# UNDERSTANDING THE ANALYSIS OF METHOD COMPARISON STUDIES WITH REPEATED MEASUREMENTS OF CLINICAL DATA

**KARINE BROUSSEAU**

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the Master of Science in Epidemiology

School of Epidemiology and Public Health

Faculty of Medicine

University of Ottawa

# TABLE OF CONTENTS


**THESIS ABSTRACT** **iv**

**CHAPTER 1. INTRODUCTION AND THESIS OVERVIEW** **1**

*1.1 BACKGROUND AND RATIONALE* *1*

*1.2 THESIS OBJECTIVES* *5*

*1.3 THESIS OVERVIEW* *6*

*1.4 REFERENCES* *7*

**CHAPTER 2. PROPOSED STATISTICAL METHODS FOR THE ANALYSIS OF METHOD COMPARISON STUDIES WITH REPEATED MEASUREMENTS OF CLINICAL DATA – A SCOPING REVIEW PROTOCOL** **12**

*2.1 PREFACE TO CHAPTER 2* *13*

*2.2 ABSTRACT* *14*

*2.3 INTRODUCTION* *16*

*2.4 METHODS* *20*

*2.5 CONCLUSION* *29*

*2.6 REFERENCES* *30*

*2.7 APPENDIX 1: SEARCH STRATEGY FOR THE MEDLINE DATABASE* *34*

**CHAPTER 3. A SCOPING REVIEW OF STATISTICAL METHODS FOR THE ANALYSIS OF METHOD COMPARISON STUDIES WITH REPEATED MEASUREMENTS OF CLINICAL DATA** **35**

*3.1 PREFACE TO CHAPTER 3* *36*

*3.2 ABSTRACT* *37*

*3.3 INTRODUCTION* *39*

*3.4 METHODS* *41*

*3.5 RESULTS* *48*

*3.6 DISCUSSION* *54*

*3.7 CONCLUSION* *58*

*3.8 REFERENCES* *59*

*3.9 TABLES* *66*

*3.10 FIGURES* *87*

*3.11 APPENDIX 1. PRISMA-SCR CHECKLIST* *88*

**CHAPTER 4. PROTOCOL FOR THE PREMISE OBSERVATIONAL STUDY: POINT-OF-CARE HEMOGLOBIN ACCURACY AND TRANSFUSION OUTCOMES IN NON-CARDIAC SURGERY AT A CANADIAN TERTIARY ACADEMIC HOSPITAL** **90**

*4.1 PREFACE TO CHAPTER 4* *91*

*4.2 ABSTRACT* *92*

*4.3 INTRODUCTION* *94*

*4.4 METHODS: PARTICIPANTS, INTERVENTIONS AND OUTCOMES* *99*

*4.5 METHODS: DATA COLLECTION, MANAGEMENT, AND ANALYSIS* *102*

*4.6 METHODS: MONITORING* *107*

*4.7 ETHICS AND DISSEMINATION* *108*

*4.8 REFERENCES* *111*

*4.9 TABLE* *122*

*4.10 FIGURE* *123*

**CHAPTER 5. STATISTICAL ANALYSIS PLAN FOR THE PREMISE STUDY** **124**

*5.1 PREFACE TO CHAPTER 5* *125*

*5.2 SECTION 1 – ADMINISTRATIVE INFORMATION* *126*

*5.3 SECTION 2 – INTRODUCTION* *128*

*5.4 SECTION 3 – STUDY METHODS* *130*

*5.5 SECTION 4 – STATISTICAL PRINCIPLES* *133*

*5.6 SECTION 5 – STUDY POPULATION* *135*

*5.7 SECTION 6 – ANALYSIS* *137*

*5.8 REFERENCES* *149*

**CHAPTER 6. INTRAOPERATIVE POINT-OF-CARE HEMOGLOBIN MEASUREMENT AND ACCURACY: A REPORT FROM THE PREMISE PROSPECTIVE OBSERVATIONAL STUDY** **153**

*6.1 PREFACE TO CHAPTER 6* *154*

*6.2 KEY POINTS* *155*

*6.3 ABSTRACT* *156*

*6.4 INTRODUCTION* *158*

*6.5 METHODS AND ANALYSIS* *160*

*6.6 RESULTS* *164*

*6.7 DISCUSSION* *167*

*6.8 OTHER INFORMATION* *170*

*6.9 REFERENCES* *171*

*6.10 TABLES* *180*

*6.11 FIGURES* *184*

*6.12 APPENDIX 1. STANDARDS FOR REPORTING DIAGNOSTIC ACCURACY STUDIES (STARD) CHECKLIST* *187*

*6.13 APPENDIX 2. SUPPLEMENTARY TABLES AND FIGURES* *189*

**CHAPTER 7. INTEGRATED DISCUSSION** **193**

**CHAPTER 8. ACKNOWLEDGEMENTS** **203**

**CHAPTER 9. FUNDING** **204**

## THESIS ABSTRACT

Method comparison studies consist of a unique study design aiming to examine agreement between two methods to measure a physiological or clinical parameter evaluated using continuous variables. Such physiological parameters are used by healthcare providers along with other clinical data to inform diagnoses and treatment decisions. When novel methods are proposed to measure a continuous physiological parameter, method comparison studies are needed to examine the agreement between this new method and an existing method that is used in standard clinical care. This standard method is generally considered as the gold standard measurement for a given physiological parameter. The issue of repeated measurements poses special challenges when conducting method comparison studies. Repeated measurements occur when a given individual included in a method comparison study has multiple measurements, which are inherently correlated with one another and are not independent (e.g., multiple glucose measurements carried out for the same patient throughout the day using a blood test, compared to a portable point-of-care device). The limits of agreement (LOA) method proposed by Bland & Altman has been adapted to adjust for the correlation between repeated measurements and is widely used for the analysis of method comparison studies that include repeated measurements. However, other statistical methods have been proposed as alternatives to LOA analysis to inform the analysis of method comparison studies with repeated measurements. There is a gap in the literature to inform this type of analysis, whereby no guidelines or synthesis of statistical methods that can be used as alternatives to the LOA method with repeated measurements have been published. Therefore, this thesis aimed to systematically review the existing literature to identify existing alternate statistical methods for the analysis of method comparison studies that include repeated measurements, using a scoping review framework. The findings of this scoping review were used to inform the analysis of the PREMISE (Point-of-care hemoglobin accuracy and

transfusion outcomes in non-cardiac surgery) study, a large prospective observational method comparison study that included repeated measurements. The aim of the PREMISE study was to examine the agreement between frequently used point-of-care devices to measure hemoglobin (POCT-Hgb) and laboratory-measured hemoglobin (lab-Hgb) in the operative setting. To further increase the understanding of the challenges associated with the analysis of method comparison studies that include repeated measurements, the analyses pertaining to agreement were performed in the context of this thesis. The findings of the PREMISE study fill an important gap in the literature pertaining to transfusion decision-making in the operative setting, where there is a paucity of evidence on the accuracy of POCT-Hgb devices, as well as from trials and transfusion guideline.

# CHAPTER 1. INTRODUCTION AND THESIS OVERVIEW

## 1.1 BACKGROUND AND RATIONALE

### 1.1.1. The challenges related to method comparison studies with repeated measurements

In healthcare, clinicians typically require a variety of patient-specific health information to secure an appropriate diagnosis, determine an optimal treatment plan, and facilitate important clinical decision-making aimed at improving patient care. Part of this crucial health information consists of physiological and clinical parameters often measured in a continuous fashion, therefore providing measurements that can take a wide range of values. An example of such a physiological parameter is hemoglobin (Hgb), a protein present in blood, more specifically in red blood cells, which delivers oxygen to all cells and tissues in the human body.[1] There are often several methods or devices for the measurement of specific physiological parameters, and new methods are constantly developed to improve efficiency. When faced with multiple existing methods to measure the same parameter, it is critical to define the method that is the most capable of providing accurate measurements in a given clinical setting. Indeed, clinicians must choose the method that is the most efficient in terms of timeliness, cost, and invasiveness without compromising the accuracy of measurements.[2–5] This is particularly important in the clinical setting, where inaccurate measurements may lead to harm or inappropriate treatment decisions.[6,7]

To help address the above question of agreement, method comparison studies were designed to examine the extent of agreement between two methods measuring the same continuous parameter.[8] Method comparison studies are typically used to determine whether a new proposed method can be used interchangeably with a currently validated method.[8] The validated method used as a comparator is usually considered the reference method, or gold standard, thus providing the closest measurement to the underlying true value.[3,6–8]

For several years, the methodological literature has emphasized the importance of including repeated measurements in method comparison studies to examine the repeatability of each method, which can have an effect on the agreement between two methods.[8–10] Therefore, it is recommended to measure a physiological parameter multiple times within a research participant using each of the methods being compared simultaneously.[8–10] This further increases comparability to the actual clinical setting where physiological parameters may be measured repeatedly.[11] An example of a clinical setting in which multiple measurements may be required is the measurement of Hgb in the operative setting, where hemodynamic parameters can potentially change rapidly due to blood loss.[12] In such case, repeated measurements of Hgb at critical time points may inform clinical decision-making in terms of the need for and timing of blood transfusions.[13] The challenge associated with the inclusion of repeated measurements is that such measurements are correlated and cannot be considered as independent, referred to as the subject by method interaction, thus requiring complex statistical adjustments to avoid erroneous conclusions concerning the agreement between two methods.[10,14]

Although including repeated measurements in method comparison studies is considered as best practice, the analysis of such data is challenging. After stressing the importance of repeated measurements in such studies, Bland & Altman have proposed statistical methods to inform the analysis of method comparison studies that include repeated measurements, called the limits of agreement (LOA).[8,10] In the analysis of method comparisons studies, the LOA method generates a range of differences between measurements from two methods being compared in which it can be expected that approximately 95% of differences will lie.[8–10] In the absence of repeated measurements, this range is obtained quite simply by computing the mean bias of the differences (i.e., average difference between methods), from which 1.96 standard deviation of the differences between methods is added or subtracted, thus providing the upper and lower point estimates of the limits of agreement.[8,9] To account for repeated measurements, the LOA may be adjusted, using a standard deviation that is adjusted for the correlation within subjects, and the

correlation between subjects.[10] They propose two similar but distinct methods to the adjustment of the LOA, depending on the underlying true value of the parameter being measured.[10] The first method is used when the true value varies, meaning that the repeated measurements obtained may have very different results, as the measured parameter has the potential to change rapidly.[10] The second method is used when the underlying true value is constant, meaning that the true value of the measured parameter should be similar for each repeated measurement taken over a specific period of observation.[10]

One limitation of the LOA method for repeated measurements is that the distribution of the data must follow certain statistical assumptions.[3,10,15,16] Those assumptions include that the differences must be normally distributed, that the agreement between methods is similar across the range of measurements, and that the variance of the repeated measurements within subjects is similar over the range of measurements.[10] When these assumptions are not met, other authors have suggested alternative statistical methods for the analysis of method comparison studies with repeated measurements.[3,15,16]

The LOA method appears to be the standard approach for the analysis of method comparison studies as it is the statistical method that is most often used in the literature[17]. Since its original proposal, a plethora of other statistical methods have been proposed for the analysis of method comparison studies with repeated measurements. However, it is clear that not all such new methods are appropriate to analyze this type of data. For instance, correlation coefficients, mean agreement, and linear regression continue to be used to examine agreement, despite serious reservations having been expressed by Bland & Altman[4,8,9,17,18]. A tutorial to inform the analysis of method comparison studies with repeated measurements has been developed and presented five different statistical methods, including the LOA approach.[19] However, to our knowledge, there has been no synthesis encompassing all described statistical methods and thus finding appropriate such methods may be challenging. This thesis therefore aimed to fill this important knowledge gap.

### 1.1.2 The PREMISE study

An example of a large prospective method comparison study that includes repeated measurements is the PREMISE (**P**oint-of-ca**re** he**m**oglob**i**n accuracy and tran**s**fusion outcom**e**s in non-cardiac surgery) study. The PREMISE study aimed to examine the accuracy of commonly utilized point-of-care testing devices (POCT) measuring Hgb (POCT-Hgb) in the operating room. Briefly, when the measurement of Hgb concentrations was required during surgery, three POCT-Hgb devices were simultaneously performed, along with a laboratory-determined Hgb (lab-Hgb) via analysis of a complete blood count (CBC). Multiple measurements of Hgb were obtained as required for standard clinical care over the course of a given operation. These were considered repeated measurements. To examine the accuracy of such devices, agreement between POCT-Hgb measurements and the lab-Hgb measurement was examined using various statistical methods, adjusting for repeated measurements when possible. In the context of the PREMISE study, 1735 blood samples collected during surgery from a total of 1139 individual participants were analyzed to examine the accuracy of POCT-Hgb devices, when compared with a lab-Hgb measurement.

The PREMISE study was designed to fill an important knowledge gap identified within the field of transfusion medicine in the operating room. Hgb measurements during surgery are an important parameter to inform red blood cell (RBC) transfusions in the operating room[20,21], with current practice guidelines suggesting to limit RBC transfusions to those with Hgb measurements between 60 and 100 g/L.[13,22] Despite the fact that a lab-Hgb measurement is considered as gold standard, POCT-Hgb devices are increasingly relied upon for the measurement of Hgb in the operating room as they are readily available for use intra-operatively, simple to use, and provide results within seconds. While POCT-Hgb devices have been validated in non-operative contexts, where Hgb values are usually static or within normal limits[23–27], there is a paucity of evidence on the accuracy of such devices in the intra-operative setting, where Hgb values have the potential

to change drastically and rapidly.[12,28] Moreover, the evidence base surrounding their use in the operating room is even more limited with Hgb measurements of 60-100 g/L, where the potential to transfuse is increased.[28–30]

Accuracy of Hgb measurements in the operative setting is of high relevance and interest, as inaccurate Hgb measurements may lead to potential mistransfusion events in the operating room. While RBC transfusions can be life-saving intervention, they also carry important peri-operative risks when they are not clinically indicated.[31,32] Thus, results pertaining to the agreement of POCT-Hgb devices from the PREMISE have the potential to further inform future transfusion decision-making in the operative setting.

### 1.2 THESIS OBJECTIVES

The global thesis objective was to understand the statistical analysis of method comparison studies that include repeated measurements of clinical continuous data to inform the analysis of the PREMISE study. There were two main research aims that guided the overall conduct of this thesis, each with specific objectives geared towards achieving the global thesis objective.

The first research aim was to identify existing analytical methods for repeated and correlated measurements of clinical data. This research aim had the objective of carrying out a systematic examination of the existing literature to identify methodological publications that present statistical methods that inform the analysis of method comparison studies that included repeated measurements, including the LOA method.

The second research aim was to understand the processes involved in the analysis of large prospective method comparison studies that include repeated measurements, using the PREMISE study as an example. The specific objectives related to this research aim were to develop a complete and comprehensive statistical analysis plan (SAP), informed by the findings of the scoping review, to understand the benefits and challenges of the execution of a SAP to

inform the analysis of a study, and to carry out the analysis of the agreement component of PREMISE

### 1.3 THESIS OVERVIEW

Two separate studies were conducted to achieve the global objective and the research aims associated with this thesis, as outlined above. In total, five manuscripts were generated in the context of this thesis, each representing a distinct chapter.

The first study consisted of a scoping review, designed to accomplish the first research aim. The protocol for the scoping review is presented in Chapter 2 and it was registered on Open Science Framework (September 8, 2023).[33] The manuscript presenting the scoping review is shown in Chapter 3. An abridged version of this manuscript will be published in a peer-reviewed journal, the length and format of of which will be adjusted to match the journal's publishing guidelines.

The second study was the PREMISE observational prospective method comparison study. The study protocol for the PREMISE study is presented in Chapter 4. It was accepted for publication in BMJ Open and is currently in press. The statistical analysis plan for the PREMISE study is presented in Chapter 5 and was submitted as an appendix to the PREMISE protocol during the submission process with BMJ Open. Finally, the manuscript presenting the main results of the PREMISE study, obtained for the statistical analyses performed in the context of this thesis, are presented in Chapter 6. This manuscript will be published in a peer-reviewed journal subsequently.

## 1.4 REFERENCES


1. Billett HH. Hemoglobin and Hematocrit. In: Walker HK, Hall WD, Hurst JW, eds. *Clinical Methods: The History, Physical, and Laboratory Examinations*. 3rd ed. Butterworths; 1990. Accessed September 16, 2023. http://www.ncbi.nlm.nih.gov/books/NBK259/

2. Mabey D, Peeling R. The Optimal Features of a Rapid Point-of-Care Diagnostic Test. In: *Revolutionizing Tropical Medicine: Point-of-Care Tests, New Imaging Technologies and Digital Health*. John Wiley & Sons, Inc; 2019:81-87. doi:10.1002/9781119282686.ch3

3. Schluter PJ. A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies. *BMC Med Res Methodol*. 2009;9:6. doi:10.1186/1471-2288-9-6

4. Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. *J R Stat Soc, B: Stat Methodol*. 1983;32(3):307-317. doi:10.2307/2987937

5. Choudhary PK. A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *J Stat Plan Inference*. 2008;138:1102-1115. doi:10.1016/j.jspi.2007.03.056

6. Celentano D, Szklo M. More on Causal Inference: Bias, Confounding, and Interaction. In: *Gordis Epidemiology*. 6th ed. Elsevier; 2019:289-306.

7. Weiss NS. Clinical Epidemiology. In: *Modern Epidemiology*. 3rd ed. Lippincott Williams & Wilkins; 2008:641-651.

8. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.

9. Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*. 1986;327:307-310. doi: 10.1016/S0140-6736(86)90837-8

10. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. *J Biopharm Stat*. 2007;17:571-582. doi:10.1080/10543400701329422

11. Hiriote S, Chinchilli VM. Matrix-based Concordance Correlation Coefficient for Repeated Measures. *Int J Biom*. 2011;67:1007-1016. doi:10.1111/j.1541-0420.2010.01549.x

12. Wu PBS, Morey TE, Harris NS, Gravenstein N, Rice MJ. Intravenous Fluids Cause Systemic Bias in a Conductivity-Based Point-of-Care Hematocrit Meter. *Anesth Analg*. 2012;114(2):314-321. doi:10.1213/ANE.0b013e31823fecbd

13. Practice Guidelines for Perioperative Blood Management: An Updated Report by the American Society of Anesthesiologists Task Force on Perioperative Blood Management. *Anesthesiology*. 2015;122(2):241-275. doi:10.1097/ALN.0000000000000463

14. Schober P, Vetter TR. Repeated Measures Designs and Analysis of Longitudinal Data: If at First You Do Not Succeed—Try, Try Again. *Anesth Analg*. 2018;127(2):569-575. doi:10.1213/ANE.0000000000003511

15. Taffé P. When can the Bland & Altman limits of agreement method be used and when it should not be used. *J Clin Epidemiol*. 2021;137:176-181. doi:10.1016/j.jclinepi.2021.04.004

16. Stevens NT, Steiner SH, MacKay RJ. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Stat Methods Med Res*. 2017;26(6):2487-2504. doi:10.1177/0962280215601133

17. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLoS ONE*. 2012;7(5):e37908. doi:10.1371/journal.pone.0037908

18. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*. 1990;20(5):337-340. doi:10.1016/0010-4825(90)90013-F

19. Parker RA, Scott C, Inácio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. *BMC Med Res Methodol*. 2020;20:154. doi:10.1186/s12874-020-01022-x

20. Bennett S, Ayoub A, Tran A, et al. Current practices in perioperative blood management for patients undergoing liver resection: a survey of surgeons and anesthesiologists. *Transfusion*. 2018;58:781-787. doi:10.1111/trf.14465

21. Jouffroy R, Baugnon T, Carli P, Orliaguet G. A survey of blood transfusion practice in French-speaking pediatric anesthesiologists. *Paediatr Anaesth*. 2011;21:385-393. doi:10.1111/j.1460-9592.2011.03531.x

22. Baker L, Park L, Gilbert R, et al. Intraoperative Red Blood Cell Transfusion Decision-making: A Systematic Review of Guidelines. *Ann Surg*. 2021;274(1):86-96. doi:10.1097/SLA.0000000000004710

23. Papadea C, Foster J, Grant S, et al. Evaluation of the i-STAT Portable Clinical Analyzer for Point-of-Care Blood Testing in the Intensive Care Units of a University Children's Hospital. *Ann Clin Lab Sci*. 2002;32(3):231-243.

24. Baart AM, de Kort WLAM, van den Hurk K, Pasker-de Jong PCM. Hemoglobin assessment: precision and practicability evaluated in the Netherlands—the HAPPEN study. *Transfusion*. 2016;56:1984-1993. doi:10.1111/trf.13546

25. Steinfelder-Visscher J, Teerenstra S, Klein Gunnewiek JMT, Weerwind PW. Evaluation of the i-STAT Point-of-Care Analyzer in Critically Ill Adult Patients. *J Extra Corpor Technol*. 2008;40:57-60.

26. Dolscheid-Pommerich RC, Dolscheid S, Grigutsch D, Stoffel-Wagner B, Graeff I. Comparability of Point-of-Care versus Central Laboratory Hemoglobin Determination in Emergency Patients at a Supra-Maximal Care Hospital. *PLoS One*. 2016;11(11):e0166521. doi:10.1371/journal.pone.0166521

27. Hiscock R, Simmons SW, Carstensen B, Gurrin LC. Comparison of Massimo Pronto-7 and Hemocue Hb 201+ with Laboratory Haemoglobin Estimation: A Clinical Study. *Anaesth Intensive Care*. 2014;42(5):608-613. doi:10.1177/0310057X1404200510

28. Johnson M, Marwick PC, Coetzee JF. Comparison of Hemoglobin Measurements by 3 Point-of-Care Devices With Standard Laboratory Values and Reliability Regarding Decisions for Blood Transfusion. *Anesth Analg*. 2020;131(2):640-649. doi:10.1213/ANE.0000000000004533

29. Morey TE, Gravenstein N, Rice MJ. Let's Think Clinically Instead of Mathematically About Device Accuracy. *Anesth Analg*. 2011;113(1):89-91. doi:10.1213/ANE.0b013e318219a290

30. Rice MJ, Gravenstein N, Morey TE. Noninvasive Hemoglobin Monitoring: How Accurate Is Enough? *Anesth Analg*. 2013;117(4):902-907. doi:10.1213/ANE.0b013e31829483fb

31. Hendrickson JE, Hillyer CD. Noninfectious Serious Hazards of Transfusion. *Anesth Analg*. 2009;108(3):759-769. doi:10.1213/ane.0b013e3181930a6e

32. Callum J, Pinkerton P, Lima A, et al. *Bloody Easy 4: Blood Transfusions, Blood Alternatives and Transfusion Reactions*. 4th ed. ORBCoN; 2016.

33. Brousseau K, McIsaac DI, Ramsay T, Fergusson DA, Martel G. Proposed statistical methods for the analysis of method comparison studies with repeated measurements of clinical data – A scoping review protocol. Published online September 8, 2023. doi:https://doi.org/10.17605/OSF.IO/4P8UT

# CHAPTER 2. PROPOSED STATISTICAL METHODS FOR THE ANALYSIS OF METHOD COMPARISON STUDIES WITH REPEATED MEASUREMENTS OF CLINICAL DATA – A SCOPING REVIEW PROTOCOL

---

Karine Brousseau[1,2,3], Daniel I. McIsaac[1,2,4], Tim Ramsay[1,2], Dean A. Fergusson[1,2], Guillaume Martel[1,2,3]

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[3]Liver and Pancreas Unit, Department of Surgery, The Ottawa Hospital, Ottawa, Ontario, Canada

[4]Department of Anesthesiology and Pain Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada

## AUTHOR CONTRIBUTIONS

Drs. Martel and Fergusson helped with the design and conception of the study. All authors critically revised the protocol and provided feedback on all study aspects, as needed.

## PUBLICATION INFORMATION

This protocol was registered on Open Science Framework (https://osf.io/4p8ut) September 8, 2023.

## 2.1 PREFACE TO CHAPTER 2

We started by designing and creating the protocol for the scoping review of proposed statistical methods for the analysis of method comparison studies with repeated measurements of clinical data. This protocol was created prior to initiating any study related procedures and was used to inform the conduct of the study. The search strategy was developed with the help of a research librarian, Sarah Visintini, who provided insight and expertise during the creation of the search strategy and adapting the search strategy to all sources of evidence. Her contribution has been acknowledged in both the study protocol and the manuscript presenting our findings, in Chapter 3.

## 2.2 ABSTRACT

**Introduction:** Treatments in medicine are frequently guided by physiologic parameters. For these continuous data to be useful, clinicians must know whether these parameters are appropriate to inform decisions. Method comparison studies can help determine the level of agreement between a new continuous diagnostic tool and the current gold standard. The Bland & Altman limits of agreement (LOA) method is the most commonly recommended and used approach to analyze method comparisons for continuous measurements. Repeated measurements in method comparison studies pose special analytic challenges, and reflect what may happen in a clinical setting where tests can be repeated in a given patient. Several statistical methods exist in the literature for their analysis, but not all are methodologically appropriate for this type of data. The aim of this paper is to present a scoping review protocol, with the objective of synthesizing statistical methods for the analysis of method comparison studies with repeated and continuous measurements, and to compare them to the Bland & Altman LOA method adapted for repeated measurements.

**Methods:** Potentially eligible articles will be identified from the literature using a peer-reviewed search strategy applied to the Medline, Embase, CINAHL, Web of Science, and ProQuest databases. The grey literature will be searched using publications in online forums, Google, as well as textbooks and book chapters identified using Google Books. Additional potentially eligible articles will be identified with forward citation chain searching of the Bland & Altman article proposing adjusted LOA methods accounting for repeated measurements. Finally, backward citation chain searching will be performed on articles that are included in the review. Abstract and full-text screening, as well as data extraction, will be done in duplicate by two independent reviewers. Any discrepancies during will be resolved by consensus. The primary outcome will be the results obtained by various statistical methods for the analysis of method comparison studies with repeated measures and how they compare to the LOA method using clinical or simulated

data. A qualitative appraisal of individual statistical methods identified and compared in eligible articles will be performed.

**Ethics and dissemination**: Ethics approval to conduct this study is not required. Results will be disseminated in scientific journals.

**ABBREVIATIONS**: JBI, Joanna Briggs Institute; LOA, limits of agreement; PRISMA-ScR, Preferred reporting items for systematic review and meta-analysis extension for scoping reviews.

## 2.3 INTRODUCTION

### Rationale

When providing care for patients, clinicians gather clinical data that help secure a diagnosis, inform decision-making, and assist in providing appropriate care and treatment. Commonly, several available instruments can provide data on the same physiological parameters, but it is noteworthy that not all methods are equally accurate or precise. When faced with multiple or new diagnostic tools to generate health data, clinicians must choose the most efficient method in terms of time, cost, and invasiveness, without compromising the accuracy and precision of the results.[1] It is particularly important that new diagnostic methods are sufficiently accurate to inform clinical decisions, as inaccurate measurements may lead to harm or inappropriate treatment decisions.[2,3] In this protocol for a scoping review, the term "methods" refers to diagnostic tools, instruments, methods, raters, or observers used to produce any clinical or physiological parameters, that may be compared in the context of a method comparison study (e.g., two different instruments to measure blood hemoglobin concentration on a given blood sample).

To determine whether a new method to generate clinical data can be used interchangeably with an existing validated method, method comparison studies can be conducted. They are designed to assess the agreement between two methods, usually by comparing a new technique with a "gold standard".[2–4] Typically, method comparison studies will refer to continuous data points such as hemoglobin or temperature[4], whereas diagnostic accuracy study will refer to clinical tests used to determine the dichotomous presence or absence of a given diagnosis[5] such as pulmonary embolism. The analysis of method comparison or diagnostic accuracy studies is challenging, and employs unique statistical concepts depending on the type of underlying data.[6,7] When assessing the agreement of continuous measurements, methods are determined to agree if they fall within a pre-specified range of differences between the two methods, that is considered close enough to avoid negatively affecting clinical decisions.[4,7,8] Given the importance of determining the

accuracy of new instruments, the design and interpretation of method comparison studies is paramount to allow the proper assessment of such accuracy.

In the clinical setting, one may need to collect repeated physiological parameters on individuals, specifically in cases where the value of the clinical data may vary widely within short periods of time or throughout the day (e.g., blood pressure or blood glucose). Repeated measurements of clinical data may very well be different from one another, but they are correlated and should not be considered as independent measurements in the analysis. It is thus important to account for repeated data in method comparison studies to replicate the clinical setting, as well as to robustly assess the repeatability and accuracy of the methods under study.[9,10]

Bland & Altman proposed an approach to analyzing agreement between two diagnostic methods measuring continuous variables in 1983.[9] They argued that statistical methods in use at the time incorrectly focused on correlation, regression analysis, or means comparison, thus not truly examining the question of agreement. They proposed the concept of limits of agreement (LOA), defined as the range within which 95% of differences in measurements between two methods should fall when measured in a given population, further emphasizing the need for repeated measures to assess the repeatability of the methods being compared.[11] This method was further described in later publications to disseminate the LOA method to medical researchers, to discourage the use of incorrect statistical methods, as well as to demonstrate adjustment techniques for repeated measures.[4,12,13]

To our knowledge, Bland & Altman were the first to propose a statistical approach to inform the analysis of repeated continuous measurements in terms of agreement, which has set a standard for such analysis.[4,9] To this date, it appears that it is the method most often used in method comparison studies.[14,15] However, additional statistical methods have been developed for method comparison studies with repeated measures, all of which have their own advantages and limitations.[14] An example of other methods include the concordance correlation coefficient[14], which is derived from correlation coefficients and adapted to assess the variation from the 45°

line through the origin, therefore assessing the agreement between two methods by means of correlation.[16] This method was further adapted to account for repeated measures[17,18], providing an appropriate alternative to the Bland & Altman LOA method. Despite the availability of suitable statistical methods to inform the analysis of this type of data, several inappropriate methods that have been criticized by Bland & Altman remain in use today, such as the use of correlation coefficients, means comparison, and linear regression.[9,15]

In addition to the above, the Bland & Altman LOA method has some underlying assumptions that may make it inappropriate for the analysis of repeated and continuous data in specific situations when the assumptions are violated.[19] Therefore, it is important to have a clear understanding and overview of all available statistical methods for this type of analysis, to help choose the most appropriate method or a combination of methods that will fit a specific dataset created from method comparison studies.

Despite the Bland & Altman LOA method being widely used and accepted by the scientific community, current recommendations and guidelines to inform the analysis of method comparison studies of continuous data that include repeated measures are limited. Parker et al have created a tutorial for practitioners on the use of five different statistical methods to analyze method comparison studies with repeated measurements.[14] Furthermore, Bland & Altman have published many articles to discuss their LOA method, emphasizing the use of repeated measures in method comparison studies, while criticizing the use of correlation coefficient for the analysis of method comparison studies.[4,11,13] To our knowledge, no systematic or scoping review has previously been conducted to present articles that quantitatively and qualitatively compare statistical methods with the LOA method adapted for repeated measurements using real-life clinical data. We believe that such a review will provide a comprehensive summary of currently available statistical methods and a critical comparison to the LOA method. More specifically, this scoping review aims to take a contemporary look at existing methods to inform the analysis of method comparison studies

that include repeated continuous measurements of clinical data and how they compare to the LOA method adapted for repeated data.

**Objectives**

The purpose of this paper is to present a protocol for a scoping review with the following objectives:

1. To identify published methodological papers discussing statistical methods to inform the analysis of method comparison studies that include repeated measures of continuous clinical data that may be compared to the LOA method adapted for repeated measurements.
2. To compare the quantitative findings of each statistical method with the LOA method adapted for repeated measures.
3. To summarize the strengths and limitations of all statistical methods that are appropriate for such analysis.

## 2.4 METHODS

This is a protocol for a scoping review of statistical methods to analyze method comparison studies with repeated continuous data that must be compared to the LOA method. A scoping review framework was chosen given the broadness of the question and that findings may include a variety of study designs. Its aim is to map and identify the available evidence in the field of statistical methods to analyze method comparison studies including repeated continuous data.[20–22] This scoping review was designed based on published best methodological practice.[20,23–25] The *Preferred Reporting Items for Systematic Reviews and Meta-Analysis* (PRISMA) statement with extension for scoping reviews (PRISMA-ScR) will be followed for the reporting of results.[24]

### 1. Identifying the research question

This scoping review aims to answer the following question: how do alternative statistical methods for analyzing agreement in method comparison studies with repeated measures of clinical data perform both quantitatively and qualitatively compared to Bland & Altman's LOA method adjusting for repeated measures?

### 2. Identifying relevant studies

#### Information sources

We will systematically search scientific databases, including Medline, Embase, CINAHL, and Web of Science using a peer-reviewed search strategy. The ProQuest databases will be searched to identify Master's and Doctoral theses that may be relevant for this review. Forward citation searching will be performed on Bland & Altman's original article published in 2007, which focused on their LOA method adapted for repeated measurements.[13]

Furthermore, we will search grey literature using Google, as well as books using Google Books, with small search strings that will cover all topics covered in our search strategy applied to scientific databases. For each search string, we will review the first five pages (50 citations) for

potentially eligible articles. If no potentially eligible article is identified in the first five pages, we will perform the next search string. If at least one article is identified in the first five pages, we will extend our search to the next page until no potentially eligible article is found in a single page. Only books that are accessible through Open Access or the author's institutional library will be considered for inclusion. This selection criterion is necessary given the anticipated limited number of books that will meet inclusion criteria and the associated expenses incurred in procuring books for screening purposes. Grey literature will also be searched by means of publishing a question on an online forum meant for discussions concerning statistical analysis strategies.

Once full-text screening is completed, we will perform backward citation searching of identified eligible articles to search for potentially eligible articles that were referenced by authors of articles included in our review.[26] All databases and information sources will be searched from inception to November 2022, the point at which the search will be conducted.

**Search strategy**

We will develop a search strategy with the help of a research librarian specialized in systematic reviews of methodological studies. Our search strategy will be peer-reviewed following Peer Review of Electronic Search Strategies (PRESS) guidelines, by an independent medical librarian that is not involved in the conduct of this review.[27] The search strategy will initially be developed in MEDLINE and translated to appropriately search all other identified databases.

