ORIGINAL ARTICLE

Accuracy and precision of transcutaneous carbon dioxide monitoring: a systematic review and meta-analysis

Aaron Conway,[1,2,3] Elizabeth Tipton,[4] Wei-Hong Liu,[1] Zachary Conway,[5] Kathleen Soalheira,[1] Joanna Sutherland,[6] James Fingleton[7]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/thoraxjnl-2017-211466).

[1]Institution of Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove, Queensland, Australia
[2]Bloomberg Faculty of Nursing, University of Toronto, Toronto, Ontario, Canada
[3]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada
[4]Teachers College, Columbia University, New York, USA
[5]School of Exercise Sciences, AustralianCatholic University, Brisbane, Queensland, Australia
[6]Coffs Harbour Health Campus and Rural Clinical School, University of New South Wales, Coffs Harbour, New South Wales, Australia
[7]Medical Research Institute of New Zealand, Wellington, New Zealand

Correspondence to
Dr Aaron Conway, Bloomberg Faculty of Nursing, University of Toronto, Toronto, Ontario, Canada ;
aaron.conway@utoronto.ca

Received 23 December 2017
Revised 31 May 2018
Accepted 13 August 2018
Published Online First
12 September 2018

To cite: Conway A, Tipton E, Liu W-H, *et al*. *Thorax* 2019;**74**:157–163.

ABSTRACT

Background Transcutaneous carbon dioxide ($TcCO_2$) monitoring is a non-invasive alternative to arterial blood sampling. The aim of this review was to determine the accuracy and precision of $TcCO_2$ measurements.

Methods Medline and EMBASE (2000–2016) were searched for studies that reported on a measurement of $PaCO_2$ that coincided with a measurement of $TcCO_2$. Study selection and quality assessment (using the revised Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2)) were performed independently. The Grading Quality of Evidence and Strength of Recommendation approach was used to summarise the strength of the body of evidence. Pooled estimates of the mean bias between $TcCO_2$ and $PaCO_2$ and limits of agreement with outer 95% CIs (termed population limits of agreement) were calculated.

Results The mean bias was −0.1 mm Hg and the population limits of agreement were −15 to 15 mm Hg for 7021 paired measurements taken from 2817 participants in 73 studies, which was outside of the clinically acceptable range (7.5 mm Hg). The lowest $PaCO_2$ reported in the studies was 18 mm Hg and the highest was 103 mm Hg. The major sources of inconsistency were sensor location and temperature. The population limits of agreement were within the clinically acceptable range across 3974 paired measurements from 1786 participants in 44 studies that applied the sensor to the earlobe using the TOSCA and Sentec devices (−6 to 6 mm Hg).

Conclusion There are substantial differences between $TcCO_2$ and $PaCO_2$ depending on the context in which this technology is used. $TcCO_2$ sensors should preferentially be applied to the earlobe and users should consider setting the temperature of the sensor higher than 42°C when monitoring at other sites.

Systematic review registration number PROSPERO; CRD42017057450.

Key messages

What is the key question?

► Transcutaneous carbon dioxide ($TcCO_2$) monitoring devices are commercially available and are being used in clinical practice, so it is vital that clinicians have a clear understanding of the accuracy of these devices to ensure they are applied in appropriate circumstances.

What is the bottom line?

► A $TcCO_2$ measurement at any single point in time could be as much as 15 mm Hg higher or lower than $PaCO_2$, meaning that an arterial blood gas sample would be required to confirm diagnosis prior to initiation (or cessation) of treatment.

Why read on?

► As clinicians would be interested in the accuracy of the type of transcutaneous monitoring device they use and for the population in which they use it, we provide population limits of agreement according to the indication for monitoring (eg, respiratory failure, surgery, intensive care unit, sedation and postoperative recovery), type of device (Sentec and TOSCA), as well as location of sensor placement and sensor temperature.

INTRODUCTION

Measurement of $PaCO_2$ in arterial blood is the reference standard for ventilation assessment.[1] Arterial puncture is painful and time-consuming, and there is risk of infection as well as tissue and nerve damage.[2] Measurement of carbon dioxide (CO_2) levels from the skin, which is known as transcutaneous carbon dioxide ($TcCO_2$) monitoring, is a non-invasive alternative to arterial blood sampling. $TcCO_2$ monitors measure $PaCO_2$ that diffuses through the skin by the application of a sensor, which is heated above body temperature (typically to between 40°C and 44°C) to achieve local arterialisation. Local arterialisation, combined with application of an algorithm that corrects the CO_2 value detected by the sensor to 37°C, is thought to provide an accurate estimate of $PaCO_2$. False-positive and false-negative indications of worsening ventilation status are both important issues to consider regarding the application of these monitors to the clinical practice context. A false-positive or false-negative indication about ventilation status from $TcCO_2$ monitoring may lead to inappropriate initiation, delay or avoidance of treatment, which could be detrimental for the patient.

