Toward standardization of statistical reporting in studies on entheseal changes

Pas, S. van der; Schrader, S.A.

Received: 6 September 2022 | Revised: 24 October 2022 | Accepted: 26 November 2022

DOI: 10.1002/oa.3188

SPECIAL ISSUE PAPER

Toward standardization of statistical reporting in studies on entheseal changes

Stéphanie van der Pas[1,2] | **Sarah Schrader**[3]

[1]Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2]Methodology, Amsterdam Public Health, Amsterdam, The Netherlands

[3]Faculty of Archaeology, Leiden University, Leiden, The Netherlands

Correspondence
Stéphanie van der Pas, Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit Amsterdam, De Boelelaan 1117, Amsterdam, The Netherlands.
Email: s.l.vanderpas@amsterdamumc.nl

Funding information
Dutch Research Council (NWO), Grant/Award Numbers: VI.Veni.192.087, VI.Vidi.201.153

Abstract

Statistical analysis, while at first glance an objective way to extract insights from data, remains at its core a human endeavor. Elements of subjectivity are introduced by the many decisions that go into the selection of a statistical method. Such subjectivity may harm the evidentiary value of results from statistical analyses. Standardization of statistical methods decreases the degrees of freedom available to researchers and may thus be seen as a way to increase the objectivity of the analysis. Here, we argue that standardization of methods is not only impossible because statistical methods rely on assumptions that need to be considered on a case-by-case basis but also undesirable because it may block innovation. We propose that the entheseal changes field is better served by standardization of reporting and discuss how reporting guidelines may be developed based on examples from biostatistics.

KEYWORDS
entheseal changes, reporting guidelines, statistics

1 | INTRODUCTION

Statistical methods for the analysis of entheseal changes are selected by humans, who often have multiple options available. Standardizing the choice of method for a particular research question—so that a particular research question is always accompanied by the same statistical method—removes a source of variation between analyses and may be seen as a way to increase objectivity of the resulting analysis. Although removing "researcher degrees of freedom" (Simmons et al., 2011) has its merits, we argue that reliability and trustworthiness of statistical results are not increased by standardizing which methods to use but rather by standardizing how statistical analyses are reported.

2 | AGAINST STANDARDIZATION OF STATISTICAL METHODS

A threat to validity of research results is presented by the many options researchers have to analyze their data. Without sufficient statistical education, this flexibility may lead a researcher down a path of cherry-picking methods that give the most desirable results (Hoffmann et al., 2021). Fully standardizing the statistical method to be used with each research question could neutralize this threat. Yet, there is no such thing as a single best statistical method to answer a particular research question. Even if statistical procedures have ostensibly the same goal, they typically differ in underlying assumptions. The extent to which such assumptions are met needs to be assessed per data set.

As a simple example, consider the choice between an unpaired *t*-test and a Mann–Whitney *U* test. Both tests are used to compare continuous outcomes between two groups. The *t*-test is designed for normally distributed data, whereas the Mann–Whitney *U* test requires no such assumption. If the normality assumption is reasonable, it is preferable in terms of power to use the *t*-test. If there are doubts about the normality, it is better to select the Mann–Whitney *U* test (Fay & Proschan, 2010). For entheseal changes, osteoarchaeologists have frequently used Mann–Whitney *U* to compare scores between two populations (see He & de Almeida Prado, 2020; Mountrakis & Manolis, 2015; Schrader, 2014; Thomas, 2014). The application of Mann–Whitney *U* is often justified because of the use of an ordinal scoring system or, sometimes, no explanation is provided.

The most appropriate method depends not only on the data structure but also the research question. Even a precise research question can leave room for interpretation, and how this room is used will depend on the available data. For example, suppose we wish to address the question: "Does Population A or Population B have higher entheseal scores?" Some aspects to consider while selecting a statistical method would include the sample size, state of preservation of both populations (is there missing data?), chronology of each population, extent of data aggregation (e.g., by joint and by laterality), age distribution, data collection method, interobserver error, and more. Depending on such aspects, the method of choice could be a Mann–Whitney *U* test but also a multilevel regression method with multiple imputations.