**3. Study selection**

**Inclusion criteria**

Eligibility criteria are defined following the PCC (Population, Concept, Context) framework[23]:

- Population: To be eligible for this review, articles must have used clinical data collected on humans or simulated clinical data, to illustrate how the statistical methods perform to

analyze method comparison studies that include repeated and continuous measurements. Therefore, human populations of all types may be included in this review.

- Concept: The concept of interest for this scoping review is the comparison of statistical methods to analyze method comparison studies with repeated continuous data by evaluating how alternate statistical methods compare in both a quantitative (e.g., statistical results with 95% confidence intervals) and a qualitative (e.g., authors' textual conclusions or interpretations on whether the methods agree based on the statistical results) manner with the LOA method for repeated measures. Our primary outcomes are the quantitative findings from each study to demonstrate how the statistical methods compare with each other when being used on the chosen dataset.
- Context: The statistical methods must be demonstrated using method comparison studies or diagnostic accuracy studies, with repeated measurements of clinical or simulated data. In the context of agreement or accuracy studies, there must be more than one instrument being compared, including at least one reference method and a gold standard. The LOA method must be one of the statistical methods being compared to help identify novel and modern analytical approaches of such data that may be compared with this method. The data used to perform the quantitative comparison of the chosen statistical methods must be repeated and continuous data collected on humans that may have been collected for the purpose of any type of study, such as observational or interventional, retrospective or prospective in nature. Data may have been collected in the context of previous clinical studies and used as a secondary purpose to illustrate the comparison of the statistical methods for such data. Alternatively, data may have been collected specifically for the purpose of comparing statistical methods informing this type of analysis without having been used in the context of a prior clinical study. Additionally, simulated datasets of clinical data will be considered for inclusion if they have simulated a correlation with a given variance between certain measurements, thereby simulating repeated measurements.

- Specifications on repeated measurements: Any publication mentioning statistical methods in the context of a dataset that contains repeated measurements are eligible, even if the statistical methods used do not account for repeated measurements (e.g., using the mean of the repeated measurements, or using only one value from the repeated measurements). Publications with a dataset that contains repeated measurements for one instrument only, but that compares all measurements for that instrument repeatedly with a gold standard or reference method that was performed only once may also be eligible (e.g., in a clinical parameter that does not vary with time).
- There will be no restriction on language or date. Articles written in other languages than English or French will be translated using Google Translate and assessed for eligibility.

**Exclusion criteria**

We will exclude studies that use the statistical methods to analyze non-clinical data or data not specifically collected in humans. We will exclude studies that perform a theoretical comparison of statistical methods without using the methods on a dataset to demonstrate its applications as we want to see how the proposed statistical methods compare directly with the LOA method.

**Selection of sources of evidence**

Articles screened and identified from selected databases with our search strategy will be imported into an online citation manager, Covidence.[28] Duplicate articles identified by our search strategy will be removed by the online citation manager.

The first step of the selection process will be titles and abstracts screening. All titles and abstracts will be screened by two independent reviewers, except for citation searching which will be conducted by one reviewer only as this has the potential to yield a high number of titles and abstracts for screening. Conflicts will be resolved by further deliberation between the two initial

reviewers followed by consultation with a senior author, if necessary. The titles and abstracts deemed as potentially eligible by the reviewers will be included in full text screening. Full text screening will also be done in duplicate, including full texts identified by citation searching, in a similar fashion as the titles and abstracts screening. Full texts that are identified as eligible to be included in this review will proceed to data charting.

## 4. Charting the data

### Data charting process

A data charting form will be created using Microsoft Excel (Microsoft Corporation, Redmond, WA) specifically for this review and will include all data to be charted from each included study. The form will first be piloted using one eligible study. The data will be extracted from that study by two reviewers independently. The two reviewers will then meet to resolve any discrepancies between the two independent data extraction forms and will adapt the data charting form to avoid discrepancies when used for other included articles. Once the data charting form is piloted, it will be used by two reviewers for independent data charting of all included studies. Discrepancies in data extraction from the reviewers will be resolved through careful deliberation between the two reviewers. During the data charting process, if we believe important data is missing in order to provide a complete summary and analysis of included articles, study authors will be contacted as needed to inquire for further information or data.

### Data items

Descriptive information about eligible articles will be charted, such as the title of the article, name of the authors, year of publication, journal of publication and any funding the authors may have received for their article or study.

We will chart descriptive statistics on the study population used to demonstrate the statistical methods. Data charted will be the study design that was conducted to create the dataset

(e.g., observational, interventional, retrospective, prospective, etc.) and the reference to the original study for which the dataset was created, if different from the included article. Other descriptive statistics charted will include sample size in terms of the number of participants included in the study and the total number of comparisons being done (i.e., how many total measurements were done (i.e., sum of all repeated measurements). Additional descriptive statistics charted will include description of the study population when available (i.e., healthy individuals, population with a specific disease, sex, age, etc.), number of methods or instruments being evaluated for agreement, presence of a "gold standard" comparator, range of repeated measures (how many repeated measures were done for each participant), and amount of missing data for each method being compared. If needed, we will refer to the original studies for more information on the population being used in the included papers, if the dataset used was created for the purpose of a prior clinical study.

Finally, we will chart the analysis results from each eligible article. To do so, we will start by listing the names of the statistical methods being compared for the analysis of method comparison studies that include repeated measures in each article. We will present the results obtained from each statistical method, for each pair of methods or instruments being compared for assessment of agreement if more than one instrument is being analyzed for agreement with a gold standard. Additionally, we will present how repeated measurements were handled for analysis (i.e., adjusting the statistical methods for the correlation between repeated measurements, choosing only one of the measures and ignored repeated measurements, taking the mean of the repeated measurements, etc.). Furthermore, we will compile all advantages and disadvantages of statistical methods being compared, as listed by the authors of the included studies, as well as the authors' conclusions and rationales about their own preference in terms of statistical methods for such analysis if available, for comparison of interpretations between authors.

## 5. Collating, summarizing, and reporting the results

### Results of individual sources of evidence

The primary outcome of our review will be the results of the measures of agreement and their 95% confidence intervals obtained from each statistical method being compared to assess agreement between two or more instruments, methods, observers, or raters, used to examine a clinical parameter in method comparison studies that include repeated and continuous measures. With this outcome, we aim to compare statistical methods in a quantitative manner to determine how each statistical method compares to the LOA method, considered as the gold standard statistical method for the analysis of such data. We will identify whether the authors performed a quantitative comparison of statistical methods in their paper and what method was used for this comparison. Results obtained by each statistical method will be compared quantitatively and qualitatively. For the quantitative comparison, the actual results, in terms of numbers, obtained by each statistical method using the same dataset will be compared. For the qualitative results, the conclusions from the authors on whether the methods being compared agree based on each statistical method will be reported and compared.

Other secondary outcomes are the strengths and limitations of each statistical method being compared, as described by the authors of each eligible article, and if available, the author's preference and rationale in terms of which statistical method to use in method comparison studies that include repeated continuous measurements. This will provide us with a comprehensive overview of the characteristics of all pertinent statistical methods for the analysis of method comparison studies with repeated measures that will help choose the appropriate statistical methods to analyse a specific dataset. We will also provide an overview of the concepts and important assumptions concerning the distribution of the data related to specific statistical tools to be used appropriately, when mentioned by the authors in eligible articles. This will provide us with a comprehensive list of assumptions that a dataset must meet when choosing a statistical analysis tool to analyse a method comparison study with repeated measures. The last two

secondary outcomes mentioned will help us provide recommendations on the circumstances in which statistical tools should be used when analyzing such studies. We will also report which method the authors believe was the most appropriate to determine whether the methods agreed or not within their chosen population.

**Synthesis of results**

To synthesize the quantitative results of included articles, we will present and plot the results obtained from each statistical method together from a given included article, when possible, but separately for each eligible article. If possible, methods reporting results as agreement limits with 95% confidence intervals (e.g. LOA) will be synthesized visually in forest plots and methods reporting correlation coefficients with 95% confidence intervals will be pooled together in a different forest plot (e.g. concordance correlation coefficient). As we are searching for statistical methods that will likely be used on different clinical datasets, it would be inappropriate to pool results.

Moreover, descriptive statistics will be computed to report the overall percentage of papers that present each of the individual statistical methods (i.e., percentage of papers having presented the concordance correlation coefficient) and to compare the concordance rates between the various statistical methods featured in the included papers. To compare the concordance rates, a dichotomous variable will be generated to determine the agreement or disagreement of the conclusions derived from each statistical method with the conclusion of agreement between instruments obtained from the LOA analysis. Within each included paper, this dichotomous variable will be utilized to compute the rates at which alternative statistical methods align with the conclusions of agreement from the LOA analysis. Additionally, this variable will be pooled across all included studies for each individual statistical method to quantify the overall rate at which each statistical method concurs with the LOA's conclusion of agreement.

Qualitative results will be presented in the form of tables for an overview of important findings, such as the statistical methods featured in each included paper, the strengths and limitations of each statistical method as reported by the authors, and the preference reported by the authors. We will determine whether a quantitative comparison was performed, and the statistical method used for comparison of the statistical methods. Furthermore, a critical appraisal to determine whether the method used to compare the statistical methods was appropriate will be performed. Finally, we will perform a qualitative overview and critical appraisal of all statistical methods included in the eligible articles by summarizing their concepts, their limitations, their assumptions, and how they compare to the LOA method. This will help inform the choice of statistical methods for future method comparison studies, and whether the quantitative comparison was appropriate.

All quantitative and qualitative comparisons will be performed for each individual study separately, and a summary of results across studies will also be provided.

**Critical appraisal of individual sources of evidence**

Given the fact that this review does not aim to inform clinical decisions and will include a wide variety of potential study designs, risk of bias assessment and critical appraisal of individual studies are not planned.[20,21]

## 2.5 CONCLUSION

To our knowledge, techniques to measure the agreement for repeated continuous measures have not been systematically synthesized. The statistical analysis of method comparison studies that include repeated measurements can be complex, and guidance is limited in selecting an optimal approach. The results from this review have the potential to address a common question faced by clinicians and researchers when evaluating new diagnostic and related instruments to support clinical decision making by helping to provide guidance in choosing the most appropriate statistical method to inform their analysis.

## ACKNOWLEDGEMENT

The authors would like to thank S. Visintini for assistance in developing the search strategies for this scoping review.

## 2.6 REFERENCES


1. Mabey D, Peeling R. The Optimal Features of a Rapid Point-of-Care Diagnostic Test. In: *Revolutionizing Tropical Medicine: Point-of-Care Tests, New Imaging Technologies and Digital Health*. 1st ed. John Wiley & Sons, Ltd; 2019:81-87. doi:10.1002/9781119282686.ch3

2. Celentano D, Szklo M. More on Causal Inference: Bias, Confounding, and Interaction. In: *Gordis Epidemiology*. 6th ed. Elsevier; 2019:289-306.

3. Weiss NS. Clinical Epidemiology. In: *Modern Epidemiology*. 3rd ed. Lippincott Williams & Wilkins; 2008:641-651.

4. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.

5. Zhou XH, Obuchowski NA, McClish DK. Introduction. In: *Statistical Methods in Diagnostic Medicine*. Second edition. John Wiley & Sons, Inc.; 2011:1-11.

6. Vetter TR, Schober P, Mascha EJ. Diagnostic Testing and Decision-Making: Beauty Is Not Just in the Eye of the Beholder. *Anesth Analg*. 2018;127(4):1085-1091. doi:10.1213/ANE.0000000000003698

7. Barnhart HX, Haber MJ, Lin LI. An Overview on Assessing Agreement with Continuous Measurements. *J Biopharm Stat*. 2007;17(4):529-569. doi:10.1080/10543400701376480

8. Anvari A, Halpern EF, Samir AE. Essentials of Statistical Methods for Assessing Reliability and Agreement in Quantitative Imaging. *Acad Radiol*. 2018;25(3):391-396. doi:10.1016/j.acra.2017.09.010

9. Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. *J R Stat Soc, B: Stat Methodol*. 1983;32(3):307-317. doi:10.2307/2987937

10. Odor PM, Bampoe S, Cecconi M. Cardiac Output Monitoring: Validation Studies–how Results Should be Presented. *Curr Anesthesiol Rep*. 2017;7(4):410-415. doi:10.1007/s40140-017-0239-0

11. Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*. 1986;327(8476):307-310. doi:10.1016/S0140-6736(86)90837-8

12. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. *Int J Epidemiol*. 1995;24(3(Suppl. 1)):S7-14. doi:10.1093/ije/24.Supplement_1.S7.

13. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat*. 2007;17(4):571-582. doi:10.1080/10543400701329422

14. Parker RA, Scott C, Inácio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. *BMC Med Res Methodol*. 2020;20:154. doi:10.1186/s12874-020-01022-x

15. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLoS ONE*. 2012;7(5):e37908. doi:10.1371/journal.pone.0037908

16. Lin LIK. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*. 1989;45:255-268. doi:10.2307/2532051

17. King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Statist Med*. 2007;26(16):3095-3113. doi:10.1002/sim.2778

18. Hiriote S, Chinchilli VM. Matrix-based Concordance Correlation Coefficient for Repeated Measures. *Biometrics*. 2011;67:1007-1016. doi:10.1111/j.1541-0420.2010.01549.x

19. Taffé P. When can the Bland & Altman limits of agreement method be used and when it should not be used. *J Clin Epidemiol*. 2021;137:176-181. doi:10.1016/j.jclinepi.2021.04.004

20. Arksey H, O'Malley L. Scoping Studies: Towards a Methodological Framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. doi:10.1080/1364557032000119616

21. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18:143. doi:10.1186/s12874-018-0611-x

22. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Baldini Soares C. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc*. 2015;13(3):141-146. doi:10.1097/XEB.0000000000000050

23. Aromataris E, Munn Z (Editors). *JBI Manual for Evidence Synthesis*. JBI; 2020. Accessed October 20, 2022. https://synthesismanual.jbi.global. https://doi.org/10.46658/JBIMES-20-01

24. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169(7):467-473. doi:10.7326/M18-0850

25. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci*. 2010;5:69. doi:10.1186/1748-5908-5-69

26. Briscoe S, Bethel A, Rogers M. Conduct and reporting of citation searching in Cochrane systematic reviews: A cross-sectional study. *Res Syn Meth*. 2020;11(2):169-180. doi:10.1002/jrsm.1355

27. *PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E)*. CADTH; 2016.

28. Veritas Health Innovation. Covidence systematic review software. www.covidence.org

29. Box 6.4.d: Cochrane Highly Sensitive Search Strategy for identifying randomized trials in MEDLINE: sensitivity- and precision-maximizing version (2008 revision); Ovid format. Accessed November 7, 2022. https://handbook-5-1.cochrane.org/chapter_6/box_6.4.d_cochrane_hsss_2008_sensprec_ovid.htm

## 2.7 APPENDIX 1: SEARCH STRATEGY FOR THE MEDLINE DATABASE

**Database:** Ovid MEDLINE(R) ALL <1946 to November 6, 2022>

**Search Strategy:**

1 (method adj3 comparison).ti,ab,kf. (9306)

2 (agreement adj3 (method* or measur* or limit* or assess*)).ti,ab,kf. (34748)

3 (diagnos* adj3 accura*).ti,ab,kf. (115373)

4 1 or 2 or 3 (158138)

5 Research Design/ (122334)

6 method?.ti,kf. (546249)

7 models, statistical/ or linear models/ (183010)

8 (statistic* adj3 (test* or analy* or method*)).ti,ab,kf. (313922)

9 5 or 6 or 7 or 8 (1113300)

10 sn.fs. (1038382)

11 9 or 10 (2077626)

12 exp "reproducibility of results"/ (458571)

13 ((repeat* or replicat* or correlat* or cluster* or multiple) adj3 measur*).ti,ab,kf. (128392)

14 ((repeat* or replicat* or correlat* or cluster* or multiple) adj1 data*).ti,ab,kf. (16035)

15 ((inter?rater or intra?rater or rater) adj3 reliab*).ti,ab,kf. (18341)

16 12 or 13 or 14 or 15 (594281)

17 (bland adj3 altman).ti,ab,kf. 19745

18 (limit* adj2 agreement).ti,ab,kf. (11534)

19 17 or 18 (25500)

20 4 and 11 and 16 and 19 (1500)

21 exp animals/ not humans.sh. [29] (5073636)

22 **20 not 21 (1440)**

# CHAPTER 3. A SCOPING REVIEW OF STATISTICAL METHODS FOR THE ANALYSIS OF METHOD COMPARISON STUDIES WITH REPEATED MEASUREMENTS OF CLINICAL DATA

---

Karine Brousseau[1,2,3], Tori Lenet[3], Daniel I. McIsaac[1,2,4], Tim Ramsay[1,2], Dean A. Fergusson[1,2], Guillaume Martel[1,2,3]

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[3]Liver and Pancreas Unit, Department of Surgery, The Ottawa Hospital, Ottawa, Ontario, Canada

[4]Department of Anesthesiology and Pain Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada

## AUTHOR CONTRIBUTIONS

Drs. Martel, Fergusson, McIsaac, and Ramsay provided feedback and recommendations during the conduct of this study and during the writing phase of this manuscript. All aspects of the search strategy and record identification was designed with the help of a research librarian, S. Visintini. Dr. Lenet assisted with all screening aspects and data charting associated with this review. Drs. Martel and Fergusson helped with resolving conflicts during the screening processes.

## PUBLICATION INFORMATION

An abridged version of this manuscript, following specific journal formatting guidelines, will be submitted for publication in a peer-reviewed scientific journal subsequently.

## 3.1 PREFACE TO CHAPTER 3

After having conceived a scoping review, we conducted this study following the protocol presented in the previous chapter. The manuscript presented here outlines all study procedures performed in the context of this review, along with the findings emerging from this scoping review. With this scoping review, we aimed to identify adequate statistical methods to inform the analysis of method comparison studies that include repeated measurements, with the goal of informing the analysis of the PREMISE study, presented in Chapters 4 to 6. The search strategy for the Medline database will be presented as an appendix to this manuscript when we will be submitting it for publication in a scientific journal. However, in the context of this thesis, it is presented in Chapter 2, with the scoping review protocol.

## 3.2 ABSTRACT

**Background:** Method comparison studies are conducted to examine the level of agreement between two instruments measuring physiological continuous parameters. The inclusion of repeated measurements in such studies poses special challenges. The Bland & Altman limits of agreement (LOA) approach has been adapted to account for the correlation between repeated measurements and is widely used in method comparison studies. Alternate statistical methods are not always appropriate for the analysis of such data and there is a paucity of evidence and guidelines pertaining to statistical methods that inform the analysis of method comparison studies that include repeated measurements.

**Objective:** This scoping review aimed to identify methodological publications that propose statistical methods to inform the analysis of method comparison studies that include repeated measurements of continuous clinical data and that may be compared with the LOA method.

**Methods:** Six online databases were searched from inception to November 2022 using a peer-reviewed search strategy. Searching of grey literature and books, as well as backward citation searching were performed to identify additional sources of evidence. Screening and data abstraction were done by two independent reviewers. Results were synthesized narratively.

**Results:** Twenty-nine publications were included in this review. Thirty-two independent statistical methods were identified from the included publications, including variants of the LOA method. Four included publications compared findings from different versions of the LOA method. Four different approaches to handling repeated measurements in the context of method comparison studies were identified and were used to group findings from the included publications. Reported strengths and limitations of the LOA method were summarized.

**Conclusion:** This scoping review provides a synthesis of existing statistical approaches to inform the analysis of method comparison studies with repeated measurements of clinical data, as well as how the various statistical methods perform when compared with various version of the LOA method. Based on the findings of this review, it is generally most advisable to consider using

adjusted LOAs or modified mixed-effect LOAs in analyzing method comparison studies with repeated measurements.

## 3.3 INTRODUCTION

Clinicians are commonly provided with a multitude of instruments to estimate physiological parameters needed to achieve a diagnosis, inform decision-making, as well as determine appropriate care and treatment plans for patients. Usually, one instrument is considered as either the true value of a physiological parameter, or the best available estimate of an unknown true value, also known as the gold standard method[1]. When new instruments are proposed to measure continuous clinical data, method comparison studies are needed to examine agreement between the new method and the gold standard.[1] In cases where a new method is demonstrated to achieve agreement with an old method, the two methods can be used interchangeably.[1] This is particularly important in situations where the new method can lead to significant savings in terms of costs, time, efficiency, or risk for the patient.

Bland & Altman were pioneers in the analysis of method comparison studies when they proposed a novel statistical method for their analyses: the limits of agreement (LOA) method.[1–4] This was among the first methods to provide an estimate of the range within which the difference in measurements from two different instruments is expected to be, therefore quantifying the extent of disagreement between methods.[1] The LOA method has since become widely accepted for the analysis of method comparison studies, and is the most commonly used method for the analysis of such studies.[5]

Ever since the introduction of the LOA method, Bland & Altman have emphasized the importance of capturing several repeated measurements within individual participants of method comparison studies to examine the repeatability of the method under study.[1–3] The inclusion of repeated measurements represents a special case of method comparison studies, whereby individual participants may have had multiple measurements (i.e., more than one measurement by each method) of the same physiological parameters over a specified period.[1] In the presence of repeated measurements, the assumption that all measurements are independent – an assumption that is required for the generation of LOAs – no longer holds due to an interaction

between the subject and the methods. This interaction needs to be accounted for when examining agreement.[1,2] While the LOA method has been well adapted for repeated measurements, other proposed statistical methods may or may not be appropriate for the analysis of method comparison studies that include repeated measurements. Several alternative methods have indeed been deemed inappropriate by Bland & Altman, such as correlation coefficients, means comparison and linear regression, and continue to be used for the analysis of such studies.[1,5]

Although the Bland & Altman approach to the analysis of method comparison studies with repeated measurements is well known and established in the literature, the identification and understanding of alternate statistical methods that are appropriate for this study design is challenging. To our knowledge, there are no guidelines or contemporary texts to inform this type of analysis. Therefore, this scoping review was conducted to synthesize the existing evidence base to inform the analysis of method comparison studies with multiple measurements per individual. A scoping review framework was chosen given the broad nature of the question, an expectation that varied sources of evidence would be identified, as well as an objective to map and identify available evidence in this field of interest.[6–8]

More specifically, this review sought to identify methodological papers discussing findings from various statistical methods to inform the analysis of method comparison studies with repeated measurements of continuous clinical data, that may be compared quantitively and qualitatively with the LOA method adjusted for repeated measurements.

## 3.4 METHODS

This scoping review is reported following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) extension for scoping reviews (PRISMA-ScR).[9] A checklist can be found in Appendix 1.

### Protocol and registration

Prior to initiating the review process, a protocol was created following the PRISMA extension for protocols (PRISMA-P)[10], as well as previously published best methodological practices for scoping reviews.[6,8,9,11,12] The protocol can be accessed on Open Science Framework (https://osf.io/4p8ut).

### Eligibility criteria

Eligibility criteria were determined following the Population, Concept, Context (PCC) framework.[11] This review aimed to identify methodological studies and reports proposing statistical methods to analyse method comparison studies with continuous clinical data that were repeated within participants. Results from an analysis using the LOA method proposed by Bland & Altman[1–3] must have been available for comparison with the results from alternate statistical methods reported within an eligible publication. Therefore, the authors may have presented LOA results, along with the results obtained from the proposed alternate statistical methods, or they may have used a dataset shared by Bland & Altman, whereby results from the LOA analysis were extracted directly from the articles published by Bland & Altman.

The dataset used to demonstrate the statistical methods must have been obtained or generated in a human population, in the context of a method comparison study with continuous and repeated measurements. Alternatively, data may have been generated by simulation of plausible clinical data (e.g. randomly generated hemoglobin values within a plausible range in humans). The dataset may have been obtained in the context of a past clinical study (secondary

use), obtained solely for the purpose of demonstrating the statistical methods (primary use), or from simulated human clinical data. Methods being examined in the context of a method comparison study may have been any type of diagnostic tool, device, rater, instrument, or observer being compared for validation of any given clinical or physiological parameter measured in a continuous fashion. The term “method” will be used exclusively in this paper to encompass all those concepts for simplicity. There were no limitations on the study design for the creation of the dataset underlying a given included study.

Any publication reporting findings from statistical methods that used a dataset containing repeated measurements within individual participant was included in this review if it met all other inclusion criteria, regardless of how repeated data was used for statistical analysis. For instance, if only one measurement was used and repeated measurements were otherwise ignored for analysis, the article was still eligible for review given that this was considered a methodological choice that was made by the authors.

Exclusion criteria consisted of datasets obtained in animals or non-clinical data, or datasets that lacked repeated measurements, as well as studies that did not report the quantitative results of statistical methods or of LOA analysis (i.e., theoretical comparison of statistical methods). There were no restrictions on language and date. Screened full-text citations in languages other than English or French were translated using Google Translate.

**Search strategy**

A comprehensive search strategy was developed with a research librarian and peer reviewed following PRESS guidelines.[13] The search strategy was initially developed on Ovid Medline and was then adapted to other databases. the Ovid Medline search strategy can be accessed in Appendix 2.

**Information sources**

The final search strategy was conducted in Ovid Medline, Ovid EMBASE, CINAHL (EBSCO), and Web of Science from inception to November 7th, 2022. The ProQuest databases were also searched using an adapted search strategy to identify Master's and Doctoral theses, from inception to November 7th, 2022. Forward citation searching was performed using Scopus to identify potentially eligible articles that have cited the Bland & Altman 2007 article focusing on LOA methods adapted for repeated measures, from the date of publication to November 25th, 2022.[2]

The grey literature and books were systematically searched using Google and Google Books, respectively. This was done using multiple small strings of key words, adapted from our main search strategy. For each string, the first five search pages, each with 10 article or webpage titles were screened for potential inclusion. If a title appeared to meet inclusion criteria, the full text was reviewed. If no eligible citation was identified within the first five pages, the next string of key words was used for screening. If at least one of the articles in the first five pages was identified as eligible, the sixth page was then screened and so forth, until no potentially eligible articles were found within a page. The grey literature was also further examined by searching and posting on Stack Exchange - Cross Validated (https://stats.stackexchange.com, published November 29, 2022), an open online forum for discussion concerning statistics.

Once all the above screening strategies were conducted, backward citation searching was conducted to identify other potentially eligible articles within the references of included sources of evidence.

**Selection of sources of evidence**

Identified eligible articles were imported in Covidence (Covidence, Melbourne, Australia), an online citation manager. Duplicate articles were removed.

Studies identified in database searches (Figure 1) were screened in duplicate by two independent reviewers (KB, TL) using titles and abstracts. Relevant citations were then sent to full-text review, which was performed by two independent reviewers (KB, TL). A similar screening process was employed for citations and reports identified from other methods, but these were screened at each step by one reviewer (KB or TL). In all cases conflicts were resolved by consensus, overseen by the senior authors (GM, DF).

**Data collection process**

A data collection form was created for this review using Microsoft Excel (Microsoft Corporation, Redmond, WA) to include all data items specified in the protocol. It was decided a priori to have data collection done independently by two reviewers, as defined in the protocol. However, after having piloted the data collection form with the first included study, a *post-hoc* decision was made to conduct the entire data collection collaboratively between two reviewers (KB, TL), given the complexity of the data items to be charted. It was reasoned that this modification would increase efficiency and accuracy, by discussing the various items and any arising conflicts while data charting.

**Data items**

The following characteristics were collected from each included publication: author names, country of origin of the corresponding author, journal, funding, and type of publication (scientific article, book chapters, etc.).

Information about the dataset used to conduct analyses and generate quantitative findings was also charted. This included information on the context in which the dataset was created (primary use, secondary use, or simulation), and the population under study (eligibility criteria for inclusion in the dataset, mean age, and sex distribution). Further information was charted to examine the distribution of data. Specifically, the sample size, the range of repeated

measurements per participant, and the total sample of pairwise comparisons (e.g., 20 participants with two measurements each for all methods being compared, yielding a total sample size of 40 pairwise method comparisons). Moreover, information about the repeated data was also charted, such as whether the repeated measurements were balanced (equal number of repeated measurements for each individual participant) or unbalanced, and whether the dataset included participants with a single comparison and participants with repeated comparisons. The final items pertaining to the dataset were the primary outcome measured by the methods being compared, the type of methods (e.g., devices, observers, questionnaires, etc.), and the name of the methods being compared, identifying both the gold standard and the index tests being examined for accuracy. Information on the degree of missing of data was also charted. Data were charted directly from included publications, and no assumptions were made to complete the above data items.

Furthermore, data concerning the findings of each statistical method being compared were charted. This included their names/descriptions and the quantitative findings from each method. When reported, 95% confidence intervals were charted to accompany point estimates generated from the statistical methods. The conclusions regarding the agreement of physiological tests under study, based on each statistical method, and the criteria used to determine agreement were extracted. Because the included papers were methodological and did not aim to inform clinical practice, but rather aimed to demonstrate the use of statistical methods, conclusions about agreement or agreement criteria were sometimes not reported. Therefore, when either the conclusion about agreement between methods or the agreement criteria used were not clearly reported in the publication, conclusions were extrapolated from the available data. For example, if agreement criteria were reported, it was possible to assume that the methods agreed based on those criteria, whereas if the conclusions about agreement between methods were reported, it was sometimes possible to make an assumption about the criteria used to determine agreement. To the extent possible, all reported data was used to make assumptions on the conclusions from

each statistical method regarding agreement given that it was an important outcome of this review to be able to compare the statistical methods.

Finally, information concerning the adjustment for correlation between repeated measurements was extracted. This included the type of data used to perform the analysis (i.e., only one measurement used, means of repeated measurements used, or all measurements used), as well as whether the results were adjusted for possible correlation between repeated measurements, and the method used for adjustment (e.g., adjusting the variance using an intra-rater variance). These data items were not always reported in publications included in this review, and thus had to be extrapolated in some cases using the available information (e.g., equations provided, figures and tables related to the dataset used, etc.). Strengths and limitations associated with each statistical method, as reported by the authors, were charted.

**Synthesis of results**

The characteristics of included articles and their underlying datasets were summarized qualitatively. Quantitative comparison between the findings of various statistical methods and their associated point estimates and 95% confidence intervals were also summarized. A list of all reported statistical methods was generated, with their associated incidence. Furthermore, the frequency with which the conclusions about inter-rater agreement from each statistical analysis matched with those from the LOA analysis was computed. Quantitative meta-analysis was not possible, owing to the variety of outcomes measured in the studies that generated the publication datasets, as well as the variety of statistical methods. To report the findings of each statistical method and examine their performance in comparison with the LOA method, included publications were grouped based on the way repeated measures were handled and how results were adjusted to account for correlation between repeated measurements. The methods for handling repeated measurements used by each publication were synthesized narratively. Finally, an overview of the strengths and limitations of the LOA method was produced.

**Critical appraisal of individual sources of evidence**

As planned, risk of bias assessment and critical appraisal of individual sources of evidence was not done, as this review does not aim to inform clinical decision-making and the included sources of evidence vary greatly in terms of study design and reporting format.[6,7]

## 3.5 RESULTS

### Selection of sources of evidence

The search strategy identified 5591 citations from major databases, in addition to 2882 from grey literature. After de-duplication and screening, the final review included 29 full-text articles meeting all inclusion criteria (Figure 1).

### Characteristics of sources of evidence

Publication characteristics and the specifications of the underlying datasets used to demonstrate findings of the proposed statistical methods are presented in Table 1.

Two types of datasets underpinned the included studies. These were clinical datasets (generated in the context of a prior clinical study) (n=19), and generated dataset (produced from a sample of individuals for the explicit purpose of testing the statistical analytic methods) (n=8). Notably, Bland & Altman generated a dataset of peak expiratory flow rate from colleagues and family. The authors provided their complete dataset in their publications, which was subsequently used by others to demonstrate novel approaches to the statistical analysis of such data.[14–23] Finally, other datasets (n=3) consisted of simulated human data, with repeated measurements added by means of a correlation factor between measurements.[24–26]

The number of participants in the included studies ranged from 6 to 1035. Given the multiplicity of measurements per participants, the number of pairwise comparisons in each study range from 10 to 1035. Observations in the context of a method comparison studies are usually the pairwise comparisons between two methods being compared, whereby if 10 participants have two measurements each, from two different methods, the total sample size consists of 10 individual participants, and 20 observations (pairwise comparison between two methods) included in the statistical analysis. The total number of comparisons per participant ranged from 1 to 60000.

The type of methods that were compared in the context of the underlying method comparison studies were most often devices (n=24). Other authors compared questionnaires

(n=4), most commonly in the context of nutritional intake examination.[24,25,27,28] Others compared observers measuring the same clinical or physiological parameters to define the agreement between health professionals (e.g., measurements on radiographs), in which case there were usually no gold standards (n=5).[29–32]

**Results of individual sources of evidence**

Comparative quantitative findings from all statistical methods presented in each included publication are summarized in Table 2. The agreement between methods being compared is presented, as well as the criteria used to adjudicate agreement. In some publications, a large array of findings from statistical analyses was presented. In those cases, a subset of findings is reported in this review to show the agreement conclusions that were obtained in comparison with LOA analysis. Among all included studies, 10 produced divergent conclusions about agreement based on the statistical methods that were used (Table 3). It is noteworthy that in the presence of divergent conclusions about agreement between LOA analysis and alternate statistical methods, the LOA analysis concluded that the underlying methods disagreed in all but one study[33], whereas alternate statistical methods yielded agreement.[3,14,17,19,27,28,30,34,35]

**Synthesis of results**

***Statistical methods presented as alternatives to the LOA method***

Three versions of the LOA method are described in the context of this scoping review. 1) The standard (Bland & Altman) LOA method refers to using the variance of the differences between methods, which can be used in the absence of repeated measurements or when only one measurement pair is used.[3] It may also be used in cases when the mean of the repeated measurements is used for the LOA analysis, thus yielding only one pairwise comparison.[1,2] 2) The adjusted LOA method refers to the Bland & Altman adjustment of the variance of the differences between methods to account for the correlated nature of the repeated measurements.[1,2] Papers

included in this review have also referred to this approach as using fixed effect principles.[35–37] 3) Modified LOA methods refer to alternate versions of the LOA method using random effects or mixed effects models to account for repeated measurements.[35–37]

Most publications included in this review have used only one version of the above LOA methods, whereas four publications compared findings obtained from multiple versions of the LOA method (Table 3).[1,35–37] Eighteen publications reported findings using the standard LOA method.[1,3,14,17–19,24,27,29–35,37–39] Three among those have also presented adjusted LOA findings for comparison[1,35,37], and three have not reported any LOA findings but instead used a dataset provided by Bland & Altman where the results of the LOA findings could be extracted.[14,18,19] Of note, one publication used the three available datasets provided by Bland & Altman to demonstrate their statistical method using three different examples.[19] For this publication, results from the LOA analysis were extracted from the publications by Bland & Altman. In this case, the results obtained from each version of the LOA method were not compared with each other as they were all used on different datasets, with one dataset using the standard version of the LOA, and the two other datasets using the adjusted version of the LOA method.