The agreement between $TcCO_2$ and $PaCO_2$ has been investigated in a large number of studies. Synthesis of the data through meta-analysis would

aid clinical decision-making regarding the appropriate circumstances in which $TcCO_2$ monitoring can be used. We aimed to determine if $TcCO_2$ has clinically acceptable accuracy and precision compared with $PaCO_2$. Accuracy is defined in this context as the average difference between $TcCO_2$ and $PaCO_2$ measurements and precision as the variance (typically reported as SD) in the differences.

METHODS

A systematic review was conducted according to a prespecified protocol (PROSPERO trial registration number: CRD42017057450).

Data sources and searches

Published studies were located by searching Medline and EMBASE from January 2000 to December 2016, 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.[3] Unpublished and ongoing studies were located by searching the International Clinical Trials Platform. Published conference abstracts were planned to be included if there was sufficient detail reported to assess study quality. Language restrictions were not imposed for the search. The Cochrane-recommended search strategy combining terms for the 'target condition' and 'index test' was used.[4] The specific search strategy for each database is in online supplementary file 1. Study selection was performed by two independent reviewers.

Studies that reported a measurement of $PaCO_2$ that coincided with a measurement of $TcCO_2$ were included. Studies conducted before the year 2000 were excluded as earlier studies evaluated outdated technology. Only studies that reported on $PaCO_2$ measured either by a point-of-care blood gas analyser or central laboratory that coincided with a measurement of the index test were included.

Data extraction and quality assessment

Information about the study characteristics (author, year of publication, country, design, sample size, clinical setting, number studied and number analysed for each outcome, number of dropouts with reason, and funding source), population characteristics (inclusion/exclusion criteria, mean/median and range of $PaCO_2$) and $TcCO_2$ characteristics (timing and methods of sampling/measurements, method of sampling/calibration) was extracted. Outcomes extracted were the mean bias (ie, accuracy) and variance or SD (ie, precision) in CO_2 between transcutaneous and arterial blood gas analyses. 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.[5]

Risk of bias assessment of the included studies was undertaken independently in duplicate using the revised Quality Assessment of Diagnositc Accuracy Studies (QUADAS-2) tool.[6] Guiding questions were used to rate the risk of bias for patient selection, conduct of the $TcCO_2$ measurement, conduct of the $PaCO_2$ measurement and flow and timing (eg, timing of $TcCO_2$ or $PcCO_2$ measurements and dropouts) as 'high', 'low' or 'unclear'. The risk of publication bias was minimised by comprehensively searching multiple databases and an international clinical trial registry.[7] Language restrictions were not imposed for the search. However, we were unable to classify four potentially eligible studies because the full text was not in English. Statistical analyses to detect reporting bias were not conducted due to lack of validated methods.[8] Although some meta-analyses of method comparison studies have used tests for detecting funnel plot asymmetry,[9] simulations have revealed that such tests will result in publication bias being incorrectly identified too often.[10]

We applied the Grading Quality of Evidence and Strength of Recommendations methodology to rate the quality of evidence.[11] Reasons used to downgrade the evidence were study limitations, inconsistency and imprecision. We did not downgrade for indirectness because this systematic review excluded studies that were not relevant. Publication bias was not formally assessed so the possibility of this bias was not excluded but not considered sufficient to require downgrading the quality of evidence.

Data synthesis and analysis

Our goal for the meta-analysis was to estimate the population limits of agreement between $TcCO_2$ and $PaCO_2$. The framework for meta-analysis of Bland-Altman method comparison studies based on a limits of agreement (LoA) approach was used.[12] We selected this method since it mirrors the approach in primary Bland-Altman studies, providing an estimate of the pooled LoAs in the population (not just the samples studied). The 'population LoA' is wider than those typically reported in meta-analyses of Bland-Altman studies.[12] Here the pooled LoAs are calculated using $\delta \pm 2\sqrt{(\sigma^2+\tau^2)}$, where δ is the average bias across studies, σ^2 is the average within-study variation in differences and τ^2 is the variation in bias across studies. We estimated δ and σ^2 using a weighted least-squares model (similar to a random-effects approach) and estimated their SEs using robust variance estimation (RVE). We used RVE instead of model-based SEs because many studies included in our review used repeated-measures designs without accounting for the correlation between measurements.[13–15] The method-of-moments estimator from ref 16 was used for the τ^2 parameter. Following ref 12 we also (1) included measures of uncertainty when interpreting the LoA estimates by calculating the outer 95% CIs for pooled LoA; and (2) adjusted repeated measurements which were not

