Noname manuscript No.
(will be inserted by the editor)

Accuracy and precision of zero-heat-flux temperature monitoring: a systematic review and meta- analysis

Aaron Conway · Âuthór 2 ·

Received: date / Accepted: date

Abstract The text of your abstract. 150 – 250 words.

Keywords key · dictionary · word ·

Mathematics Subject Classification (2000) MSC code 1 · MSC code 2 ·

1 Introduction

Core temperature monitoring is an important intervention within perioperative and intensive care settings. Thermoregulatory dysfunction is commonly associated with the induction of anesthesia and can lead to adverse outcomes including cardiac dysrhythmias, altered hemostasis and increased risk of surgical site infection (Frank et al. 1997; Kurz, Sessler, and Lenhardt 1996; Michelson et al. 1994; Rohrer and Natale 1992). The pulmonary artery catheter is the reference standard for continuous core temperature monitoring($ref - troublefindingastatementregardingthisintheliterature$), however, the invasive nature of this method renders an increased risk of bloodstream infections and damage to surrounding tissue(Hadian and Pinsky 2006). Although clinically accurate relative to pulmonary artery temperature measurements, surro-

Grants or other notes about the article that should go on the front page should be placed here. General acknowledgments should be placed at the end of the article.

Aaron Conway
Peter Munk Cardiac Centre, UHN & Lawrence S. Bloomberg Faculty of Nursing, University of Toronto
E-mail: aaron.conway@utoronto.ca

Âuthór 2
Department of ZZZ, University of WWW
E-mail: djf@wef

gate measures of core temperature at esophageal, rectal, and bladder locations remain mildly invasive interventions(ref).

Zero-heat-flux (ZHF) thermometry is an alternative non-invasive method that allows for continuous monitoring of core temperature. Originally developed in the 1970s, ZHF technology was recently implemented in the 3M SpotOn (3M, St Paul, MN), a single-use, disposable sensor (ref). In practice, the ZHF sensor, such as the 3M SpotOn, is placed on the lateral surface of the forehead and initially warmed to equilibrate the temperature of the skin surface to the underlaying core tissues(ref). Equipped with a thermal insulator, the ZHF sensor eliminates heat loss to the environment to allow for changes in core temperature to be directly reflected by a change in skin surface temperature (ref).

The agreement between ZHF and established core temperature devices has been investigated in several studies conducted within the clinical setting, each with varied results. Appraisal of these studies and synthesis of the results would aid clinicians in deciding the appropriate circumstances in which ZHF may be used.

We aimed to determine if ZHF has clinically acceptable accuracy and precision relative to established core temperature measurement devices. Accuracy is defined as the average difference between temperature measurements from the ZHF and comparator device and precisions as the variance (typically reported as the SD) in the differences.

2 Methods

A systematic review was conducted in accordance with a predetermined protocol (PROSPERO trial registration number: ################).

2.1 Data sources and searches

Published studies were found by searching Medline and EMBASE from January 2000 to July 2019, >as well as the reference lists of articles identified to be relevant to the review. This search strategy is an efficient approach for systematic reviews of diagnostic test accuracy studies (Preston et al. 2015). Unpublished and ongoing studies were located by searching the International Clinical Trials Platform.

Published conference abstracts were included if there was enough information reported to appraise the quality of the study. There were no language restrictions applied during the search. >The Cochrane-recommended search strategy combining terms for the 'target condition' and 'index test' was used (Mann and Gilbody 2012).

The search strategies used for each data base can be located in file 1.

The selection of studies was undertaken by two independent reviewers. Studies were included if reported measures from a ZHF and comparator device coincided. Studies involving a case control design were excluded due to

potential for overestimation of the intervention performance. Studies were excluded if conducted on non-human subjects or outside of a clinical healthcare setting. Studies were excluded if conducted earlier than the year 2000 as prior studies evaluated outdated technology that is no longer relevant to current clinical practice.

2.2 Data extraction and quality assessment

Information was extracted regarding study characteristics (author, year of publication, country, design, sample size, clinical setting, numbers studied and analyses for each outcome), population characteristics (inclusion and exclusion criteria) and ZHF characteristics (placement of sensor, timing and methods of measurements). The outcomes that were extracted included the mean bias (eg, accuracy) and variance (eg, SD, precision) in temperature measurement between the ZHF sensor and comparator device. >Information was extracted about how repeated measurements were handled: (1) analysed each pair of data separately; (2) treated each pair of data as independent; or (3) used either analysis of variance or a random effects model as a way to control for the dependent nature of the repeated measures data (Myles and Cui 2007).

