Accuracy and precision of zero-heat-flux temperature monitoring

A systematic review and meta-analysis

Aaron Conway PhD1,2,3,✉, Megan Bittner BSc2, Dan Phan BSc2, Kristina Chang MScN1, Navpreet Kamboj BScN2, Elizabeth Tipton PhD4, and Matteo Parotto MD, PhD5,6

1 Peter Munk Cardiac Centre, University Health Network, Toronto, Canada  
2 Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, Canada  
3 School of Nursing, Queensland University of Technology (QUT), Brisbane, Australia  
4 Department of Statistics, Northwestern University, Evanston, IL, United States  
5 Department of Anesthesia and Pain Management, Toronto General Hospital, UHN, Toronto, Canada  
6 Department of Anesthesia and Interdepartmental Division of Critical Care Medicine, University of Toronto, Toronto, Canada

✉ Correspondence: [Aaron Conway PhD <[aaron.conway@utoronto.ca](mailto:aaron.conway@utoronto.ca)> | +1 416 946 7112 | 585 University Ave | Toronto, ON M5G 2N2](mailto:aaron.conway@utoronto.ca)

## Declarations

Clinical trial number: Not applicable

Prior presentations: Not applicable

Abstract word count: 300 words

Total word count: 3075 words

Introduction word count: 303 words

Discussion word count: 781 words

Number of figures: 2

Number of tables: 1

Abbreviated title: Accuracy of zero-heat-flux thermometers

Funding statement: Support was provided solely from institutional and/or departmental sources. MP is supported by an Early Investigator Merit Award from the Department of Anesthesia, University of Toronto and Toronto General Hospital

Conflicts of interest: The authors declare no competing interests.

### Abstract

**Background**

Zero-heat-flux thermometers provide clinicians with the ability to continuously and non-invasively monitor body temperature. These devices are increasingly being used to substitute for invasive core temperature measurement during surgery and in critical care. The aim of this review was to determine the accuracy and precision of zero-heat-flux temperature measurements.

**Methods**

Medline and EMBASE were searched for studies that reported on a measurement of core or peripheral temperature that coincided with a measurement from the zero-heat-flux device. Study selection and quality assessment (using QUADAS-2) was performed independently. The GRADE approach was used to summarize the strength of the evidence. Pooled estimates of the mean bias and limits of agreement with outer 95% confidence intervals (population limits of agreement) were calculated.

**Results**

Sixteen studies were included. The primary meta-analysis of zero-heat-flux versus core temperature consisted of 20 comparisons from 14 individual studies. Data from 675 participants with 181,167 paired measurements were included. The pooled estimate for the mean bias was 0.03°C. Population limits of agreement, which take into consideration the between-study heterogeneity and sampling error, were wide, spanning from -1°C to 1.06°C. The GRADE evidence quality rating was downgraded to moderate due to concerns about study limitations. Population limits of agreement for the sensitivity analysis restricted to studies rated as having low risk of bias across all the domains of the QUADAS-2 were similar to the primary analysis.

**Conclusion**

The range of uncertainty in the accuracy of the zero-heat-flux thermometer should be taken into account when using this device to inform clinical decision-making. Clinicians should therefore consider the potential that a temperature measurement from a zero-heat-flux thermometer could be as much as 1°C higher or lower than core temperature. Use of this device may not be appropriate in situations where a difference in temperature of less than 1°C is important to detect.

### Key points

**Question:** How accurate are temperature measurements from zero-heat-flux temperature monitors?

**Findings:** Population limits of agreement, which take into consideration the between-study heterogeneity and sampling error, were wide, spanning from -1°C to 1.06°C.

**Meaning:** Zero-heat-flux temperature monitoring would not be appropriate if a difference in temperature of 1°C is important to detect for the given clinical situation.

## Introduction

Continuous core temperature monitoring is commonly used within perioperative and intensive care settings. Thermoregulatory dysfunction is commonly associated with the induction of anesthesia and can lead to adverse outcomes including cardiac arrhythmia, altered hemostasis and increased risk of surgical site infection.1–4 The pulmonary artery catheter is the reference standard for continuous core temperature monitoring, however, the invasive nature of this method renders an increased risk of bloodstream infections and damage to surrounding tissue.5 Although clinically accurate relative to pulmonary artery temperature measurements, surrogate measures of core temperature at esophageal, rectal, and bladder locations remain mildly invasive interventions.