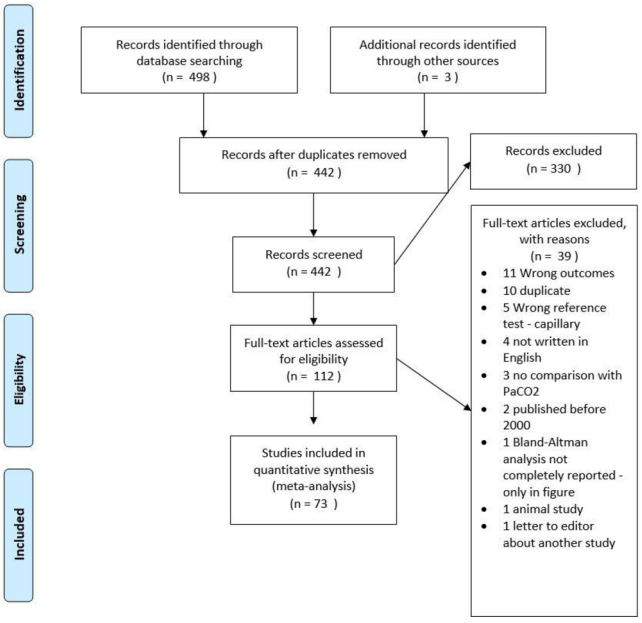

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram.

properly adjusted in individual studies (by using weights proportional to the number of samples not the total number of measurements). Formulas for these calculations from ref 12 are provided in online supplementary file 1. All analyses were conducted in the R statistical program.[17] The R code (provided in ref 12) and all data used in the meta-analyses are available at https://doi.org/10.6084/m9.figshare.6244058.v2.

The results from the individual studies were converted into a standard format to conduct meta-analyses, with bias meaning $PaCO_2$-$TcCO_2$ measured in mm Hg. In two of the studies, the results were reported for two separate groups of participants, so these were treated in the meta-analyses as separate 'studies'. Other studies reported separate results for analyses conducted using different $TcCO_2$ device types or sensor locations performed on the same patients. Only the result with the largest number of paired measurements between $PaCO_2$ and $TcCO_2$ was selected for inclusion in the main analysis, with others included in subgroup meta-analyses where appropriate. For the studies that reported results for patients while receiving both two-lung and one-lung ventilation during thoracic surgery, we used the result for two-lung ventilation in the main analysis.

The conventionally cited clinically acceptable agreement between $TcCO_2$ and $PaCO_2$ is 7.5 mm Hg (or 1 kPa).[18] We deemed that outer confidence bounds for 95% LoA between transcutaneous and arterial CO_2 measurements (termed as 'population limits of agreement') outside of these bounds would not be clinically acceptable.

We performed sensitivity analysis for the primary meta-analysis based on risk of bias (eg, treating 'unclear risk of bias' as 'high risk' and removing 'high risk of bias' studies from the analyses). As clinicians would be interested in the accuracy of the type of transcutaneous monitoring device they use and for the population in which they use it, we conducted subgroup analyses according to the indication for monitoring (eg, volunteer study, respiratory failure, surgery, intensive care unit (ICU), sedation and postoperative recovery), type of device (Sentec and TOSCA), as well as location of sensor placement and sensor temperature.

RESULTS

Study selection and description

There were 73 studies eligible for inclusion (figure 1). The characteristics of each study are in online supplementary file 1. The 73 studies enrolled 2817 participants predominantly from Europe, USA and UK. Sixteen (22%) studies included adult participants in ICUs, 6 (7%) studies included paediatric participants in ICUs or having surgery, 7 (10%) studies included neonates, 13 (18%) studies were conducted with adults undergoing surgery with general anaesthesia, 13 (18%) studies were focused on acute respiratory failure, 9 (12%) studies included participants with

Table 1 Results of meta-analysis of agreement between transcutaneous and arterial carbon dioxide

									Population limits of agreement	
	Studies	N	n	Bias	SD	τ^2	LoA_L	LoA_U	CI_L	CI_U
Main analysis	73	7021	2817	−0.1	1.9	8.9	−7.1	6.9	−15.1	14.9
Low risk only	23	1600	842	0.2	1.8	3.8	−5.2	5.5	−9.2	9.6
Device type										
Sentec	30	3585	1256	0.1	1.9	3.5	−5.2	5.5	−8.5	8.7
TOSCA	45	3313	1561	−0.4	1.9	2.8	−5.4	4.6	−7.2	6.4
Sensor location										
Earlobe	44	3974	1786	−0.1	1.8	1.9	−4.6	4.5	−5.7	5.5
Chest	13	1041	471	1.0	2.0	3.9	−4.6	6.7	−10.6	12.7
Arm	7	247	157	−1.5	1.8	1.5	−5.9	2.9	−8.3	5.3
Other monitoring site	16	2156	448	−0.5	1.9	30.1	−12.1	11.1	−134.0	133.0
Sensor temperature										
42°C	43	3635	1634	0.1	1.8	12.0	−7.7	8.0	−24.9	25.2
More than 42°C	23	1471	768	0.1	1.9	2.5	−4.9	4.9	−7.2	7.2
Clinical setting										
ICU	16	2128	467	−0.6	2.0	1.9	−5.4	4.2	−7.3	6.1
Neonates	7	1298	263	−2.9	2.2	5.9	−9.4	3.6	−25.9	20.1
Acute respiratory failure	14	993	614	1.7	2.0	3.2	−3.7	7.1	−7.8	11.2
Surgery with general anaesthesia	13	707	348	−0.2	1.7	22.1	−10.1	9.8	−95.9	95.6
Surgery with one-lung ventilation	4	129	74	−1.1	1.7	3.4	−6.2	3.9	−18.4	16.1
Paediatric ICU and surgery	6	501	172	−0.4	1.8	0.2	−4.1	3.3	−5.1	4.4
Sedated and spontaneously breathing	5	403	160	−0.4	1.7	5.5	−6.2	5.4	−27.0	26.2
Chronic respiratory failure	9	322	286	−0.1	1.8	2.4	−4.8	4.6	−8.5	8.3
Outpatients requiring lung function tests	4	555	535	−0.1	1.6	1.4	−4.0	3.9	−7.3	7.3