Using the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), each reviewer independently assessed the risk of bias for the included studies (Whiting et al. 2011). Using guiding questions, reviewers rated the risk of bias for patient selection, conduct of the ZHF measurements, conduct of the comparator device measurements, and timing and flow (eg, timing of ZHF and established core temperature measurements, dropouts) as “high”, “low” or “unclear”. We worked to minimize the risk of publication bias by conducting a comprehensive search of multiple databases as well as an international clinical trial registry (Glasziou et al. 2001). >Statistical analyses to detect reporting bias were not conducted due to lack of validated methods (Begg 2005). Although some meta-analyses of method comparison studies have used tests for detecting funnel plot asymmetry (Niven et al. 2015), simulations have revealed that such tests will result in publication bias being incorrectly identified too often (Deeks, Macaskill, and Irwig 2005).

In order to rate the quality of evidence, we applied the Grading Quality of Evidence and Strength of Recommendation methodology (Schünemann et al. 2008). Evidence was downgraded in accordance with study limitations, inconsistency and imprecision. There were no circumstances in which evidence was downgraded for indirectness as this systematic review only included relevant studies. Although the possibility of publication bias was not excluded, this bias was not formally assessed as it was not considered sufficient enough to reason downgrading the quality of evidence.

2.3 Data synthesis and analysis

The objective for the meta-analysis was to estimate the population limits of agreement between temperature measurements from the ZHF and established comparator devices. >The framework for meta-analysis of Bland-Altman method comparison studies based on limits of agreement (LoA) approach was used (Tipton and Shuster 2017).

This method was selected because it parallels the approach used in primary Bland-Altman studies, whereby an estimate is generated for the pooled LoAs in the population (not just in the samples studied). The 'population LoA' is more broad than the LoA commonly reported in the meta-analyses of Bland-Altman studies (Tipton and Shuster 2017). In these instances, >the pooled LoAs are calculated using $\delta \pm 2\sqrt{\sigma 2 + \tau 2}$, where δ is the average bias across studies, $\sigma 2$ is the average within-study variation in differences and $\tau 2$ is the variation in bias across studies.

Estimations of δ and $\sigma 2$ were made using a weighted least-squares model (similar to a random-effects approach), with the associated SEs estimated using robust variance estimation (RVE). RVE was used alternatively to model-based SEs as several studies included in the systematic review used repeated-measures designs without accommodating for the correlation between measurements (Hedges, Tipton, and Johnson 2010; Tanner-Smith, Tipton, and Polanin 2016; Tipton 2015). We used the >method-of-moments estimator from ref 16 for the $\tau 2$ parameter.

In accordance with ref 12, (1) measures of uncertainty were included >when interpreting the LoA estimates by calculating the outer 95% CIs for pooled LoA;

and (2) repeated measurements were adjusted if improperly adjusted for in >individual studies (by using weights proportional to the number of participants, not the total number of measurements).

The formulas for the above calculations from ref 12 can be found in file 1. The R statistical program was used to conduct all analyses (Team 2017). All data and R code (provided in ref 12) used in the meta-analyses can be located here.

Prior to conducting the meta-analyses, the results from each study were converted into a standard format, with bias meaning $comparator - ZHF$ measured in $^{\circ}C$. In several studies, results were reported for multiple groups of participants, therefore in the meta-analysis each of these groups was treated as a separate 'comparison'. Other studies reported multiple sets of results, whereby analyses were conducted between ZHF and various comparator devices used on the same participant. These instances were also treated as a separate 'comparison' if the comparator devices were apart of separate meta-analyses groups. One study reported intraoperative, postoperative and overall results for the same participants. Only the paired measurements from the overall results were included in the main and low risk bias analyses, leaving paired measurements exclusively from the intraoperative and postoperative timepoints to be included in respective meta-analyses subgroups. The conven-

