samples of one million women. The *P* was < 0.05 and the confidence interval was ± 0.3%, for a real difference of 0.15% (9). We should always apply the clinical judgment and sound critical sense when reading results. Numbers are important, but **a difference is only a difference… if it makes a difference.**

MC: The same facts can be presented in a more or less "appealing" way, which would lead to a different impact on the viewer.

SMB: Exactly. Another thing that we should pay attention to is to the way the data are presented. Relative risk estimates are often more impressive than absolute risk estimates. A questionnaire presented to respondents gave the same results in two different ways: as an absolute change in outcome rates (overall mortality reduction from 7.8% to 6.3% meaning a difference of 1.5%) or a relative reduction (of 20.3%). Both differences were stated as statistically significant, but almost half of the respondents perceived them as different and almost 90% of those were inclined to change their clinical practice based on the relative change in the outcome rate (10).

MC: Are there other factors that would lead to the overestimation of the importance of a certain treatment, except for our personal interpretation of data?

SMB: We can think of publication bias. There is a higher tendency of reporting statistically significant effects and there might be also a delay in publishing negative results. There is also the risk of duplicate publication of data from trials demonstrating larger treatment effects. A classic example is ondasetron. A metaanalysis including duplicated data overestimated the efficacy of ondasetron by 23%

MC: What about studies that show a strong correlation between two variables?

SMB: Correlation does not necessary imply causality. Some statistical results would support hilarious conclusions, if we would take correlations as causality. There is a significant correlation between stork populations and human birth rates across Europe (P=0.008) (11) and between the divorce rate and margarine consumption in Maine, USA (correlation 0.9925, P<0.0001) (12). Clearly, these are not meaningful findings.

MC: How should we read and understand a positive trial? Could it have a false positive conclusion?

SMB: There could be sources of errors in positive trials. Multiple comparisons due to multiple endpoints, repeated looks and subgroup analysis could impair the chances of a correct result in the absence of adequate statistics. Bias coming from post-randomization exclusion or other causes should be avoided. Even if the trial is randomized, there could still be biases introduced by, say, differences in diagnostic intensity or in supportive care, that cannot simply be attributed to the treatment itself.

MC: What about negative trials?

SMB: Absence of proof is not proof of absence. Except for the sample size, there are other factors that would lead to a type II error – failing to detect a real difference between two interventions. We can mention heterogeneity (in patient populations, diagnostic, quality of care and follow-up procedures), short follow-up or high rate of loss of patients from follow-up.

MC: Is it mandatory to have a randomized controlled trial to support the implementation of a new treatment or technique?

SMB: Randomized controlled phase III trials are the gold standard for assessing differential benefits of two therapies. However, not all the current clinical practice is based on results of such trials. Leaving aside the anecdotic lack of evidence of the benefit of the use of parachutes (13) when jumping out of airplanes, there is no evidence from RCTs demonstrating the clinical benefit of simulators, CT-based

treatment planning, mega-voltage radiotherapy or other elements from our daily practice. However, if the aim is to show a better outcome from a technological innovation the need for randomized controlled trials remains. We are used to this in the case of a novel medication meant to improve outcome. This is equally valid for technology, but the immediate goal of a technical innovation is often to improve treatment quality, which may be demonstrable without randomization. We may fell that improved quality will impact the outcomes. But since survival outcomes are influenced by multiple factors, using them as indicators of treatment quality may have a low sensitivity and specificity. When deciding if a RCT is really needed and ethical, we should take into account the principle of equipoise, a balance of benefits and risks in both arms of the study. There are situations when we have reasons to believe that treatments would not be equivalents in terms of safety, so it would be unethical to enroll patients. Non-randomized or "observational" studies should be seen as a complement to RCTs (14).

MC: Do we still need to critically read a paper which has already gone through peer-review?