τ^2, variation in bias between studies; bias, pooled estimate of mean differences calculated as $PaCO_2$-$TcCO_2$ in mm Hg; CI_L, outer confidence bound for lower 95% limit of agreement; CI_U, outer confidence bound for upper 95% limit of agreement; ICU, intensive care unit; LoA_L, lower 95% limit of agreement calculated from pooled estimates of bias and SD of differences with robust variance estimation; LoA_U, upper 95% limit of agreement calculated from pooled estimates of bias and SD of differences with robust variance estimation; N, number of paired measurements; n, number of participants; SD, pooled SD of differences; $TcCO_2$, transcutaneous carbon dioxide.

chronic respiratory failure, 5 (7%) studies included patients who were sedated but spontaneously breathing either during or after surgery, and 4 (5%) studies were conducted with patients undergoing lung function testing. Two studies intentionally manipulated the range of $PaCO_2$ by inducing hypoventilation and hyperventilation. The lowest $PaCO_2$ reported in the studies was 18 mm Hg and the highest was 103 mm Hg.

Several different $TcCO_2$ monitors were evaluated across the studies included in this review, including the TCM3 (n=12), TCM4 (n=11), TOSCA 500 with Sensor 92 (n=7), TOSCA not otherwise classified (n=13), Sentec with V-Sign sensor (n=27), Sentec with V-Sign 2 sensor (n=2), Fastrac (n=1), Microgas (n=2) and PeriFlux (n=1). All studies reported that device manufacturer instructions were followed regarding calibration and stabilisation of the sensor prior to undertaking assessments. Most studies reported that the temperature of the $TcCO_2$ sensor was less than 43°C (n=49; 67%). The earlobe was the most common sensor location site evaluated (n=45). Other sensor location sites included the chest, upper arm, abdomen, forehead, cheek and palmar surface of the forearm.

There was a high risk of bias associated with patient selection for 14 (19%) studies, conduct of $TcCO_2$ and $PaCO_2$ measurements in 7 (10%) and 9 (12%) studies, respectively (mostly due to $PaCO_2$ measurements being taken with knowledge of the $TcCO_2$ measurement and vice versa), and participant flow for 7 (10%) studies. The authors declared that they either had a conflict of interest or had received equipment or funding from the manufacturers of the device being evaluated in 19 (26%) studies.

Agreement between transcutaneous and arterial CO_2 measurements

Table 1 presents the results of the primary meta-analysis, sensitivity analysis and subgroup analyses. Data from all 73 studies were included in the primary meta-analysis. Although the pooled estimate of the mean bias between $PaCO_2$ and $TcCO_2$ was small (0.1 mm Hg), the variation in these differences was large (figure 2), resulting in the two methods differing from −15 mm Hg to 15 mm Hg across all patients studied. These population limits of agreement were not in the clinically acceptable range. A summary of findings is presented in table 2. We downgraded the quality of evidence for the primary outcome 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 were also outside of the clinically acceptable range (−9 mm Hg to 10 mm Hg; 1600 paired measurements from 842 participants in 23 studies).

The population limits of agreement for the TOSCA device were within the clinically acceptable range (−7 to 6 mm Hg; 3313 paired measurements from 1561 participants in 45 studies) but not for the Sentec device (−9 to 9 mm Hg; 3585 paired measurements from 1256 participants in 30 studies). However, population limits of agreement differed according to the location that the sensor was applied and the temperature of the sensor. $TcCO_2$ monitoring was accurate to a clinically acceptable degree in a meta-analysis of 44 studies (3974 paired measurements from 1786 participants) where the sensor was applied to the earlobe with either the TOSCA (20 studies) or Sentec (24 studies) device. The population limits of agreement were −6 to 6 mm Hg. In contrast, population limits of agreement were outside the clinically acceptable range where $TcCO_2$ monitoring was conducted with the sensor on the chest (−11 to 12.7 mm Hg; 1041 paired measurements from 471 participants in 13 studies) and the arm (−8 mm Hg to 5.3 mm Hg; 247 paired measurements from 157 participants in 7 studies). There was a large amount of variation in bias between the 16 studies where $TcCO_2$ sensors were located at other sites (τ^2=30.1), resulting in extremely wide estimates for population limits of agreement (−134 to 133 mm Hg). Of note, studies that applied the sensor to the earlobe set the temperature of the device to 42°C, whereas studies that applied sensors to the chest or other sites used a variety of different temperature settings. Population limits of agreement were wider in meta-analysis of studies which set the temperature of the sensor to 42°C (−25 to 26 mm Hg) compared with studies where the sensor temperature was higher (−7 to 7 mm Hg). There was large variation in bias between these studies (τ^2=12.0), which was likely due to the location of sensor placement (earlobe in 35 studies, chest in 4 studies and arm/forearm in 3 studies).

