European Spine Journal (2018) 27:977–979
https://doi.org/10.1007/s00586-018-5558-4

LETTER TO THE EDITOR

Generalizing the results: how can we improve our reports?

Mikhail Saltychev[1] · **Merja Eskola**[2]

Received: 22 January 2018 / Accepted: 22 March 2018 / Published online: 26 March 2018

The most common goal of clinical research is to measure the probable effect of the intervention in clinical practice. Unfortunately, reports are often limited to describing only a studied sample and not an entire population itself. We were interested to investigate how well reports on high-end randomized controlled studies in the field of surgery present information regarding a population outside the studied sample.

The search on PubMed was conducted in January 2018 using the clause: "Eur Spine J"[Journal] AND ((Controlled Clinical Trial[ptyp] OR Randomized Controlled Trial[ptyp]) AND ("2017/01/01"[PDAT]: "3000/12/31"[PDAT])). The search resulted in nine randomized controlled trials published in European Spine Journal in 2017 [1–9]. The main outcome measures and statistics reported in the identified trials are shown in Table 1. Of the nine trials, only two employed 95% confidence intervals (95% CIs) to describe the results obtained from ordinal or continuous outcomes [6, 8]. When dealing with rate of event (reoperation, adverse effect, etc.), only one study used a relative risk ratio (RR) and a 95% CI to describe the findings [3]. Otherwise, continuous and ordinal outcomes were reported as means, standard deviations (SDs), and *p* values, and in some cases, additionally, as ranges, absolute numbers, and percentages. The rates of event were mostly reported as absolute numbers and/or percentages. Only one study used minimal clinically important difference (MCID) to evaluate the clinical significance of the results [6].

The use of only descriptive statistics and omitting the clinical significance of the results may leave clinically important implications unnoticed. We would like to draw attention tothe following statistical considerations:

- Absolute number of event, mean, standard deviation, and percentage describe precisely a studied sample. However, they fail to predict estimates that may occur if a different sample is drawn randomly from a targeted population.
- *p* value describes if the null hypothesis ('no difference between groups') is true or not. However, 'statistically significant' *p* value only tells that difference does exist. It does not state how much different the two groups are and how that difference behaves if different samples would randomly be selected from a population [10].
- The same problem arises from the use of absolute rates or percentage of event—no information can be obtained from these estimates regarding the overall situation in an entire population.
- Statistical significance does not necessarily mean that the results are significant clinically (in fact, they rarely are). The observed statistically significant result should be compared against MCID for a particular outcome (if available) whenever possible.

All the reports that are examples in this study were well conducted randomized controlled trials with important objectives and impressive samples. However, they reported the characteristics of the samples instead of the populations they represent. This may leave an uncertainty if the effect observed in a sample works on an entire population as well. Such a situation leaves the responsibility of generalization and practical use of the results to the readers—a task that may be overwhelming for many clinicians who do not possess a substantial statistical know-how.

To illustrate our point, let us introduce an example. Gibson et al. [3] reported that "…affected side leg pain was lower in the TED group [intervention group] at 2 years (1.9 ± 2.6 vs 3.5 ± 3.1, $p = 0.002$)…" and later "…pain in the affected leg was significantly better in the TED group at 2 years…". These statistics tell that there was a statistically significant difference of 1.6 points observed between the average scores of two particular groups (70 subjects in each one). However, these statistics do not tell if the result will persist if groups with different individuals will be randomly

✉ Mikhail Saltychev
mikhail.saltychev@gmail.com

[1] Department of Physical and Rehabilitation Medicine, Turku University Hospital and University of Turku, PO Box 52, 20521 Turku, Finland

[2] Expert Services, Turku University Hospital and University of Turku, Turku, Finland