Zero-heat-flux thermometry is an alternative non-invasive method that allows for continuous monitoring of core temperature. Originally developed in the 1970s, zero-heat-flux technology was recently implemented in the 3M SpotOn temperature monitoring system (3M, St Paul, MN), as a single-use, disposable sensor. In practice, the zero-heat-flux sensor, such as the 3M SpotOn, is placed on the lateral surface of the forehead and is initially warmed to equilibrate the temperature of the skin surface to the underlying core tissues. Equipped with a thermal insulator, the zero-heat-flux 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.6

The agreement between zero-heat-flux thermometers and core as well as other peripheral thermometers has been investigated in multiple studies over the last five years. Appraisal of these studies and synthesis of the results would aid clinicians in deciding the appropriate circumstances in which zero-heat-flux thermometers may be used. We aimed to determine if zero-heat-flux thermometers have clinically acceptable accuracy and precision relative to established core and peripheral temperature measurement devices. Accuracy is defined as the average difference between temperature measurements from the zero-heat-flux and comparator device and precision as the variance (standard deviation) in the differences.

## Methods

A systematic review was conducted in accordance with a predetermined protocol. The primary comparison for this review was temperature measured from a zero-heat-flux thermometer versus temperature measured from a core site, which we defined as temperature measured at either an arterial, esophageal, bladder or rectal site. Secondary comparisons were made between zero-heat-flux thermometers and temperatures taken at peripheral sites.

### Inclusion criteria

Observational studies that reported temperature measurements from a zero-heat-flux thermometer and comparator thermometer were included. 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. No publication date restrictions were applied. 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.

### Data sources and searches

Published studies were found by searching Medline and EMBASE to July 2019. The Cochrane-recommended search strategy combining terms for the ‘target condition’ and ‘index test’ was used.7 This search strategy is an efficient approach for systematic reviews of diagnostic test accuracy studies.8 We also conducted forward citation searching, by using Google Scholar to search the citations of the first article published on the accuracy of zero-heat-flux thermometers. The search strategies used for each data base can be located at this [link](https://aconway.dev/zhf). Selection of studies was undertaken independently by two reviewers using [Covidence](https://www.covidence.org).

### 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 temperature measurement 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 zero-heat-flux and comparator thermometers. We also extracted information about how repeated measurements were handled. In particular we assessed whether studies: (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.9

Two reviewers independently assessed the risk of bias for the included studies using the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).10 Reviewers rated the risk of bias for patient selection, conduct of the zero-heat-flux measurements, conduct of the comparator thermometer measurements, and timing and flow (eg, timing of zero-heat-flux and established core temperature measurements, dropouts) as ‘high’, ‘low’ or ‘unclear’ risk of bias. 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.11 Statistical approaches for detection of reporting bias were not conducted due to lack of validated methods.12 Simulations have revealed that tests for detecting funnel plot asymmetry will result in publication bias being incorrectly identified too often.13

In order to rate the quality of evidence, we applied the Grading Quality of Evidence and Strength of Recommendation methodology.14 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.

### Data synthesis and analysis

The objective for the meta-analysis was to estimate the population limits of agreement between temperature measurements from the zero-heat-flux and established comparator thermometers. A framework for meta-analysis of Bland-Altman method comparison studies based on limits of agreement approach was used.15 This method was selected because it parallels the approach used in primary Bland-Altman studies, whereby an estimate is generated for the pooled limits of agreement in the population (not just in the samples studied). The ‘population limits of agreement’ is more broad than the limits of agreement commonly reported in the meta-analyses of Bland-Altman studies.15 Pooled limits of agreement were calculated using , where is the average bias across studies, is the average within-study variation in differences and is the variation in bias across studies.

Estimations of and were made using a weighted least-squares model (similar to a random-effects approach), with the associated standard errors estimated using robust variance estimation. Robust variance estimation was used alternatively to model-based standard errors as some studies included in the systematic review used repeated-measures designs without accommodating for the correlation between measurements.16–18 We used the method-of-moments estimator from DerSimonian & Laird 19 for the parameter.

Measures of uncertainty were included in our meta-analyses by calculating the outer 95% confidence intervals for pooled limits of agreement. We also accounted for repeated measurements if they were not properly adjusted for in individual studies. This was achieved by using weights proportional to the number of participants, not the total number of measurements. The R statistical program was used to conduct all analyses.20 All data and R code used in the meta-analyses can be located at this [link](https://github.com/awconway/zhf-review).