The multitude of ways in which data sets and research questions can be different from each other, sometimes even subtly so, is a key reason why strict standardization of methods is impossible. And even if these sources of variation are taken away, the varying assumptions of statistical methods still allow for multiple approaches to a research question. This phenomenon is illustrated by a study where 29 research teams were given the same data set and were asked to answer the same research question: "Are soccer teams more likely to give red cards to dark-skin-toned players than to light-skin-toned players?" The 29 teams used 29 unique analyses with 21 unique combinations of covariates and found a statistically significant effect in 20 cases (Silberzahn et al., 2018).

Even when some standardization is possible, there is a risk in fully standardizing a statistical approach. Standardization of methods may block innovation, as it could slow the adoption of new statistical methods. For example, in orthopedic research, it is common to study the time to revision of hip prosthesis—this is critical research as surgeries, such as hip replacements, are a burden to the often elderly patients, and it is preferred to select prostheses and surgical approaches so that the prosthesis lasts as long as possible. This is a survival outcome ("How long until …?"), and traditionally, this question is studied by Kaplan–Meier (Kaplan & Meier, 1958). Kaplan–Meier is only designed for one terminal outcome (in this case, revision surgery), and in this typically elderly, population, a second terminal outcome is relevant, namely, death (Gillam et al., 2011; Van der Pas et al., 2017, 2018). There are statistical methods available that can include both death and revision surgery as an outcome, such as the Fine–Gray competing risks estimator (Fine & Gray, 1999) or multistate models (e.g., Andersen et al., 2002), the latter of which has the additional advantage of being able to include multiple prostheses per patient (Gillam et al., 2012). Adoption of these methods has turned out to be slow and quite controversial (Lacny et al., 2015, 2021; Ranstam & Robertsson, 2017; Sayers et al., 2018; Van der Pas et al., 2018), one commonly given reason is that it is harder to compare results across studies when methodologies different from Kaplan–Meier are used.

3 | STANDARDIZATION OF REPORTING

Rather than standardizing analytical methods, we propose that the entheseal changes field is better served by standardizing reporting of statistical analyses. This is the approach advocated in medical statistics, where guidelines for various study designs already exist and new ones continue to be developed (see the EQUATOR network for a large collection of guidelines, https://www.equator-network.org). Such guidelines do not prescribe what method to use but list which issues to consider and transparently report in the paper. In addition, the field could benefit from the STRengthening Analytical Thinking for Observational Studies (STRATOS) (Sauerbrei et al., 2014) initiative, which aims to guide authors through their choice of method and produces guidance documents focusing on issues like missing data and measurement error. Following such guidelines helps in striking a balance between avoiding research (mal)practices where too many researcher degrees of freedom are exercised (Simmons et al., 2011) while still leaving room for innovation in statistical methods.

As an illustration, we consider the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guideline (Von Elm et al., 2007). STROBE is meant for observational studies and has specifically been designed for cohort, case–control, and cross-sectional studies. It consists of a checklist and an elaboration and explanation document (Vandenbroucke et al., 2007). The checklist briefly summarizes 22 statistical aspects to consider and report. For example, item 6 ("Participants") requests to give eligibility criteria of participants, the sources and methods of selecting participants, and methods of follow-up. Item 12 ("Statistics") consists of five subitems, including a request to describe how missing data were handled and to describe any methods used to examine subgroups and interactions. Some journals, such as *Medicine*, request a filled-in copy of the checklist if appropriate for the study design, indicating on which page each item can be found. The elaboration and explanation document offers further guidance to researchers, giving options to fulfill each criterion and suggestions on what language to use.

STROBE was created as a collaborative initiative of researchers including epidemiologists, methodologists, statisticians, and journal editors. It is an ongoing process, and STROBE is revised based on comments and new insights. For studying entheseal changes, existing guidelines like STROBE could be taken as the basis for a new guideline. Ideally, researchers with experience in statistical analysis of entheseal changes data will come together to discuss which specific

issues they encountered and how they handled them. As many statistical issues are common to both medical statistics and the study of entheseal changes, perhaps an "add-on" to STROBE can be developed with some extra items on the checklist and an "explanation and elaboration" file for the issues which are common and specific to the study of entheseal changes. Other guidelines, such as the TRIPOD statement for prediction models, could be relevant as well.