$TcCO_2$ monitoring was accurate to a clinically acceptable degree for only a minority of the subgroup meta-analyses conducted according to clinical indication. Population limits of agreement were within the clinically acceptable range for studies that enrolled adults in ICU (16 studies), children undergoing surgery or in ICU (6 studies) and adults undergoing lung function testing (4 studies).

DISCUSSION

It is vital that clinicians have a clear understanding of the accuracy of $TcCO_2$ monitoring devices to ensure they are applied in appropriate circumstances. Both the primary meta-analysis and sensitivity analysis, including only studies at low risk of bias, revealed population limits of agreement outside of the clinically acceptable range. Clinicians using transcutaneous monitoring to assess ventilation status in patients across the broad range of populations included in our systematic review should therefore determine baseline $PaCO_2$ and the $TcCO_2$-$PaCO_2$ gradient and to confirm the diagnosis of hypercapnoea prior to initiation (or cessation) of treatment.

The results from our subgroup analyses have important implications for how $TcCO_2$ monitoring should be applied. No specific

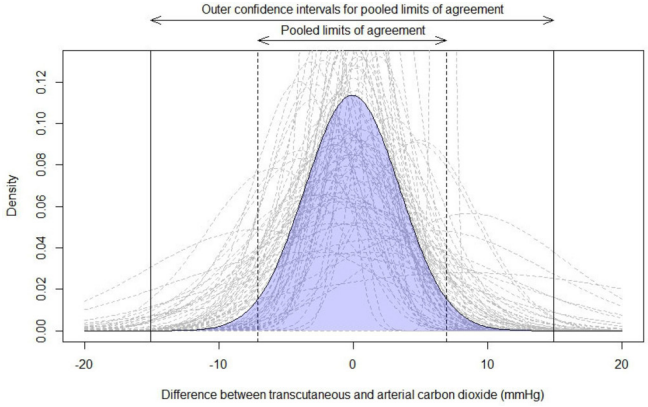

Figure 2 Comparisons within and across studies. Dotted curves are distributions of the differences between arterial and transcutaneous carbon dioxide (CO_2) in individual studies. Solid curve filled with blue is the distribution of the pooled estimate of the difference between arterial and transcutaneous CO_2. Dotted vertical lines indicate bounds for the pooled estimates for limits of agreement between arterial and transcutaneous CO_2. Solid vertical lines indicate bounds for the outer 95% CIs for the pooled estimates of limits of agreement between arterial and transcutaneous CO_2 (ie, population limits of agreement).

Table 2 Summary of findings for accuracy and precision of transcutaneous carbon dioxide monitoring