Prior to conducting the meta-analyses, the results from each study were converted into a standard format, with bias meaning comparator thermometer minus zero-heat-flux thermometer measured in degrees Celsius (°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 zero-heat-flux and various comparator devices used on the same participant. These instances were also treated as a separate ‘comparison’ if the comparator devices were a part of separate meta-analysis 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 conventionally cited clinically acceptable agreement between zero-heat-flux and comparator devices is 0.5°C.6 It was deemed that outer confidence bounds for 95% limits of agreement between zero-heat-flux and core temperature measurements (termed as ‘population limits of agreement’) outside of these bounds would be clinically unacceptable.

A sensitivity analysis for the primary comparison (zero-heat-flux versus temperature measurement at core site) 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. We also conducted a sensitivity analysis for the primary comparison (zero-heat-flux versus temperature measurement at core site) excluding studies that received funding from industry. As clinicians would be interested in the accuracy of zero-heat-flux relative within the clinical setting in which they use it, we conducted subgroup analyses for the primary comparison according to clinical setting of the study (either intraoperative or intensive care unit).

## Results

### Study selection and description

Sixteen studies were included (Figure 1). Two studies reported only in abstract form were not included and assigned as ‘studies awaiting classification’ because there was insufficient information provided.

The characteristics of included studies can be viewed at this [link](https://aconway.dev/zhf). The primary comparison of zero-heat-flux versus core temperature measurements (eg, arterial, bladder, esophageal or rectal) consisted of 20 comparisons from 14 individual studies. In total, data from 675 participants with 181,167 paired measurements were included in this comparison (two studies did not report the total number of measurements included in their analysis). The sensitivity analysis for the primary comparison with only studies that were judged as low risk of bias across all domains included 11 comparisons from 6 studies that enrolled 372 participants with 105,640 paired measurements (two studies did not report the total number of measurements included in their analysis). There were 5 studies that compared zero-heat-flux to core temperature measurements in ICU patients, comprising 7 comparisons with 155,598 measurements from 246 participants. Another 10 studies were included that compared zero-heat-flux to core temperature measurements in patients undergoing surgery, comprising 14 comparisons with 25,569 measurements from 532 participants. Nasopharyngeal thermometers were used as the comparator devices in 4 studies, with 109,819 paired measurements from 109,819 participants. Sublingual thermometers were used as the comparator devices in 2 studies that reported 22,731 paired measurements from 107 participants.

All studies included in this systematic review evaluated the zero-heat-flux temperature monitoring system manufactured by 3M. Previously known as the SpotOn Temperature Monitoring System, the 3M zero-heat-flux device is now referred to commercially as the Bair Hugger Temperature Monitoring System. All comparisons reported adherence to the zero-heat-flux device manufacturer instructions and placed the sensor on the forehead of the participants. One included study also reported results for comparisons where the zero-heat-flux thermometer was not placed on the forehead. We did not include these comparisons in our meta-analysis.

High risk of bias was associated with patient selection for 13 (43%) comparisons from 9 (56%) studies, conduct of zero-heat-flux and comparator measurements in 10 (33%) comparisons from 8 (50%) studies and 12 (40%) comparisons 9 (30%) studies, respectively (mostly due to zero-heat-flux 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 zero-heat-flux device under evaluation.

### Primary comparison: zero-heat-flux thermometer versus core thermometers

A graphical overview of the findings for the primary analysis is presented in Figure 2. The pooled estimate for the mean bias between zero-heat-flux and core temperature measurements was 0.03°C. The population limits of agreement, which take into consideration the between-study heterogeneity and sampling error, were wide, spanning from -1°C to 1.06°C (181,167 measurements; 675 participants; 14 studies). The quality of evidence for the primary comparison was downgraded to low quality due to concerns about study limitations and inconsistency. Population limits of agreement for the sensitivity analysis restricted to studies rated as having low risk of bias across all the domains of the QUADAS-2 were similar to the primary analysis (105,640 measurements; 372 participants; 6 studies). The mean bias was again 0.02°C with population limits of agreement spanning from -0.95°C to 0.98°C. A further sensitivity analysis excluding studies that received industry funding revealed population limits of agreement that were considerably wider than the primary analysis (-1.29°C to 1.36°C). The GRADE rating for this sensitivity was downgraded to low quality though, due to concerns about study limitations, inconsistency and imprecision.