In the workshop from which this special issue stems and later discussions, some topics of specific interest that came up and which could be included in entheseal changes-specific reporting guidelines were the following:

- Multiple observations from one person and, depending on the method of choice (e.g., Henderson et al., 2016), one enthesis (dependent data);
- The method used for examining entheseal changes (e.g., Hawkey & Merbs, 1995; Henderson et al., 2016; Karakostis & Harvati, 2021);
- Multiplicity correction in case many tests were carried out on the same data;
- Data aggregation (e.g., by side and joint) calculation methods and whether nonaggregated as well as aggregate scores were available;
- Giving effect estimates and quantifying uncertainty, rather than relying on *p*-values only (Smith, 2020);
- Missing data; in particular patterns in the missing data (e.g., which entheses had the highest percentage of missing data) and whether any missing data were removed and in what granularity (entheses or individual and pairwise or listwise);
- Age-at-death estimation: How was age considered when examining entheseal changes? If age at death could not be estimated, how were the data for that individual handled?;
- Similarly, sex estimation: What technique was used and how was this considered when analyzing entheseal data, if at all? If sex could not be estimated, how were the data for that individual handled?; and
- Open science practices (preregister; share code or specific steps taken in statistical analysis; sharing data when possible).

Creating a guideline specifically for entheseal changes requires significant community effort, which does not end at creating the guideline. To be able to implement the guidelines, researchers will also need sufficient training or collaborations with statisticians. This is a topic we feel can receive more attention in graduate training programs.

4 | CONCLUSION

Standardization of methods does not do justice to the many (subtle) ways in which research questions and data availability can differ and has the potential to block innovation. The *Adaptive Tools for Resilient Bones* workshop from which this special issue stemmed highlighted the variability of research questions for which entheseal changes are applied, but also, the wide array of statistical tests that can be applied to these data. Although we foresee the standardization of statistical tests as unrealistic and potentially detrimental to the field, we also wish to close the door to practices where one is essentially fishing for results (Hoffmann et al., 2021; Simmons et al., 2011). Instead, we advocate for standardization when reporting data and consulting guidelines to select statistical methods. Standardization of reporting increases transparency and ensures that aspects relevant to assessing the evidentiary value of an analysis are included. The effort required from the community to design and maintain reporting guidelines is significant but may be reduced by building on existing guidelines like STROBE.

ACKNOWLEDGMENT

This research received funding from the Dutch Research Council (NWO) under grant agreements VI.Veni.192.087 and VI.Vidi.201.153.

CONFLICT OF INTEREST

None.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Stéphanie van der Pas https://orcid.org/0000-0002-2448-5378
Sarah Schrader https://orcid.org/0000-0003-0424-6748

REFERENCES

Andersen, P. K., Abildstrom, S. Z., & Rosthøj, S. (2002). Competing risks as a multistate model. *Statistical Methods in Medical Research*, *11*(2), 203–215. https://doi.org/10.1191/0962280202sm281ra

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. https://doi.org/10.1214/09-SS051

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*(446), 496–509. https://doi.org/10.1080/01621459.1999.10474144

Gillam, M. H., Ryan, P., Salter, A., & Graves, S. E. (2012). Multistate models and arthroplasty histories after unilateral total hip arthroplasties: Introducing the summary notation for arthroplasty histories. *Acta Orthopaedica*, *83*(3), 220–226. https://doi.org/10.3109/17453674.2012.684140

Gillam, M. H., Salter, A., Ryan, P., & Graves, S. E. (2011). Different competing risks models applied to data from the Australian Orthopaedic Association National Joint Replacement Registry. *Acta Orthopaedica*, *82*(5), 513–520. https://doi.org/10.3109/17453674.2011.618918

Hawkey, D. E., & Merbs, C. F. (1995). Activity-induced musculoskeletal stress markers (MSM) and subsistence strategy changes among ancient Hudson Bay Eskimos. *International journal of Osteoarchaeology*, *5*(4), 324–338. https://doi.org/10.1002/oa.1390050403

He, L. R., & de Almeida Prado, P. S. (2020). An evaluation of the relationship between the degree of entheseal changes and the severity of osteodegenerative processes at fibrocartilaginous entheses. *The Anatomical Record*, *304*, 1255–1265. https://doi.org/10.1002/ar.24541