Outcome	Summary accuracy and precision (mm Hg)	Participants	Study characteristics	Quality	Implications
Main analysis	► Mean bias −0.1. ► Population LoA −15.1 to 14.9.	7021 paired measurements from 2817 participants in 73 studies.	► Low risk of bias (23). ► Large between-study heterogeneity τ^2=8.9.	Low (due to study limitations and inconsistency).	Considering both the primary meta-analysis and results of the sensitivity analysis identified population LoA outside of the clinically acceptable range (7.5 mm Hg), $TcCO_2$ monitoring could not be substituted for $PaCO_2$ across the populations and circumstances in which these devices were studied.
Low-risk studies only	► Mean bias 0.2. ► Population LoA −9.2 to 9.6.	1600 paired measurements from 842 participants in 23 studies.	► Sensor temperature: 42°C (14), not reported (9). ► Sensor on: earlobe (16), not reported (7). ► Clinical setting: ICU (6), chronic respiratory failure (1), PICU (1), neonatal ICU (2), surgery (5), lung function testing (1), acute respiratory failure (6).		
Sentec	► Mean bias 0.1. ► Population LoA −8.5 to 8.7.	3585 paired measurements from 12 565 participants in 30 studies.	► Low risk of bias (10). ► Sensor temperature: 42°C (24), not reported (6). ► Clinical setting: ICU (8), sedation (3), PICU (3), neonates (1), surgery with general anaesthesia (5), lung function testing (2), acute respiratory failure (6), chronic respiratory failure (2).	Moderate (due to study limitations).	The differences in the population LoA identified between $TcCO_2$ device types were small and likely caused by the circumstances in which they are used. The Sentec device sensor was set at 42°C for monitoring regardless of sensor location (eg, earlobe or chest). In contrast, the TOSCA devices more commonly had sensor temperatures set higher than 42°C when monitoring at sites other than the earlobe.
TOSCA	► Mean bias −0.4. ► Population LoA −7.2 to 6.4.	3313 paired measurements from 1561 participants in 45 studies.	► Low risk of bias (17). ► Sensor temperature: 42°C (17), not reported (28). ► Clinical setting: ICU (7), sedation (2), PICU (4), neonates (4), surgery (9), lung function testing (2), acute respiratory failure (8), paediatric surgery (1), healthy participants (1), cardiopulmonary exercise stress testing (1), chronic respiratory failure (6).		
Earlobe	► Mean bias −0.1. ► Population LoA −5.7 to 5.5.	3974 paired measurements from 1786 participants in 45 studies.	► Low risk of bias (19). ► Sensor temperature: 42°C (37), not reported (8). ► Clinical setting: ICU (10), sedation (3), PICU (3), neonates (2), surgery (4), lung function testing (3), acute respiratory failure (11), paediatric surgery (1), healthy participants (1), chronic respiratory failure (5), cardiopulmonary exercise stress testing (1).	Moderate (due to study limitations).	The earlobe should be considered the preferred site for monitoring. Sensor temperature should be set higher than 42°C if not monitoring at the earlobe.
Chest	► Mean bias 1.0. ► Population LoA −10.6 to 12.7.	1041 paired measurements from 471 participants in 13 studies.	► Low risk of bias (2). ► Sensor temperature: 42°C (2), 43°C (7), 44°C (3), not reported (1). ► Clinical setting: ICU (3), neonates (1), surgery (3), lung function testing (1), acute respiratory failure (2), chronic respiratory failure (3).		
Arm	► Mean bias −1.3. ► Population LoA −8.3 to 5.3.	247 paired measurements from 157 participants in 7 studies.	► Low risk of bias (2). ► Sensor temperature: 42°C (2), 43°C (1), 44°C (2), 45°C (2). ► Clinical setting: ICU (1), sedation (1), surgery (4), chronic respiratory failure (1).		
Other monitoring site	► Mean bias −0.8. ► Population LoA −134 to 133.	2156 paired measurements from 448 participants in 16 studies.	► Low risk of bias (4). ► Sensor temperature: 38°C–42°C (1), 42°C (4), 43°C (1), 43.5°C (3), 44°C (2), 45°C (1), not reported (4). ► Clinical setting: ICU (2), sedation (1), PICU (4), neonates (4), surgery (4), acute respiratory failure (1).		
Sensor temperature: 42°C	► Mean bias 0.2. ► Population LoA −24.9 to 25.2.	3635 paired measurements from 1634 participants in 43 studies.	► Low risk of bias (17). ► Sensor on: earlobe (35), chest (4), arm/forearm (3), not reported (1). ► Clinical setting: ICU (9), sedation during surgery (3) PICU (3), neonatal ICU (2), surgery with general anaesthesia (8), lung function testing (3), acute respiratory failure (8), chronic respiratory failure (4), paediatric surgery (1), cardiopulmonary exercise stress testing (1), volunteer (1).	Moderate (due to study limitations).	
Sensor temperature: >42°C	► Mean bias 0.1. ► Population LoA −7.2 to 7.2.	1471 paired measurements from 768 participants in 23 studies.	► Low risk of bias (4). ► Sensor on: earlobe (2), chest (12), arm/forearm (5), not reported (4). ► Clinical setting: ICU (3), sedation during surgery (1) PICU (3), neonatal ICU (2), surgery (6), lung function testing (1), acute respiratory failure (2), chronic respiratory failure (5).		

(), number of studies; ICU, intensive care unit; LoA, limits of agreement; PICU, paediatric intensive care unit; $TcCO_2$, transcutaneous carbon dioxide.

recommendations for a preferred site or sites are provided by manufacturers. Similarly, guidelines on transcutaneous monitoring from the American Association for Respiratory Care do not provide a recommendation for the optimal site to place a $TcCO_2$ sensor.[18] Our analysis indicates that TOSCA and Sentec $TcCO_2$ device sensors should preferentially be placed on the earlobe because the population limits of agreement were within the bounds of the clinically acceptable range (<7.5 mm Hg). Monitoring on the earlobe had similar LoAs to those reported in a meta-analysis of capillary blood gas LoAs, where the mean bias was −0.1 mm Hg and the SD of bias was 2.9 mm Hg.[19] If $TcCO_2$ monitoring is feasible from the earlobe, it should be considered the preferable solution to use for ventilation assessment over capillary blood gas because of its non-invasiveness and ability to provide a continuous assessment of ventilation. Likewise, the results of our meta-analyses suggest that $TcCO_2$ measurements would be more accurate than estimations of $PaCO_2$ derived from venous blood gas samples. The mean bias between $PaCO_2$ and venous blood CO_2 measurements was estimated to range from −10.7 mm Hg and 2.4 mm Hg in a meta-analysis of 16 studies.[20] Pooled estimates of the LoAs between venous and arterial measurements of CO_2 were not reported.[20] It should be noted though that not all patients who may benefit from continuous ventilation assessment will be suitable for application of a $TcCO_2$ monitoring sensor to their earlobes. For example, the earlobes of a neonate requiring ventilation assessment may not be large enough to accommodate a $TcCO_2$ sensor. Adult patients with multiple piercings, undergoing surgery to or with trauma around the head and neck would also preclude the application of a $TcCO_2$ sensor to the earlobe.