Table 1 presents results of the primary, sensitivity and all subgroup analyses for the primary comparison as well as secondary comparisons. We conducted two subgroup analyses for the primary comparison according to the clinical setting in which the study was conducted. In the subset of studies conducted in the ICU, the mean bias was -0.04°C with population limits of agreement between -1.47°C and 1.39 °C. The GRADE rating for quality of evidence was downgraded to very low quality due to study limitations, inconsistency and imprecision. In the subset of studies that evaluated the use of zero-heat-flux thermometers during surgery found a mean bias of 0.03°C with population limits of agreement from -1.02°C to 1.07°C. The quality of evidence was rated as moderate quality due to concerns about study limitations.

### Secondary comparisons: zero-heat-flux thermometer versus peripheral thermometers

Zero-heat-flux thermometers were compared with sublingual thermometers and nasopharyngeal thermometers in the studies included in this review. The mean bias between zero-heat-flux and sublingual temperature measurements in meta-analysis of results from 2 studies was -0.21°C. Due to the limited number of studies and measurements, population limits of agreement were extremely wide, spanning from -17.95°C to 17.53°C. The quality of evidence for this comparison was rated as very low quality due to serious concerns about imprecision.

The mean bias between zero-heat-flux and nasopharyngeal temperature measurements in meta-analysis of results from 4 studies was 0°C. Population limits of agreement were -0.99°C to 0.98°C. We downgraded the quality of evidence to low, again due to concerns about imprecision and study limitations.

## Discussion

This systematic review showed that a temperature measurement from a zero-heat-flux thermometer could be as much as 1°C higher or lower than core temperature. These results may have important implications for practice that clinicians should consider. It was reassuring that results of our sensitivity analysis restricted to studies judged to be at low risk of bias using the QUADAS-2 tool were similar. As such, it is vital for healthcare professionals considering using this device to first determine if differences in temperature smaller than this magnitude would be important for the given clinical situation. If so, then it may not be appropriate to substitute the zero-heat-flux device in place of a core thermometer.

Our estimates of the accuracy of zero-heat-flux thermometers are similar to results from a previous meta-analysis that compared other peripheral thermometers, such as sublingual and temporal artery devices, with core temperature measurements.21 However, this previous meta-analysis used a statistical approach that did not incorporate the magnitude of heterogeneity in results between studies or sampling error. As such, it is possible that the zero-heat-flux thermometer may still be more precise than other peripheral thermometers.

Subgroup analyses revealed insights to direct future research. There were fewer participants included in the subgroup of studies conducted in the Intensive Care Unit compared to those conducted during surgery. As a result, population limits of agreement for the Intensive Care Unit subgroup were broad. Additional studies are required to increase confidence in the accuracy of zero-heat-flux device in this setting, where continuous temperature monitoring is often required.

Although comparisons with core temperature are the most important to evaluate, we included studies that compared zero-heat-flux and peripheral temperature monitoring devices in our review because clinicians may be interested to determine if this new technology is sufficiently accurate to replace other ways to non-invasively monitor temperature. Also, there may be situations where it is standard practice for only peripheral temperature monitoring to be used. In these circumstances, it is not practical to expose research participants to the potential risks of invasive core temperature monitoring for the purpose of evaluating the new temperature monitoring device. Whether or not the zero-heat-flux thermometer is sufficiently accurate to be used in place of nasopharyngeal thermometers is unclear. Population limits of agreement calculated from four studies were broad, spanning from 1°C less than zero-heat-flux temperature to one degree higher. Further studies are therefore required.

It should be noted that we did not include studies that used the Temple Touch Pro because it is not strictly a zero-heat-flux device. This is a new thermometer that is similar to the zero-heat-flux device in that it is placed cutaneously, but the underlying technology is different. Far fewer studies have evaluated the Temple Touch Pro compared to the zero-heat-flux device.22,23

Many studies in this review analysed a large number of measurements of temperature with relatively small sample sizes. Importantly, the approach we used for our meta-analysis takes this into account. By using robust variance estimation, weights for pooling estimates in the meta-analysis become proportional to the number of patients, not the total number of measurements.15

The accuracy of zero-heat-flux temperature monitoring has been evaluated in various intraoperative contexts as well as in the intensive care unit setting. Our subgroup analyses did not indicate that the accuracy of this temperature monitoring device was any more effective in a particular setting. Only one study included in our systematic review included pediatric patients. Additional studies to evaluate the accuracy of this device in children may be warranted to increase confidence.