Twelve publications reported findings using the adjusted LOA method.[1,15,16,19–23,25,35–37] Of note, six of the 12 studies did not report LOA findings within their publication, but these were extracted from the Bland & Altman publications as the authors had used datasets provided by Bland & Altman.[15,19-23] Three publications presented new variations of the LOA method (modified LOA), using random effects models to adjust for the correlated nature of the repeated measurements.[35–37] Three publications did not report on the method used to generate LOA findings, nor could it be deduced from the full-text.[26,28,40]

***Adjustments for correlation with repeated measurements***

Several techniques for handling repeated measurements in method comparison studies have been proposed in included publications (Table 4). Overall, 17 publications have presented

statistical methods to adjust for the correlation between repeated measurements, while five have used the means of repeated measurements in at least one statistical method under study. Three publications used all measurements within their datasets but considered them all as independent and did not adjust for correlation. Twelve publications used only one measurement and excluded the repeated measurements from analysis. Two of those did not report an LOA analysis, but used a dataset provided by Bland & Altman. Therefore, results were not extracted from the publications, but rather from the publication by Bland & Altman.[14,18] One publication did not report how the repeated measurements were used for any of the statistical analyses.[28]

The included publications were separated into five different groups (Table 4). These were based on the methods used to adjust for repeated measurements, given that the findings of statistical analyses may vary based on this adjustment due to the within-subject variability of repeated measurements.[1,2,35] This effect was demonstrated in the first group of publications, that that presented their findings based on different methods to compute LOAs.[1,35–37] Three of the four publications included results from a standard LOA analysis that considered all data as independent measures, and all three reported wider limits when using the adjusted LOA methods.[1,35,37] The authors of three of the four publications from this group proposed novel or modified approaches to the LOA method, using a random effects model, reporting narrower limits of agreement with those, when compared with "fixed effect" adjusted LOAs.[35–37] Reported advantages of using a random effects model to generate LOA estimates included allowing for adjustment based on other exploratory covariates[37], better generalizability to a wider population from which the sample population was identified[36], and an easier way to identify outliers.[36]

The second group of publications consisted of those that adjusted for the correlated nature of repeated measurements for all statistical methods presented, either by adjusting the variance used to generate estimates, or by using random or mixed effects models.[15,16,19–23,25] All eight publications proposed vastly different statistical methods for the analysis of method comparison studies with repeated measurements. Findings that resulted from each of the statistical methods

are reported in Table 2. For six of the eight publications, LOA findings were extracted from the Bland & Altman publications, given that the authors of those publications used datasets provided by Bland & Altman and did not report LOA findings.[15,19–23]

The third group of publications included those that presented an adjusted statistical approach for the analysis of method comparison studies with repeated measurements, but where LOA findings did not appear to be adjusted (standard LOA).[14,18,26,33,38,40] Of note, two did not directly report LOA findings and were extracted elsewhere[14,18], while two did not report whether LOA findings were adjusted and this could not be deduced from the full-text.[26,40]

The fourth group of publications included those that generated means of repeated measurements, providing only one average measurement per participant, per method being compared.[24,27,29,31] Two of those publications utilized this approach in the context of daily nutrient intake measurements.[24,27]

The final group of publications consisted of those that used a dataset containing repeated measurements, but did not adjust the statistical analyses for repeated measurements.[3,17,30,32,34,39] In many of those papers, repeated measurements were used to examine intra-rater reliability (agreement within measurements) and the repeated measurements were excluded from the analysis of inter-rater reliability, using only one of the measurements per individual.[3,30,32] Others did not intend to propose statistical methods to adjust for repeated measurements, but nevertheless used a dataset with repeated measurements. In those cases, only one measurement was chosen, and the repeated measures were excluded from the analysis[17,39], or the repeated measurements were all included in the analyses independently without considering their correlation.[30,34]

***Reported strengths and limitations of the LOA method***

Reported qualitative findings pertaining to the overall strengths and limitations associated with the LOA method, all versions combined, are presented in Table 5. The most commonly

reported strengths (n=9/29) were its simplicity, as well as ease of use and interpretation.[1,3,14,19,22,24,26,31,35] Others have reported that LOA findings provide useful information on the bias and the magnitude of differences between methods being compared[27,28,32], and that it provides a visual comparison between methods.[32,38] One group also suggested that the results are based on the original unit of measurement, which allows for direct comparison with a clinically-acceptable difference between the two methods being compared.[35]

The most commonly reported limitation (n=7/29) was the parametric nature of this statistical method, as it relies on assumptions about the normality of differences between the methods being compared, as well as homoscedasticity.[1,25,26,34–36,38] Others have suggested that limits of agreement are subjective, are based on clinical judgement, and are not based on statistical factors.[14,27–29] Moreover, others have indicated that LOA results do not provide information about which method is most accurate[14,22,35], which can be a limitation specifically in cases where the gold standard provides worst estimates than the newer method being examined for agreement.[14] Finally, it has been reported that there is uncertainty in the point estimates of the LOAs, which can lead to biased interpretations of agreement if confidence limits are not presented alongside LOA point estimates.[24,35] In fact, Batterham et al. have reported that when looking at the point estimates of their LOA analysis only, they would be interpreted as agreement between the two methods based on the acceptable clinical difference. However, the interpretation of the 95% confidence intervals suggests that the limits could potentially fall outside of the clinically acceptable difference with resampling.[24]

## 3.6 DISCUSSION

This scoping review has identified and presented multiple approaches to inform the analysis of method comparison studies that include repeated measurements of continuous clinical data. Among 29 methodological publications included in this work, five different approaches to handle repeated data were identified, including a variety of statistical methods that were directly compared and reviewed for their strengths and limitations.

Three main variants of the LOA method were identified, including the standard unadjusted Bland & Altman LOA method[1,3], the Bland & Altman LOA method adjusted for repeated measurements (fixed effect)[1,2], and novel approaches to the LOA method using mixed or random effects to allow for adjustment of other covariates and better generalizability to a bigger population of interest.[35–37] Several authors compared the findings from each LOA method on the same dataset and emphasized the importance of adjusting the variance when computing LOAs, considering the within-subject variability of repeated measurements and thus yielding wider LOA estimates.[1,35–37] The importance of the assumptions of normality and homoscedasticity when using LOAs has been singled out as a major limitation of LOAs in several papers. Bland & Altman indeed recommend various data visualization techniques to verify that data distributions respect those conditions.[1,2] Other authors have proposed alternative statistical methods when those conditions are not met, such as Bayesian statistics[20], bias and precision plots[23,26], a probability of agreement approach[21], and log-transformation of data.[31]

A total of 32 individual statistical methods were presented and identified from the included publications, of which six generate agreement intervals that use the same unit of measurement as the outcome being measured. This can facilitate interpretation and comparison to the clinically acceptable difference.[35] However, these methods come with limitations, such as the need for clinically acceptable agreement to be defined a priori, leading to subjectivity in the interpretation of results given the lack of statistical factors to inform the findings in terms of agreement.[34,35] The other statistical methods generate findings that do not have a unit of measurement, such as

coefficients, probabilities of agreement, and percentages of agreement, which may lead to greater difficulty in interpretation.[35] While Bland & Altman have consistently recommended against the use of correlation coefficients and comparison of means[1,3], this review found that these methods are often proposed for the analysis of method comparison studies, as shown in Table 3. Some have promoted the use of correlation coefficients as an appropriate method to examine reliability[29] or to detect bias[30], whereas others do not recommend the use of this approach given the fact that it measures association rather than agreement[3,27,28], and does not provide sufficient information to examine the level of agreement between methods.[27,28,32]

When planning the statistical analysis of a method comparison study with repeated measurements, it is critical to examine the dataset and type of repeated measurements to choose appropriate statistical approaches. This review has grouped included publications into five different categories to help readers identify those that may be of value based on planned approach to handling repeated measurements (Table 4). The four publications included in the first group, in which different versions of the LOA method are compared, provide important data to inform the analysis of a method comparison study with repeated measurements. The publications included in the second and third groups may assist in choosing alternate statistical methods that adjust for the correlation between repeated measurements and the data included in table 2 provides information on the performance of these methods when compared with the LOA approach. However, prudence is recommenced when comparing the results obtained from each statistical method in the third group, as the LOA findings were not adjusted for repeated measurements.

The fourth and fifth groups of publications presented in table 4 are alternative approaches that have been suggested for the analysis of method comparison studies. In the fourth group of publications, means of the repeated measurements were used. These cannot be recommended for general use, except perhaps in cases where this approach represents standard clinical practice, as in the case of daily food intake records.[40,41] However, it is important to note that this approach is at risk of omitting important information about the distribution of the differences

between methods.[22] Finally, the fifth group has ignored repeated measurements by excluding them from analysis. Given the wide range of more robust statistical methods available to adjust for repeated measurements in the context of a method comparison study, the exclusion of repeated measurements from analysis is not recommended. It is noteworthy that the exclusion of repeated data was not specifically recommended by the authors of the included publications, as some were presenting statistical methods for the analysis of method comparison studies with single measurements but used a dataset that contained repeated measurements.

There are no other systematic reviews of methodological publications aiming at identifying and comparing statistical methods to inform the analysis of method comparison studies with repeated measurements. A systematic review conducted by Zaki and colleagues examined 210 method comparison studies and reported that the Bland & Altman LOA method is the statistical method that is most often used in the context of method comparison studies (85%), with the inappropriate use of correlation coefficients remaining in use by many authors (27%).[5] Zaki and colleagues synthesized statistical methods that were used to analyse method comparison studies in the published literature, in contrast to the current review where methodological publications were examined for statistical methods specifically pertaining to repeated measurements.

There are limitations to the current scoping review. The LOA approach proposed by Bland & Altman to inform the analysis of method comparison studies with repeated measurements was used as an anchor for search, whereby only publications referring to “limits of agreement” or “Bland-Altman” within their abstracts were considered for inclusion. This limit was specifically set to allow for the comparison of findings between the LOA method and other statistical methods. It is possible that this approach may have prevented the identification of other statistical methods that would have been appropriate for the analysis of method comparison studies with repeated measurements. This scenario could have arisen where authors would have presented such methods without comparing them with the LOA method, or without using a dataset provided by Bland & Altman. Furthermore, multiple versions of the LOA method exist in the literature, and the

method adapted for repeated measurements was not always used in included publications, thus limiting the quantitative comparisons of findings between different statistical methods. Moreover, the specific LOA method used to report findings was not always reported in the included publications. Finally, since non-clinical methodological publications were specifically sought for, the aim of those papers was usually not to comment on the agreement between the methods in the context of the dataset used to examine the statistical methods. There are thus limitations pertaining to the conclusions about agreement between methods, as reported in table 3.

## 3.7 CONCLUSION

This scoping review provides a synthesis of existing statistical approaches to inform the analysis of method comparison studies with repeated measurements of clinical data, as well as how the various statistical methods perform when compared with various version of the LOA method. Based on the findings of this review, it is generally most advisable to consider using adjusted LOAs or modified mixed-effect LOAs in analyzing method comparison studies with repeated measurements.


## FUNDING

This scoping review was conducted in the context of a Master's thesis. No funds were received or used for this scoping review.


## AUTHOR CONTRIBUTIONS

KB designed this study, with the help of GM, DF, DM, and TR. All aspects of the search strategy and record identification was designed with the help of a research librarian, S. Visintini. KB and TL performed all screening aspects and data charting associated with this review. KB wrote the original manuscript draft. All authors contributed to the review and editing of the final manuscript.


## ACKNOWLEDGEMENTS

The authors would like to thank Sarah Visintini for helping with the search strategy and identification of sources of evidence.

## 3.8 REFERENCES


1. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.

2. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. *J Biopharm Stat*. 2007;17:571-582. doi:10.1080/10543400701329422

3. Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*. 1986;327(8476):307-310. doi:10.1016/S0140-6736(86)90837-8

4. Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. *J R Stat Soc, B: Stat Methodol*. 1983;32(3):307-317. doi:10.2307/2987937

5. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLoS ONE*. 2012;7(5):e37908. doi:10.1371/journal.pone.0037908

6. Arksey H, O'Malley L. Scoping Studies: Towards a Methodological Framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. doi:10.1080/1364557032000119616

7. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18:143. doi:10.1186/s12874-018-0611-x

8. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Baldini Soares C. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc*. 2015;13:141-146. doi:10.1097/XEB.0000000000000050

9. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169:467-473. doi:10.7326/M18-0850

10. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*. 2015;349:g7647. doi:10.1136/bmj.g7647

11. Aromataris E, Munn Z (Editors). *JBI Manual for Evidence Synthesis*. JBI; 2020. Accessed October 20, 2022. https://synthesismanual.jbi.global. doi:10.46658/JBIMES-20-01

12. Tricco AC, Lillie E, Zarin W, et al. A scoping review on the conduct and reporting of scoping reviews. *BMC Med Res Methodol*. 2016;16:15. doi:10.1186/s12874-016-0116-4

13. *PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E)*. CADTH; 2016.

14. Alanen E. Everything all right in method comparison studies?. *Stat Methods Med Res*. 2012;21(4):297-309. doi:10.1177/0962280210379365

15. Choudhary PK. A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *J Stat Plan Inference*. 2008;138:1102-1115. doi:10.1016/j.jspi.2007.03.056

16. Francq BG, Govaerts B. How to regress and predict in a Bland-Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. *Stat Med*. 2016;35:2328-2358. doi:10.1002/sim.6872

17. Kim J, Lee JH. A novel graphical evaluation of agreement. *BMC Med Res Methodol*. 2022;22:51. doi:10.1186/s12874-022-01532-w

18. Quiroz J, Burdick RK. Assessment of Individual Agreements with Repeated Measurements Based on Generalized Confidence Intervals. *J Biopharm Stat*. 2009;19(2):345-359. doi:10.1080/10543400802622576

19. Roy A. An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *J Biopharm Stat*. 2009;19(1):150-173. doi:10.1080/10543400802535141

20. Schluter PJ. A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies. *BMC Med Res Methodol*. 2009;9:6. doi:10.1186/1471-2288-9-6

21. Stevens NT, Steiner SH, MacKay RJ. Comparing heteroscedastic measurement systems with the probability of agreement. *Stat Methods Med Res*. 2018;27(11):3420-3435. doi:10.1177/0962280217702540

22. Stevens NT, Steiner SH, MacKay RJ. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Stat Methods Med Res*. 2017;26(6):2487-2504. doi:10.1177/0962280215601133

23. Taffé P. Assessing bias, precision, and agreement in method comparison studies. *Stat Methods Med Res*. 2020;29(3):778-796. doi:10.1177/0962280219844535

24. Batterham MJ, Van Loo C, Charlton KE, Cliff DP, Okely AD. Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement. *Br J Nutr*. 2016;115:1273-1280. doi:10.1017/S0007114516000040

25. Haghayegh S, Kang HA, Khoshnevis S, Smolensky MH, Diller KR. A comprehensive guideline for Bland-Altman and intra class correlation calculations to properly compare two methods of

measurement and interpret findings. *Physiol Meas*. 2020;41:055012. doi:10.1088/1361-6579/ab86d6

26. Taffé P. When can the Bland & Altman limits of agreement method be used and when it should not be used. *J Clin Epidemiol*. 2021;137:176-181. doi:10.1016/j.jclinepi.2021.04.004

27. Ambrosini GL, de Klerk NH, Musk AW, Mackerras D. Agreement between a brief food frequency questionnaire and diet records using two statistical methods. *Public Health Nutr*. 2001;4(2):255-264.

28. Lombard MJ, Steyn NP, Charlton KE, Senekal M. Application and interpretation of multiple statistical tests to evaluate validity of dietary intake assessment methods. *Nutr J*. 2015;14:40. doi:10.1186/s12937-015-0027-y

29. Anvari A, Halpern EF, Samir AE. Essentials of Statistical Methods for Assessing Reliability and Agreement in Quantitative Imaging. *Acad Radiol*. 2018;25(3):391-396. doi:10.1016/j.acra.2017.09.010

30. Armstrong SM, Miller JM, Benson K, et al. Revisiting reliability of quantified perineal ultrasound: Bland and Altman analysis of a new protocol for the rectangular coordinate method. *Neurourol Urodyn*. 2006;25:731-738.

31. Euser AM, Dekker FW, le Cessie S. A practical approach to Bland-Altman plots and variation coefficients for log transformed variables. *J Clin Epidemiol*. 2008;61:978-982. doi:10.1016/j.jclinepi.2007.11.003

32. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil*. 1998;12:187-199. doi:10.1191/026921598672178340

33. Cecconi M, Rhodes A, Poloniecki J, Della Rocca G, Grounds RM. Bench-to-bedside review: The importance of the precision of the reference technique in method comparison studies – with specific reference to the measurement of cardiac output. *Crit Care*. 2009;13:201. doi:10.1186/cc7129

34. Lorne E, Diouf M, de Wilde RBP, Fischer MO. Assessment of interchangeability rate between 2 methods of measurements: An example with a cardiac output comparison study. *Med*. 2018;97(7):e9905. doi:10.1097/MD.0000000000009905

35. Parker RA, Scott C, Inácio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. *BMC Med Res Methodol*. 2020;20:154. doi:10.1186/s12874-020-01022-x

36. Parker RA, Weir CJ, Rubio N, et al. Application of Mixed Effects Limits of Agreement in the Presence of Multiple Sources of Variability: Exemplar from the Comparison of Several Devices to Measure Respiratory Rate in COPD Patients. *PLoS ONE*. 2016;11(12):e0168321. doi:10.1371/journal.pone.0168321

37. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. *Br J Anaesth*. 2007;99(3):309-311.

38. Lai D, Shiao SYPK. Comparing two clinical measurements: A linear mixed model approach. *J Appl Stat*. 2005;32(8):855-860. doi:10.1080/02664760500080157

39. Saugel B, Grothe O, Nicklas JY. Error Grid Analysis for Arterial Pressure Method Comparison Studies. *Anesth Analg*. 2018;126(4):1177-1185. doi:10.1213/ANE.0000000000002585

40. Choudhary PK, Nagaraja HN. *Measuring Agreement: Models, Methods, and Applications*. John Wiley & Sons; 2017:336. doi:10.1002/9781118553282

41. Farshchi HR, Macdonald I, Madjd A, Taylor MA. Benefits and Limitations of Traditional Self-Report Instruments. In: *Advances in the Assessment of Dietary Intake.* CRC Press; 2017.

42. Anvari A, Dhyani M, Stephen AE, Samir AE. Reliability of Shear-Wave Elastography Estimates of the Young Modulus of Tissue in Follicular Thyroid Neoplasms. *AJR Am J Roentgenol*. 2016;206:609-616. doi:10.2214/AJR.15.14676

43. Altman D, Bland J. The analysis of blood pressure data. In: *Blood Pressure Measurement*. Elsevier; 1991:287-314.

44. Bowling LS, Sageman WS, O'Connor SM, Cole R, Amundson DE. Lack of agreement between measurement of ejection fraction by impedance cardiography versus radionuclide ventriculography. *Crit Care Med*. 1993;21(10):1523-1527.

45. Euser AM, Finken MJJ, Keijzer-Veen MG, Hille ETM, Wit JM, Dekker FW. Associations between prenatal and infancy weight gain and BMI, fat mass, and fat distribution in young adulthood: a prospective cohort study in males and females born very preterm. *Am J Clin Nutr*. 2005;81:480-487. doi:10.1093/ajcn.81.2.480

46. Shiao SYP. Functional versus Fractional Oxygen Saturation Readings: Bias and Agreement Using Simulated Solutions and Adult Blood. *Biol Res Nurs*. 2002;3(4):210-221. doi:10.1177/10900402003004006

47. De Wilde RBP, Schreuder JJ, Van Den Berg PCM, Jansen JRC. An evaluation of cardiac output by five arterial pulse contour techniques during cardiac surgery. *Anaesthesia*. 2007;62:760-768. doi:10.1111/j.1365-2044.2007.05135.x

48. Myles PS, Mcrae R, Ryder I, Hunt JO, Buckland MR. Association between Oxygen Delivery and Consumption in Patients Undergoing Cardiac Surgery. is There Supply Dependence? *Anaesth Intens Care*. 1996;24:651-657. doi:10.1177/0310057X9602400603

49. Rubio N, Parker RA, Drost EM, et al. Home monitoring of breathing rate in people with chronic obstructive pulmonary disease: observational study of feasibility, acceptability, and change after exacerbation. *Int J Chron Obstruct Pulmon Dis*. 2017;12:1221-1231. doi:10.2147/COPD.S120706

50. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG. Open-access MIMIC-II database for intensive care research. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE; 2011:8315-8318. doi:10.1109/IEMBS.2011.6092050

### 3.9 TABLES

**Table 1.** Study and dataset characteristics.

| Study author(s), year, Country, publication type | Type of dataset used[a] | Sample size (participants/ comparisons)[c] | Measurements per rater, per participant / total number of comparisons | Outcome(s) measured (units) | Number and type of raters[d] | Reference standard (RS) | Index test(s) (IT) |
|---|---|---|---|---|---|---|---|
| Alanen[14], 2012, Finland, Scientific article | Secondary use of a generated dataset[3] | 17 / 34 | 2 / 34 | Peak expiratory flow rate (PEFR) (L/min) | 2 devices | Wright peak flow meter | Mini Wright peak flow meter |
| Ambrosini et al.[27], 2001, Australia, Scientific article | Original and generated dataset | 76 / 76 | 2 IT, 4 RS / 152 | Dietary beta-carotene (mg) / Dietary retinol (mcg) | 2 questionnaires | Mean of four 1-week diet records | Two Food Frequency Questionnaires (FFQ) |
| Anvari et al.[29], 2017, USA, Scientific article | Secondary use of a clinical dataset[42] | 35 / 35 | 2 / 70 | Thyroid nodule tissue stiffness (kPa) | 2 observers | Not applicable (NA) | NA |
| Armstrong et al.[30], 2006, USA, Scientific article | Secondary use of a clinical dataset[b] | 6 / 24 | 4 / 24 | Urethrovesical junction position and mobility, using a rectangular | 2 observers (R1, R2) | NA | NA |
| | | 24 / 24 | 4 / 96 | | 2 observers (R1, R3) | NA | NA |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | coordinate method (mm) | | | |
| Batterham et al.[24], 2016, Australia, Scientific article | Original and simulated dataset | 80 / 80 | 1 IT, 3 RS / 240 | Mean Iodine intake (mcg) | 2 questionnaires | Mean of three 24-hour dietary recall | FFQ |
| Bland & Altman, 1999[1], UK, Scientific article | Secondary use of a clinical dataset[43] | 85 / 255 | 3 / 255 | Systolic blood pressure (SBP) (mmHg) | 2 devices | Sphygmomanometer (by observer "J") | Machine "S" |
| Bland & Altman, 1986[3], UK, Scientific article | Original and generated dataset | 17 / 17 | 2 / 34 | PEFR (L/min) | 2 devices | Wright peak flow meter | Mini Wright peak flow meter |
| Cecconi et al.[33], 2009, Italy, Scientific article | Not reported (NR) | 20 / 20 | 1 RS, 4 IT / 80 | Cardiac output (L/min) | 2 devices | Thermodilution | Intermittent thermodilution |
| Choudhary[15], 2008, USA, Scientific article | Secondary use of a clinical dataset[1,44] | 12 / 60 | 3-6 / 60 | Left ventricular ejection fraction (LVEF) (%) | 2 devices | Radionuclide ventriculography | Impedance cardiography |
| Choudhary and Nagaraja[40], 2017, USA, | Secondary use of a clinical dataset[b] | 61 / 177 | 1-3 / 177 | Blood oxygen saturation (%) | 2 devices | Co-oxymetry | Pulse oxymetry |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Book chapter | | | | | | | |
| Euser et al[31], 2008, Netherlands, Scientific article | Secondary use of a clinical dataset[45] | 4 / 4 | 2 / 8 | Biceps skinfold thickness (mm) | 13 observers | NA | NA |
| Francq and Govaerts[16], 2015, Belgium, Scientific article | Secondary use of a clinical dataset[1,43] | 85 / 255 | 3 / 255 | SBP (mmHg) | 2 devices | Sphygmomanometer (by observer "J") | Machine "S" |
| Haghayegh et al.[25], 2020, USA, Scientific article | Original and simulated dataset | 80 / 80 | 3 RS, 1 IT / 240 | Mean intake of Iodine (mcg) | 2 questionnaires | Mean of three 24-hour dietary recall | FFQ |
| Kim and Lee[17], 2022, USA, Scientific article | Secondary use of a generated dataset[3] | 17 / 17 | 2 / 34 | PEFR (L/min) | 2 devices | Wright peak flow meter | Mini Wright peak flow meter |
| Lai and Shiao[38], 2005, USA, Scientific article | Secondary use of a clinical dataset[46] | 2 / 625 | NR / 625 | Foetal haemoglobin percentage levels (%) | 2 devices | Percent fractional oxyhemoglobin ($HbO_2$) | Percent functional oxyhemoglobin ($SO_2$) |
| Lombard et al.[28], 2015, South Africa, | Original and | 47 / 47 | 4 RS, 1 IT / 188 | Energy intake (kJ) | 2 questionnaires | Mean of four 24-hour dietary recall | FFQ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Scientific article | generated dataset | | | | | | |
| Lorne et al.[34], 2018, France, Scientific article | Secondary use of a clinical dataset[47] | 24 /199 | NR / 199 | Cardiac output (L/min) | 6 devices | Pulmonary artery thermodilution | 5 arterial pulse contour techniques (Wesseling method,LiDCO, PiCCO, Hemac method, Modelflow) |
| Myles and Cui[37], 2007, Australia, Editorial | Secondary use of a clinical dataset[48] | 20 / 144 | 7-8 / 144 | Oxygen consumption (mL/min) | 2 devices | Inspired gas analysis ($GVO_2$) | Reverse Fick method ($FVO_2$) |
| Parker et al.[35], 2020, Canada, Scientific article | Secondary use of a clinical dataset[36,49] | 21 / 385 | 15-19 / 385 | Respiratory rate (breaths/ minute) | 2 devices | Standard respiratory rate monitor (Oxycon mobile) | Chest-band |
| Parker et al.[36], 2016, Scotland, Scientific article | Secondary use of a clinical dataset[49] | 21 / 385 | 9-19 / 385 | Respiratory rate (breaths/ minute) | 6 devices | Standard respiratory rate monitor (Oxycon mobile) | Five commercially available respiratory rate monitors: Camera, photoplethysmography , Impedance, Accelerometer, Chest-band |
| Quiroz and Burdick[18], 2009, USA, Scientific article | Secondary use of a generated dataset[3] | 17 / 34 | 2 / 34 | PEFR (L/min) | 2 devices | Wright peak flow meter | Mini Wright peak flow meter |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Rankin and Stokes[32], 1998, UK, Scientific article | Original and generated dataset | 10 / 10 | 4 GR, 1 MS / 40 | Cross-section area of anterior tibial muscle group ($cm^2$) | 2 observers (GR, MS) | NA | NA |
| Roy[19], 2009, USA, Scientific article | Secondary use of a clinical dataset[1,43] | 85 / 255 | 3 / 255 | SBP (mmHg) | 2 devices | sphygmomanometer (by observer "J") | Machine "S" |
| | Secondary use of a generated dataset[3] | 17 / 34 | 2 / 34 | PEFR (L/min) | 2 devices | Wright peak flow meter | Mini Wright peak flow meter |
| | Secondary use of a clinical dataset[1,44] | 12 / 60 | 3-6 / 60 | LVEF (%) | 2 devices | Radionuclide ventriculography | Impedance cardiography |
| Saugel et al.[39], 2018, Germany, Scientific article | Secondary use of a clinical dataset[50] | 1035 / 1035 | 1-60,000 / NR | Arterial systolic blood pressure | 2 devices | arterial catheter | oscillometry |
| | | | | Mean arterial blood pressure | 2 devices | arterial catheter | oscillometry |
| Schluter[20], 2009, Australia, Scientific article | Secondary use of a clinical dataset[1,43] | 85 / 255 | 3 / 255 | SBP (mmHg) | 2 devices | sphygmomanometer (by observer "J") | Machine "S" |
| Stevens et al.[21], 2018, USA, | Secondary use of a clinical dataset[1,43] | 85 / 255 | 3 / 255 | SBP (mmHg) | 2 devices | sphygmomanometer (by observer "J") | Machine "S" |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Scientific article | | | | | | | |
| Stevens et al.[22], 2017, Canada, Scientific article | Secondary use of a clinical dataset[1,43] | 85 / 255 | 3 / 255 | SBP (mmHg) | 2 devices | sphygmomanometer (by observer "J") | Machine "S" |
| Taffé[26], 2021, Switzerland, Scientific article | Original and simulated dataset | 100 / NR | 10-20 / NR | "Simulating data from a uniform distribution having values between 20 and 40." | 2 simulated devices | Simulated device "Y1" | Simulated device "Y2" |
| Taffé[23], 2020, Switzerland, Scientific article | Secondary use of a clinical dataset[1,43] | 85 / 255 | 3 / 255 | SBP (mmHg) | 2 devices | sphygmomanometer (by observer "J") | Machine "S" |

Abbreviations: FFQ, Food Frequency Questionnaires; IT, Index tests; LVEF, Left ventricular ejection fraction; NA, Not applicable; NR, Not reported; PEFR, Peak expiratory flow rate; RS, Reference standard; SBP, Systolic blood pressure.

[a]Datasets may be original, meaning that it was created for the purpose of the presented paper, or the authors could have used an existing dataset for secondary use. The dataset may be clinical, generated, or simulated. A clinical dataset was created for clinical research (i.e., method comparison study in a given population) and is now used in the context of the presented article. A generated dataset was collected within a random set of individuals for the purpose of demonstrating statistical methods as opposed to aiming to examine rater accuracy. A simulated dataset was not captured in humans, but generated via simulation as though they were clinical measurements done on humans.

[b]Reference to original dataset not reported.

[c]Sample size represents the total number of participants and observations used for analysis, which may not be equal to the total number of comparisons, depending on how they handle the repeated measurements.

[d]This includes the raters that were used for analysis. More raters may have been measured in the dataset, but not analyzed within the presented study.