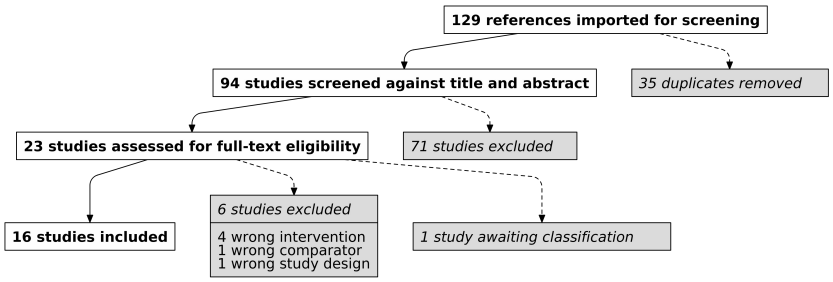

Fig. 1 PRISMA Flow Diagram

tionally cited clinically acceptable agreement between ZHF and comparator devices is 0.5 $°C$ ($ref - mostofthestudiessitethisrangefromEshragi...$). >It was deemed that outer confidence bounds for 95% LoA between ECT and ZHF (termed as 'population limits of agreement') outside of these bounds would

be clinically unacceptable. Sensitivity analysis for the primary meta-analysis was performed based on risk of bias, whereby 'unclear risk of bias' was treated as 'high risk' and 'high risk of bias' studies were excluded from analyses. As clinicians would be interested in the accuracy of ZHF relative to the thermometer devices they use, and within the clinical setting in which they use it we conducted subgroup analyses according to the comparator device used (eg, core, peripheral, nasopharangeal) and clinical setting (eg, intraoperative, intensuve care unit).

3 Results

Sixteen studies were eligible for inclusion in this systematic review (Figure 1).

The characteristics of each study are in online supplementary file 1.

Using these studies, we conducted 6 different meta-analyses. The first group compared ZHF to core (eg, arterial, bladder, esophageal, rectal) temperature measurements and included 22 (73%) comparisons from 14 (%) studies, and enrolled 607 participants with 179,821 paired measurements. The second group compared ZHF to core temperature measurements in low risk bias studies studies that scored a low risk of bias in all categories) and included 10 (%) comparisons from 5 (%) studies, and enrolled 273 participants with 104,294 paired measurements. The third group compared ZHF to core temperature measurements in ICU patients and included 7 (23%) comparisons from 5 (31%) studies, and enrolled 246 participants with 155,598 paired measurements. The fourth group compared ZHF to core temperature measurement in intraoperative patients and included 16 (53%) comparisons from 10 (63%) studies, and enrolled 464 participants with 24,223 paired measurements. The fifth group compared ZHF to nasopharyngeal temperature measurements and

included 4 (13%) comparisons from 4 (25%) studies, and enrolled 297 participants with 109,819 paired measurements. The sixth group compared ZHF to sublingual temperature measurement and included 2 (7%) comparisons from 2 (13%) studies, and enrolled 107 participants with 22731 paired measurements. The studies included in this systematic review evaluated a ZHF temperature monitoring system manufactured by 3M. Previously known as the SpotOn Temperature Monitoring System, the 3M ZHF device is now referred to as the Bair Hugger Temperature Monitoring System. All comparisons reported adherence to the ZHF device manufacturer instructions and placed the sensor on the forehead of the participants. High risk of bias was associated with patient selection for 13 (43%) comparisons from 9 (56%) studies, conduct of ZHF and comparator measurements in 10(33%) comparisons from 8 (50%) studies and 12 (40%) comparisons 9 (30%) studies, respectively (mostly due to ZHF measurements being taken with knowledge of the comparator measurements and vice versa) and participant flow for 13 (43%) studies. In 19 (50%) comparisons from 8 (50%) studies, the authors had declared a conflict of interest or receipt of funding or equipment from the manufacturer of the ZHF device under evaluation.

3.1 Study selection and description

3.2 Agreement between ZHF and core temperature measurements

4 Discussion

4.1 Limitations

5 Conclusion

References

Begg, Colin B. 2005. “Systematic Reviews of Diagnostic Accuracy Studies Require Study by Study Examination: First for Heterogeneity, and Then for Sources of Heterogeneity.” *Journal of Clinical Epidemiology* 58 (9). Elsevier Limited: 865.

Deeks, Jonathan J, Petra Macaskill, and Les Irwig. 2005. “The Performance of Tests of Publication Bias and Other Sample Size Effects in Systematic Reviews of Diagnostic Test Accuracy Was Assessed.” *Journal of Clinical Epidemiology* 58 (9). Elsevier: 882–93.