We found large differences in population limits of agreement in subgroup analyses focusing on sensor temperature. Meta-analysis restricted to studies that used a sensor temperature of 42°C exhibited worse agreement with $PaCO_2$ in comparison with meta-analysis of studies that used higher temperatures. The difference in these results can be explained by the large between-study variation in bias in the subgroup analysis of studies that used a sensor temperature of 42°C but applied the sensor to either the earlobe, chest or arm. Of note, the majority of studies where the sensor was applied to participants' earlobes set the temperature of the sensor to 42°C but still yielded clinically acceptable population limits of agreement. Together, the findings from both subgroup analyses indicate that if monitoring on the earlobe is not possible, the sensor temperature should be set higher than 42°C.

A strength of this analysis is the incorporation of the variation in bias between studies, the bias associated with repeated measures not accounted for in the analysis of individual studies, as well as measures of uncertainty (CIs) into our estimates of the agreement between transcutaneous and arterial CO_2 measurements. Standard meta-analysis approaches focused on providing the average bias and the average precision separately may have erroneously led clinicians to believe that the agreement between $TcCO_2$ and $PaCO_2$ measurements is acceptable. The risk here is clearly evident for the primary meta-analysis where the estimate of the population limits of agreement were within clinically acceptable bounds but the outer 95% CIs were far wider. There was a larger difference between the population limits of agreement and the CIs because the sampling variation in each component of the LoA is taken into account when calculating the CIs (the mean bias, SD and variation in bias between studies). By incorporating the between-study heterogeneity in bias and sampling variation, it is clear that the LoAs in the population are much broader and not clinically acceptable for interchangeable use across the range of situations in which $TcCO_2$ monitoring was tested in the main analysis.

The trending ability of $TcCO_2$ monitors is an important information for clinicians to consider when using $TcCO_2$ in practice. This is because $PaCO_2$ continuously changes in response to a variety of factors. In addition, evaluating trends in $TcCO_2$ may be useful in clinical practice for evaluating the effectiveness of interventions employed to improve ventilation status. Conclusions about trending ability can be drawn from the accuracy and precision of absolute measurements (ie, LoAs) by making a qualitative judgement about whether or not the index test is sufficiently precise. We did not identify strong evidence to support the trending ability of $TcCO_2$ for ventilation assessment in this systematic review because the population limits of agreement for the primary meta-analysis were wide. If part of the imprecision relates to patient-specific characteristics, such as vascularity, it is possible that there will be a systematic measurement error for within-patient readings, and therefore the within-patient trend may have tighter LoAs than individual measurements. Therefore, methods other than the Bland-Altman approach may be more suited to quantitatively assess trending ability. For example, the 4Q, polar analysis and clinical concordance methods have recently been recommended for consideration in method comparison studies evaluating the validity of cardiac output monitors.[21] Similar to $TcCO_2$ monitoring, cardiac output monitors provide a continuous estimate of a dynamic physiological parameter, which changes rapidly from various influences. As such, further research aiming to examine the trending ability of $TcCO_2$ monitoring should consider incorporating such assessments.

The results of subgroup analyses according to the indication for monitoring identified specific clinical areas where further research into the accuracy of $TcCO_2$ monitoring would be beneficial because population limits of agreement were outside of the clinically acceptable range. These include acute respiratory failure, thoracic surgery with single lung ventilation and for assessment of ventilation in patients who are sedated during or after surgery. However, it should be noted that studies within these subgroups applied $TcCO_2$ monitoring using different devices (Sentec and TOSCA), sensor locations and temperatures, which may explain the large variation in bias between studies and resulting imprecise estimates of LoAs. Clinicians who use $TcCO_2$ monitoring in these areas should be confident that $TcCO_2$ measurements would be within clinically acceptable agreements if applied to the earlobe or another monitoring site at a temperature above 42°C using either the Sentec or TOSCA device.

Transcutaneous CO_2 monitoring is commonly used to reduce the frequency of arterial blood gas analysis in neonates due to the limitations of other methods to estimate $PaCO_2$ in this population, such as end-tidal CO_2 and capillary blood gas analysis.[22] Yet we identified weak evidence for accuracy and precision with population limits of agreement that were far outside the clinically acceptable range. It should be noted though that we only included studies that compared $TcCO_2$ with $PaCO_2$. Several articles were excluded due to $TcCO_2$ being compared instead with capillary blood gas analysis.[23]

Limitations

Data were not extracted on adverse events related to transcutaneous monitoring. We cannot rule out the possibility of publication bias. However, this may not be as serious a problem for diagnostic test accuracy studies as it is for randomised trials.[8] We did not use meta-regression or tests for interaction between subgroups to investigate for sources of heterogeneity because of our focus on the

population limits of agreement, which incorporated the variation in bias between studies into the estimates. This systematic review did not assess the clinical utility of these monitors. Therefore, the evidence to be derived from this systematic review should only be considered within the context of other information about the clinical utility, reliability and ease of use of these devices during normal clinical practice. Of note, the level we set as the limit for clinically acceptable agreement between $PaCO_2$ and $TcCO_2$ (7.5 mm Hg) was chosen based on recommendations for ventilation monitoring made by the American Association for Respiratory Care therapists. If a difference in repeated measurements of $PaCO_2$ less than this magnitude would be important for a given situation in clinical practice, then an arterial blood sample should be drawn to confirm diagnosis when a change in $TcCO_2$ is observed during monitoring.