**Table 2.** Quantitative results extracted from each included source of evidence.

| Study author(s) | Statistical methods used | Quantitative results[b] | Conclusion about agreement | Agreement criteria used |
|---|---|---|---|---|
| Alanen | Standard Limits of agreement (LOA)[3] | $LOA_L$: -79.7 (95%CI -114.3, 45.1)<br>$LOA_U$: 75.5 (95%CI 40.9, 110.1)[3] | No agreement | Limits within ±10 L/min[3] |
| | Structural equations (SEM) approach | Fit of the final model: $\chi^2$ = 7.50<br>Degrees of freedom (df) = 11 | Agreement | Acceptance of the null hypothesis |
| Ambrosini et al. | Standard LOA | β-carotene: $LOA_L$: 50%, $LOA_U$: 447% | No agreement | 100% represents ideal agreement |
| | | Retinol: $LOA_L$: 11%, $LOA_U$: 349% | No agreement | |
| | Pearson correlation coefficient (PCC) | β-carotene: 0.36 (95%CI: 0.14 -0.54) | Agreement | 95% CI not including the value "0" is statistically significant |
| | | Retinol: 0.51 (95%CI: 0.32 - 0.66) | Agreement | |
| | Mean agreement | β-carotene: 149% (95%CI: 132-170) | No agreement | 95% CI not including the value "100%" is statistically significant, 100% represents ideal agreement |
| | | Retinol: 63% (95%CI: 52-77) | No agreement | |
| Anvari et al. | Standard LOA | $LOA_L$: -9.9, $LOA_U$: 11 | Agreement | NR |
| | Intraclass correlation coefficient (ICC) | Single measurement:<br>0.872 (95% CI: 0.763–0.933), | Agreement | ICC > 0.70 |
| | | Average measurements:<br>0.932 (95% CI: 0.865-0.965) | Agreement | |
| Armstrong et al.[a] | Standard LOA | R1/R2:<br>Kegel $D_x$: -9, 8, Kegel $D_y$: -6, 14 | No agreement on both axes | Limits within ±2 mm |
| | | R1/R3:<br>Kegel $D_x$: -5, 5, Kegel $D_y$: -4, 5 | No agreement on both axes | |
| | PCC | R1/R2:<br>Kegel $D_x$: 0.87, Kegel $D_y$: 0.70 | Agreement for $D_x$, no agreement for $D_y$ | NR (Assumed PCC>0.7) |

| | | | | |
|---|---|---|---|---|
| | | R1/R3: Kegel $D_x$: 0.96, Kegel $D_y$: 0.93 | Agreement on both axes | |
| | T-test | R1/R2:<br>Kegel $D_x$: *P* value not significant (NS)<br>Kegel $D_y$: *P*<0.001 | Agreement for $D_x$, no agreement for $D_y$ | *P* value < 0.05 |
| | | R1/R3: Kegel $D_x$: *P* value NS<br>Kegel $D_y$: *P* value NS | Agreement on both axes | |
| Batterham et al. | Standard LOA | $LOA_L$: -88.38 (95%CI -122.24, -54.52)<br>$LOA_U$: 83.82 (95%CI 49.96, 117.96) | Agreement based on point estimates only. No agreement based on 95%CI | NR (assumed limits within 90 to 100 mcg) |
| | Two one-sided t-test (TOST) | Equivalence margin (SEM) of 5 mcg:<br>*t* upper: 0.55 (p=0.291)<br>*t* lower: -1.48 (p=0.071) | No agreement | Both TOST *P*-values < 0.05 with 79 degrees of freedom, with a 90% CI of the mean difference within -10.49 to 5.89 used for equivalence testing |
| | | SEM of 10 mcg:<br>*t* upper: 1.57 (p=0.060)<br>*t* lower: -2.50 (p=0.007) | No agreement | |
| | | SEM of 15 mcg:<br>*t* upper: 2.69 (p=0.006)<br>*t* lower: -3.52 (p=0.000) | Agreement | |
| | | SEM of 10%:<br>*t* upper: 2.06 (p=0.021)<br>*t* lower: -2.99 (p=0.002) | Agreement | |
| | Paired t-test | *t* -0.465 (p=0.643) | NR (assumed no agreement) | Assumed based on *P*-value |
| Bland & Altman, 1999 | Standard LOA | $LOA_L$: -54.7 (95%CI -61.9, -47.5)<br>$LOA_U$: 22.1 (95%CI 14.9, 29.3) | No agreement | Limits within ±10 mmHg |
| | Adjusted LOA | $LOA_L$: -56.68 (95%CI -63.5, -49.9)<br>$LOA_U$: 25.44 (95%CI 18.7, 32.2) | No agreement | |

| | | | | |
|---|---|---|---|---|
| | Non-parametric approach | Grade D (16% within 5 mmHg, 35% within 10 mmHg, 49% within 15 mmHg - fails to achieve Grade C) | No agreement | To achieve grade C: at least 40% measurements ≤5 mmHg, 65% ≤10 mmHg, 85% ≤15 mmHg |
| Bland & Altman, 1986 | Standard LOA | $LOA_L$: 79.7 (95%CI -114.3, -45.1)<br>$LOA_U$: 75.5 (95%CI 40.9, 110.1) | No agreement | Limits within ±10 L/min |
| | Correlation coefficient | $r$ = 0.94 (p<0.001) | "Measurements are linearly related" | Test of significance ($P$<0.05) |
| Cecconi et al. | Standard LOA | $LOA_L$: -2.7, $LOA_U$: 2.3 | Agreement | NR |
| | Coefficient of error (CE) | CE = 4%<br>CV = 15%<br>PE = 30% | No agreement, based on CE and CV alone, agreement based on PE | PE within ±30% |
| Choudhary, 2008 | Adjusted LOA[1] | $LOA_L$: -1.3521, $LOA_U$: 2.7705[1] | NR[1] | NR[1] |
| | Tolerance interval | $U_x$=2.33 | *Poor agreement* | NR |
| Choudhary and Nagaraja, 2017 | LOA (version not reported) | $LOA_L$: -9.6, $LOA_U$: 14.6 | NR | NR |
| | Concordance correlation coefficient (CCC) | 0.85 (lower one-sided 95% confidence bound: 0.80) | Weak agreement | NR |
| | Total deviation index (TDI) (0.90) | 10.91 (upper one-sided 95% confidence bound: 12.14) | Weak agreement | NR |
| Euser et al. | Standard LOA | Log-transformed measurements:<br>$LOA_L$: -0.392, $LOA_U$: 0.392<br>Ratio of 2 measurements:<br>$LOA_L$: 0.400, $LOA_U$: 2.499<br>Difference between 2 measurements as a function of the mean X:<br>$LOA_L$: -0.85X, $LOA_U$: 0.85X | NR | NR |

| | | | | |
|---|---|---|---|---|
| | Coefficient of variation (CV) | Inter-observer CV=12.5%<br>Inter-observer CV of log-transformed data= 33.1% | No agreement, based on log-transformed data | NR |
| Francq and Govaerts | Adjusted LOA | Agreement interval (AI):<br>$LOA_L$: -25.440, $LOA_U$: 56.679<br>XL-AI: $LOA_L$: -29.871, $LOA_U$: 61.110 | No agreement | Limits within ±10 mmHg |
| | Correlated-errors-in-variables model (CEIV) | $\beta$ tolerance interval: -6.030, 37.269<br>$\beta\gamma$ tolerance interval: -6.757, 37.996 | No agreement | |
| Haghayegh et al. | Adjusted LOA | Simulation 1:<br>$LOA_L$: -1127.14 (95%CI -1462.45, -791.84)<br>$LOA_U$: 55.36 (95%CI -279.95, 390.67)<br>Simulation 2:<br>$LOA_L$: -2.869 (95%CI -5.652, -0.086)<br>$LOA_U$: 7.629 (4.846, 10.412) | NR | "when MDC (minimal detectable change) < MCIC (minimal clinically important change) and bias ≈ 0" |
| | ICC | Simulation 1:<br>$ICC_{Absolute}$: 0.855 (95%CI 0.60, 0.95)<br>$ICC_{Consistency}$: 0.99 (95%CI 0.997, 1.000)<br>Simulation 2:<br>$ICC_{absolute}$: 0.977<br>$ICC_{consistency}$: 0.990 | NR | 95%CI of ICC:<br><0.5=poor reliability<br>0.5-0.75=moderate reliability<br>0.75-0.90=good reliability<br>>90=excellent reliability |
| Kim and Lee | Standard LOA | $LOA_L$: -84.30, $LOA_U$: 80.06 | No agreement | NR |
| | CCC | 0.943 | Agreement | CCC>0.75 |
| | PCC | 0.943 | NR | NR |
| | Bias correction factor | 0.999 | NR | NR |
| Lai and Shiao | Standard LOA | $LOA_L$: -2.48, $LOA_U$: 4.54 | No agreement | NR – visual inspection of the Bland & Altman plot |
| | Linear mixed model | Fixed effect parameter intercept: 2.5056 (p=0.0054) | No agreement | *P*-value<0.05 associated with the estimated parameters |

| | | | | |
|---|---|---|---|---|
| | | Estimated parameter of the coefficient of fetal hemoglobin percent: -0.0263 (p<0.0001)<br>Random effect first-order autocorrelation parameter: $\rho$=0.6978 (p<0.0001) | | |
| Lombard et al. | LOA (version not reported) | $LOA_L$: -13406, $LOA_U$: 10694 (p<0.0001) | "Biased agreement" | p<0.05=no agreement |
| | Spearman Correlation | *r*=0.26 (p<0.0001) | Acceptable agreement based on *r* value, biased agreement based on *p* value | Good: ≥ 0.50; Acceptable: 0.20 – 0.49; Poor < 0.20 |
| | Wilcoxon Signed Rank test | P>0.05 | Good agreement | Good: p > 0.05; Poor: ≤ 0.05 |
| | Percentage difference (%) | -9.8% | Good agreement | Good: 0.0 – 10.9%; Acceptable: 11.0 – 20.0%; Poor: > 20.0% |
| | Cross classification (% in same tertile) | 46.8% | Poor agreement | Good: ≥ 50%; Poor: < 50% |
| | Cross classification (% in opposite tertile) | 19.2% | Poor agreement | Good: ≤ 10%, Poor: > 10% |
| | Weighted kappa statistic | 0.2 | Acceptable agreement | Good: ≥ 0.61; Acceptable: 0.20 – 0.59; Poor: < 0.20 |
| Lorne et al. | Standard LOA[47] | Wesseling: $LOA_L$: -0.80, $LOA_U$: 1.26 | No agreement | Limits within ±0.5 |
| | | LiDCO: $LOA_L$: -1.55, $LOA_U$: 1.20 | No agreement | |
| | | PiCCO: $LOA_L$: -1.60, $LOA_U$: 1.89 | No agreement | |
| | | Hemac: $LOA_L$: -0.81, $LOA_U$: 1.89 | No agreement | |
| | | Modelflow: $LOA_L$: -0.74, $LOA_U$: 0.74 | No agreement | |
| | Interchangeability rate | Wesseling: Inclusion rate 76% | No agreement | ≥95% = excellent<br>≥ 90% = good |
| | | LiDCO: Inclusion rate 73% | No agreement | |

| | | | | |
|---|---|---|---|---|
| | | PiCCO: Inclusion rate 62% | No agreement | 75-90% = poor<br><75% = not clinically relevant |
| | | Hemac: Inclusion rate 86% | No agreement | |
| | | Modelflow: Inclusion rate 93%, Interchangeability cut-off 4.80 | Agreement | |
| Myles and Cui | Standard LOA | $LOA_L$: -128, $LOA_U$: 88 | NR | NR |
| | Adjusted LOA | $LOA_L$: -154, $LOA_U$: 95 | Unacceptable agreement | NR |
| | Modified (random effects) LOA | $LOA_L$: -116, $LOA_U$: 57 | NR | NR |
| Parker et al., 2020 | Standard LOA | -6.40 to 3.19 | No agreement | Limits within ±5 |
| | Adjusted LOA | -9.99 to 6.78 | No agreement | |
| | Modified (mixed effect) LOA | -11.57 to 8.38 | No agreement | |
| | Modified (mixed effect) LOA, with interactions | -11.86 to 9.30 | No agreement | |
| | CCC | 0.68 (95% CI 0.60 to 0.72) | "Slight" agreement | 95%CI does not include the value of zero or negative values |
| | TDI | 10.9 (95% CI 9.4 to 12.7) | No agreement | Clinically acceptable difference of ±5 |
| | Coverage probability (CP) | 0.63 (95% CI 0.56 to 0.70) | No agreement | ≥95% |
| | Coefficient of individual agreement (CIA) | 0.68 (95% CI 0.57 to 0.75) | No agreement | ≥80% |
| | PCC | 0.74 (95%CI 0.69, 0.78) | NR | NR |
| | Unadjusted CCC | 0.72 (95%CI 0.67, 0.76) | NR | NR |
| Parker et al., 2016[a] | Adjusted LOA | Camera (rate/ second): $LOA_L$: -13.35, $LOA_U$: 6.72 | No agreement | Limits within ±10 |
| | | Accelerometer: $LOA_L$: -8.74, $LOA_U$: 4.38 | Agreement | |
| | | Camera (rate/ second): | No agreement | |

| | | | | |
|---|---|---|---|---|
| | Modified (mixed effect) LOA | $LOA_L$: -12.71 (95%CI -14.84, -11.42)<br>$LOA_U$: 6.30 (95%CI 5.00, 8.39) | | |
| | | Accelerometer:<br>$LOA_L$: -8.63 (-9.45, -7.96)<br>$LOA_U$: 4.27 (95%CI 3.62, 5.21) | Agreement | |
| | Modified (mixed effect) LOA, after removing outliers | Camera (rate/ second):<br>$LOA_L$: -11.54 (95%CI -13.20, -10.40)<br>$LOA_U$: 5.24 (95%CI 4.10, 6.80) | No agreement | |
| | | Accelerometer:<br>$LOA_L$: -7.91 (95%CI -8.75, -7.20)<br>$LOA_U$: 3.63 (95%CI 3.01, 4.64) | Agreement | |
| Quiroz and Burdick | Standard LOA[3] | $LOA_L$: -79.7 (95%CI -114.3, -45.1)<br>$LOA_U$: 75.5 (95%CI 40.9, 110.1) | No agreement[3] | Limits within ±10 L/min[3] |
| | 90% Generalized confidence intervals (GCI) | $GCI_L$: -79.14 , $GCI_U$: 85.34 | No agreement | At least 90% of the absolute difference is less than 10 |
| Rankin and Stokes | Standard LOA | $LOA_L$: -2.73, $LOA_U$: 1.53 | "Reasonable" Agreement | Limits within ±1.5 |
| | ICC | 0.92 (95% CI 0.72 to 0.98) | Agreement | NR |
| Roy | Standard LOA | PEFR data[3]:<br>$LOA_L$: 79.7 (95%CI -114.3, -45.1)<br>$LOA_U$: 75.5 (95%CI 40.9, 110.1) | No agreement[3] | Limits within ±10 L/min[3] |
| | Adjusted LOA | SBP data[1]:<br>$LOA_L$: -56.68 (95%CI -63.5, -49.9)<br>$LOA_U$: 25.44 (95%CI 18.7, 32.2) | No agreement[1] | Limits within ±10 mmHg[1] |
| | | LVEF data[1]: $LOA_L$: -1.3521, $LOA_U$: 2.7705 | NR[1] | NR[1] |
| | Linear mixed effect model (LME), based on Bonferroni adjusted p-values for variabilities | SBP data: Bias: p<0.0001<br>Between-subject: p=1.0<br>Within-subject: p<0.0001 | No agreement | p-values≥0.05 |
| | | PEFR data: Bias: p=1.0 | Agreement | |

| | | | | |
|---|---|---|---|---|
| | | Between-subject: p=1.0<br>Within-subject: p=0.8199 | | |
| | | LVEF data: Bias: p=0.0612<br>Between-subject: p=1.0<br>Within-subject: p=1.0 | Agreement | |
| Saugel et al. | Standard LOA | Systolic arterial pressure:<br>$LOA_L$: -43.9, $LOA_U$: 56 | No agreement | Bias ±5 mmHg and SD ±8 mmHg |
| | | Mean arterial pressure:<br>$LOA_L$: -17.1, $LOA_U$: 73.6 | No agreement | |
| | Error grid analysis | Systolic arterial pressure:<br>risk levels A-E: 78%, 14%, 6%, 1%, and 1% | NR | NR |
| | | Mean arterial pressure:<br>risk levels A-E: 18%, 54%, 20%, 7%, and 8% | NR | NR |
| Schluter | Adjusted LOA[1] | $LOA_L$: -56.68 (95%CI -63.5, -49.9)<br>$LOA_U$: 25.44 (95%CI 18.7, 32.2) | No agreement[1] | Limits within ±10 mmHg[1] |
| | Exchangeable multivariate hierarchical Bayesian model (HB1) | $LOA_L$: -56.2, $LOA_U$: 25.1 | No agreement | NR |
| | Non-exchangeable multivariate hierarchical Bayesian model (HB2) | $LOA_L$: -55.9, $LOA_U$: 25.0 | No agreement | NR |
| Stevens et al., 2018 | Adjusted LOA[1] | $LOA_L$: -56.68 (95%CI -63.5, -49.9)<br>$LOA_U$: 25.44 (95%CI 18.7, 32.2)[1] | No agreement[1] | Limits within ±10 mmHg[1] |
| | Probability of agreement approach, using a clinically acceptable difference ±10 (heteroscedastic) | Probability of agreement never exceeds 0.3 | No agreement | $\theta$>0.95 |

| | | | | |
|---|---|---|---|---|
| Stevens et al., 2017 | Adjusted LOA[1] | $LOA_L$: -56.68 (95%CI -63.5, -49.9)<br>$LOA_U$: 25.44 (95%CI 18.7, 32.2)[1] | No agreement[1] | Clinically acceptable difference ±10 mmHg[1] |
| | Probability of agreement approach, using a clinically acceptable difference ±10 (homoscedastic) | Unconditional probability of agreement ($\theta$) 0.799 (95%CI 0.61, 0.98) | No agreement | $\theta$>0.95 |
| Taffé, 2021 | LOA (version not reported) | $LOA_L$: -19.6, $LOA_U$: 15.3 | NR | NR |
| | Bias, agreement, and precision plots | Assessment of agreement is done by visual inspection of a series of plots. | NR | Assessment of confidence bands in precision plot. |
| Taffé, 2020 | Adjusted LOA[1] | $LOA_L$: -56.68 (95%CI -63.5, -49.9)<br>$LOA_U$: 25.44 (95%CI 18.7, 32.2)[1] | No agreement[1] | Limits within ±10 mmHg[1] |
| | Bias, agreement, and precision plots | Assessment of agreement is done by visual inspection of a series of plots. | No agreement | Visual inspection of the precision plot (confidence bands do not overlap) |

Abbreviations: CCC, Concordance correlation coefficient; CE, Coefficient of error; CI, Confidence intervals; CV, Coefficient of variation; df, Degrees of freedom; GCI, Generalized confidence intervals; ICC, Intraclass correlation coefficient; LOA, Limits of agreement; LVEF, Left ventricular ejection fraction; NR, Not reported; PCC, Pearson correlation coefficient; PE, Percentage error; PEFR, Peak expiratory flow rate; SBP, Systolic blood pressure; TDI, Total deviation index; TOST, Two one-sided t-test.

[a]Not all results from this paper are presented if many had similar conclusions. Results presented were chosen to show variety in agreement conclusions.

[b]When a reference is provided with LOA results, LOA results were extracted from a different source.

**Table 3.** Overview of all statistical methods presented in the included publications.

| Statistical methods | Number of publications presenting the statistical methods |
|---|---|
| **Limits of agreement methods** | |
| Unadjusted LOA | 18 |
| LOA adjusted for repeated measures (“fixed effect”) | 12 |
| Mixed/random effect LOA | 3 |
| Not reported | 3 |
| **Agreement ranges** | |
| Total deviation index (TDI)[35,40] | 2 |
| Tolerance interval[15] | 1 |
| Correlated-errors-in-variables model (CEIV)[16] | 1 |
| 90% Generalized confidence intervals (GCI)[18] | 1 |
| Exchangeable multivariate hierarchical Bayesian model (HB1)[20] | 1 |
| non-exchangeable multivariate hierarchical Bayesian model (HB2)[20] | 1 |
| **Correlation coefficients, probabilities, percentages** | |
| Pearson correlation coefficient (PCC) [3,17,27,30,35] | 5 |
| Intraclass correlation coefficient (ICC) [25,29,32] | 3 |
| Concordance correlation coefficient (CCC)[35,40] | 2 |
| Unadjusted CCC[17,35] | 2 |
| T-test[24,30] | 2 |
| Spearman Correlation[28] | 1 |
| Bias correction factor[17] | 1 |
| Weighted kappa statistic[28] | 1 |
| Structural equations (SEM) approach[14] | 1 |
| Two one-sided t-test (TOST)[24] | 1 |
| Linear mixed model[38] | 1 |
| Linear mixed effect model (LME), with Bonferroni adjusted p-values[19] | 1 |
| Wilcoxon Signed Rank test[28] | 1 |

| | |
|---|---|
| Coverage probability (CP)[35] | 1 |
| Coefficient of individual agreement (CIA)[35] | 1 |
| Probability of agreement approach, heteroscedastic distribution[21] | 1 |
| Probability of agreement approach, homoscedastic distribution[22] | 1 |
| Mean agreement[27] | 1 |
| Non-parametric approach[1] | 1 |
| Coefficient of error (CE)[33] | 1 |
| Coefficient of variation (CV)[31] | 1 |
| Percentage difference (%)[28] | 1 |
| Cross classification[28] | 1 |
| Interchangeability rate[34] | 1 |
| Error grid analysis (EGA)[39] | 1 |
| **Other methods** | |
| Bias, agreement, and precision plots[23,26] | 2 |

Abbreviations: CCC, Concordance correlation coefficient; LOA, Limits of agreement.

**Table 4.** Overview of proposed methods for handling repeated measurements.

| Study authors | Using one measurement only, excluding the repeated measures from the analyses | Considering all data as independent, ignoring correlated nature of the repeated measures | Using the means of the repeated measurements | Adjusting for correlation between repeated measurements | Not reported |
|---|---|---|---|---|---|
| **1. Comparison between various versions of the LOA** | | | | | |
| Bland & Altman, 1999[a] | Standard LOA | | | Adjusted LOA | |
| Myles and Cui[a] | Standard LOA | | | Adjusted LOA, Modified LOA | |
| Parker et al., 2020[a] | Standard LOA | | | Adjusted LOA, Modified LOA | |
| Parker et al., 2016 | | | | Adjusted LOA, Modified LOA | |
| **2. Alternate statistical methods adjusted for repeated measurements versus adjusted LOA** | | | | | |
| Choudhary, 2008[b] | | | | Adjusted LOA, Alternate statistical methods | |
| Francq and Govaerts | | | | Adjusted LOA, Alternate statistical methods | |
| Haghayegh et al. | | | | Adjusted LOA, Alternate statistical methods | |
| Roy[b] | Standard LOA | | | Adjusted LOA, Alternate statistical methods | |

| | | | | | |
|---|---|---|---|---|---|
| Schluter[b] | | | | Adjusted LOA, Alternate statistical methods | |
| Stevens et al., 2018[b] | | | | Adjusted LOA, Alternate statistical methods | |
| Stevens et al., 2017[b] | | | | Adjusted LOA, Alternate statistical methods | |
| Taffé, 2020[b] | | | | Adjusted LOA, Alternate statistical methods | |
| **3. Alternate statistical methods adjusted for repeated measurements versus standard LOA** | | | | | |
| Alanen[b] | Standard LOA | | | Alternate statistical methods | |
| Cecconi et al. | | | Standard LOA | Alternate statistical methods | |
| Choudhary and Nagaraja, 2017 | | | | Alternate statistical methods | LOA (version not reported) |
| Lai and Shiao | | Standard LOA | | Alternate statistical methods | |
| Quiroz and Burdick[b] | Standard LOA | | | Alternate statistical methods | |
| Taffé, 2021 | | | | Alternate statistical methods | LOA (version not reported) |
| **4. Alternate statistical methods adjusted using the mean of repeated measurements versus standard LOA** | | | | | |
| Ambrosini et al. | | | Standard LOA, Alternate statistical methods | | |

| | | | | | |
|---|---|---|---|---|---|
| Anvari et al. | Standard LOA | | Standard LOA, Alternate statistical methods | | |
| Batterham et al. | | | Standard LOA, Alternate statistical methods | | |
| Euser et al. | | | Standard LOA, Alternate statistical methods | | |
| **5. Alternate statistical methods unadjusted for repeated measurements versus standard LOA** | | | | | |
| Armstrong et al. | Standard LOA, Alternate statistical methods | Standard LOA, Alternate statistical methods | | | |
| Bland & Altman, 1986 | Standard LOA, Alternate statistical methods | | | | |
| Kim and Lee | Standard LOA, Alternate statistical methods | | | | |
| Lorne et al. | | Standard LOA, Alternate statistical methods | | | |
| Rankin and Stokes | Standard LOA, Alternate statistical methods | | | | |
| Saugel et al. | Standard LOA, Alternate statistical methods | | | | |
| **6. Adjustments for repeated measurements not reported for LOA estimates and alternate statistical methods** | | | | | |
| Lombard et al. | | | | | LOA, Alternate statistical methods |

Abbreviations: LOA, Limits of agreement.

[a]Both unadjusted LOA and adjusted LOA for repeated measurements are presented in these papers for comparison purposes.

[b]LOA data extracted from Bland & Altman articles.

**Table 5.** Summary of strengths and limitations of the LOA methods, as reported by the authors of the included publications.

| Strengths | Limitations |
|---|---|
| Simple / straightforward / easy to do and interpret[1,3,14,19,22,24,26,31,35] | Assumptions about the normality and the homoscedasticity of the distribution of the data need to be met[1,25,26,34,35,36,38] |
| Provides information about the bias and magnitude of differences between methods[27,28,32] | Agreement limits are subjective and based on clinical judgement, not based on statistical factors[14,27,28,29] |
| Provides a visual representation of the comparison between methods[32,38] | Does not provide information about which method is superior[14,22,35] |
| Results are based on the original unit of measurement[35] | There is uncertainty in point estimates that can lead to biased conclusions, 95% CI should be presented[24,35] |

Abbreviations: CI, Confidence intervals; LOA, Limits of agreement.

## 3.10 FIGURES

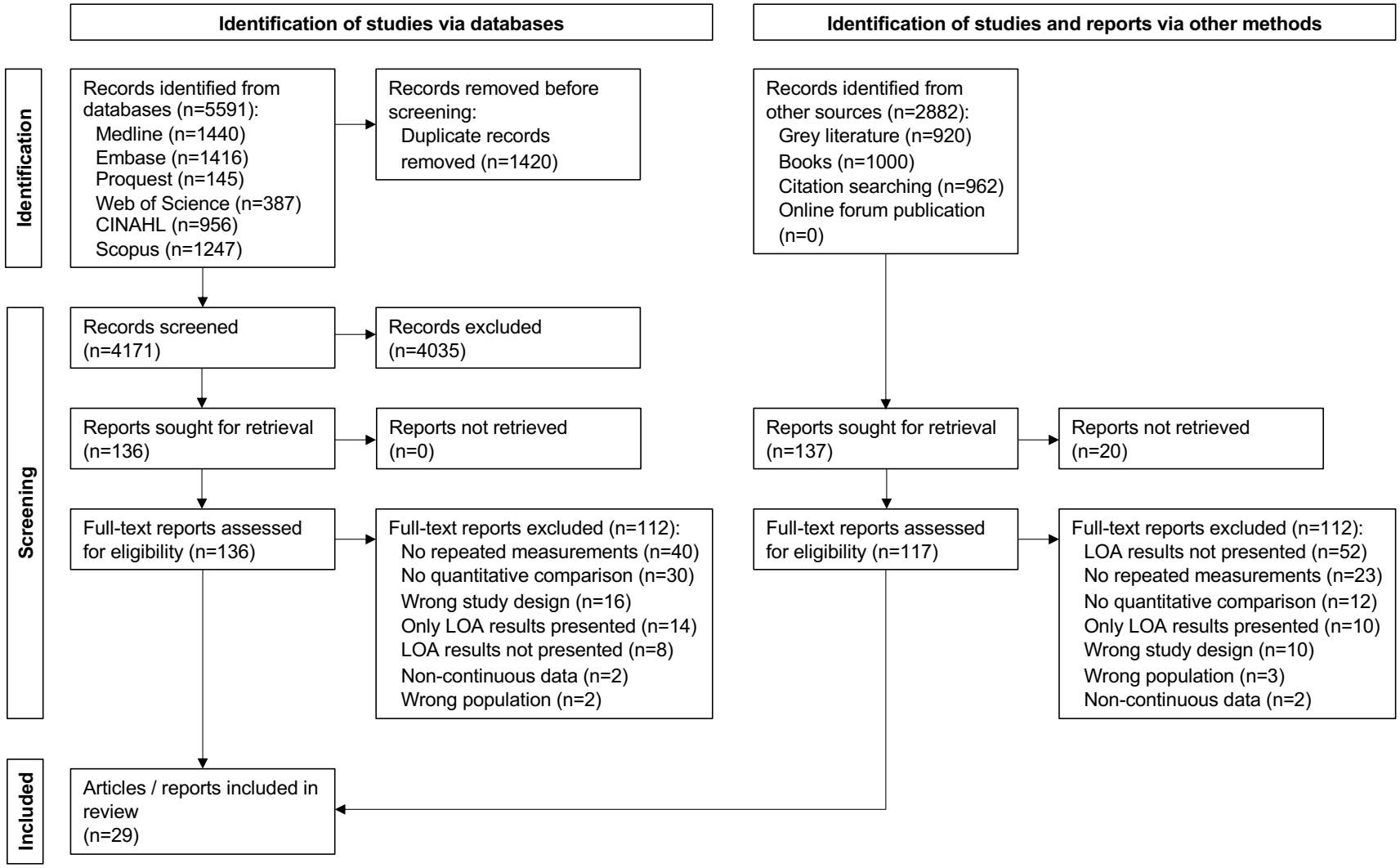



**Figure 1.** PRISMA flow diagram.
LOA, Limits of agreement.

## 3.11 APPENDIX 1. PRISMA-SCR CHECKLIST

**Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist**

| SECTION | ITEM | PRISMA-ScR CHECKLIST ITEM | REPORTED ON PAGE # |
|---|---|---|---|
| **TITLE** | | | |
| Title | 1 | Identify the report as a scoping review. | 43 |
| **ABSTRACT** | | | |
| Structured summary | 2 | Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives. | 45, 46 |
| **INTRODUCTION** | | | |
| Rationale | 3 | Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach. | 47, 48 |
| Objectives | 4 | Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives. | 48 |
| **METHODS** | | | |
| Protocol and registration | 5 | Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number. | 49 |
| Eligibility criteria | 6 | Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale. | 49, 50 |
| Information sources | 7 | Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed. | 51 |
| Search | 8 | Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated. | 42 |
| Selection of sources of evidence | 9 | State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review. | 51, 52 |
| Data charting process | 10 | Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators. | 52 |
| Data items | 11 | List and define all variables for which data were sought and any assumptions and simplifications made. | 52-54 |
| Critical appraisal of individual sources of evidence | 12 | If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate). | 55 |

| SECTION | ITEM | PRISMA-ScR CHECKLIST ITEM | REPORTED ON PAGE # |
|---|---|---|---|
| Synthesis of results | 13 | Describe the methods of handling and summarizing the data that were charted. | 54 |
| **RESULTS** | | | |
| Selection of sources of evidence | 14 | Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram. | 56, Figure 1 |
| Characteristics of sources of evidence | 15 | For each source of evidence, present characteristics for which data were charted and provide the citations. | 56, 57, Table 1 |
| Critical appraisal within sources of evidence | 16 | If done, present data on critical appraisal of included sources of evidence (see item 12). | NA |
| Results of individual sources of evidence | 17 | For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives. | 57, Table 2 |
| Synthesis of results | 18 | Summarize and/or present the charting results as they relate to the review questions and objectives. | 57-61, Table 3-5 |
| **DISCUSSION** | | | |
| Summary of evidence | 19 | Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups. | 62-64 |
| Limitations | 20 | Discuss the limitations of the scoping review process. | 64-65 |
| Conclusions | 21 | Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps. | 66 |
| **FUNDING** | | | |
| Funding | 22 | Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review. | 66 |

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

*From:* Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.

# CHAPTER 4. PROTOCOL FOR THE PREMISE OBSERVATIONAL STUDY: POINT-OF-CARE HEMOGLOBIN ACCURACY AND TRANSFUSION OUTCOMES IN NON-CARDIAC SURGERY AT A CANADIAN TERTIARY ACADEMIC HOSPITAL

---

Karine Brousseau[1,2], Leah Monette[1,2], Daniel McIsaac[2,3], Aklile Workneh[1,2], Alan Tinmouth[2], Julie Shaw[4], Tim Ramsay[2], Ranjeeta Mallick[2], Justin Presseau[2], Christopher Wherrett[2,3], François Martin Carrier[5], Dean Fergusson[2], Guillaume Martel[1,2*]

[1]Liver and Pancreas Unit, Department of Surgery, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

[3]Department of Anesthesiology, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

[4]Division of Biochemistry, Eastern Ontario Regional Laboratories Association and Department of Pathology and Laboratory Medicine, University of Ottawa.

[5]Department of Anesthesiology and Department of Medicine, Critical Care Division, Centre Hospitalier de l'Université de Montréal, Université de Montréal, Montreal, QC, Canada

## PUBLICATION INFORMATION

This manuscript was accepted for publication in *BMJ Open* and is currently in press, as of September 26, 2023.

## 4.1 PREFACE TO CHAPTER 4

This chapter presents the protocol for the PREMISE (point-of-care hemoglobin accuracy and transfusion outcomes in non-cardiac surgery) prospective method comparison study. This protocol was strictly followed during the data collection phase. This chapter helps with the appraisal of the statistical analysis plan associated with this study and the manuscript presenting findings from this study, reported in Chapters 5 and 6.

## 4.2 ABSTRACT

**Introduction**: Transfusions in surgery can be life-saving interventions, but inappropriate transfusions may lack clinical benefit and cause harm. Transfusion decision-making in surgery is complex and frequently informed by hemoglobin (Hgb) measurement in the operating room. Point-of-care testing for hemoglobin (POCT-Hgb) is increasingly relied upon given its simplicity and rapid provision of results. POCT-Hgb devices lack adequate validation in the operative setting, particularly for Hgb values within the transfusion zone (60-100 g/L). This study aims to examine the accuracy of intraoperative POCT-Hgb instruments in non-cardiac surgery, and the association between POCT-Hgb measurements and transfusion decision-making.

**Methods and analysis**: PREMISE is an observational prospective method comparison study. Enrolment will occur when adult patients undergoing major non-cardiac surgery require POCT-Hgb, as determined by the treating team. Three concurrent POCT-Hgb results, considered as index tests, will be compared to a laboratory analysis of Hgb (lab-Hgb), considered the gold standard. Participants may have multiple POCT-Hgb measurements during surgery. The primary outcome is the difference in individual Hgb measurements between POCT-Hgb and lab-Hgb, primarily among measurements that are within the transfusion zone. Secondary outcomes include POCT-Hgb accuracy within the entire cohort, post-operative morbidity, mortality, and transfusion rates. The sample size is 1750 POCT-Hgb measurements to obtain a minimum of 652 Hgb measurements <100 g/L, based on an estimated incidence of 38%. The sample size was calculated to fit a logistic regression model to predict instances when POCT-Hgb are inaccurate, using 4 g/L as an acceptable margin of error.

**Ethics and dissemination**: Institutional ethics approval has been obtained by the Ottawa Health Science Network – Research Ethics Board (OHSN-REB) prior to initiating the study. Findings from this study will be published in peer-reviewed journals and presented at relevant scientific conferences. Social media will be leveraged to further disseminate the study results and engage with clinicians.

**Keywords**: hemoglobin; hematocrit; point-of-care test; transfusion; red blood cell; perioperative medicine; surgery; anesthesiology.

## STRENGTHS AND LIMITATIONS OF THIS STUDY

- PREMISE is pragmatic and will include diverse participants undergoing major non-cardiac operations at risk of red blood cell transfusions.
- The large sample size will allow the generation of a prediction model to identify conditions under which POCT-Hgb devices are at a particularly high risk of failure.
- Tabletop blood gas analyzers were not included in this study given resource limitations and that it is not commonly used as a point-of-care device to determine hemoglobin concentrations in the setting where this study is taking place.

**Abbreviations**: CBC, Complete blood count; EDCS, Electronic data capture system; GCP, Good Clinical Practice; Hct, Hematocrit; Hgb, Hemoglobin; IQMH, Institute for Quality Management in Healthcare; Lab, Laboratory; LoA, Limits of Agreement; OHSN-REB, Ottawa Health Science Network Research Ethics Board; POCT, Point-of-care-testing; RA, Research assistant; RBC, Red blood cell; REB, Research Ethics Board; SD, Standard deviation.