### Limitations

We did not extract data on adverse events due to zero-heat-flux temperature monitoring. The possibility of publication bias cannot be ruled out, although the evidence suggests this may not be as serious of a problem for studies that are not randomized controlled trials.12 Our focus for the meta-analysis was on calculating population limits of agreement, which incorporate the variation in bias between studies into the estimates. For this reason, we did not use meta-regression or tests for interaction between subgroups as a way to investigate sources of heterogeneity. It is important to note that this review did not assess the clinical utility of temperature monitoring using the zero-heat-flux device. The evidence from this review should be considered in the context of other information about the reliability and ease of use of this device.

## Conclusion

Substantial differences between core and zero-heat-flux temperature measurements were identified in this meta-analysis. Clinicians should consider the range of uncertainty in the accuracy of the zero-heat-flux thermometer when using this device to inform their decision-making. As such, there may be circumstances where use of this device would not be appropriate because a difference in temperature of 1°C is important to detect for the given clinical situation.

# References

1. Frank SM, Fleisher LA, Breslow MJ, Higgins MS, Olson KF, Kelly S, Beattie C: Perioperative maintenance of normothermia reduces the incidence of morbid cardiac events: A randomized clinical trial. JAMA 1997; 277:1127–34

2. Kurz A, Sessler DI, Lenhardt R: Perioperative normothermia to reduce the incidence of surgical-wound infection and shorten hospitalization. New England Journal of Medicine 1996; 334:1209–16

3. Michelson AD, MacGregor H, Barnard MR, Kestin AS, Rohrer MJ, Valeri CR: Reversible inhibition of human platelet activation by hypothermia in vivo and in vitro. Thrombosis and Haemostasis 1994; 72:633–40

4. Rohrer MJ, Natale AM: Effect of hypothermia on the coagulation cascade. Critical Care Medicine 1992; 20:1402–5

5. Hadian M, Pinsky MR: Evidence-based review of the use of the pulmonary artery catheter: Impact data and complications. Critical Care 2006; 10:S8

6. Eshraghi Y, Nasr V, Parra-Sanchez I, Van Duren A, Botham M, Santoscoy T, Sessler DI: An evaluation of a zero-heat-flux cutaneous thermometer in cardiac surgical patients. Anesthesia & Analgesia 2014; 119:543–9

7. Mann R, Gilbody SM: 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 2012; 1:9

8. Preston L, Carroll C, Gardois P, Paisley S, Kaltenthaler E: 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 2015; 4:82

9. Myles PS, Cui J: I. Using the bland–altman method to measure agreement with repeated measures. BJA: British Journal of Anaesthesia 2007; 99:309–11

10. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM: QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine 2011; 155:529–36

11. Glasziou P, Irwig L, Bain C, Colditz G: Systematic reviews in health care: A practical guide. Cambridge University Press, 2001

12. Begg CB: Systematic reviews of diagnostic accuracy studies require study by study examination: First for heterogeneity, and then for sources of heterogeneity. Journal of Clinical Epidemiology 2005; 58:865

13. Deeks JJ, Macaskill P, Irwig L: 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 2005; 58:882–93

14. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams JW, Kunz R, Craig J, Montori VM, others: Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336:1106–10

15. Tipton E, Shuster J: A framework for the meta-analysis of bland–altman studies based on a limits of agreement approach. Statistics in Medicine 2017; 36:3621–35

16. Hedges LV, Tipton E, Johnson MC: Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 2010; 1:39–65

17. Tanner-Smith EE, Tipton E, Polanin JR: Handling complex meta-analytic data structures using robust variance estimates: A tutorial in r. Journal of Developmental and Life-Course Criminology 2016; 2:85–112

18. Tipton E: Small sample adjustments for robust variance estimation with meta-regression. Psychological Methods 2015; 20:375

19. DerSimonian R, Laird N: Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7:177–88

20. R Core Team: R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2018 at <<https://www.R-project.org/>>

21. Niven DJ, Gaudet JE, Laupland KB, Mrklas KJ, Roberts DJ, Stelfox HT: Accuracy of peripheral thermometers for estimating temperature: A systematic review and meta-analysis. Annals of Internal Medicine 2015; 163:768–77

22. Evron S, Weissman A, Toivis V, Shahaf DB, You J, Sessler DI, Ezri T: Evaluation of the temple touch pro, a novel noninvasive core-temperature monitoring system. Anesthesia & Analgesia 2017; 125:103–9

23. Sastre JA, Pascual MJ, Lopez T: Evaluation of the novel non-invasive zero-heat-flux tcore™ thermometer in cardiac surgical patients. Journal of Clinical Monitoring and Computing 2019; 33:165–72

# Figure legend

Figure 1: PRISMA Flow Diagram

Figure 2: Graphical overview of findings for the primary analysis