CONCLUSION

This meta-analysis has identified that there may be substantial differences between $TcCO_2$ and $PaCO_2$ depending on the context in which this technology is used in clinical practice. Measuring $TcCO_2$ from the earlobe with either the TOSCA or Sentec device would yield clinically acceptable accuracy. As such, this monitoring site is recommended for use in clinical practice. For optimal accuracy and precision, users should set the temperature of the sensor higher than 42°C when monitoring at sites other than the earlobe.

Contributors All authors have agreed on the final version. AC: study design, data extraction and quality assessment, statistical analysis, drafted the manuscript, and is responsible for the overall content as the guarantor. ET: study design, statistical analysis, revised the manuscript for important intellectual content. W-HL: study design, data extraction and quality assessment, revised the manuscript for important intellectual content. ZC: data extraction and quality assessment, revised the manuscript for important intellectual content. KS: data extraction and quality assessment, revised the manuscript for important intellectual content. JS: study design, data extraction and quality assessment, revised the manuscript for important intellectual content. JF: study design, data extraction and quality assessment, revised the manuscript for important intellectual content.

Funding Department of Health, Australian Government, National Health and Medical Research Council (APP1091657).

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data used for the study are available as online supplementary information and at https://doi.org/10.6084/m9.figshare.6244058.v2.

REFERENCES

1 Huttmann SE, Windisch W, Storre JH. Techniques for the measurement and monitoring of carbon dioxide in the blood. *Ann Am Thorac Soc* 2014;11:645–52.
2 Williams AJ. ABC of oxygen: assessing and interpreting arterial blood gases and acid-base balance. *BMJ* 1998;317:1213–6.
3 Preston L, Carroll C, Gardois P, *et al*. 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. *Syst Rev* 2015;4:82.
4 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. *Syst Rev* 2012;1:9.
5 Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. *Br J Anaesth* 2007;99:309–11.
6 Whiting PF, Rutjes AW, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.
7 Glasziou P, Irwig L, Bain C, *et al*. *Systematic Reviews in Health Care: A Practical Guide*: Taylor & Francis, 2002.
8 Begg CB. Systematic reviews of diagnostic accuracy studies require study by study examination: first for heterogeneity, and then for sources of heterogeneity. *J Clin Epidemiol* 2005;58:865–6.
9 Niven DJ, Gaudet JE, Laupland KB, *et al*. Accuracy of peripheral thermometers for estimating temperature: a systematic review and meta-analysis. *Ann Intern Med* 2015;163:768–77.
10 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. *J Clin Epidemiol* 2005;58:882–93.
11 Schünemann HJ, Schünemann AH, Oxman AD, Brozek J, *et al*. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106–10.
12 Tipton E, Shuster J. A framework for the meta-analysis of Bland-Altman studies based on a limits of agreement approach. *Stat Med* 2017;36:3621–35.
13 Tanner-Smith EE, Tipton E, Polanin JR. Handling complex meta-analytic data structures using robust variance estimates: a tutorial in R. *J Dev Life Course Criminol* 2016;2:85–112.
14 Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods* 2010;1:164–5.
15 Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods* 2015;20:375–93.
16 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
17 Team RC Core Team R. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for statistical computing, 2017. http://www.R-project.org/.
18 Restrepo RD, Hirst KR, Wittnebel L, *et al*. AARC clinical practice guideline: transcutaneous monitoring of carbon dioxide and oxygen: 2012. *Respir Care* 2012;57:1955–62.
19 Zavorsky GS, Cao J, Mayo NE, *et al*. Arterial versus capillary blood gases: a meta-analysis. *Respir Physiol Neurobiol* 2007;155:268–79.
20 Byrne AL, Bennett M, Chatterji R, *et al*. Peripheral venous and arterial blood gas analysis in adults: are they comparable? A systematic review and meta-analysis. *Respirology* 2014;19:168–75.
21 Montenij LJ, Buhre WF, Jansen JR, *et al*. Methodology of method comparison studies evaluating the validity of cardiac output monitors: a stepwise approach and checklist. *Br J Anaesth* 2016;116:750–8.
22 Mukhopadhyay S, Maurer R, Puopolo KM. Neonatal transcutaneous carbon dioxide monitoring--effect on clinical management and outcomes. *Respir Care* 2016;61:90–7.
23 Karlsson V, Sporre B, Ågren J. Transcutaneous pco2 monitoring in newborn infants during general anesthesia is technically feasible. *Anesth Analg* 2016;123:1004–7.