## 4.3 INTRODUCTION

Worldwide, close to 85 million red blood cell (RBC) units are transfused annually, with surgical patients accounting for up to 44% of transfusions.[1-3] Clinically indicated RBC transfusions in surgery can be life-saving interventions, while inappropriate transfusions provide no clinical benefit and can cause harm.[4] Transfusions carry perioperative risks including transfusion-related acute lung injury and transfusion-associated circulatory overload.[5] They may also lead to transfusion-related immunomodulation which is associated with worsened perioperative and long-term oncologic outcomes in surgical patients.[6,7] While these risks exist, they must be balanced against the risk of end-organ ischemia due to anemia resulting from under-transfusion.[8]

There is widespread evidence of significant variation in transfusion practice in surgical patients.[9,10] Although some variation may be expected, major differences in perioperative RBC transfusion events that cannot be explained by disease severity or patient preference likely reflects unwarranted variations in clinical care.[11,12] Intraoperative RBC transfusions are generally defined as unwarranted when given in the absence of an acceptable clinical indication.[13,14] Published estimates suggest that 19% to 49% of RBC transfusions in surgery may be inappropriate, including in the field of liver surgery where a majority of transfusion events were unwarranted and 87% received too many RBC units.[15-18] These findings emphasize the need to prevent unwarranted RBC transfusions in surgery given the cost of blood, associated risks, limited supply, and the nature of altruistic donations.[19,20]

Intraoperative transfusion decision-making is a complex and dynamic process, and is unique given the risk of acute and rapid blood loss, and of wide variations in hemodynamic parameters in surgical patients.[21] While transfusion should be primarily informed by clinical parameters, it also often relies on clinician judgement and non-clinical behavioural factors.[22] Although transfusion practices outside of the operating room are guided by high-quality evidence supporting restrictive transfusion triggers, there is no consensus supporting restrictive transfusion practices during surgery.[23,24] Most guidelines informing intraoperative transfusions are sub-

specialty specific, of poor quality, fail to provide specific recommendations, and are primarily extrapolated from the non-operative setting.[25] Moreover, few clinical trials have compared intraoperative transfusion strategies, the majority of which had limited sample sizes and notable variability in chosen transfusion triggers.[26] As such, transfusion decision-making in the operating room represents a significant evidence and knowledge gap.

Current practice guidelines suggest that the decision to transfuse RBCs in surgery should be generally limited to those with a hemoglobin (Hgb) of 60-100 g/L and informed by important clinical parameters.[27] Implicit in this recommendation is the emphasis on measuring Hgb at critical times during surgery. Recent surveys suggest that intraoperative Hgb measurement is the most important parameter in deciding whether to transfuse and that 89% of pediatric anesthesiologists report measuring the Hgb prior to initiating a transfusion.[28,29] Other publications support the critical relationship between Hgb measurement during surgery and transfusion decisions, including our recent systematic review in which eight out of 10 published guidelines provided transfusion recommendations based on Hgb/hematocrit(Hct) measurement.[25,30,31]

Traditionally, Hgb measurement during surgery was done through submission of a complete blood count (CBC) specimen to the main laboratory. More recently, point-of-care testing (POCT) instruments capable of measuring or calculating hemoglobin (POCT-Hgb) have become commonplace in surgery due to their ease of use and rapid delivery of a result. Two main classes of POCT-Hgb methods are currently in use in operating rooms, with a third class growing in utilization. The first class uses a modified azide-methemoglobin measurement method and yields a hemoglobin value from 10 µL of whole blood in 10-60 seconds.[32] A commonly used device in this category is HemoCue (HemoCue AB, Ängelholm, Sweden). The second class can yield a panel of results for several analytes from 65-100 µL of whole blood in 120 seconds using the conductometric methods and calculates the Hgb using the measured Hct (Hgb (g/L) = Hct (%) x 3.4).[33,34] Commonly used devices in this category include i-STAT (Abbott Laboratories, Abbott Park, IL, USA) and Epoch (Siemens Healthineers, Erlagen, Germany). The third class of POCT-

Hgb methods involves non-invasive monitoring of capillary hemoglobin through multiwavelength sensors and pulse co-oximetry. The Rad-67 Pulse CO-Oximeter (Masimo, Irvine, CA, USA) is one such example.

To date, POCT-Hgb instruments have been validated with static and normal Hgb values, such as in healthy blood donors.[35] In clinical care, they have been examined primarily in the non-operative setting, such as the outpatient clinic or emergency department, reporting conflicting results in relation to bias and accuracy of POCT-Hgb devices.[33,36-38] Their value in surgery, where Hgb can change rapidly due to bleeding and hemodilution from concurrent intravenous fluid administration, are relatively untested.[39] More importantly, there is also a paucity of published validation data based on intraoperative Hgb values within the transfusion zone of 60-100 g/L used for clinical decision making, highlighting a major gap in existing evaluations.[40-42] According to a recent systematic review, several comparison studies have examined HemoCue against a laboratory standard, with a minority including surgical patients.[43] Only a small number of intraoperative evaluations of HemoCue have been published, concluding that while this POCT-Hgb device is often more accurate than other POCT-Hgb devices, it should not be used interchangeably with a lab-Hgb value.[40,44] Comparison studies evaluating i-STAT in surgery are even more limited in terms of numbers, sample sizes, and methodology.[45-47]

Despite the lack of adequate validation in surgery, POCT-Hgb results are clearly relied upon to inform intraoperative transfusion decisions. Unpublished data from the authors' institution reveals that POCT was the sole Hgb measurement technique (96%) in a granular cohort of liver, pancreas, and sarcoma resections (n=304 patients), of which 50% had at least one POCT-Hgb done. Among patients with an intraoperative POCT-Hgb <100 g/L, 54% went on to receive at least one perioperative RBC transfusion. Data from a separate liver resection cohort identified that 22% of patients who had a POCT-Hgb had an intraoperative transfusion, with a mean trigger Hgb of 88 ± 21 g/L.[48] The importance of POCT-Hgb in driving and supporting transfusion decisions has also been emphasized by multiple other authors.[30,40,44,45,49,50]

Routine use of POCT-Hgb may not be warranted without rigorous and definitive evidence to inform intraoperative practice. Investigators have previously emphasized that POCT-Hgb technologies should be examined in clinical context and not solely as laboratory comparisons, as only 13% of POCT-Hgb evaluations have done so in the past.[51,52] The following proposed investigational plan is crucial and urgently needed as it will fill multiple evidence and knowledge gaps in perioperative medicine, aligning with several national and international groups research priorities on transfusion medicine.[27,53-56] This study will generate a comprehensive and robust examination of three commonly used POCT-Hgb devices in the operating room. The proposed work is pragmatic, including all types of non-cardiac procedures within a wide spectrum of surgical patients of all sexes, ages, and ethnicities. It will generate a critical link between POCT-Hgb measurements and RBC transfusion. As such, this study will examine POCT-Hgb accuracy in clinical context by generating a sample that is adequately powered, with special emphasis on tests conducted within the Hgb transfusion zone of 60-100 g/L.[27] We intend to leverage the wealth of data it will generate to inform future interventional trials on transfusion practices in the operating room.

**Objectives**

The overarching objective of this study is to determine the accuracy of intraoperative POCT-Hgb instruments in non-cardiac surgery, as well as its association with transfusion decision-making.

***Primary objective:*** To determine the accuracy of POCT-Hgb instruments in non-cardiac surgery when compared to lab-Hgb, and more specifically the accuracy of POCT-Hgb measurements when lab-Hgb values are within the Hgb transfusion zone of 60-100 g/L.

***Secondary Objectives:***

(1) To examine the accuracy of POCT-Hgb measurements when compared to lab-Hgb, within the entire Hgb range.

(2) To build clinical prediction models to predict the agreement failure of POCT-Hgb and lab-Hgb measurements, including for measurements where lab-Hgb values are within the transfusion zone of 60-100 g/L.

(3) To examine the association between POCT-Hgb measurements and intraoperative RBC transfusion.

(4) To examine intraoperative RBC transfusion appropriateness.

## 4.4 METHODS: PARTICIPANTS, INTERVENTIONS AND OUTCOMES

### Study Design and Settings

PREMISE is a method comparison study comparing different POCT-Hgb technologies in a prospective cohort of patients undergoing elective or urgent non-cardiac surgery. The study will take place at an academic health science center, comprised of two inpatient campuses (total 1,117 beds), in Ottawa, Ontario, Canada. Results will be reported following the STARD 2015 guidelines.[57]

### Eligibility Criteria

***Inclusion criteria:*** All adult patients undergoing major non-cardiac operations requiring invasive hemodynamic monitoring will be considered for inclusion. Non-cardiac surgery includes general surgery (gastrointestinal tract, splenic, abdominal wall, retroperitoneal, and breast operations), non-obstetrical gynecology, head and neck surgery, neurosurgery, orthopedic surgery, plastic surgery, spine surgery, thoracic surgery, urologic surgery, and vascular surgery. We will include both elective and urgent operations.

***Exclusion criteria*:** This study is intent on being pragmatic and generalizable, such that there are few exclusion criteria. Patients will be excluded if they are having cardiac surgery, if they are not undergoing surgery under general or neuraxial anesthesia (e.g. conscious sedation), or if they do not have an arterial catheter or central venous catheter. Furthermore, patients will be excluded if a blood sample is being collected outside of recruitment hours (Monday to Friday, 10AM-7PM), or due to a lack of personnel during recruitment hours (i.e., vacation or sick days).

### Study procedures

***General approach*:** Each time an intraoperative Hgb measurement is undertaken by an anesthesiologist, the patient will undergo three concurrent Hgb measurements using the same blood sample. The anesthesiologist will draw a standard three mL of whole blood from the

arterial/central line and run the first POCT-Hgb device of their choosing for clinical purposes (HemoCue or i-STAT). The exact time of blood draw will be recorded. The remainder of the sample (approximately 2.5mL, which is usually discarded) will be given to a research assistant who will concurrently submit the central lab-Hgb sample and run the other POCT-Hgb device. A third POCT-Hgb will be obtained concurrently using a finger sensor (Rad-67 Pulse CO-Oximeter). All POCT-Hgb analyses will be performed within five minutes of the blood being drawn and the lab-Hgb analysis will be performed the same day, by the end of the day. None of the research Hgb results will be disclosed to clinicians. Figure 1 shows a flow chart of the study procedures.

***Index tests*:** The three types of POCT-Hgb measurements will represent the commonest available POCT-Hgb method classes. They will be considered index tests, while the central lab-Hgb will be considered the reference (gold standard) test. For each device, the manufacturer's instructions and institutional procedures will be strictly adhered to, including device handling, training, storage, calibration, and quality control.

***Reference (gold standard) test*:** The central lab-Hgb sample will be collected in an ethylenediaminetetraacetic acid (EDTA) vacuum collection tube and delivered to the laboratory for processing in the usual fashion. A Sysmex XE-2100 analyzer (Sysmex Corporation, Kobe, Japan) will then be used to measure Hgb using the standard hemiglobincyanide method.

**Outcomes**

***Primary outcome:*** The primary outcome will be the Hgb concentration differences between POCT-Hgb and lab-Hgb, with a lab-Hgb within the transfusion zone of 60-100 g/L, given its clinical importance. The primary outcome will lead to measures of accuracy (mean difference with 95% limits of agreement) and incidence of agreement failure based on an acceptable difference of ± 4 g/L (as recommended by the Institute for Quality Management in Healthcare definition (IQMH)).[58] Outcomes that will be measured are presented in Table 1.

***Secondary outcomes:***

(1) Accuracy and agreement failure of POCT-Hgb devices for the entire Hgb range

(2) Intraoperative transfusion events

(3) Post-operative transfusion events

(4) Perioperative morbidity, following the classification of surgical complications proposed by Dindo et al, and mortality up to 30 days, determined by chart review.[59]

**Participant Timeline**

***Patient selection*:** During surgery, all patients meeting the above inclusion and exclusion criteria will be considered potentially eligible. Enrollment and data collection will occur when there is a clinical need for intraoperative Hgb measurement, at the discretion of the treating anesthesiologist. A patient may have multiple Hgb measurement events during one operation, all of which will be eligible for inclusion in this study.

**Sample size**

The sample size was calculated based on the primary outcome of agreement failure incidence, assuming an acceptable difference of ± 4 g/L, within the subset of samples that are in the transfusion zone of 60-100 g/L. More specifically, the sample size was chosen to allow the fitting of a robust clinical prediction model for agreement failure. Fifteen predictors will be considered for the prediction model. Assuming a c-statistic of 0.75, a prevalence of 0.5 whereby at least 50% of i-STAT measurements are expected to fall outside of the acceptable ± 4 g/L margin of error, a total of 652 measurements will be needed to fit a stable logistic regression model with 15 independent variables, based on methods described by Riley et al.[60] As such, patients will be enrolled until at least 652 lab-Hgb measurements are below 100 g/L. Given that preliminary data from our institution suggest that 38% of POCT-Hgb values are below 100 g/L, it is expected that a total of approximately 1,750 measurements will be required.

## 4.5 METHODS: DATA COLLECTION, MANAGEMENT, AND ANALYSIS

### Recruitment

Research assistants (RA) will be present in the operating rooms on a daily basis during study hours to enroll eligible patients. Anesthesiologists will be encouraged to call the RA when they are using the POCT-Hgb devices. The RA will also monitor and identify all requests for POCT-Hgb devices and accompany the devices to the relevant operating room.

### Data Collection Methods

***Assessment and collection of outcomes:*** Study data will be collected using an electronic data capture system (EDCS). All relevant clinical variables for POCT-Hgb events will be recorded on a standardized web-based case report form by the research staff. All transfusion events will be recorded in real time in the operating room, as well as using the transfusion medicine information system following transfer to the recovery room.

***Training:*** The research team will be trained on the use of the POCT devices and on study-related procedures. Devices calibration and quality control will be conducted in accordance with our existing institutional policies and the manufacturer's instructions.

***Loss to follow-up:*** We expect no loss to follow-up up to 30 days post-surgery. All such patients are carefully monitored in hospital, followed by one or two clinic visits in the first three weeks after discharge. In the event of a postoperative death within 30 days of surgery, patients will be considered to have completed follow-up.

### Data Management

Participants will be identified only by an identification number on all study documents and electronic databases. For each participant, relevant data will be collected and entered into the

EDCS by the research staff. Any hard copy study documents will be kept in locked and secure drawers within the study offices.

**Statistical Methods**

***Analytic plan:*** A detailed Statistical Analysis Plan has been generated for this study (Thesis Chapter 3). Data will be analyzed using SAS 9.4 (SAS Institute, Cary, NC, USA) and R version 4.0.2 (The R Foundation for Statistical Computing). The primary unit of analysis will be individual Hgb measurements. For each event, the three POCT-Hgb results will be compared to the gold standard lab-Hgb value. When required, repeated measures will be accounted for, given the correlated nature of multiple Hgb tests conducted serially in the same patient during one operation, using different methods developed for each of the proposed analysis. All analyses will be reported both for the transfusion zone cohort of Hgb 60-100 g/L and for the entire Hgb range (all observations).

***Continuous agreement*:** Continuous agreement between POCT-Hgb and lab-Hgb will be compared using the Limits of Agreement (LoA) method proposed by Bland and Altman, adjusted for multiple measurements per individual.[61,62] Limits of agreement will be presented with 95% confidence intervals.[61,63] Mean differences (y-axis) will be generated and plotted against the means of POCT-Hgb and lab-Hgb values (x-axis). LoA will be calculated as ±1.96 standard deviation (SD) around the mean difference, with a SD adjusted for multiple measurements. Calculations will also be repeated using mean differences expressed as a percentage difference ((lab-Hgb – POCT-Hgb)/lab-Hgb). Interchangeability between the POCT-Hgb and the lab-Hgb will be examined by plotting the two limits of agreement and determining visually whether these fall within the IQMH allowable difference of ± 4 g/L.[58] As a complement to the Bland-Altman analysis, folded empirical distribution plots will be constructed.[40] This visual representation of the distribution of actual differences complements the Bland-Altman plots by providing an easier visual assessment of the tails, or largest differences.

***Error grid analysis*:** The method of Morey et al will be utilized to further examine agreement between POCT-Hgb and lab-Hgb, in the context of clinical relevance.[41] POCT-Hgb (y-axis) will be plotted against lab-Hgb (x-axis). Three zones will be defined within the surface area of the plot, based on clinical risk of RBC transfusion. These will be modified from Morey, as this group used an allowable difference of 10 g/L rather than the 4 g/L defined by IQMH.[41,58] The uppermost and lowermost regions of the plot will include much of the green zone, where measurement errors have little influence on RBC transfusion decisions. A narrow band around the unity line completes the green zone. The yellow zone is mainly situated in portions of the plot where lab-Hgb or POCT-Hgb lies between 60-100 g/L. In this zone, measurement discrepancies have the potential to yield transfusion errors. Finally, the red zone will be defined as those areas on the plot where major clinical mistakes can be expected to occur. In these areas, discrepancies between measurements are very wide and at least one of the values is below 60 g/L, implying a guaranteed situation of over- or under-transfusion. The proportion of patients that fall within the yellow zone of transfusion risk will be calculated.

***Dichotomous agreement*:** The primary outcome will also be analyzed by dichotomizing the results of individual pairwise comparisons between POCT-Hgb and lab-Hgb. The absolute values of Hgb differences will be examined and any difference greater than 4 g/L (lab-Hgb <100 g/L) or greater than 5% (lab-Hgb ≥100 g/L) will be considered to lie outside of the allowable test difference based on the IQMH definition.[58] These comparisons will be coded as agreement failure. The proportion of agreement failure and 95% confidence interval adjusted for repeated measurements will be calculated.[64] This will also be done using two other reference standards for agreement as a secondary analysis, being an allowable difference of ±10 g/L often used in similar studies based on expert opinion, and an allowable difference of ±7%, based on guidelines by the Clinical Laboratory Improvement Amendments (CLIA) of 1988.[41,42,65-71]

***Transfusion thresholds agreement*:** One must also consider that the purpose of measuring Hgb is to determine whether a patient needs an RBC transfusion. The latter decision

is a binary event and thus it is useful to examine agreement between measurements based on the potential to transfuse RBC at different Hgb thresholds. For this analysis, actual transfusions will not be examined, but rather their hypothetical potential based solely on Hgb thresholds. To do so, Cohen's kappa statistics will be calculated at different thresholds, comparing the proportion of patients that would potentially be transfused at each Hgb value with each measurement method beyond chance alone.[41,72]

***Predictive modelling*:** This will be done to examine whether specific patient- and sample-level covariates predict agreement failure (IQMH definition) for measurements where lab-Hgb values are within the transfusion zone of 60-100 g/L.[58] This analysis will potentially identify a subset of patients for whom POCT-Hgb should not be used to guide transfusion decisions, rather than making conclusions based on agreement alone. Generalized linear mixed models-logistic regression models (to account for repeated measures within individual patients), restricted to samples that are laboratory-identified as being in the transfusion zone, will evaluate the predictive capacity of the following variables: sex, weight, height, elective vs emergency operation, preoperative Hgb, duration of surgery up to POCT, intravenous fluid infusion volume, blood loss, arterial vs. venous blood sample, and patient temperature at the time of blood sampling. These models will be repeated three times for each POCT test, one to predict overall unacceptable differences (absolute difference greater than 4 g/L), one to predict unacceptable low POCT measurements, and one to predict unacceptable high POCT measurements. Given the exploratory nature of this objective, no adjustment will be made for multiple comparisons.

***Transfusion appropriateness*:** Anesthesiologists will document reasons for transfusions. Intraoperative transfusion events will be adjudicated for appropriateness in terms of the decision to initiate transfusion, as well as the number of RBC units administered. Existing transfusion appropriateness instruments will be applied prospectively.[13,14,73] Adjudication will be conducted independently by two individuals trained in the use of the adjudication instruments. The correlational relationship between transfusion events appropriateness and POCT-Hgb

dichotomous agreement (only the last test prior to transfusion) will be examined using 2x2 contingency tables and Chi-squared tests. This relationship will help identify situations when the potential for mistransfusion is at a higher risk based on the POCT-Hgb value, while acknowledging the fact that transfusion decision-making is related to a variety of factors and does not rely solely on Hgb measurements.

## 4.6 METHODS: MONITORING

### Data monitoring

Given the observational nature of the study, a Data and Safety Monitoring Board will not be needed. A monitoring plan will be put in place to ensure proper conduct of the study.

### Harms

There is no potential for harm in this study, as all blood sampling is conducted for clinical purposes only.

### Auditing

The study may be monitored or audited in accordance with the current approved protocol, Good Clinical Practice (GCP), and relevant regulations and standard operating procedures.

### Patient and public involvement

No patients were involved in the development of this study protocol.

## 4.7 ETHICS AND DISSEMINATION

### Research Ethics Approval

The protocol and all other study material have been submitted and approved by the Ottawa Health Science Network – Research Ethics Board (OHSN-REB) prior to study initiation (protocol ID 20200731-01H). The principal investigator will ensure that this study is conducted in accordance with relevant regulations, GCP, and Tri-Council Policy Statement 2.

### Protocol Amendments

If any protocol amendment or modification takes place, approval from the OHSN-REB will be obtained before implementing any change. Once REB approval is obtained, the PI will brief all relevant parties on the amendments and a copy of the amended protocol will be provided.

### Consent

The current study is observational and requires no patient participation. Given its observational nature, absence of risk, and focus on quality, this method comparison study represents one of the circumstances under which waived consent is permitted.

### Confidentiality

The study staff will ensure that the participants' anonymity is maintained. Personal health information (PHI) and personal identifying information (PII) recorded for this study will be maintained in a secure, password-protected database accessible only to the investigators and the research staff. This file will be linked to other study documents, forms, and databases by way of a unique study identification number (study ID). The study list will be maintained exclusively on secure servers at The Ottawa Hospital. Any data derived from the study and released to the study statistician will be completely de-identified.

**Declaration of Interests**

The investigators have no conflicts of interest with, have not received funding from, and have no association with any of the device manufacturers described in the protocol.

**Access to Data**

All collected forms will be computerized into a dataset and delivered to the study statistician. The PI will not have access to these data until study completion.

**Ancillary and Post-Study Care**

There is no ancillary and post-study care to report.

**Dissemination Policy**

The Investigators will be involved in reviewing drafts of the manuscripts, abstracts, press releases and any other publications arising from the study. Authors will acknowledge that the study was funded by the Canadian Blood Services (CBS). Authorship will be determined in accordance with the International Committee of Medical Journal Editors (ICMJE) guidelines and other contributors will be acknowledged.

**Authors' contributions**

GM is the principal investigator for this study. GM, DM, JS, TR, and DF designed the study. AT, DM, JS, JP, CW, FMC, TR, and DF provided expertise during the study design phase and protocol writing. AW, LM, KB implemented and are managing the study. TR and RM provided statistical expertise. KB and RM will be conducting the primary statistical analysis with the help of GM, TR, and DF.

**Funding statement**

This work was supported by the Intramural Research Grant Program from Canadian Blood Services grant number IG2020-GM.

**Competing interests statement**

The authors declare no conflict of interest.

**Study status**

Protocol version 2.0, dated 30 November 2021

Date data accrual began: 18 March 2021

Estimated date of study completion: fourth quarter of 2023.

## 4.8 REFERENCES


1. García-Roa M, del Carmen Vicente-Ayuso M, Bobes AM, et al. Red blood cell storage time and transfusion: current practice, concerns and future perspectives. *Blood Transfus*. 2017;15(3):222-231. doi:10.2450/2017.0345-16

2. Shehata N, Forster A, Lawrence N, et al. Changing trends in blood transfusion: an analysis of 244,013 hospitalizations. *Transfusion*. 2014;54(10pt2):2631-2639. doi:10.1111/trf.12644

3. Whitaker BI, Rajbhandary S, Harris A. *The 2013 AABB Blood Collection, Utilization, and Patient Blood Management Survey Report*. AABB; 2015.

4. Hendrickson JE, Hillyer CD. Noninfectious Serious Hazards of Transfusion. *Anesthesia & Analgesia*. 2009;108(3):759-769. doi:10.1213/ane.0b013e3181930a6e

5. Callum J, Pinkerton P, Lima A, et al. *Bloody Easy 4: Blood Transfusions, Blood Alternatives and Transfusion Reactions*. 4th ed. ORBCoN; 2016.

6. Bennett S, Baker LK, Martel G, et al. The impact of perioperative red blood cell transfusions in patients undergoing liver resection: a systematic review. *HPB*. 2017;19(4):321-330. doi:10.1016/j.hpb.2016.12.008

7. Aquina CT, Blumberg N, Becerra AZ, et al. Association Among Blood Transfusion, Sepsis, and Decreased Long-term Survival After Colon Cancer Resection. *Annals of Surgery*. 2017;266(2):311-317. doi:10.1097/SLA.0000000000001990

8. Shander A, Javidroozi M, Ozawa S, Hare GMT. What is really dangerous: anaemia or transfusion? *BJA: British Journal of Anaesthesia*. 2011;107(suppl_1):i41-i59. doi:10.1093/bja/aer350

9. Ejaz A, Spolverato G, Kim Y, Frank SM, Pawlik TM. Identifying Variations in Blood Use Based on Hemoglobin Transfusion Trigger and Target among Hepatopancreaticobiliary Surgeons. *Journal of the American College of Surgeons*. 2014;219(2):217-228. doi:10.1016/j.jamcollsurg.2014.02.033

10. Qian F, Osler TM, Eaton MP, et al. Variation of Blood Transfusion in Patients Undergoing Major Noncardiac Surgery. *Annals of Surgery*. 2013;257(2):266-278. doi:10.1097/SLA.0b013e31825ffc37

11. Partington A, Chew DP, Ben-Tovim D, Horsfall M, Hakendorf P, Karnon J. Screening for important unwarranted variation in clinical practice: a triple-test of processes of care, costs and patient outcomes. *Australian Health Review*. 2017;41(1):104-110. doi:10.1071/AH15101

12. Harrison R, Manias E, Mears S, Heslop D, Hinchcliff R, Hay L. Addressing unwarranted clinical variation: A rapid review of current evidence. *Journal of Evaluation in Clinical Practice*. 2019;25(1):53-65. doi:10.1111/jep.12930

13. Bennett S, Tinmouth A, McIsaac DI, et al. Ottawa Criteria for Appropriate Transfusions in Hepatectomy: Using the RAND/UCLA Appropriateness Method. *Annals of Surgery*. 2018;267(4):766-774. doi:10.1097/SLA.0000000000002205

14. Spradbrow J, Cohen R, Lin Y, et al. Evaluating appropriate red blood cell transfusions: a quality audit at 10 Ontario hospitals to determine the optimal measure for assessing appropriateness. *Transfusion*. 2016;56(10):2466-2476. doi:10.1111/trf.13737

15. Ejaz A, Spolverato G, Kim Y, Frank SM, Pawlik TM. Variation in triggers and use of perioperative blood transfusion in major gastrointestinal surgery. *British Journal of Surgery*. 2014;101(11):1424-1433. doi:10.1002/bjs.9617

16. Ranganathan P, Ahmed S, Kulkarni AP, Divatia JV. Appropriateness of perioperative blood transfusion in patients undergoing cancer surgery: A prospective single-centre study. *Indian J Anaesth*. 2012;56(3):234-237. doi:10.4103/0019-5049.98763

17. Baker L, Lenet T, Vered M, et al. An audit and evaluation of appropriateness of perioperative allogenic red blood cell transfusion in liver surgery. *HPB (Oxford, England)*. 2020;22:S79-S79. doi:10.1016/j.hpb.2020.04.552

18. Lenet T, Baker L, Vered M, et al. Red blood cell transfusions in liver surgery. *British Journal of Surgery*. 2023;110(2):269-270. doi:10.1093/bjs/znac389

19. Shander A, Hofmann A, Ozawa S, Theusinger OM, Gombotz H, Spahn DR. Activity-based costs of blood transfusions in surgical patients at four hospitals. *Transfusion*. 2010;50(4):753-765. doi:10.1111/j.1537-2995.2009.02518.x

20. Stokes EA, Wordsworth S, Staves J, et al. Accurate costs of blood transfusion: a microcosting of administering blood products in the United Kingdom National Health Service. *Transfusion*. 2018;58(4):846-853. doi:10.1111/trf.14493

21. Ackland GL, Brudney CS, Cecconi M, et al. Perioperative Quality Initiative consensus statement on the physiology of arterial blood pressure control in perioperative medicine. *British Journal of Anaesthesia*. 2019;122(5):542-551. doi:10.1016/j.bja.2019.01.011

22. Salem-Schatz SR, Avorn J, Soumerai SB. Influence of Clinical Knowledge, Organizational Context, and Practice Style on Transfusion Decision Making: Implications for Practice Change Strategies. *JAMA*. 1990;264(4):476-483. doi:10.1001/jama.1990.03450040072034

23. Holst LB, Petersen MW, Haase N, Perner A, Wetterslev J. Restrictive versus liberal transfusion strategy for red blood cell transfusion: systematic review of randomised trials with

meta-analysis and trial sequential analysis. *BMJ : British Medical Journal (Online)*. 2015;350. doi:10.1136/bmj.h1354

24. Hagaman D, Pilla MA, Ehrenfeld JM. Intraoperative Transfusion Guidelines: Promoting Clinician Adherence in the Operating Room. *Anesthesia & Analgesia*. 2018;127(3):596-597. doi:10.1213/ANE.0000000000003472

25. Baker L, Park L, Gilbert R, et al. Intraoperative Red Blood Cell Transfusion Decision-making: A Systematic Review of Guidelines. *Annals of Surgery*. 2021;274(1):86-96. doi:10.1097/SLA.0000000000004710

26. Lenet T, Baker L, Park L, et al. A Systematic Review and Meta-analysis of Randomized Controlled Trials Comparing Intraoperative Red Blood Cell Transfusion Strategies. *Annals of Surgery*. 2022;275(3):456-466. doi:10.1097/SLA.0000000000004931

27. Practice Guidelines for Perioperative Blood Management: An Updated Report by the American Society of Anesthesiologists Task Force on Perioperative Blood Management*. *Anesthesiology*. 2015;122(2):241-275. doi:10.1097/ALN.0000000000000463

28. Bennett S, Ayoub A, Tran A, et al. Current practices in perioperative blood management for patients undergoing liver resection: a survey of surgeons and anesthesiologists: TRANSFUSION IN LIVER RESECTION. *Transfusion*. 2018;58(3):781-787. doi:10.1111/trf.14465

29. Jouffroy R, Baugnon T, Carli P, Orliaguet G. A survey of blood transfusion practice in French-speaking pediatric anesthesiologists. *Pediatric Anesthesia*. 2011;21(4):385-393. doi:10.1111/j.1460-9592.2011.03531.x

30. Giraud B, Frasca D, Debaene B, Mimoz O. Comparison of haemoglobin measurement methods in the operating theatre. *BJA: British Journal of Anaesthesia*. 2013;111(6):946-954. doi:10.1093/bja/aet252

31. Barker SJ, Shander A, Ramsay MA. Continuous Noninvasive Hemoglobin Monitoring: A Measured Response to a Critical Review. *Anesthesia & Analgesia*. 2016;122(2):565-572. doi:10.1213/ANE.0000000000000605

32. Sanchis-Gomar F, Cortell-Ballester J, Pareja-Galeano H, Banfi G, Lippi G. Hemoglobin Point-of-Care Testing: The HemoCue System. *J Lab Autom*. 2013;18(3):198-205. doi:10.1177/2211068212457560

33. Papadea C, Foster J, Grant S, et al. Evaluation of the i-STAT Portable Clinical Analyzer for Point-of-Care Blood Testing in the Intensive Care Units of a University Children's Hospital. *Ann Clin Lab Sci*. 2002;32(3):231-243.

34. Ng WL, Short TG, Gunn KN, Fuge GS, Slon B. Accuracy and Reliability of the i-STAT Point-of-Care Device for the Determination of Haemoglobin Concentration before and after Major Blood Loss. *Anaesth Intensive Care*. 2014;42(4):495-499. doi:10.1177/0310057X1404200411

35. Baart AM, de Kort WLAM, van den Hurk K, Pasker-de Jong PCM. Hemoglobin assessment: precision and practicability evaluated in the Netherlands—the HAPPEN study. *Transfusion*. 2016;56(8):1984-1993. doi:10.1111/trf.13546

36. Steinfelder-Visscher J, Teerenstra S, Klein Gunnewiek JMT, Weerwind PW. Evaluation of the i-STAT Point-of-Care Analyzer in Critically Ill Adult Patients. *J Extra Corpor Technol*. 2008;40(1):57-60.