SMB: Peer review is a quality stamp, and not a warranty. We could give an example: Godlee *et al* introduced 8 deliberate errors or weaknesses in a manuscript ready for publication in British Medical Journal and sent it out for review. More than 200 potential reviewers responded, but the median number of spotted errors was 2, nobody commented on more than 5, and 1 in 6 reviewers did not comment on a single one. The rate of error detection was not modified by asking them to sign the reports or blinding them to the authors and origin of the paper (15).

MC: How can study quality or strength of evidence be objectively assessed?

SMB: The simple answer is: it cannot! There are multiple quality instruments and guidelines. Twenty years ago there were already 20 systems for evaluation of systematic reviews, 49 for RCTs and 40 for grading the strength of a body of evidence. Quality scores can be applied when evaluating the results of a trial. But although many of these instruments make some sense, there is not a single silver bullet that works across the whole range of trials. As Andrew S. Tannenbaum, an American-Dutch computer scientist, said: **The nice thing about standards is that there are so many of them to choose from.**

Abbreviations:
QUANTEC - Quantitative Analyses of Normal Tissue Effects in the Clinic
ASTRO - American Society for Radiation Oncology
RTOG - Radiation Therapy Oncology Group
ESTRO - European Society for Radiation Oncology
RCT - randomized controlled trial

Statements:
Authors' contribution: MEC drafted the text and SB reviewed and edited the text.
Conflict of interest: None
Funding: None

References:

1. University of Maryland School of Medicine. Bentzen S. [Internet]. Available from: https://www.medschool.umaryland.edu/profiles/Bentzen-Soren/

2. Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. JAMA. 1999;281(12):1110-1111. doi: 10.1001/jama.281.12.1110

3. Evans JT, Nadjari HI, Burchell SA. Quotational and reference accuracy in surgical journals. A continuing peer review problem. JAMA. 1990;263(10):1353-1354.

4. Vittrup AS, Kirchheiner K, Fokdal LU, et al. Reporting of Late Morbidity After Radiation Therapy in Large Prospective Studies: A Descriptive Review of the Current Status. Int J Radiat Oncol Biol Phys. 2019;105(5):957-967. doi: 10.1016/j.ijrobp.2019.08.040

5. Chan AW, Hróbjartsson A, Jørgensen KJ, Gøtzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. BMJ. 2008;337:a2299. doi: 10.1136/bmj.a2299

6. Marsh MJ, Hawkins BS. Publications from multicentre clinical trials: statistical techniques and the accessibility to the reader. Stat Med. 1994;13:2393–2406.

7. U.S. Bureau of Labor Statistics. Fastest Growing Occupations [Internet]. Available from: https://www.bls.gov/ooh/fastest-growing.htm

8. Bentzen SM. Towards evidence based radiation oncology: improving the design, analysis, and reporting of clinical outcome studies in radiotherapy. Radiother Oncol. 1998;46(1):5-18. doi: 10.1016/s0167-8140(97)00226-0

9. American Society for Radiation Oncology (ASTRO). ASTROnews 2016 Annual Meeting Guide [Internet]. Available from:https://www.astro.org/uploadedFiles/MAIN_SITE/News_and_Publications/Magazine(ASTROnews)/Volumes/ASTROnews_%20AMG16.pdf

10. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. Am J Med. 1992;92(2):121-124. doi: 10.1016/0002-9343(92)90100-p

11. Matthews R. Storks Deliver Babies. Teaching Statistics. 2000;22:36-28. doi: 10.1111/1467-9639.00013

12. Business Insider. Spurious Correlations by Tyler Vigen [Internet]. Available from: https://www.businessinsider.com/spurious-correlations-by-tyler-vigen-2014-5

13. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ. 2003;327(7429):1459-1461. doi: 10.1136/bmj.327.7429.1459

14. Bentzen SM. Randomized controlled trials in health technology assessment: overkill or overdue? Radiother Oncol. 2008;86(2):142-147. doi: 10.1016/j.radonc.2008.01.012

15. Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. JAMA. 1998;280(3):237-240. doi: 10.1001/jama.280.3.237