Henderson, C. Y., Mariotti, V., Pany-Kucera, D., Villotte, S., & Wilczak, C. (2016). The new "Coimbra method": A biologically appropriate method

for recording specific features of fibrocartilaginous entheseal changes. *International Journal of Osteoarchaeology*, *26*, 925–932. https://doi.org/10.1002/oa.2477

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A. L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925. https://doi.org/10.1098/rsos.201925

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Karakostis, F. A., & Harvati, K. (2021). New horizons in reconstructing past human behavior: Introducing the "Tübingen University Validated Entheses-based Reconstruction of Activity" method. *Evolutionary Anthropology*, *30*(3), 185–198. https://doi.org/10.1002/evan.21892

Lacny, S., Faris, P., Bohm, E., Woodhouse, L. J., Robertsson, O., & Marshall DAl. (2021). Competing risks methods are recommended for estimating the cumulative incidence of revision arthroplasty for health care planning purposes. *Orthopedics*, *44*(4), e549–e555. https://doi.org/10.3928/01477447-20210618-16

Lacny, S., Wilson, T., Clement, F., Roberts, D. J., Faris, P. D., Ghali, W. A., & Marshal, D. A. (2015). Kaplan-Meier survival analysis overestimates the risk of revision arthroplasty: A meta-analysis. *Clinical Orthopaedics and Related Research*, *473*, 3431–3442. https://doi.org/10.1007/s11999-015-4235-8

Mountrakis, C., & Manolis, S. K. (2015). Entheseal changes of the upper limb in a Mycenaean population from Athens. *Mediterranean Archaeology and Archaeometry*, *15*, 209–220. https://doi.org/10.5281/ZENODO.15054

Ranstam, J., & Robertsson, O. (2017). The Cox model is better than the Fine and Gray model when estimating relative revision risks from arthroplasty register data. *Acta Orthopaedica*, *88*(6), 578–580. https://doi.org/10.1080/17453674.2017.1361130

Sauerbrei, W., Abrahamowicz, M., Altman, D. G., Cessie, S., & Carpenter, J. on behalf of the STRATOS initiative(2014). STRengthening Analytical Thinking for Observational Studies: The STRATOS initiative. *Statistics in Medicine*, *33*, 5413–5432. https://doi.org/10.1002/sim.6265

Sayers, A., Evans, J. T., Whitehouse, M. R., & Blom, A. W. (2018). Are competing risks models appropriate to describe implant failure? *Acta Orthopaedica*, *89*(3), 256–258. https://doi.org/10.1080/17453674.2018.1444876

Schrader, S. A. (2014). Elucidating inequality in Nubia: An examination of entheseal changes at Kerma (Sudan). *American Journal of Physical Anthropology*, *156*, 192–202. https://doi.org/10.1002/ajpa.22637

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, R. J. (2020). *P* > .05: The incorrect interpretation of "not significant" results is a significant problem. *American Journal of Biological Anthropology*, *172*, 521–527. https://doi.org/10.1002/ajpa.24092

Thomas, A. (2014). Bioarchaeology of the middle Neolithic: Evidence for archery among early European farmers. *American Journal of Physical Anthropology*, *154*, 279–290. https://doi.org/10.1002/ajpa.22504

van der Pas, S. L., Nelissen, R. G. H. H., & Fiocco, M. (2017). Patients with staged bilateral total joint arthroplasty in registries. *The Journal of Bone and Joint Surgery*, *99*(15), e82. https://doi.org/10.2106/JBJS.16.00854

van der Pas, S. L., Nelissen, R. G. H. H., & Fiocco, M. (2018). Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events. *Acta Orthopaedica*, *89*(2), 145–151. https://doi.org/10.1080/17453674.2018.1427314

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., & Egger, M. (2007). STROBE initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *Epidemiology*, *18*(6), 805–835. https://doi.org/10.1097/EDE.0b013e3181577511 PMID: 18049195.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). STROBE initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Epidemiology*, *18*(6), 800–804. https://doi.org/10.1097/EDE.0b013e3181577654 PMID: 18049194.

How to cite this article: van der Pas, S., & Schrader, S. (2023). Toward standardization of statistical reporting in studies on entheseal changes. *International Journal of Osteoarchaeology*, *33*(3), 475–478. https://doi.org/10.1002/oa.3188