37. Dolscheid-Pommerich RC, Dolscheid S, Grigutsch D, Stoffel-Wagner B, Graeff I. Comparability of Point-of-Care versus Central Laboratory Hemoglobin Determination in Emergency Patients at a Supra-Maximal Care Hospital. *PLoS One*. 2016;11(11):e0166521. doi:10.1371/journal.pone.0166521

38. Hiscock R, Simmons SW, Carstensen B, Gurrin LC. Comparison of Massimo Pronto-7 and Hemocue Hb 201+ with Laboratory Haemoglobin Estimation: A Clinical Study. *Anaesth Intensive Care*. 2014;42(5):608-613. doi:10.1177/0310057X1404200510

39. Wu P, Morey TE, Harris NS, Gravenstein N, Rice MJ. Intravenous Fluids Cause Systemic Bias in a Conductivity-Based Point-of-Care Hematocrit Meter. *Anesthesia & Analgesia*. 2012;114(2):314-321. doi:10.1213/ANE.0b013e31823fecbd

40. Johnson MF (SA), Marwick PCF (SA), Coetzee JF. Comparison of Hemoglobin Measurements by 3 Point-of-Care Devices With Standard Laboratory Values and Reliability Regarding Decisions for Blood Transfusion. *Anesthesia & Analgesia*. 2020;131(2):640-649. doi:10.1213/ANE.0000000000004533

41. Morey TE, Gravenstein N, Rice MJ. Let's Think Clinically Instead of Mathematically About Device Accuracy. *Anesthesia & Analgesia*. 2011;113(1):89-91. doi:10.1213/ANE.0b013e318219a290

42. Rice MJ, Gravenstein N, Morey TE. Noninvasive Hemoglobin Monitoring: How Accurate Is Enough? *Anesthesia & Analgesia*. 2013;117(4):902-907. doi:10.1213/ANE.0b013e31829483fb

43. Hiscock R, Kumar D, Simmons SW. Systematic review and meta-analysis of method comparison studies of Masimo pulse co-oximeters (Radical-7™ or Pronto-7™) and HemoCue® absorption spectrometers (B-Hemoglobin or 201+) with laboratory haemoglobin

estimation. *Anaesthesia and Intensive Care*. 2015;43(3):341-350. doi:10.1177/0310057x1504300310

44. Lamhaut L, Apriotesei R, Combes X, Lejay M, Carli P, Vivien B. Comparison of the Accuracy of Noninvasive Hemoglobin Monitoring by Spectrophotometry (SpHb) and HemoCue® with Automated Laboratory Hemoglobin Measurement. *Anesthesiology*. 2011;115(3):548-554. doi:10.1097/ALN.0b013e3182270c22

45. Maslow A, Bert A, Singh A, Sweeney J. Point-of-Care Hemoglobin/Hematocrit Testing: Comparison of Methodology and Technology. *Journal of Cardiothoracic and Vascular Anesthesia*. 2016;30(2):352-362. doi:10.1053/j.jvca.2015.11.010

46. Kolotiniuk NV, Manecke GR, Pinsky MR, Banks D. Measures of Blood Hemoglobin and Hematocrit During Cardiac Surgery: Comparison of Three Point-of-Care Devices. *Journal of Cardiothoracic and Vascular Anesthesia*. 2018;32(4):1638-1641. doi:10.1053/j.jvca.2017.11.022

47. Myers GJ, Browne J. Point of care hematocrit and hemoglobin in cardiac surgery: a review. *Perfusion*. 2007;22(3):179-183. doi:10.1177/0267659107080826

48. Rekman J, Workneh A, Rao P, et al. Point-of-care hemoglobin testing and liver surgery. *HPB*. 2018;20:S490. doi:10.1016/j.hpb.2018.06.2966

49. Morey TE, Gravenstein N, Rice MJ. Assessing Point-of-Care Hemoglobin Measurement: Be Careful We Don't Bias with Bias. *Anesthesia & Analgesia*. 2011;113(6):1289-1291. doi:10.1213/ANE.0b013e31822906b2

50. Broderick AJ. Point-of-care haemoglobin measurement – state of the art or a bleeding nuisance? *Anaesthesia*. 2015;70(11):1225-1229. doi:10.1111/anae.13231

51. Pecoraro V, Germagnoli L, Banfi G. Point-of-care testing: where is the evidence? A systematic survey. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2014;52(3):313-324. doi:10.1515/cclm-2013-0386

52. Hildyard C, Curry N. Point-of-care testing: a standard of care? *Anaesthesia*. 2015;70(10):1113-1118. doi:10.1111/anae.13225

53. Callum JL, Waters JH, Shaz BH, Sloan SR, Murphy MF. The AABB recommendations for the Choosing Wisely campaign of the American Board of Internal Medicine. *Transfusion*. 2014;54(9):2344-2352. doi:10.1111/trf.12802

54. Canadian Society for Transfusion Medicine. Transfusion Medicine: Ten Things Physicians and Patients Should Question by Canadian Society for Transfusion Medicine. Choosing Wisely Canada. Published 2018. Accessed July 4, 2019. https://choosingwiselycanada.org/wp-content/uploads/2019/04/Transfusion-medicine2.pdf

55. *Proceedings from the National Summit on Overuse Organized by The Joint Commission and the American Medical Association Convened Physician Consortium for Performance Improvement® (PCPI®)*.; 2012.

56. Yip PM, Venner AA, Shea J, et al. Point-of-care testing: A position statement from the Canadian Society of Clinical Chemists. *Clinical Biochemistry*. 2018;53:156-159. doi:10.1016/j.clinbiochem.2018.01.015

57. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. doi:10.1136/bmjopen-2016-012799

58. Johnston A, Bourner G, Martin T, et al. Guidance for quality control practices and precision goals for CBCs based on IQMH patterns-of-practice survey. *International Journal of Laboratory Hematology*. 2019;41(1):15-22. doi:10.1111/ijlh.12915

59. Dindo D, Demartines N, Clavien PA. Classification of Surgical Complications: A New Proposal With Evaluation in a Cohort of 6336 Patients and Results of a Survey. *Annals of Surgery*. 2004;240(2):205-213. doi:10.1097/01.sla.0000133083.54934.ae

60. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441

61. Bland JM, Altman DG. Measuring agreement in method comparison studies. 1999;8:135-160.

62. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. *Journal of Biopharmaceutical Statistics*. 2007;17(4):571-582. doi:10.1080/10543400701329422

63. Zou GY. Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. *Stat Methods Med Res*. 2013;22(6):630-642. doi:10.1177/0962280211402548

64. Shan G. Accurate confidence intervals for proportion in studies with clustered binary outcome. *Stat Methods Med Res*. 2020;29(10):3006-3018. doi:10.1177/0962280220913971

65. Herman J, Park B, Awsare B, et al. Point-of-care versus central testing of hemoglobin during large volume blood transfusion. *BMC Anesthesiol*. 2019;19(1):1-7. doi:10.1186/s12871-019-0916-2

66. Man J, Zielinski MD, Das D, Wutthisirisart P, Pasupathy KS. Improving non-invasive hemoglobin measurement accuracy using nonparametric models. *Journal of Biomedical Informatics*. 2022;126:103975. doi:10.1016/j.jbi.2021.103975

67. Marques NR, Kramer GC, Voigt RB, Salter MG, Kinsky MP. Trending, Accuracy, and Precision of Noninvasive Hemoglobin Monitoring During Human Hemorrhage and Fixed Crystalloid Bolus. *Shock*. 2015;44(Supplement 1):45-49. doi:10.1097/SHK.0000000000000310

68. Macknet MR, Allard M, Applegate RL, Rook J. The Accuracy of Noninvasive and Continuous Total Hemoglobin Measurement by Pulse CO-Oximetry in Human Subjects Undergoing Hemodilution. *Anesthesia & Analgesia*. 2010;111(6):1424-1426. doi:10.1213/ANE.0b013e3181fc74b9

69. Medicare, Medicaid and CLIA programs; regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA)--HCFA. Final rule with comment period. *Federal register*. 1992;57(40):7002-7186.

70. Center for Devices and Radiological Health. Clinical Laboratory Improvement Amendments (CLIA). *FDA*. Published online September 13, 2021. Accessed September 20, 2022. https://www.fda.gov/medical-devices/ivd-regulatory-assistance/clinical-laboratory-improvement-amendments-clia

71. Clinical Laboratory Improvement Amendments (CLIA) | CDC. Published August 12, 2022. Accessed September 20, 2022. https://www.cdc.gov/clia/index.html

72. Rigby AS. Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. *Disability and Rehabilitation*. 2000;22(8):339-344. doi:10.1080/096382800296575

73. Frank SM, Resar LMS, Rothschild JA, Dackiw EA, Savage WJ, Ness PM. A novel method of data analysis for utilization of red blood cell transfusion. *Transfusion*. 2013;53(12):3052-3059. doi:10.1111/trf.12227

### 4.9 TABLE

**Table 1**. Outcome measures and their evaluation time-points.

| **Outcome Measures** | **Time-points of evaluation** |
|---|---|
| POCT-Hgb (HemoCue, i-STAT, Rad-67) and Lab-Hgb measurements | Day 0 (intraoperatively) |
| Intraoperative red blood cell transfusion events | Day 0 (intraoperatively) |
| POCT-Hgb time-point parameters at the time of blood sampling: duration of surgery, intravenous infusion fluid volumes, estimated blood loss, type of blood sample (arterial vs. venous), patient temperature, intraoperative physiologic parameters (blood pressure, etc.), use of intraoperative blood thinners or blood loss preventative methods (e.g., tranexamic acid, Cell Saver, prothrombin complex concentrate, heparin, protamine, etc.), and use of intraoperative vasopressors (e.g., norepinephrine, phenylephrine, epinephrine, etc.) | Day 0 (time of intraoperative blood sample for POCT-Hgb measurement) |
| Serial hemoglobin values (pre-operative, post-transfusion when applicable, post-operative up to 30 days post-surgery, and at 30days ± 7days) | Preoperatively<br>Day 0-30<br>Day 30 |
| Post-operative transfusion rates | Day 30 |
| Morbidity | Day 30 |
| Mortality | Day 30 |

## 4.10 FIGURE

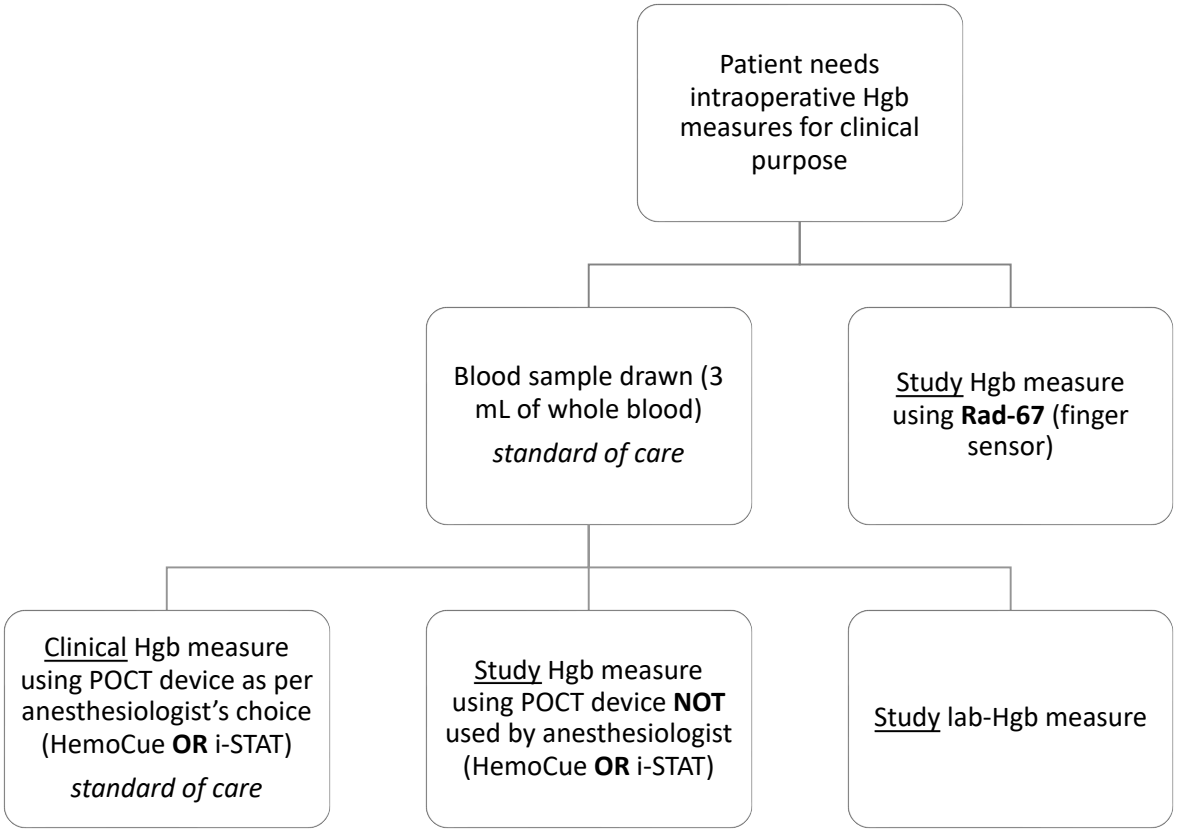



**Figure 1** Flow chart of study procedures. Hgb, Hemoglobin; POCT, Point-of-Care Testing.

# CHAPTER 5. STATISTICAL ANALYSIS PLAN FOR THE PREMISE STUDY

Karine Brousseau[1,2], Daniel McIsaac[2,3], Tim Ramsay[2], Ranjeeta Mallick[2], Dean Fergusson[2], Guillaume Martel[1,2]

[1]Liver and Pancreas Unit, Department of Surgery, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

[3]Department of Anesthesiology, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

## AUTHOR CONTRIBUTIONS

All authors assisted with the development of this statistical analysis plan and provided feedback and recommendations for all planned statistical analyses.

## PUBLICATION INFORMATION

This statistical analysis plan was submitted as an appendix to the PREMISE protocol presented in Chapter 4. The PREMISE protocol and statistical analysis plan as an appendix were accepted for publication in BMJ Open and is currently in press, as of September 26, 2023.

## 5.1 PREFACE TO CHAPTER 5

This chapter presents a comprehensive statistical analysis plan (SAP) for the PREMISE study. This SAP serves as an appendix to the PREMISE protocol (Chapter 4). Findings from the scoping review helped inform the statistical analyses set out in this SAP. In the original appendix, a table was included in section *5.4 SECTION 3 – STUDY METHODS*. However, for the purpose of this thesis, this table was removed as it was already presented in Chapter 4, in the PREMISE protocol. The SAP informed the analysis of the PREMISE study, for which the findings are presented in Chapter 5.

## 5.2 SECTION 1 – ADMINISTRATIVE INFORMATION

### Roles and Responsibilities

| Name | Role | Institution |
|---|---|---|
| Karine Brousseau | MSc candidate | OHRI – Clinical Epidemiology Program |
| Dr. Ranjeeta Mallick | Study Statistician | OHRI – Ottawa Methods Center |
| Dr. Guillaume Martel | Study Principal Investigator | OHRI – Clinical Epidemiology Program |
| Dr. Dean Fergusson | Study Co-Investigator | OHRI – Ottawa Methods Centre |
| Dr. Tim Ramsay | Study Statistician | OHRI – Ottawa Methods Centre |
| Dr. Daniel McIsaac | Study Co-Investigator | OHRI – Clinical Epidemiology Program |

### Contributions

Karine Brousseau developed the statistical analysis plan (SAP), based on the analyses set out in the study protocol. Dr. Guillaume Martel is the study Principal Investigator and helped develop this SAP. Dr. Dean Fergusson is a senior investigator and provided mentorship and advice regarding the design, conduct, and analysis of this study. Dr. Tim Ramsay and Dr. Ranjeeta Mallick are the study statisticians and assisted in the development of the SAP. Dr. Guillaume Martel, Dr. Tim Ramsay, Dr. Ranjeeta Mallick, Dr. Dean Fergusson, and Dr. Daniel McIsaac reviewed and approved the SAP.

**Abbreviations and Definitions**

| | |
|---|---|
| **AE** | **Adverse events** |
| **ASA** | **American Society of Anesthesiologists** |
| **BMI** | **Body mass index** |
| **GLMM** | **Generalized linear mixed models** |
| **Hgb** | **Hemoglobin** |
| **IQMH** | **Institute for Quality Management in Healthcare definition** |
| **IV** | **Intravenous** |
| **Lab** | **Laboratory** |
| **LoA** | **Limits of Agreement** |
| **OHRI** | **Ottawa Hospital Research Institute** |
| **POCT** | **Point-of-care testing** |
| **POD** | **Post-operative day** |
| **PREMISE** | **Point-of-care hemoglobin accuracy and transfusion outcomes in non-cardiac surgery** |
| **RA** | **Research Assistant** |
| **RBC** | **Red blood cell** |
| **SAP** | **Statistical analysis plan** |
| **TOH** | **The Ottawa Hospital** |

## 5.3 SECTION 2 – INTRODUCTION

### Background and Rationale

The PREMISE (Point-of-care hemoglobin accuracy and transfusion outcomes in non-cardiac surgery) study will provide a comprehensive and definitive examination of the most commonly used point-of-care testing (POCT) instruments capable of measuring hemoglobin (POCT-Hgb) in Canadian operating rooms. It will address the critical issue of accuracy in hemoglobin (Hgb) measurement with sufficient statistical power to generate robust findings. Previous investigators have emphasized that POCT-Hgb technologies should be examined in clinical context and not solely as ex vivo laboratory comparisons, as only 13% of POCT evaluations have done so in the past[1,2]. As such, this work will examine POCT-Hgb accuracy in clinical context by generating a sample that is sufficiently powered to study tests within the critical Hgb transfusion zone of 60-100 g/L, as defined by the practice guidelines of the American Society of Anesthesiologists[3]. POCT-Hgb devices being evaluated in terms of accuracy for the purpose of this study are HemoCue (HemoCue AB, Ängelholm, Sweden), i-STAT (Abbott Laboratories, Abbott Park, IL, USA), and Rad-67 Pulse CO-Oximeter (Masimo, Irvine, CA, USA). The first two devices rely on a small blood sample, while the third uses a non-invasive finger probe. Hemoglobin measured in the central laboratory (lab-Hgb) will be considered as the gold standard and will serve as the comparator to examine the accuracy of the POCT-Hgb.

The proposed work includes all types of non-cardiac operations, within a wide spectrum of surgical patients of all sexes, ages, and ethnicities. It will generate a critical link between POCT-Hgb measurements and RBC transfusion. It is expected to inform our understanding of appropriate and inappropriate transfusions, as well as to produce data on transfusion triggers in the operating room.

This statistical analysis plan (SAP) provides our complete and comprehensive plan for the analysis of the PREMISE study, following guidance proposed by Gamble et al[4].

**Objectives**

The overall objective of this study is to determine the accuracy of intraoperative POCT-Hgb instruments in non-cardiac surgery, as well as its association with transfusion decision-making.

**Primary objective:** To determine the accuracy of POCT-Hgb instruments in non-cardiac surgery when compared to lab-Hgb, and more specifically the accuracy of POCT-Hgb measurements when lab-Hgb values are within the Hgb transfusion zone of 60-100 g/L.

**Secondary Objectives:**

1. To examine the accuracy of POCT-Hgb measurements when compared to lab-Hgb, within the entire Hgb range.
2. To build a clinical prediction model to predict the agreement failure of POCT-Hgb and lab-Hgb measurements, including for measurements where lab-Hgb values are within the transfusion zone of 60-100 g/L.
3. To examine the association between POCT-Hgb measurements and intraoperative RBC transfusion.
4. To examine intraoperative RBC transfusion appropriateness.

## 5.4 SECTION 3 – STUDY METHODS

### 3.1 Study Design

This is a prospective observational method comparison study of POCT-Hgb technologies against the gold standard of lab-Hgb. It takes place in the operating rooms of The Ottawa Hospital (TOH) while patients are undergoing elective or urgent non-cardiac surgery. Patients are enrolled into this study when an anesthesiologist determines that it is clinically indicated to verify the patient's Hgb using a POCT-Hgb device. For each patient enrolled into the study, concurrent Hgb values measured using three POCT devices will be compared to a Hgb value measured in the main laboratory, all using the same blood sample. Hgb values may be analyzed more than once (i.e., repeated measures) during a given operation, depending on the anesthesiologist's need to assess Hgb.

Screening will take place on the day of surgery. All patients having non-cardiac surgery that meet the inclusion criteria will be considered as potentially eligible. When an anesthesiologist determines that a patient's Hgb needs to be tested during surgery, further screening will be performed by asking inclusion-specific questions to the anesthesiologist, prior to any testing being done to ensure patients are eligible for enrolment into the study. Once the patient is deemed eligible, they will be enrolled into the study and their Hgb may be tested. All consecutive participants meeting eligibility will be enrolled.

Baseline data will be collected by chart review of the patient's health history after they have been enrolled into the study. Patients will be followed for a period of 30 days after the date of their surgery to monitor for post-operative transfusions, adverse events, vital status, and laboratory values.

### 3.2 Sample Size

The sample size was calculated based on the primary outcome of agreement failure incidence, assuming an acceptable difference of ± 4 g/L, within the subset of samples that are in

the transfusion zone of 60-100 g/L. More specifically, the sample size was chosen to allow the fitting of a robust clinical prediction model for agreement failure. Fifteen predictors will be considered for the prediction model. Assuming a c-statistic of 0.75, a prevalence of 0.5, whereby at least 50% of i-STAT measurements are expected to fall outside of the acceptable ± 4 g/L margin of error, a total of 652 measurements will be needed to fit a stable logistic regression model with 15 independent variables, based on methods described by Riley et al[5]. As such, patients will be enrolled until at least 652 lab-Hgb measurements are below 100 g/L. Given that preliminary data from our institution suggest that 38% of POCT-Hgb values are below 100 g/L, it is expected that a total of approximately 1,750 measurements will be required.

### 3.3 Statistical interim analyses and stopping guidance (if applicable)

No statistical interim analyses planned.

### 3.4 Timing of final analysis

The study is due to finish approximately 2.5 years after the start of accrual, once the total sample size is achieved. Participants will be followed up to 30 days after the day of surgery for adverse events, considered as a secondary outcome. Once data collection is completed, the data will be cleaned, verified and locked for analysis. All data will be extracted from the electronic data capture system and analyzed using a statistical software.

### 3.5 Timing of outcome assessments

The primary outcome is the difference between POCT-Hgb and lab-Hgb measurements for all individual blood samples collected for this study. This data will be measured and collected simultaneously during the participant's surgery, when the anesthesiologists determines that the Hgb needs to be assessed and may be done repeatedly during the surgery.

Data pertaining to secondary outcomes of interest will be collected from patient's electronic medical records, intra- and post-operatively, up to one month after the participant's surgery (Table 1). All outcome variables will be pulled for final analysis only once data collection will be complete, one month after the last participant has been enrolled and the dataset is ready for analysis.

## 5.5 SECTION 4 – STATISTICAL PRINCIPLES

### 4.1 Confidence Intervals and *P* values

All results will be presented with 95% confidence intervals, with a threshold for *P* values of 0.05 to determine statistical significance, thus allowing for a type 1 error of up to 5%. No corrections for multiple statistical testing will be done as we have one primary outcome that is pre-specified.

### 4.2 Adherence and protocol deviations

All protocol deviations will be tracked and logged. Protocol deviations and non-adherence will be compiled and synthesized for frequency. There is a small possibility that data collected during a protocol deviation may be of different quality or biased. Therefore, sensitivity analyses will be conducted when feasible to assess the effect on our data. The main types of deviations that may occur are:

1. The Rad-67 finger probe POCT-Hgb may be measured more than five minutes after the blood draw used for other POCT-Hgb and lab-Hgb results, due to having limited access to fingers. A sensitivity analysis will be conducted to determine whether the accuracy was different than in samples for which measurement was performed within five minutes of blood draw.
2. A brief survey consisting of a few questions needs to be conducted with the anesthesiologist at each POCT-Hgb timepoints to capture intraoperative data up to that point. The anesthesiologist may not be available to answer the questions, due to fast-paced emergency clinical care. In this case, the research assistant (RA) will estimate the intraoperative data at that point by visually inspecting the operative room and looking at the patient chart. This may influence the estimate of blood loss and IV fluid infusion. A sensitivity analysis will be performed to determine whether anesthesiologist availability influenced data integrity at those specific POCT-Hgb measurement time-points.

3. The clinical blood sample may be insufficient to perform all required tests. A sensitivity analysis will be performed to determine whether this may have an effect on the accuracy of the devices to measure Hgb.
4. Finally, there may be limited access to the fingers of participants during surgery. In those cases, Rad-67 measurement may not be possible at all leading to missing data from this POCT-Hgb device. The management of missing data will be addressed in section 6.3 of this protocol.

### 4.3 Analysis populations

The primary analysis will be conducted on blood samples with a lab-Hgb measurement between 60-100 g/L. Secondary and sensitivity analyses will be conducted using the full population, at the individual blood sample level, and at the participant-level, considering repeated blood sample measurements may have been done during a participant's surgery. Finally, sensitivity analyses will also be conducted using only blood samples that were tested without any protocol deviation.

## 5.6 SECTION 5 – STUDY POPULATION

### 5.1 Eligibility

Inclusion criteria are very broad in order to include a variety of participants with different characteristics. Therefore, all individuals at least 18 years of age undergoing major non-cardiac surgery under general or neuraxial anesthesia at The Ottawa Hospital (TOH) General or Civic campuses will be considered for inclusion. Non-cardiac surgery includes general surgery, non-obstetrical gynecologic surgery, head and neck surgery, neurosurgery, orthopedic surgery, plastic surgery, thoracic surgery, urologic surgery, and vascular surgery. Patients will need to have an arterial catheter or a central venous catheter to be eligible. Samples collected outside of study hours will be excluded as there will not be any staff available for enrolment and data collection during those times.

### 5.2 Withdrawal/Follow-up

We expect no loss to follow-up in this study as participants are followed up until 30 days after their surgery. Patients are usually carefully monitored for at least three to seven days in hospital following surgery and have a prescribed clinical follow-up with the surgeon within a month of discharge. Thus, data should be available for the entirety of the follow-up period. In the unlikely situation of a patient death within 30 days of surgery, these will be considered as having completed their full follow-up. If a participant was enrolled, but it is assessed that they did not meet the inclusion criteria after enrolment (e.g., under 18 years of age or not under general or spinal anesthesia), participants and the accompanying data will be withdrawn from the study.

### 5.3 Baseline Patient Characteristics

Baseline patient characteristics include:

- Sex
- Age

- Weight (kg)
- Height (cm)
- Body mass index (BMI) ($kg/m^2$)
- Medical comorbidities to derive the Charlson Comorbidity Index[5]: history of smoking, history of alcoholism, hypertension, dyslipidemia, diabetes mellitus, liver disease, malignancy (solid tumors, lymphoma, leukemia), metastatic disease, acquired immunodeficiency syndrome (AIDS), moderate to severe chronic kidney disease, congestive heart failure, myocardial infarction, chronic obstructive pulmonary disease, peripheral vascular disease, cardiovascular accident/transient ischemic attack, dementia, hemiplegia, connective tissue disorder, peptic ulcer disease, other cardiac comorbidities.
- Pre-operative chemotherapy (within six months of surgery)
- Pre-operative radiation treatments to the surgical area
- Pre-operative laboratory studies (up to 90 days before surgery): hemoglobin (g/L), hematocrit (L/L), platelet count ($x10^9/L$), international normalized ratio (INR), creatinine (μmol/L), sodium (mmol/L), bilirubin (μmol/L), and troponin T.
- Campus where the surgery takes place
- Baseline surgical parameters: type of surgery, surgery priority (elective or emergent), surgery start and end times, American Society of Anesthesiologists (ASA) score, and type of anesthesia (general or spinal).

## 5.7 SECTION 6 – ANALYSIS

### 6.1 Outcome Definition

Primary outcome: The primary outcome is the difference in Hgb measurements between each individual POCT-Hgb device and the lab-Hgb within the transfusion zone of 60-100 g/L. This outcome will be analyzed both in a continuous and a dichotomous fashion. For the primary outcome, accuracy of POCT-Hgb devices is defined as a reference standard of $\pm$ 4 g/L from the lab-Hgb value for agreement. The absolute value of the difference between POCT-Hgb and lab-Hgb will be examined and any difference greater than 4 g/L in blood samples having a lab-Hgb <100 g/L will be considered to lie outside of the allowable test difference based on the Institute for Quality Management in Healthcare definition (IQMH)[6] and will be coded as agreement failure for dichotomous analyses.

Secondary outcomes: Dichotomous outcomes will be considered for failure rates of the devices using three different reference standards. Failure rates will be measured in the complete population and within the transfusion threshold. First, we will use the definition for agreement based on the IQMH as for the primary outcome, coding blood samples as agreement failure if the difference between POCT-Hgb and lab-Hgb values is greater than 4 g/L when the lab-Hgb is <100 g/L and the difference is greater than 5% in samples with a lab-Hgb $\geq$100 g/L[6]. Second, a difference of 10 g/L in Hgb concentration between POCT-Hgb and lab-Hgb values as a reference will be used to determine agreement, as it is a threshold often used in agreement studies concerning POCT-Hgb devices based on expert opinions[7–12]. In this case, any difference greater than 10 g/L from the lab-Hgb value will be coded as agreement failure in all blood samples. Third, a difference greater than 7% between the POCT-Hgb and lab-Hgb will be coded as agreement failure, based on the guidelines from Clinical Laboratory Improvement Amendments (CLIA) of 1988, as supported by the Center for Disease Control and Prevention (CDC) and the U.S. Food & Drug Administration (FDA)[7,13–16]. Using a percentage of difference will allow for a smaller

difference in lower values of Hgb, where POCT-Hgb device accuracy is critical for important transfusion decisions. Finally, the total difference between POCT-Hgb results and lab-Hgb results measured in a continuous fashion in the full population will also be considered as a secondary outcome.

Other secondary outcomes will be assessed at two different time-points, intra-operatively and one month post-operatively. Secondary outcomes measured intra-operatively include transfusion rates, and the amount, type (red blood cells, fresh frozen plasma, platelets, cryoprecipitate/fibrinogen, albumin (5% or 25%)), and number of units of blood transfusions. Secondary outcomes measured one month post-operatively include blood transfusion rates, serial Hgb measurements, length of stay, type and quantity of post-operative blood transfusions, re-hospitalization, re-operation, mortality, intensive care unit (ICU) admission, and new occurrence of myocardial infarction, stroke or cerebrovascular accident, deep vein thrombosis, pulmonary embolism, or acute kidney injury.

**6.2 Analysis Methods**

Data will be analyzed using SAS 9.4 (SAS Institute, Cary, NC, USA) and R (The R Foundation). Prior to any analysis being initiated, a careful exploratory data analysis will be undertaken to understand trends in the data using data visualization techniques for each variable. All analyses will be adjusted for repeated measurements of Hgb within participants when possible.
There are two units of analysis considered:

- For the primary outcome and secondary outcomes pertaining to agreement where POCT-Hgb results are compared to lab-Hgb results, the unit of analysis is the individual Hgb measurements from each blood sample done during a patient's surgery, which can be repeated for a same participant during surgery.
- For other secondary outcomes such as transfusion rates, adverse events, and post-operative outcomes, the unit of analysis is the participant.

Descriptive statistics:

- Descriptive statistics will be generated for the baseline variables (see section 5.3) to define the study population using counts and percentages for categorical variables, and means and standard deviations for continuous data, or median and interquartile range as needed for variables with skewed distributions.
- Descriptive statistics will be presented for both units of analysis. In Table 1, baseline characteristics will be presented at the patient level, for the full population. Differences between sex will be explored and a stratification by sex will be presented in Table 1 if there are statistical differences in baseline characteristics. In Table 2, baseline characteristics will be presented at the blood sample level, stratified according to transfusion thresholds, presenting data for the samples with a lab-Hgb <100 g/L and for the samples with a lab-Hgb ≥100 g/L. Further stratification will present data by agreement failure (based on the IQMH recommendations), at the blood sample level.

Repeated measurements:

- Repeated measurements will be accounted for when possible, when the analysis is at the blood sample level (i.e, samples clustered within participants).
- For the primary analysis examining the accuracy of POCT-Hgb devices, the limits of agreement will be computed using the Limits of Agreement (LoA) method that is adapted for repeated measurements.
- For predictive modelling, we will account for repeated measurements and random effects using a generalized linear mixed model framework.
- Participants may have had multiple different operations, making them eligible for this study more than once. Multiple operations for the same participant will be considered as independent, given that the surgical parameters may vary widely between operations, therefore affecting the accuracy of POCT-Hgb devices. A sensitivity analysis will be

conducted to examine whether the multiple operations per participant are truly independent.

**Primary analysis**

Limits of Agreement:

- We will use Bland & Altman's LoA method adapted for repeated measurements to analyze our data in a continuous fashion. We will use the method where the true value varies, as Hgb values have the potential to change rapidly and drastically in the operative setting[17]. The limits of agreement generated for each POCT-Hgb device are expected to contain 95% of the paired comparisons between POCT-Hgb measurement and the lab-Hgb measurement for the study population, being adults undergoing major non-cardiac surgery[17]. To evaluate the precision of the limits of agreement, 95% confidence intervals will be generated for both point estimates of the LoA[18–20].
- For the primary analysis, the LoA will be computed for Hgb measurements with a lab-Hgb value <100 g/L, within the transfusion zone.
- The limits of agreement will be computed for each POCT-Hgb device separately.
- Due to the repeated measures included in our dataset, we will adjust for two different types of variances, being within subjects (or intra-subject variance) and between subjects (or inter-subject variance) using one-way analysis of variance[17].
- If the limits of agreement are within -4 and 4 g/L, the POCT-Hgb device will be considered as accurate, having good agreement with the lab-Hgb, according to the standards of the IQMH.[6] The POCT-Hgb device that has the narrowest LoA will be considered as the most precise method to determine Hgb in the operative setting, when compared with a lab-Hgb value.
- We will also generate scatter plots of the difference between the POCT-Hgb and lab-Hgb against the mean value of the POCT-Hgb and lab-Hgb values for each blood sample. This

scatterplot will allow visualization of the limits of agreement by adding horizontal lines of the mean and the upper and lower limits of agreement, ensuring that the computed LoA fits the data well for each POCT-Hgb device.

- Furthermore, data visualization techniques have been proposed by Bland and Altman to assess whether the data respects the underlying assumptions required when using this type of statistical method[17]. First, we will create a scatter plot of the difference between the POCT-Hgb measurement and the lab-Hgb measurement on the y-axis against the mean of both methods on the x-axis. This will allow visualization of any subject by method interaction[17]. Second, we will create scatter plots of the standard deviation of repeated methods against the lab-Hgb value and this will allow verification of any relationship between the variability of the differences in the repeated measures and the magnitude of the true value[17]. Third, we will visually assess if our data respects the independence assumption by plotting the difference between the POCT-Hgb and lab-Hgb values against the order in which the measurements were recorded[17].

**Secondary analyses**

*Continuous agreement*

Limits of agreement:

- The limits of agreement, as described under the primary analysis section above, will be computed for the full population, as a secondary analysis. Further analysis will be done for each sex, to examine whether sex may have an impact on the accuracy of POCT-Hgb devices.