Frank, Steven M, Lee A Fleisher, Michael J Breslow, Michael S Higgins, Krista F Olson, Susan Kelly, and Charles Beattie. 1997. “Perioperative Maintenance of Normothermia Reduces the Incidence of Morbid Cardiac Events: A Randomized Clinical Trial.” *Jama* 277 (14). American Medical Association: 1127–34.

Glasziou, Paul, Les Irwig, Chris Bain, and Graham Colditz. 2001. *Systematic Reviews in Health Care: A Practical Guide*. Cambridge University Press.

Hadian, Mehrnaz, and Michael R Pinsky. 2006. "Evidence-Based Review of the Use of the Pulmonary Artery Catheter: Impact Data and Complications." *Critical Care* 10 (3). BioMed Central: S8.

Hedges, Larry V, Elizabeth Tipton, and Matthew C Johnson. 2010. "Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates." *Research Synthesis Methods* 1 (1). Wiley Online Library: 39–65.

Kurz, Andrea, Daniel I Sessler, and Rainer Lenhardt. 1996. "Perioperative Normothermia to Reduce the Incidence of Surgical-Wound Infection and Shorten Hospitalization." *New England Journal of Medicine* 334 (19). Mass Medical Soc: 1209–16.

Mann, Rachel, and Simon M Gilbody. 2012. "Should Methodological Filters for Diagnostic Test Accuracy Studies Be Used in Systematic Reviews of Psychometric Instruments? A Case Study Involving Screening for Postnatal Depression." *Systematic Reviews* 1 (1). BioMed Central: 9.

Michelson, Alan D, Hollace MacGregor, Marc R Barnard, Anita S Kestin, Michael J Rohrer, and C Robert Valeri. 1994. "Reversible Inhibition of Human Platelet Activation by Hypothermia in Vivo and in Vitro." *Thrombosis and Haemostasis* 72 (05). Schattauer GmbH: 633–40.

Myles, PS, and James Cui. 2007. "I. Using the Bland–Altman Method to Measure Agreement with Repeated Measures." Oxford University Press.

Niven, Daniel J, Jonathan E Gaudet, Kevin B Laupland, Kelly J Mrklas, Derek J Roberts, and Henry Thomas Stelfox. 2015. "Accuracy of Peripheral Thermometers for Estimating Temperature: A Systematic Review and Meta-Analysis." *Annals of Internal Medicine* 163 (10). American College of Physicians: 768–77.

Preston, Louise, Christopher Carroll, Paolo Gardois, Suzy Paisley, and Eva Kaltenthaler. 2015. "Improving Search Efficiency for Systematic Reviews of Diagnostic Test Accuracy: An Exploratory Study to Assess the Viability of Limiting to Medline, Embase and Reference Checking." *Systematic Reviews* 4 (1). BioMed Central: 82.

Rohrer, Michael J, and ANITA M Natale. 1992. "Effect of Hypothermia on the Coagulation Cascade." *Critical Care Medicine* 20 (10): 1402–5.

Schünemann, Holger J, Andrew D Oxman, Jan Brozek, Paul Glasziou, Roman Jaeschke, Gunn E Vist, John W Williams, et al. 2008. "Grading Quality of Evidence and Strength of Recommendations for Diagnostic Tests and Strategies." *Bmj* 336 (7653). British Medical Journal Publishing Group: 1106–10.

Tanner-Smith, Emily E, Elizabeth Tipton, and Joshua R Polanin. 2016. "Handling Complex Meta-Analytic Data Structures Using Robust Variance Estimates: A Tutorial in R." *Journal of Developmental and Life-Course Criminology* 2 (1). Springer: 85–112.

Team, R Core. 2017. "R Core Team (2017). R: A Language and Environment for Statistical Computing." *R Found. Stat. Comput. Vienna, Austria. URL Http://Www. R-Project. Org/., Page R Foundation for Statistical Computing.*

Tipton, Elizabeth. 2015. “Small Sample Adjustments for Robust Variance Estimation with Meta-Regression.” *Psychological Methods* 20 (3). American Psychological Association: 375.

Tipton, Elizabeth, and Jonathan Shuster. 2017. “A Framework for the Meta-Analysis of Bland–Altman Studies Based on a Limits of Agreement Approach.” *Statistics in Medicine* 36 (23). Wiley Online Library: 3621–35.

Whiting, Penny F, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, and Patrick MM Bossuyt. 2011. “QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies.” *Annals of Internal Medicine* 155 (8). Am Coll Physicians: 529–36.