Table 1 Main outcomes and statistics reported in the trials

Study	Comparison	Main outcomes	Main statistics
Bono [1]	2-week vs. 6-week restrictions after lumbar discectomy	VAS, ODI	Mean, SD, *p* value
		Reherniation rate	Event %
Fukui [2]	Two different devices for hemostasis	Blood loss, surgery duration	Mean, SD, range, *p* value
Gibson [3]	Transforaminal endoscopic discectomy vs. microdiscectomy	VAS, ODI length of stay, SF-36	Mean, SD, range, *p* value
		Complication rate	RR, 95% CIs
Høy [4]	Two lumbar fusion techniques	DPQ, SF-36, LBPQ, ODI	Mean, SD, *p* value
		Reoperation rate, global satisfaction	Event %, *p* value
Hung [5]	Two wound drainage techniques in lumbar fusion	VAS, ODI, blood loss, timing of ambulation, LOS	Mean, SD, *p* value
		Fusion achievement rate	Event %, *p* value
Krappel [6]	Discectomy alone vs. discectomy combined with spinal stabilization system	VAS, ODI, SF-36, blood loss, surgery duration, LOS	Mean, SD, 95% CI, and *p* value MCID pain
		Adverse events	Event %
Sturesson [7]	SI joint fusion vs. conservative management	VAS, ODI, SLR, EQ-5D-3L, walking distance, global satisfaction	Mean, SD, *p* value, absolute number, %
		Adverse events	Absolute number, mean
Sundseth [8]	Cervical arthroplasty vs. fusion	VAS, NDI, surgery duration	Mean, 95% CI, SD, *p* value
		Reoperation rate	Absolute number
Yilmaz [9]	Two physiotherapy programs	VAS, TKS, ODI, TUG, 6-MWT, NHP	Mean, SD, *p* value

VAS pain visual analogue scale, *ODI* Oswestry disability index, *LOS* length of stay in hospital, *SF-36* short form health survey, *NDI* Neck disability index, *TUG* time-up-go test, *6-MWT* 6-Minute walk test, *NHP* Nottingham health profile, *LBPQ* Low Back Pain Questionnaire, *DPQ* Dallas Pain Questionnaire, *TKS* Tampa kinesiophobia scale, *SD* standard deviation, *95% CI* 95% confidence interval, *RR* relative risk ratio

drawn from the population of interest. For other groups, the difference will probably be sometimes below and sometimes over 1.6 points. The highest possible and the lowest possible limits of this variation form a confidence interval. Confidence interval can be calculated from raw data if available or approximated from the reported means and standard deviations. For the math enthusiasts, it would be like the formulas below. Fortunately, such statistical methods are included by almost every statistical software or even simple online calculators:

$$\text{Difference in means} = \text{Mean}_1 - \text{Mean}_2$$

$$95\%\ \text{CI} = (\text{Mean}_1 - \text{Mean}_2) \pm 1.96 \times \sqrt{\frac{(n_1 - 1) \times \text{SD}_1^2 + (n_2 - 1) \times \text{SD}_2^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the number of subjects in two groups

In our example, the result would be 1.6 (95% 0.64–2.56, *p* value 0.001) points. This is the effect size of the study. It tells us that even if we randomly select limitless number of additional samples from the studied population, the mean difference of each additional sample will be (with 95% certainty) between 0.64 and 2.56 points of pain VAS. Even further, our next step would be to compare the obtained figures with MCID for pain VAS (which is usually set at least 1.5 points). In our example, the lower limit of 95% CI is 0.6—way below the level of MCID. Thus, we must conclude that even if the difference between groups was statistically significant (p value < 0.05), it was insignificant clinically.