Folded empirical distribution plots:

- As a complement to the Bland & Altman plots, we will build folded empirical distribution function curves, also called mountain plots[21,22]. This type of plot will illustrate the median and range of the difference between each POCT-Hgb device and lab-Hgb value and will

allow comparison of the distribution of differences of all POCT-Hgb devices by superimposing them on the same graph, which will also indicate which device appears to be more precise by assessing the narrowness of the plots[21]. This plot will also allow us to determine whether there is a systematic bias in any of the POCT-Hgb devices if the median difference is not equal to zero[22].

Error grid analysis:

- We will conduct an error grid analysis for each POCT-Hgb device, consisting of three zones, based on practice guidelines. Three different error grids will be designed, one for each reference guideline as described in *section 6.1 Outcome definition*. The green zone will show acceptable agreement between methods where the difference is lower than 10% for lab-Hgb values under 100 g/L, the yellow zone indicates a significant difference between methods that may lead to inappropriate transfusions, and the red zone indicates differences between methods that may lead to major errors in transfusion decisions[9]. Furthermore, from error grid analyses, we will be able to generate percentages of measurements that are categorized within each zone with 95% confidence intervals[9].
- Three separate error grids will be created, representing all three thresholds described in section *6.1 Outcome definition*, above.

*Dichotomous agreement*

Transfusion thresholds agreement using Cohen's kappa statistic:

- We will determine transfusion thresholds agreement using Cohen's $\kappa$ statistic to evaluate the probabilities of agreement beyond chance alone at important Hgb thresholds[9,23]. To do so, we will examine agreement between measurements on the basis of the potential to transfuse RBC at different Hgb thresholds for which there is a hypothetical potential for transfusion.
- We will generate Cohen's $\kappa$ statistics with 95% confidence limits at Hgb thresholds of <100, <90, <85, <80, <75, and <70 g/L, comparing the proportion of patients that would

potentially be transfused at each Hgb value with each measurement method beyond chance alone[9,22,23].

- All Cohen's $\kappa$ statistics generated for each threshold will provide us with a value between 0 (poor agreement) and 1 (perfect agreement)[9].

Individual pairwise comparisons:

- Individual pairwise comparisons between POCT-Hgb measurements and lab-Hgb results will be dichotomized and coded as agreement failure according to three different reference standards, as defined in section *6.1 Outcome definition*. Descriptive statistics will be used to demonstrate agreement failure for each device, using counts and percentages. The proportion of agreement failure and 95% confidence intervals will be examined for each POCT-Hgb device. Proportions of agreement failure of the various POCT-Hgb devices will be compared using a chi-square test of independence to determine whether there is a significant difference between the proportions of agreement failure of POCT-Hgb devices.

Predictive modelling:

- We will build generalized linear mixed models (GLMM) to examine what clinical parameters increase the odds of agreement failure in blood samples having a lab-Hgb measurement in the transfusion zone (60-100 g/L). The GLMM framework for dichotomous variables will account for the correlation in repeated measurements within a participant and will assume a compound symmetry correlation structure.
- The outcome being examined is agreement failure, described as results from the POCT-Hgb measurements falling outside of the accepted +/- 4 g/L from the lab-Hgb value, based on the IQMH definition[6].
- In total, three predictive models will be built, one model for each POCT-Hgb device will be generated to predict overall unacceptable differences $\pm$ 4 g/L.
- Relationships between parameters will be examined in an exploratory data analysis prior to starting predictive modelling. This will help examine if there is correlation or collinearity

between parameters being considered in the models. We will estimate variance inflation factors (VIF) to evaluate the presence of collinearity between variables for each group of variables listed below[24]. Relationships between variables that are particularly important to explore for collinearity are:

  - Weight, height, and body mass index
  - Duration of surgery and estimated blood loss
  - Type of surgery and estimated blood loss
  - Type of surgery and type of anesthesia
  - Type of surgery and use of vasopressors or blood loss preventative methods
  - Systolic, diastolic and mean arterial blood pressure

- Predictors being examined are:
  - Baseline characteristics: age, sex, weight, height, body mass index, smoking history, significant alcohol consumption, medical comorbidities, previous chemotherapy, previous radiation therapy to the surgical site, pre-operative Hgb
  - Type of surgery
  - ASA score
  - Arterial or venous blood sample
  - Type of anesthesia (general or neuraxial)
  - Patient physiological parameters at the time of blood sampling (temperature, blood pressure, mean arterial pressure)
  - Clinical parameters up to blood sampling (duration of surgery, intravenous fluid infusion volume, estimated blood loss, use of vasopressors, use of anticoagulants or blood preventative methods (e.g., tranexamic acid), intra-operative blood transfusion, and ST-segment changes).
- Predictors will be primarily examined using univariate models and will be considered as important confounders if the effect on the outcome is significant at a 20% level ($p<0.2$).

- Clinically important predictors will be kept in the final models even if they are not significantly associated with the outcome ($p>0.1$) in univariate models. Clinically important predictors include estimated blood loss, volume of IV fluid infusions, duration of surgery, and body temperature at time of POCT.
- We will explore whether sex is an effect modifier in the models.
- Final models should include a maximum of 15 predictors, as the sample size was powered to fit predictive regression models of this size. In models where the total number of outcomes is less than our sample size, there will need to be at least 10 events (failure of POCT-Hgb device) per predictor kept in the model to ensure stability of the model[25].
- Discrimination of the predictive model will be assessed using the c-statistic[26].
- Calibration of the predictive model will be assessed using the calibration slope[27].
- The internal validation of the predictive models will be assessed using optimism-corrected bootstrap methods[26].
- No adjustment will be made for multiple comparisons.

Subgroup analyses

- All subgroup analyses will be exploratory. All planned analyses described in this SAP will be performed on the whole population and within pre-specified subgroups. Comparisons between subgroups will be descriptive only, no statistical analysis will be performed to assess whether there are significant differences between groups. Subgroup analyses will be performed on the following variables:
    - Lab-Hgb value at POCT timepoint: blood samples with a lab-Hgb <100 g/L versus blood samples with a lab-Hgb ≥100 g/L
    - Sex
    - Duration of surgery up to POCT, intravenous fluid infusion volume up to POCT, estimated blood loss up to POCT, patient temperature at time of POCT: subgroups

for each variable will be categorized into different thresholds (e.g., quartiles or quintiles)

- Source of blood sample: arterial versus venous

Sensitivity analyses

- Sensitivity analyses will be performed to examine the effect of protocol deviations and outliers on our results. A sensitivity analysis will be performed to examine whether multiple surgeries per patient are correlated or independent. Further sensitivity analyses will be performed to examine the effect of missing data on the predictive regression models.

Secondary and tertiary analyses not pertaining to agreement:

- This statistical analysis plan is designed to analyze the agreement between the individual POCT-Hgb devices and the lab-Hgb measurements. Therefore, secondary analyses not pertaining to agreement between POCT-Hgb and lab-Hgb, as well as tertiary analyses will be described in future statistical analysis plans. These will include data on transfusion and postoperative outcomes.

**6.3 Missing Data**

Missing data is expected with all four methods of testing Hgb. In very rare instances, there may be missing data for the lab-Hgb value due to a clotted blood sample or an insufficient blood sample. As the lab-Hgb value is the reference standard to assess the precision of POCT-Hgb devices, any Hgb measurement missing data will be excluded from analysis. Furthermore, during the data collection phase of the study, blood samples with missing lab-Hgb values will not be considered in the total sample size.

Missing data for the i-STAT or HemoCue devices should be rare, but might occur in instances of device failure (e.g., if the i-STAT cartridge is not filled properly or contains air

bubbles), unavailability of the device that was being used in another operating room, or due to an insufficient quantity of blood to run all tests. Considering the very rare instance of missing data in those devices, they will be excluded from analysis. Therefore, for the analysis of HemoCue and i-STAT, the analysis will be done as a complete case analysis[28]. Sensitivity analyses will be performed to evaluate the effect of missing data on the predictive regression models.

We expect to have more frequent missing data related to our primary outcome for Hgb values provided by the RAD-67 device. Two different types of missing data can be considered. First, there are instances where there is no access to the patient's fingers, which is necessary for this device. This may occur in emergency situations or when the patient's arms are tucked in for clinical purposes. Second, there are instances when the RAD-67 technology cannot provide a Hgb value, due to a low perfusion index in the fingers, which is considered as device failure. Imputation techniques will not be used to produce values in blood samples as this is an important study outcome. An exploratory analysis will be performed to assess the effect of excluding this data from predictive modelling and we will consider creating a dummy variable for the missing data when building the regression models for this device. For secondary analyses exploring the proportion of agreement failure of this device, missing data will be excluded, therefore providing us with the agreement properties of this device in a subset of the population. Further exploratory analyses will be performed to evaluate the cases under which data was missing for this device. Given that preliminary data indicates that the frequency of missing data of any type from the Rad-67 is approximately 20%, completely ignoring this data may lead to biased results[28].

In predictive modelling, when missing data from predictors of interest are less than 20%, we will impute the missing data using simple imputation. We will not use multiple imputation due to the complexity related to bootstrapping and multiple imputation methods. Missing data from predictive variables will be considered as missing at random. If missing data is greater than 20%, the predictor will not be considered in the predictive models.

### 6.4 Harms

As this is an observational study, there is no intervention related to this study. As a result, no patient harm can result from inclusion in this study. That said, patients will experience adverse events related to their operation and/or anesthesia in keeping with their clinical care. Therefore, there are no harms expected as a result of this study and adverse events (AE) experienced by participants are not related to study-related procedures. However, we will capture AE for observational purposes, and they will be reported as secondary outcomes, as defined in *section 6.1.*

## 5.8 REFERENCES


1. Pecoraro V, Germagnoli L, Banfi G. Point-of-care testing: Where is the evidence? A systematic survey. *Clinical Chemistry and Laboratory Medicine*. 2014;52(3):313-324. doi:10.1515/cclm-2013-0386

2. Hildyard C, Curry N. Point-of-care testing: a standard of care? *Anaesthesia*. 2015;70(10):1113-1118. doi:10.1111/anae.13225

3. American Society of Anesthesiologists. *Practice Guidelines for Perioperative Blood Management: An Updated Report by the American Society of Anesthesiologists Task Force on Perioperative Blood Management: An Updated Report by the American Society of Anesthesiologists Task Force on Perioperative Blood Manajement*. Vol 122. Lippincott Williams and Wilkins; 2015. doi:10.1097/ALN.0000000000000463

4. Gamble C, Krishan A, Stocken D, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA*. 2017;318(23):2337-2343. doi:10.1001/jama.2017.18556

5. Charlson ME, Carrozzino D, Guidi J, Patierno C. Charlson Comorbidity Index: A Critical Review of Clinimetric Properties. *Psychother Psychosom*. 2022;91(1):8-35. doi:10.1159/000521288

6. Johnston A, Bourner G, Martin T, et al. Guidance for quality control practices and precision goals for CBCs based on IQMH patterns-of-practice survey. *International Journal of Laboratory Hematology*. 2019;41(1):15-22. doi:10.1111/ijlh.12915

7. Herman J, Park B, Awsare B, et al. Point-of-care versus central testing of hemoglobin during large volume blood transfusion. *BMC Anesthesiol*. 2019;19(1):240. doi:10.1186/s12871-019-0916-2

8. Man J, Zielinski MD, Das D, Wutthisirisart P, Pasupathy KS. Improving non-invasive hemoglobin measurement accuracy using nonparametric models. *Journal of Biomedical Informatics*. 2022;126:103975. doi:10.1016/j.jbi.2021.103975

9. Morey TE, Gravenstein N, Rice MJ. Let's Think Clinically Instead of Mathematically About Device Accuracy. *Anesthesia & Analgesia*. 2011;113(1):89-91. doi:10.1213/ANE.0b013e318219a290

10. Rice MJ, Gravenstein N, Morey TE. Noninvasive Hemoglobin Monitoring: How Accurate Is Enough? *Anesthesia & Analgesia*. 2013;117(4):902-907. doi:10.1213/ANE.0b013e31829483fb

11. Marques NR, Kramer GC, Voigt RB, Salter MG, Kinsky MP. Trending, Accuracy, and Precision of Noninvasive Hemoglobin Monitoring During Human Hemorrhage and Fixed Crystalloid Bolus. *Shock*. 2015;44(Supplement 1):45-49. doi:10.1097/SHK.0000000000000310

12. Macknet MR, Allard M, Applegate RL, Rook J. The Accuracy of Noninvasive and Continuous Total Hemoglobin Measurement by Pulse CO-Oximetry in Human Subjects Undergoing Hemodilution. *Anesthesia & Analgesia*. 2010;111(6):1424-1426. doi:10.1213/ANE.0b013e3181fc74b9

13. Whitehead RD, Mei Z, Mapango C, Jefferds MED. Methods and analyzers for hemoglobin measurement in clinical laboratories and field settings. *Annals of the New York Academy of Sciences*. 2019;1450(1):147-171. doi:10.1111/nyas.14124

14. Medicare, Medicaid and CLIA programs; regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA)--HCFA. Final rule with comment period. *Federal register*. 1992;57(40):7002-7186.

15. Center for Devices and Radiological Health. Clinical Laboratory Improvement Amendments (CLIA). *FDA*. Published online September 13, 2021. Accessed September 20, 2022. https://www.fda.gov/medical-devices/ivd-regulatory-assistance/clinical-laboratory-improvement-amendments-clia

16. Clinical Laboratory Improvement Amendments (CLIA) | CDC. Published August 12, 2022. Accessed September 20, 2022. https://www.cdc.gov/clia/index.html

17. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. *Journal of Biopharmaceutical Statistics*. 2007;17(4):571-582. doi:10.1080/10543400701329422

18. Zou GY. Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. *Stat Methods Med Res*. 2013;22(6):630-642. doi:10.1177/0962280211402548

19. Olofsen E, Dahan A, Borsboom G, Drummond G. Improvements in the application and reporting of advanced Bland-Altman methods of comparison. *J Clin Monit Comput*. 2015;29(1):127-139. doi:10.1007/s10877-014-9577-3

20. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing Methods of Clinical Measurement: Reporting Standards for Bland and Altman Analysis: *Anesthesia & Analgesia*. 2000;90(3):593-602. doi:10.1097/00000539-200003000-00018

21. Monti KL. Folded Empirical Distribution Function Curves-Mountain Plots. *The American Statistician*. 1995;49(4):342-345. doi:10.2307/2684570

22. Johnson M, Marwick PC, Coetzee JF. Comparison of Hemoglobin Measurements by 3 Point-of-Care Devices With Standard Laboratory Values and Reliability Regarding Decisions for Blood Transfusion. *Anesthesia & Analgesia*. 2020;131(2):640-649. doi:10.1213/ANE.0000000000004533

23. Rigby AS. Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. *Disability and Rehabilitation*. 2000;22(8):339-344. doi:10.1080/096382800296575

24. García CB, García J, López Martín MM, Salmerón R. Collinearity: revisiting the variance inflation factor in ridge regression. *Journal of Applied Statistics*. 2015;42(3):648-661. doi:10.1080/02664763.2014.980789

25. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. 1996;49(12):1373-1379. doi:10.1016/S0895-4356(96)00236-3

26. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res*. 2017;26(2):796-808. doi:10.1177/0962280214558972

27. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

28. Naaktgeboren CA, de Groot JAH, Rutjes AWS, Bossuyt PMM, Reitsma JB, Moons KGM. Anticipating missing reference standard data when planning diagnostic accuracy studies. *BMJ : British Medical Journal (Online)*. 2016;352:i402. doi:10.1136/bmj.i402

# CHAPTER 6. INTRAOPERATIVE POINT-OF-CARE HEMOGLOBIN MEASUREMENT AND ACCURACY: A REPORT FROM THE PREMISE PROSPECTIVE OBSERVATIONAL STUDY

Karine Brousseau[1,2], Leah Monette[1,2], Daniel McIsaac[2,3], Christopher Wherrett[2,3], Ranjeeta Mallick[2], Aklile Workneh[1,2], Tim Ramsay[2], Alan Tinmouth[2,4], Julie Shaw[5,6], Justin Presseau[2], François Martin Carrier[7,8], Dean Fergusson[2,4], Guillaume Martel[1,2]*

[1]Department of Surgery, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

[3]Department of Anesthesiology & Pain Medicine, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

[4]Department of Medicine, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada

[5]Eastern Ontario Regional Laboratories Association, Ottawa, ON, Canada

[6]Department of Pathology and Laboratory Medicine, University of Ottawa, Ottawa, ON, Canada

[7]Département d'anesthésie, Centre Hospitalier de l'Université de Montréal, Université de Montréal, Montreal, QC, Canada

[8]Division de soins critiques, Département de médecine, Centre Hospitalier de l'Université de Montréal, Université de Montréal, Montreal, QC, Canada

## AUTHOR CONTRIBUTIONS

Dr. Martel is the principal investigator for this study. Drs. Martel, Fergusson, McIsaac, Wherrett, Tinmouth, Shaw, Presseau, and Carrier contributed to the design of this study. Dr. McIsaac, Mallick, Ramsay, and Carrier provided feedback and recommendations for the statistical analyses. Ms. Monette and Ms. Workneh participated in the data collection.

## PUBLICATION INFORMATION

This manuscript will be submitted to a peer-reviewed scientific journal at a later date.

## 6.1 PREFACE TO CHAPTER 6

Following the elaboration of a detailed protocol and statistical analysis plan, the PREMISE study was conducted, and the data collected in the context of this study was analyzed. Findings from the statistical analysis of the PREMISE study are presented in this chapter.

## 6.2 KEY POINTS

**Question:** Are point-of-care testing devices to measure hemoglobin concentrations (POCT-Hgb) accurate when used in major non-cardiac surgery?

**Findings:** In this observational method comparison study, limits of agreement (LOA) were generated to examine the agreement level between POCT-Hgb and laboratory-determined hemoglobin measurements. None of the examined POCT-Hgb devices had LOA within the acceptable difference of four g/L, as defined by the Institute for Quality Management in Healthcare.

**Meaning:** While POCT-Hgb devices are interesting for rapid determination of hemoglobin in dynamic environments like the operating room, the accuracy of such devices should be taken into consideration before using them as sole clinical parameters to inform transfusion decision-making.

## 6.3 ABSTRACT

**Importance:** Point-of-care testing devices to measure hemoglobin concentrations (POCT-Hgb) are frequently utilized to inform transfusion decision-making in surgery despite the conflicting evidence regarding their accuracy.

**Objective:** The aim of this study was to examine the accuracy of three commonly used POCT-Hgb devices in the operating room: the HemoCue (HemoCue AB, Ängelholm, Sweden), the i-STAT (Abbott Laboratories, Abbott Park, IL, USA), and the Rad-67 (Masimo, Irvine, CA, USA). A particular attention was dedicated to Hgb measurements within 60-100 g/L, with a higher potential for transfusions.

**Design, setting, and participants:** The PREMISE study is an observational and prospective method comparison study. A consecutive convenience sample of patients undergoing major non-cardiac surgery at a tertiary-care academic hospital and requiring POCT-Hgb measurements during pre-defined data collection hours were eligible for this study. The hemoglobin (Hgb) concentrations were measured using three POCT-Hgb devices, serving as index tests, and compared with a laboratory-determined Hgb concentration (lab-Hgb), considered as the reference standard.

**Main outcomes and measures:** The main outcome measured was the individual pairwise comparisons of Hgb between POCT-Hgb and lab-Hgb. Agreement was determined based on differences of ±4 g/L between POCT-Hgb and lab-Hgb measurements. Alternate agreement thresholds of ±10 g/L and ±7% g/L were also explored.

**Results:** A total of 1735 intraoperative blood samples collected from 1139 participants were analyzed using POCT-Hgb devices between March 2021 and July 2022. The primary analysis consisted of computing limits of agreement (LOA) adjusted for repeated measurements, with 95% confidence intervals, as recommended by Bland & Altman. Main LOA estimates were -15.2 to 9.0 g/L for i-STAT, -9.6 to 7.5 g/L for HemoCue, and -21.0 to 36.2 g/L for Rad-67. Secondary analyses

consisting of error grid analysis, folded empirical distribution function plots, and Cohen's $\kappa$ statistics were also conducted and their findings are reported.

**Conclusions and relevance:** Of the three POCT-Hgb devices analyzed, HemoCue appears to be the most accurate, with the narrowest LOA estimates falling within the alternate agreement threshold of $\pm$10 g/L. Clinicians are encouraged to use prudence when using POCT-Hgb devices in surgical patients and to complement transfusion decisions with other pertinent clinical parameters.

## 6.4 INTRODUCTION

Red blood cell (RBC) transfusions are common during surgery and can be life-saving interventions.[1–3] On the other end of the spectrum, RBC transfusions administered without an acceptable clinical indication provide no clinical benefit, carry perioperative risks, and can potentially cause harm.[4–10] Transfusion decision-making in the operative setting is thus a complex and dynamic process[11], that is further hampered by limited available evidence and practice guidelines for surgical patients.[12,13] Not surprisingly, there is widespread evidence of variability in intraoperative transfusion practice.[14–17]

Previous studies have reported a critical association between hemoglobin (Hgb) or hematocrit (Hct) measurement and intraoperative transfusion decision-making.[18–20] Practice guidelines recommend measuring Hgb/Hct as part of an approach to monitoring for anemia in surgery[12] and suggest that RBC transfusions are rarely indicated with hemoglobin levels greater than 100 g/L.[21] While Hgb concentrations were traditionally measured using a blood sample sent to a central laboratory (lab-Hgb), point-of-care testing devices measuring Hgb (POCT-Hgb) are now most commonly used in the operating room environment, owing to ease of use and rapid results delivery of a few seconds to a few minutes.

There is currently limited evidence on the accuracy of POCT-Hgb devices in the intraoperative setting, where Hgb concentrations have the potential to change drastically and rapidly.[22] Most studies examining the accuracy of POCT-Hgb devices were conducted on small sample sizes and specific surgical populations with conflicting conclusions, and paucity of data has specifically been highlighted at Hgb concentrations of 60-100 g/L where RBC transfusions are most probable.[18,23–34] Despite this knowledge gap, POCT-Hgb devices are often relied upon to inform transfusion decisions given their rapidity in providing Hgb measurements.[35]

In this context, the primary objective of this study was to fill the abovementioned knowledge gaps by determining the accuracy of three commonly-used intraoperative POCT-Hgb

instruments in non-cardiac surgery, more specifically within the range of Hgb of 60-100 g/L. The accuracy of POCT-Hgb measurements within the full Hgb range was also examined as a secondary objective.

## 6.5 METHODS AND ANALYSIS

### Study design and participants

This was a prospective observational method comparison study conducted at a tertiary-care academic hospital between March 2021 and July 2022. Briefly, all intraoperative POCT blood draws for clinical purposes were concurrently submitted for central laboratory analysis, as well as analysis using the other two POCT devices. The first POCT result was used by the requesting anesthesiologist for clinical purposes, while all other results were research/observational data and were not shared with the clinical team. The study protocol was accepted for publication and is currently in press. Ethical and institutional approval was obtained (OHSN-REB 20200731-01H). Study results are reported following STARD 2015 guidelines[36], and reporting standards for Bland-Altman agreement analyses.[37] A STARD checklist is presented in Appendix 1.

The study population consisted of consecutive patients undergoing major non-cardiac surgery during specific recruitment hours. All adult surgical patients requiring intraoperative blood analysis using a POCT-Hgb device were considered for enrolment. Patients were excluded if they were undergoing cardiac or obstetrical surgery, if they did not have general or neuraxial anesthesia, if they did not have an arterial or a central venous catheter, or if POCT-Hgb analysis was performed outside of recruitment hours (Monday-Friday, 10 AM to 7 PM). Repeated measurements during an operation were allowable.

### Test methods

***Index tests:*** Three POCT-Hgb devices were examined, including HemoCue (HemoCue AB, Ängelholm, Sweden), i-STAT (Abbott Laboratories, Abbott Park, IL, USA), and Rad-67(Masimo, Irvine, CA, USA). Approximately three mL of whole blood was drawn from the arterial or central venous catheter for use with HemoCue and i-STAT. Rad-67 measured Hgb concentrations using a non-invasive finger sensor. All index tests were conducted within five

minutes of the blood draw. The manufacturers' instructions on handling, training, storage, calibration, and quality control, as well as institutional policies were followed for each device.

***Reference standard***: The remaining whole blood sample was collected using an ethylenediaminetetraacetic acid (EDTA) vacuum collection tube and processed by trained staff from the central laboratory with a Sysmex XE-2100 analyzer (Sysmex Corporation, Kobe, Japan), using the standard hemglobincyanide method. Laboratory staff did not have access to the POCT-Hgb results. The lab-Hgb value was considered the reference standard, being the closest to the true value of Hgb per current standards.

The individual differences between POCT-Hgb and lab-Hgb measurements were used to examine agreement. The Institute for Quality Management in Healthcare (IQMH) definition was used, namely an allowable difference of ±4 g/L from lab-Hgb values under 100 g/L and of ±5% for lab-Hgb values ≥100 g/L.[38] Although the IQMH definition represents the laboratory standard in Canada, additional agreement thresholds of ±10 g/L and of ±7% were also examined in sensitivity analyses, as these have been respectively used in the literature[39-44] or recommended in guidelines by the Clinical Laboratory Improvement Amendments.[44-48]

**Statistical analysis**

Statistical analysis was performed using SAS 9.4 (SAS Institute, Cary, NC, USA) following a statistical analysis plan submitted for publication with the study protocol. The primary analysis consisted of examining continuous agreement of individual pairwise comparisons for each POCT-Hgb device using the limits of agreement (LOA) method adjusted for repeated measurements, with 95% confidence intervals (95%CI) around each LOA estimate.[49-51] Adjustment for repeated measurements was done following the method where the true value varies[49], given that Hgb measurements can vary drastically within a surgery. Therefore, when computing LOA estimates, the standard deviation of the differences between methods was adjusted to account for the within-

subject variance and the between-subject variance, using the mean square error and the mean square for subjects.[49]

Visual examination of the pairwise comparisons were done using plots to ensure that assumptions about the distribution of the data were met for the LOA analysis.[49] Assumptions checked included that the distribution of the differences between methods were normally distributed, that the agreement was similar over the range of agreement as well as over time, and that there were no relationship between the variability of the differences within each subject and the magnitude of the Hgb measurements.[49] Sensitivity analyses were conducted to examine the influence of outliers, repeated measurements, and protocol deviations on the LOA estimates. Subgroup analyses were conducted to examine the effect of sex on LOA estimates. Bland-Altman plots of the difference between POCT-Hgb and lab-Hgb measurements against the lab-Hgb measurement were generated.[49,50] While it is recommended to use the average of the two measurements for the x-axis to account for possible variability in the reference test when the true value is unknown[50,52], it was decided to use lab-Hgb on the x-axis given that it is widely accepted as the gold standard.[53] Folded empirical distribution function curves were also plotted, as a non-parametric method for complementary visualization of the distribution of the differences.[24,54]

Error grid analysis was conducted as a secondary analysis of continuous agreement.[39,55] The results of each POCT-Hgb device was plotted against the lab-Hgb measurement on a grid consisting of three different zones representing low (green), moderate (yellow), and high (red) risks for potentially clinically-relevant transfusion errors.[39] A modification to the grid suggested by Morey et al. was done to adjust the narrow band around the unity line within lab-Hgb values of 60-100 g/L to allow a difference of $\pm$4 g/L[38] instead of $\pm$10%.[39] The rates of pairs of measurements falling within each zone were computed.

Crude rates of disagreement in individual pairwise comparisons as a dichotomous outcome were generated, following the three different agreement thresholds described above.

Nine-five percent CI were generated to test the null hypothesis of equality of proportions, adjusted for repeated measurements, using the Wilson interval method with continuity correction.[56]

Cohen's kappa coefficients with 95% CI were computed to further analyze the pairwise comparisons as a dichotomous outcome. Hgb thresholds were chosen to create hypothetical transfusion thresholds and examine the probabilities of agreement between POCT-Hgb and lab-Hgb based on the potential to transfuse at various Hgb thresholds beyond chance alone.[39,57,58] The strength of agreement was determined using the ranges of agreement defined by Landis and Koch.[57,58]

Indeterminate index tests or reference tests were considered as missing data. Missing lab-Hgb measurements (i.e., reference standard) were completely excluded from analysis. Indeterminate POCT-Hgb measurements (i.e., index tests) were removed from the analysis for that specific POCT-Hgb device, whereas other POCT-Hgb measurements for that same blood sample were included for analysis.

The sample size was 1750 individual Hgb measurements which was expected to yield at least 652 Hgb measurements with a lab-Hgb <100 g/L, as preliminary data revealed that approximately 38% of POCT-Hgb blood samples in the operative setting had a lab-Hgb measurement <100 g/L (unpublished data). The sample size of 652 Hgb measurements was determined to be needed to fit a stable predictive logistic regression model with up to 15 independent predictors, assuming a c-statistic of 0.75, a prevalence of 0.5 (i.e. approximately 50% of POCT-Hgb obtained with i-STAT would fall outside of the acceptable difference of $\pm$4 g/L[38]), based on methods by Riley et al.[59] The predictive model will be presented as a standalone publication.

## 6.6 RESULTS

### Participants

A total of 1135 participants undergoing surgery met inclusion criteria and were included in the study, with 1735 individual Hgb measurements collected. Enrolment was thus considered complete as 680 individual Hgb measurements were collected with a lab-Hgb value <100 g/L. A flow diagram of participants and individual Hgb measurements can be found in Appendix 2 (Figure 3). A total of 732 participants had a single measurement done, whereas 994 participants had multiple measurements done during surgery, ranging from two to nine measurements each. Nine Hgb (0.5%) measurements were excluded from analysis due to missing lab-Hgb results, providing a final sample of 1726 individual Hgb measurements for analysis ranging from 49 to 167 g/L. After accounting for missing data, 1711 (99.1%), 1725 (99.9%), and 1312 (76.0%) POCT-Hgb measurements from i-STAT, HemoCue, and Rad-67, respectively, were included for analysis. Baseline patient characteristics and intraoperative characteristics are shown in Table 1, with descriptive statistics computed at the individual participant level (n=1135). No adverse event resulted from performing the index tests or reference standard.

### Test results

The LOA and 95% CI obtained for each POCT-Hgb device are presented in Table 2 for the full population and for those at a higher risk of blood transfusions (lab-Hgb <100 g/L). Visual representations of the LOA results using Bland-Altman plots are presented in Figure 1. Mean differences were slightly negative for HemoCue and i-STAT, those POCT-Hgb devices yielding slightly lower Hgb measurements than the lab-Hgb on average, whereas Rad-67 provided higher Hgb measurements than the lab-Hgb, on average, with a slightly higher mean difference within the population of lab-Hgb<100 g/L.

For all three POCT devices, none of the LOA estimates were within the allowable difference of ±4 g/L. When considering other agreement thresholds, HemoCue demonstrated LOA

estimates that were within ±10 g/L from lab-Hgb for both the full populations and those with lab-Hgb <100g/L. HemoCue has the narrowest LOA estimates and Rad-67 has the widest LOA estimates of the three POCT-Hgb devices, within both populations.

For Rad-67, the width of LOA estimates was similar for both study populations. However, in those lab-Hgb measurements <100 g/L, LOA estimates were slightly more positive, given the associated increased positive bias in Hgb measurements. None of the LOA estimates for i-STAT and Rad-67 fell within the pre-determined agreement thresholds. Sensitivity analyses examining the influence of repeated measurements and protocol deviations did not change the conclusions of agreement obtained for each device, within both populations (data not shown). Results of the sensitivity analysis examining outliers are presented in Appendix 2 (Table 4).

Folded empirical distribution function curves are presented in Appendix 2 (Figure 4), providing a non-parametric distribution of 95% of the differences for each device. The findings from this figure are similar to those drawn from the LOA analysis.

Unadjusted disagreement rates between each POCT-Hgb device and lab-Hgb, based on the three pre-determined agreement thresholds, are presented for the full population and for the population at highest risk of blood transfusions (lab-Hgb <100 g/L) in Figure 3. HemoCue had the lowest unadjusted disagreement rates when compared to i-STAT and Rad-67, for all populations, and for all agreement thresholds. Disagreement rates were similar for i-STAT based on the agreement threshold of ±10 g/L and for HemoCue using the IQMH definition, within all populations. Disagreement rates for Rad-67 were all higher than 50%, using the IQMH definition and an agreement threshold of ±7%, whereby more than half of Rad-67 measurements disagreed with the lab-Hgb results for all populations. Using an agreement threshold of ±10%, the null hypothesis of equality of proportions could not be rejected within the full population for Rad-67.

Error grid analyses were done for each POCT-Hgb device. The proportion of observations within each risk zone is reported in Figure 2, for the full population and for lab-Hgb values <100

g/L. Zones B1 represent the potential for under-transfusion, whereby the POCT-Hgb device provides unacceptably high Hgb values when compared with the lab-Hgb measurement, and zones B2 represents the potential for over-transfusion. HemoCue had the highest proportion of observations within the green zone (Zone A), and Rad-67 had the lowest proportion of observations in that zone. Of note, HemoCue was the only POCT-Hgb device with an observation within the red zone, the zone with the highest risk of clinically-significant transfusion error. There was a higher potential for under-transfusion than over-transfusion with Hemocue and i-STAT, whereas the potential for under-transfusion was higher with Rad-67.

Finally, Cohen's kappa statistics were computed and are presented in Appendix 2 (Table 5). These provide an estimate of the proportion of agreement between POCT-Hgb and lab-Hgb measurements, assuming different triggers to initiate RBC transfusion, after accounting for the probability of agreement by chance alone. HemoCue showed the best agreement beyond chance alone, with kappa statistics that can be classified as substantial to almost perfect agreement. i-STAT had kappa statistics slightly lower than HemoCue, classified as moderate to substantial agreement, whereas Rad-67 had the lowest kappa statistics, with results classified as fair to moderate agreement.