Another common outcome in surgical and other medical research is the rate of some events (reoperation, adverse effect, death, etc.). As another example, Sundseth et al. reported that "…one patient in the fusion group and eight in the arthroplasty group had undergone index level reoperation, $p = 0.03$…" [8]. The absolute proportion looks very high 1/8, but, once more, it describes only these two particular groups. To define the possible situation in the entire population (if different random samples will be drawn), additional statistical methods should be employed—in this case, the most common would probably be a relative risk ratio (RR) along with its confidence interval. The re-calculation done by us ended up with the following results: RR = 7.7 (95% 0.98–59.8, *p* value 0.052) and NNT (here, number needed to harm) 10.5 (95% 58.1–5.77). Thus, even though the proportion 1/8 seemed to be high, there was not a statistically (and, therefore, also clinically) significant risk on reoperation with number of patients 'needed to harm' (this is a term for NNT when harmful event is involved) been 58—meaning there is a chance that even 58 patients could be treated before one will need a reoperation. In other words, there is a 95% probability that arthroplasty does not have a higher risk of reoperations compared to a fusion.

Descriptive statistics is a good starting point of analyzing the results obtained in a trial. With rare exceptions, it should not, however, be the endpoint of the analysis. We strongly advice reporting effect sizes (that are a measure of clinical effectiveness) including 95% CIs additionally to *p* values and RRs additionally to percentages. This is what clinicians look for, not only information on subjects in a particular study, but practical and clinically meaningful suggestions to deal with any patient entering their clinic.

Compliance with ethical standards

Funding None to declare.

Conflict of interest None of the authors has any potential conflict of interest.

References

1. Bono CM, Leonard DA, Cha TD, Schwab JH, Wood KB, Harris MB et al (2017) The effect of short (2-weeks) versus long (6-weeks) post-operative restrictions following lumbar discectomy: a prospective randomized control trial. Eur Spine J 26(3):905–912
2. Fukui D, Kawakami M, Nakao SI, Miyamoto E, Morishita S, Matuoka T et al (2017) Reduced blood loss and operation time in lumbar posterolateral fusion using a bipolar sealer. Eur Spine J 26(3):726–732
3. Gibson JNA, Subramanian AS, Scott CEH (2017) A randomised controlled trial of transforaminal endoscopic discectomy vs microdiscectomy. Eur Spine J 26(3):847–856
4. Hoy K, Truong K, Andersen T, Bunger C (2017) Addition of TLIF does not improve outcome over standard posterior instrumented fusion. 5–10 years long-term follow-up: results from a RCT. Eur Spine J 26(3):658–665
5. Hung PI, Chang MC, Chou PH, Lin HH, Wang ST, Liu CL (2017) Is a drain tube necessary for minimally invasive lumbar spine fusion surgery? Eur Spine J 26(3):733–737
6. Krappel F, Brayda-Bruno M, Alessi G, Remacle JM, Lopez LA, Fernandez JJ et al (2017) Herniectomy versus herniectomy with the DIAM spinal stabilization system in patients with sciatica and concomitant low back pain: results of a prospective randomized controlled multicenter trial. Eur Spine J 26(3):865–876
7. Sturesson B, Kools D, Pflugmacher R, Gasbarrini A, Prestamburgo D, Dengler J (2017) Six-month outcomes from a randomized controlled trial of minimally invasive SI joint fusion with triangular titanium implants vs conservative management. Eur Spine J 26(3):708–719
8. Sundseth J, Fredriksli OA, Kolstad F, Johnsen LG, Pripp AH, Andresen H et al (2017) The Norwegian Cervical Arthroplasty Trial (NORCAT): 2-year clinical outcome after single-level cervical arthroplasty versus fusion-a prospective, single-blinded, randomized, controlled multicenter study. Eur Spine J 26(4):1225–1235
9. Yilmaz Yelvar GD, Cirak Y, Dalkilinc M, Parlak Demir Y, Guner Z, Boydak A (2017) Is physiotherapy integrated virtual walking effective on pain, function, and kinesiophobia in patients with non-specific low-back pain? Randomised controlled trial. Eur Spine J 26(2):538–545
10. Ranstam J (2012) Why the *P*-value culture is bad and confidence intervals a better alternative. Osteoarthr Cartil 20(8):805–808