## 6.7 DISCUSSION

The PREMISE prospective observational method comparison study has successfully examined 1726 blood samples collected in the operating room over 17 months, encompassing a broad range of elective and urgent operations. The principal analyses presented in this work have demonstrated that none of the three POCT-Hgb devices under study can be considered interchangeable with central laboratory Hgb measurements, based on the IQMH definition.[38] Of the three POCT-Hgb devices under investigation, HemoCue was most accurate, whereby 95% of its measurements were within ±10 g/L from the lab-Hgb measurement, this being an alternate agreement threshold used in the literature. On the other hand, LOA estimates for i-STAT and Rad-67 did not land within any of the three pre-specified agreement thresholds.

Although it is advisable to keep outlier data points within LOA analyses, it is also recommended to examine their effect on LOA estimates.[2] Despite the POCT-Hgb result being very different from lab-Hgb measurement among outliers, those POCT-Hgb measurements were still within plausible ranges of Hgb. Thus, such Hgb measurements could have been used by clinicians for potential transfusion decision-making, as clinicians are ultimately not aware of the lab-Hgb when using POCT devices in the operating room. Despite the noted differences in LOA estimates when removing outliers, the conclusion about interchangeability of devices is not affected by this observation. Therefore, it was decided to maintain those observations within the analysis.

The current work is consistent with previous findings in the literature, such that HemoCue was found to be generally more accurate than pulse co-oximetry devices.[60,61] However, the current study provided more accurate point estimates of the LOA given its large sample size of 1135 individual participants and 1726 pairwise comparisons, compared to previous studies with relatively small sample sizes and rarely including repeated measurements.[60,61] Moreover, this study focused primarily on the range of Hgb with a higher risk for transfusions (Hgb <100 g/L), as opposed to previous studies, which aligns with the important knowledge gap and specifically

called for large adequately powered studies to address this question, as highlighted by previous authors.[24,39,40]

The Rad-67 device was chosen as a pulse co-oximetry device for examination of accuracy for its ability to provide spot-check estimates of Hgb concentrations to be comparable with the HemoCue and i-STAT. To our knowledge, this study is the first to examine the accuracy of the Rad-67 within the operating room, with previous studies on this device was on postpartum women[62] and in a preoperative context.[63] In the intraoperative setting, previous authors have instead examined the accuracy of the Radical-7 (Masimo, Irvine, CA), a non-invasive pulse co-oximetry POCT-Hgb device that provides a continuous reading of Hgb, with conflicting conclusions, and of the Pronto-7 (Masimo, Irvine, CA), an alternate pulse co-oximetry device that provides intermittent readings of Hgb.[61] Some of the studies also evaluated the trends in Hgb readings and has shown that the device performs well in monitoring and detecting trends in Hgb, which has been argued to be as important as accurate Hgb measurements.[64,65]

One of the primary strengths of this study is its large sample size, which provides more precise point estimates, as evidenced by the narrow 95% CI associated with these analyses.[66] Perhaps most importantly, this study is unique as it addresses the knowledge gap[24,39,40] pertaining to POCT-Hgb accuracy within Hgb zone of 60-100 g/L where RBC transfusion is most probable. Second, the inclusion of robust statistical methods to account for repeated measures when multiple POCT-Hgb measurements were done within an operation is an important element of this work.[49,50,67] Third, multiple statistical methods were used to examine the outcomes, all of which were complementary and supported the identification of patterns in Hgb measurements differences. Finally, this study was designed to be pragmatic, inclusive, and non-interventional so as to have no influence on concurrent clinical care. This strategy made it possible to enroll most eligible participants, thus increasing generalizability.

This study has several limitations. First, data points are missing in about one quarter of Rad-67 measurements, which is unlikely to be random. Indeed, these were missing because no

finger was available during surgery, or because the device failed to register a Hgb measurement due to low peripheral perfusion. Given that Hgb measurement was the main study outcome, statistical techniques for data imputation were not used. Second, study hours were designed to capture as most POCT-Hgb utilizations, but many measurements were done outside of study hours. Although it is possible that certain types of patients may be more likely to have surgery after hours or on weekends (e.g., penetrating trauma patients), this is unlikely to have significantly affected results as the study captured a large proportion of various emergency cases. Third, only three types of POCT devices were studied. These were chosen as they represent the commonest devices in the North American setting and due to their availability and/affordability within the study budget. Table-top blood gas analyzers are also frequently used in operating rooms, with little published accuracy data.[60] This is unfortunately a limitation of this work. Finally, although the lab-Hgb results obtained from a central laboratory are generally considered a gold standard by clinicians, it is noteworthy that these are not perfect measures of Hgb as some measurement error is acceptable under IQMH standards.[38] This error increases the potential for measurement bias, a type of information bias.[68]

In conclusion, this study has shown that a wide range of Hgb measurements can be obtained from POCT-Hgb devices, sometimes being considerably different than lab-Hgb. Therefore, none of the three studied POCT-Hgb devices can be considered interchangeable with lab-Hgb under IQMH standards.[38] Under conditions of higher risks of transfusion (Hgb <100 g/L), pulse co-oxymetry was significantly less accurate when compared with other POCT-Hgb devices.

## 6.8 OTHER INFORMATION

### Registration number

This study was not registered in a public registry, due to its observational nature. This study was approved by the Ottawa Health Science Network – Research Ethics Board (OHSN-REB 20200731-01H).

### Reference to study protocol

The study protocol was accepted for publication in *BMJ Open* and is currently in press.


### Funding

This study was funded by the Intramural Research Grant Program from Canadian Blood Services, grant number IG2020-GM.

### Acknowledgements

The authors would like to thank the Departments of Anesthesia and Pain Medicine, the Department of Surgery, the operating room nurses and staff, and the staff from the Eastern Ontario Regional Laboratory Association (EORLA) laboratories at The Ottawa Hospital for their ongoing support during the data collection phase. The principal investigator, G. Martel, takes responsibility for the integrity of the data and the accuracy of analyses. The authors declare no conflicts of interest.

## 6.9 REFERENCES


1. Dhindra N. Estimate blood requirements – search for a global standard. WHO. Accessed September 8, 2020. https://www.who.int/bloodsafety/transfusion_services/estimation_presentations.pdf?ua=1

2. Shehata N, Forster A, Lawrence N, et al. Changing trends in blood transfusion: an analysis of 244,013 hospitalizations. *Transfusion*. 2014;54:2631-2639. doi:10.1111/trf.12644

3. Whitaker BI, Rajbhandary S, Harris A. *The 2013 AABB Blood Collection, Utilization, and Patient Blood Management Survey Report*. AABB; 2015.

4. Hendrickson JE, Hillyer CD. Noninfectious Serious Hazards of Transfusion. *Anesth Analg*. 2009;108:759-769. doi:10.1213/ane.0b013e3181930a6e

5. Madjdpour C, Spahn DR. Allogeneic red blood cell transfusions: efficacy, risks, alternatives and indications. *Br J Anaesth*. 2005;95:33-42. doi:10.1093/bja/aeh290

6. Callum J, Pinkerton P, Lima A, et al. *Bloody Easy 4: Blood Transfusions, Blood Alternatives and Transfusion Reactions*. 4th ed. ORBCoN; 2016.

7. Vamvakas EC, Blajchman MA. Transfusion-related immunomodulation (TRIM): An update. *Blood Rev*. 2007;21:327-348. doi:10.1016/j.blre.2007.07.003

8. Bennett S, Baker LK, Martel G, et al. The impact of perioperative red blood cell transfusions in patients undergoing liver resection: a systematic review. *HPB*. 2017;19(4):321-330. doi:10.1016/j.hpb.2016.12.008

9. Rohde JM, Dimcheff DE, Blumberg N, et al. Health Care–Associated Infection After Red Blood Cell Transfusion: A Systematic Review and Meta-analysis. *JAMA*. 2014;311(13):1317-1326. doi:10.1001/jama.2014.2726

10. Aquina CT, Blumberg N, Becerra AZ, et al. Association Among Blood Transfusion, Sepsis, and Decreased Long-term Survival After Colon Cancer Resection. *Ann Surg*. 2017;266(2):311-317. doi:10.1097/SLA.0000000000001990

11. Ackland GL, Brudney CS, Cecconi M, et al. Perioperative Quality Initiative consensus statement on the physiology of arterial blood pressure control in perioperative medicine. *Br J Anaesth*. 2019;122(5):542-551. doi:10.1016/j.bja.2019.01.011

12. Baker L, Park L, Gilbert R, et al. Intraoperative Red Blood Cell Transfusion Decision-making: A Systematic Review of Guidelines. *Ann Surg*. 2021;274(1):86-96. doi:10.1097/SLA.0000000000004710

13. Hagaman D, Pilla MA, Ehrenfeld JM. Intraoperative Transfusion Guidelines: Promoting Clinician Adherence in the Operating Room. *Anesth Analg*. 2018;127(3):596-597. doi:10.1213/ANE.0000000000003472

14. Hutton B, Fergusson D, Tinmouth A, McIntyre L, Kmetic A, Hébert PC. Transfusion rates vary significantly amongst Canadian medical centres. *Can J Anesth*. 2005;52(6):581-590. doi:10.1007/BF03015766

15. Ejaz A, Spolverato G, Kim Y, Frank SM, Pawlik TM. Identifying Variations in Blood Use Based on Hemoglobin Transfusion Trigger and Target among Hepatopancreaticobiliary Surgeons. *J Am Coll Surg*. 2014;219:217-228. doi:10.1016/j.jamcollsurg.2014.02.033

16. Snyder-Ramos SA, Möhnle P, Weng YS, et al. The ongoing variability in blood transfusion practices in cardiac surgery. *Transfusion*. 2008;48:1284-1299. doi:10.1111/j.1537-2995.2008.01666.x

17. Qian F, Osler TM, Eaton MP, et al. Variation of Blood Transfusion in Patients Undergoing Major Noncardiac Surgery. *Ann Surg*. 2013;257(2):266-278. doi:10.1097/SLA.0b013e31825ffc37

18. Giraud B, Frasca D, Debaene B, Mimoz O. Comparison of haemoglobin measurement methods in the operating theatre. *Br J Anaesth*. 2013;111(6):946-954. doi:10.1093/bja/aet252

19. Turgeon AF, Fergusson DA, Doucette S, et al. Red blood cell transfusion practices amongst Canadian anesthesiologists: a survey. *Can J Anesth*. 2006;53(4):344-352. doi:10.1007/BF03022497

20. Barker SJ, Shander A, Ramsay MA. Continuous Noninvasive Hemoglobin Monitoring: A Measured Response to a Critical Review. *Anesth Analg*. 2016;122(2):565-572. doi:10.1213/ANE.0000000000000605

21. Practice Guidelines for Perioperative Blood Management: An Updated Report by the American Society of Anesthesiologists Task Force on Perioperative Blood Management. *Anesthesiology*. 2015;122(2):241-275. doi:10.1097/ALN.0000000000000463

22. Wu PBS, Morey TE, Harris NS, Gravenstein N, Rice MJ. Intravenous Fluids Cause Systemic Bias in a Conductivity-Based Point-of-Care Hematocrit Meter. *Anesth Analg*. 2012;114(2):314-321. doi:10.1213/ANE.0b013e31823fecbd

23. Frasca D, Mounios H, Giraud B, Boisson M, Debaene B, Mimoz O. Continuous monitoring of haemoglobin concentration after in-vivo adjustment in patients undergoing surgery with blood loss. *Anaesthesia*. 2015;70:803-809. doi:10.1111/anae.13028

24. Johnson M, Marwick PC, Coetzee JF. Comparison of Hemoglobin Measurements by 3 Point-of-Care Devices With Standard Laboratory Values and Reliability Regarding Decisions for

Blood Transfusion. *Anesth Analg*. 2020;131(2):640-649. doi:10.1213/ANE.0000000000004533

25. Kolotiniuk NV, Manecke GR, Pinsky MR, Banks D. Measures of Blood Hemoglobin and Hematocrit During Cardiac Surgery: Comparison of Three Point-of-Care Devices. *J Cardiothorac Vasc Anesth*. 2018;32(4):1638-1641. doi:10.1053/j.jvca.2017.11.022

26. Lardi AM, Hirst C, Mortimer AJ, McCollum CN. Evaluation of the HemoCue ® for measuring intra-operative haemoglobin concentrations: a comparison with the Coulter Max-M ®. *Anaesthesia*. 1998;53:349-352. doi:10.1046/j.1365-2044.1998.00328.x

27. Mariño Z, Carrión JA, Bedini JL, et al. Evaluation of a portable hemoglobinometer (HemoCue) to control anemia in hepatitis C liver transplant recipients undergoing antiviral therapy. *Eur J Gastroenterol Hepatol*. 2011;23:942-947. doi:10.1097/MEG.0b013e328348f9c2

28. Miller RD, Ward TA, Shiboski SC, Cohen NH. A Comparison of Three Methods of Hemoglobin Monitoring in Patients Undergoing Spine Surgery. *Anesth Analg*. 2011;112:858-863. doi:10.1213/ANE.0b013e31820eecd1

29. Richards NA, Boyce H, Yentis SM. Estimation of blood haemoglobin concentration using the HemoCue® during caesarean section: the effect of sampling site. *Int J Obstet Anesth*. 2010;19:67-70. doi:10.1016/j.ijoa.2009.05.010

30. Rippmann CE, Nett PC, Popovic D, Seifert B, Pasch T, Spahn DR. Hemocue®, an accurate bedside method of hemoglobin measurement? *J Clin Monit.* 1997;13:373-377. doi: 10.1023/A:1007451611748.

31. Skelton VA, Wijayasinghe N, Sharafudeen S, Sange A, Parry NS, Junghans C. Evaluation of point-of-care haemoglobin measuring devices: a comparison of Radical-7™ pulse co-

oximetry, HemoCue® and laboratory haemoglobin measurements in obstetric patients*. *Anaesthesia*. 2013;68:40-45. doi:10.1111/anae.12039

32. Maslow A, Bert A, Singh A, Sweeney J. Point-of-Care Hemoglobin/Hematocrit Testing: Comparison of Methodology and Technology. *J Cardiothorac Vasc Anesth.* 2016;30(2):352-362. doi:10.1053/j.jvca.2015.11.010

33. Ng WL, Short TG, Gunn KN, Fuge GS, Slon B. Accuracy and Reliability of the i-STAT Point-of-Care Device for the Determination of Haemoglobin Concentration before and after Major Blood Loss. *Anaesth Intensive Care*. 2014;42:495-499. doi:10.1177/0310057X1404200411

34. Riess ML, Pagel PS. Noninvasively Measured Hemoglobin Concentration Reflects Arterial Hemoglobin Concentration Before but Not After Cardiopulmonary Bypass in Patients Undergoing Coronary Artery or Valve Surgery. *J Cardiothorac Vasc Anesth*. 2016;30(5):1167-1171. doi:10.1053/j.jvca.2016.03.148

35. Rekman J, Workneh A, Rao P, et al. Point-of-care hemoglobin testing and liver surgery. *HPB*. 2018;20:S490. doi:10.1016/j.hpb.2018.06.2966

36. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6:e012799. doi:10.1136/bmjopen-2016-012799

37. Gerke O. Reporting Standards for a Bland–Altman Agreement Analysis: A Review of Methodological Reviews. *Diagnostics*. 2020;10(5):334. doi:10.3390/diagnostics10050334

38. Johnston A, Bourner G, Martin T, et al. Guidance for quality control practices and precision goals for CBCs based on IQMH patterns-of-practice survey. *Int J Lab Hem*. 2019;41:15-22. doi:10.1111/ijlh.12915

39. Morey TE, Gravenstein N, Rice MJ. Let's Think Clinically Instead of Mathematically About Device Accuracy. *Anesth Analg*. 2011;113(1):89-91. doi:10.1213/ANE.0b013e318219a290

40. Rice MJ, Gravenstein N, Morey TE. Noninvasive Hemoglobin Monitoring: How Accurate Is Enough? *Anesth Analg*. 2013;117(4):902-907. doi:10.1213/ANE.0b013e31829483fb

41. Man J, Zielinski MD, Das D, Wutthisirisart P, Pasupathy KS. Improving non-invasive hemoglobin measurement accuracy using nonparametric models. *J Biomed Inform* 2022;126:103975. doi:10.1016/j.jbi.2021.103975

42. Marques NR, Kramer GC, Voigt RB, Salter MG, Kinsky MP. Trending, Accuracy, and Precision of Noninvasive Hemoglobin Monitoring During Human Hemorrhage and Fixed Crystalloid Bolus. *Shock*. 2015;44(Supplement 1):45-49. doi:10.1097/SHK.0000000000000310

43. Macknet MR, Allard M, Applegate RL, Rook J. The Accuracy of Noninvasive and Continuous Total Hemoglobin Measurement by Pulse CO-Oximetry in Human Subjects Undergoing Hemodilution. *Anesth Analg*. 2010;111(6):1424-1426. doi:10.1213/ANE.0b013e3181fc74b9

44. Herman J, Park B, Awsare B, et al. Point-of-care versus central testing of hemoglobin during large volume blood transfusion. *BMC Anesthesiol*. 2019;19:240. doi:10.1186/s12871-019-0916-2

45. Whitehead RD, Mei Z, Mapango C, Jefferds MED. Methods and analyzers for hemoglobin measurement in clinical laboratories and field settings: Methods for hemoglobin measurement. *Ann NY Acad Sci*. 2019;1450:147-171. doi:10.1111/nyas.14124

46. Center for Devices and Radiological Health. Clinical Laboratory Improvement Amendments (CLIA). *FDA*. Published online September 13, 2021. Accessed September 20, 2022.

https://www.fda.gov/medical-devices/ivd-regulatory-assistance/clinical-laboratory-improvement-amendments-clia

47. Medicare, Medicaid and CLIA programs; regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA)--HCFA. Final rule with comment period. *Federal register*. 1992;57(40):7002-7186.

48. Clinical Laboratory Improvement Amendments (CLIA) | CDC. Published August 12, 2022. Accessed September 20, 2022. https://www.cdc.gov/clia/index.html

49. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. *J Biopharm Stat*. 2007;17:571-582. doi:10.1080/10543400701329422

50. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.

51. Zou GY. Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. *Stat Methods Med Res*. 2013;22(6):630-642. doi:10.1177/0962280211402548

52. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*. 1995;346:1085-1087. doi:10.1016/S0140-6736(95)91748-9

53. Krouwer JS. Why Bland–Altman plots should use X, not (Y+X)/2 when X is a reference method. *Statist Med*. 2008;27:778-780. doi:10.1002/sim.3086

54. Monti KL. Folded Empirical Distribution Function Curves-Mountain Plots. *Am Stat*. 1995;49(4):342-345. doi:10.2307/2684570

55. Clarke WL. The Original Clarke Error Grid Analysis (EGA). *Diabetes Technol Ther*. 2005;7(5):776-779. doi:10.1089/dia.2005.7.776

56. Shan G. Accurate confidence intervals for proportion in studies with clustered binary outcome. *Stat Methods Med Res*. 2020;29(10):3006-3018. doi:10.1177/0962280220913971

57. Rigby AS. Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. *Disabil Rehabil*. 2000;22(8):339-344. doi:10.1080/096382800296575

58. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33:159-174. doi:10.2307/2529310

59. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441

60. Ahn H (San), Lenet T, Gilbert R, et al. Accuracy of point-of-care testing devices for hemoglobin in the operating room: a systematic review and metaanalysis. *Can J Surg*. 2022;65:S45.

61. Hiscock R, Kumar D, Simmons SW. Systematic review and meta-analysis of method comparison studies of Masimo pulse co-oximeters (Radical-7™ or Pronto-7™) and HemoCue® absorption spectrometers (B-Hemoglobin or 201+) with laboratory haemoglobin estimation. *Anaesth Intensive Care*. 2015;43(3):341-350. doi:10.1177/0310057x1504300310

62. Mills K, Vermeer JM, Berry WE, Karreman E, Lett CD. Determining the validity of non-invasive spot-check hemoglobin co-oximetry testing to detect anemia in postpartum women at a tertiary care centre, a prospective cohort study. *BMC Pregnancy Childbirth*. 2023;23:479. doi:10.1186/s12884-023-05783-3

63. Ke YH, Hwang KY, Thin TN, Sim YE, Abdullah HR. The usefulness of non-invasive co-oximetry haemoglobin measurement for screening pre-operative anaemia. *Anaesthesia*. 2021;76:54-60. doi:10.1111/anae.15171

64. Colquhoun DA, Forkin KT, Durieux ME, Thiele RH. Ability of the Masimo pulse CO-Oximeter to detect changes in hemoglobin. *J Clin Monit Comput*. 2012;26:69-73. doi:10.1007/s10877-012-9335-3

65. Gamal M, Abdelhamid B, Zakaria D, et al. Evaluation of Noninvasive Hemoglobin Monitoring in Trauma Patients with Low Hemoglobin Levels. *Shock*. 2018;49(2):150-153. doi:10.1097/SHK.0000000000000949

66. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol*. 2003;22:85-93. doi:10.1002/uog.122

67. Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*. 1986;327:307-310.

68. Porta M. Measurement bias. In: *A Dictionary of Epidemiology*. 6th ed. Oxford University Press; 2016.

## 6.10 TABLES

**Table 1.** Baseline patient characteristics.

| | **Count (%) / Mean (SD) (n=1135)** |
|---|---|
| **Baseline patient characteristics** | |
| Age (years) | 64 (15)[a] |
| Female Sex | 489 (43.1) |
| Body mass index (kg/m$^2$) (n=1108) | 28.3 (7.2)[a] |
| Preoperative hemoglobin (g/L) (n=1111) | 120.7 (23.5)[a] |
| Charlson Comorbidity Index (CCI) | 2 (1-4)[b] |
| Localized solid tumor | 467 (41.2) |
| Peripheral vascular disease | 238 (21.0) |
| Myocardial infarction | 231 (20.4) |
| Uncomplicated diabetes mellitus | 228 (20.1) |
| CCI score of 0 | 155 (13.7) |
| Preoperative chemotherapy | 156 (13.8) |
| Preoperative radiation treatment | 95 (8.4) |
| Type of surgery | |
| General surgery | 229 (20.2) |
| Vascular surgery | 228 (20.1) |
| Orthopedic surgery | 226 (19.9) |
| Neurosurgery | 198 (17.4) |
| Other[c] | 254 (22.4) |
| Priority of surgery | |
| Elective | 678 (59.7) |
| Urgent (i.e., non-elective) | 457 (40.3) |
| **Intraoperative characteristics** | |
| Anesthesia type | |
| General anesthesia | 891 (78.5) |
| Neuraxial anesthesia | 27 (2.4) |
| General and neuraxial anesthesia | 212 (18.7) |
| Estimated blood loss (mL) (n=1086) | 550 (300-1000)[b] |
| Crystalloid infusion (mL) (n=1091) | 2687.5 (1900-3545)[b] |
| Intraoperative RBC transfusion | 346 (30.5) |
| Cell salvage | 75 (6.6) |
| Medication use | |
| Intravenous heparin | 249 (21.9) |
| Prophylactic (subcutaneous) heparin | 199 (17.6) |
| Tranexamic acid | 612 (53.9)[d] |
| Protamine | 203 (17.9) |
| Vasopressor use | 1069 (94.2) |
| Phenylephrine | 913 (80.4) |
| Ephedrine | 565 (49.8) |
| Norepinephrine | 324 (28.6) |
| Other[e] | 96 (8.5) |

Abbreviations: ASA, American Society of Anesthesiologists; BMI, Body Mass Index; NA, Not applicable; SD, Standard deviation.
Data for each variable are complete unless otherwise indicated.
[a] Mean and standard deviation presented for continuous measurements.
[b] Median and inter-quartile range presented for continuous measurements.
[c] Other types of surgery include urology (9.4%), thoracic surgery (6.2%), ears, nose, and throat (ENT) surgery (3.0%), plastic surgery (2.0%), non-obstetrical gynecology (1.8%).
[d] Of those having received TXA, 69 (11.3%) participants were co-enrolled in a blinded RCT with a 1:1 allocation of TXA versus placebo. Of those, approximately 50% should have received TXA. The study allocation was still blinded at the time of analysis.
[e] The numbers of each individual vasopressors do not add up to the group total as patients may have received more than one type of vasopressor during surgery.
[f] Other vasopressors used included vasopressin (4.6%), epinephrine (2.6%), and milrinone (1.2%).

**Table 2.** Limits of agreement adjusted for repeated measurements between POCT-Hgb devices and lab-Hgb for the full population and for samples with a lab-Hgb <100 g/L.

| POCT-Hgb device | Mean difference (g/L) | Lower limit (95% CI) (g/L) | Upper limit (95% CI) (g/L) |
|---|---|---|---|
| **Full population** | | | |
| **HemoCue (n=1725)** | -1.1 | **-9.6** (-10.0, -9.2) | **7.5** (7.1, 7.8) |
| **i-STAT (n=1711)** | -3.1 | **-15.2** (-15.8, -14.7) | **9.0** (8.4, 9.5) |
| **Rad-67 (n=1312)** | 7.6 | **-21.0** (-22.5, -19.5) | **36.2** (34.7, 37.7) |
| **Lab-Hgb < 100 g/L (higher potential for RBC transfusions)** | | | |
| **HemoCue (n=680)** | -0.8 | **-9.5** (-10.1, -8.9) | **8.0** (7.3, 8.6) |
| **i-STAT (n=671)** | -2.4 | **-16.2** (-17.2, -15.2) | **11.5** (10.5, 12.5) |
| **Rad-67 (n=490)** | 12.9 | **-14.7** (-17.0, -12.4) | **40.5** (38.2, 42.8) |

Abbreviations: CI, Confidence interval; Lab-Hgb, laboratory-determined hemoglobin; POCT-Hgb, Point-of-care testing of hemoglobin; RBC, Red blood cell.

**Table 3.** Crude rates of disagreement between POCT-Hgb devices and lab-Hgb (gold standard), based on three different agreement thresholds.

| POCT-Hgb device | Full population (n=1726) | | Lab-Hgb <100 g/L (n=680) | |
|---|---|---|---|---|
| | n | % (95% CI)[a] | n | % (95% CI) |
| **Disagreement threshold: ± 4 g/L when lab-Hgb <100 g/L and ± 5% g/L when lab-Hgb ≥100 g/L** | | | | |
| **HemoCue** | 1725 | 11.9 (10.5, 13.6) | 680 | 12.8 (10.4, 15.6) |
| **i-STAT** | 1711 | 40.9 (38.6, 43.3) | 671 | 51.3 **(47.4, 55.1)**[b] |
| **Rad-67** | 1312 | 72.7 (70.2, 75.1) | 490 | 81.8 (78.1, 85.1) |
| **Disagreement threshold: ± 7% g/L** | | | | |
| **HemoCue** | 1725 | 5.0 (4.0, 6.2) | 680 | 6.3 (4.7, 8.5) |
| **i-STAT** | 1711 | 26.6 (24.5, 28.8) | 671 | 35.2 (31.6, 38.9) |
| **Rad-67** | 1312 | 63.3 (60.7, 65.9) | 490 | 76.5 (72.5, 80.2) |
| **Disagreement threshold: ± 10 g/L** | | | | |
| **HemoCue** | 1725 | 2.0 (1.4, 2.8) | 680 | 2.4 (1.4, 3.9) |
| **i-STAT** | 1711 | 10.6 (9.2, 12.2) | 671 | 11.0 (8.8, 13.7) |
| **Rad-67** | 1312 | 51.9 **(49.2, 54.6)** | 490 | 60.6 (56.1, 64.9) |

Abbreviations: CI, Confidence interval; Lab-Hgb, laboratory-determined hemoglobin; POCT-Hgb, Point-of-care testing of hemoglobin.

[a] 95% confidence intervals were calculated using the Wilson score method with a continuity correction to adjust for the correlated nature of the repeated measurements.

[b] 95% CI results that are in bold characters are statistically not significant; we cannot reject the null hypothesis that the probability of disagreement is 0.5.

## 6.11 FIGURES

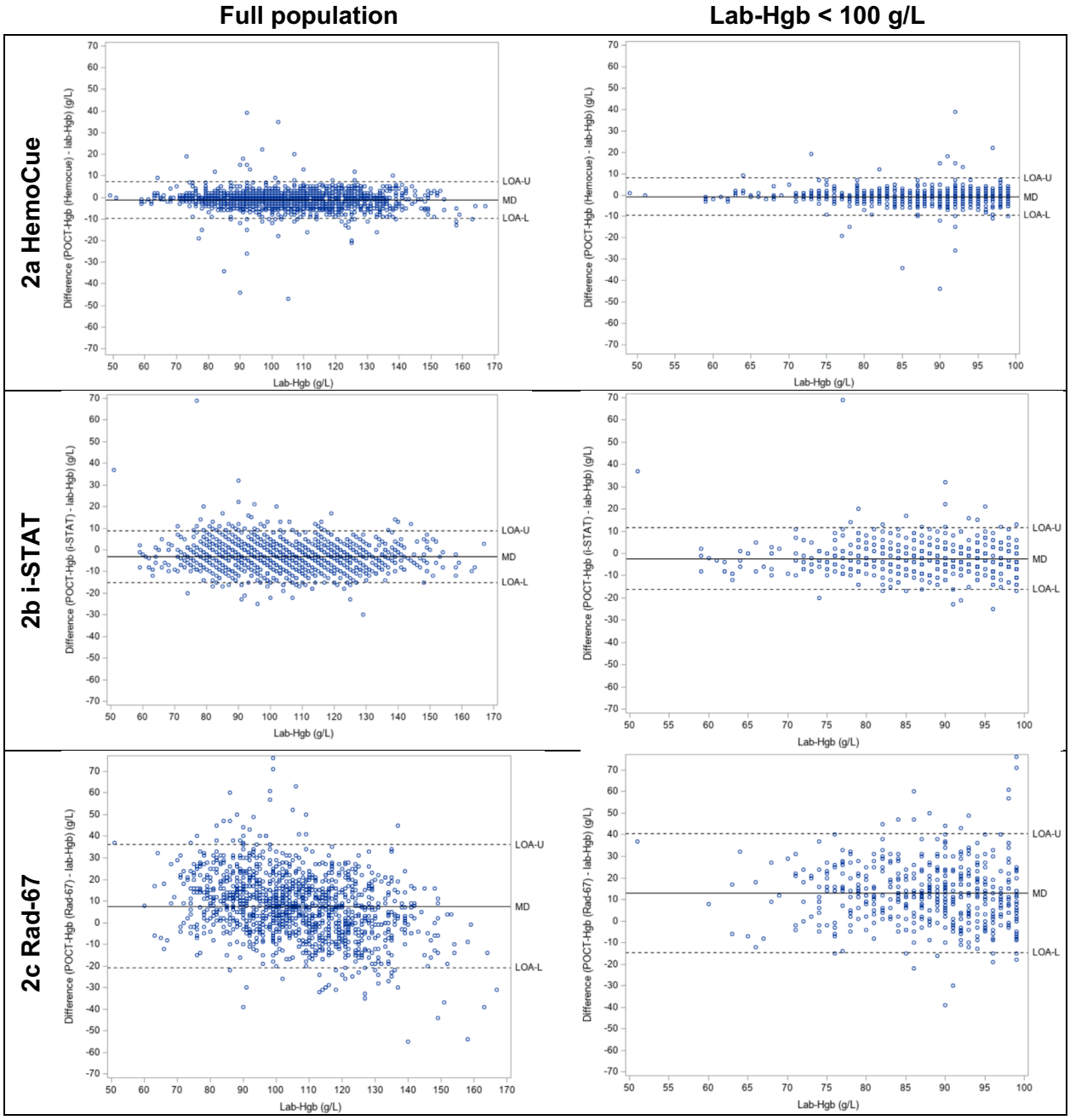
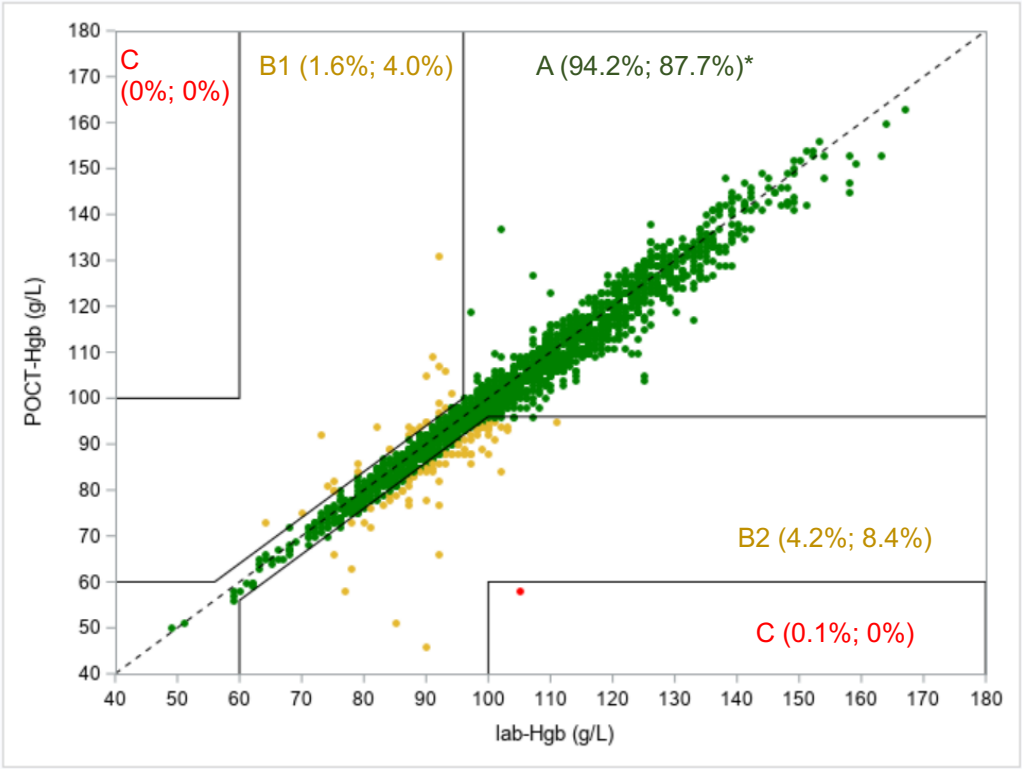
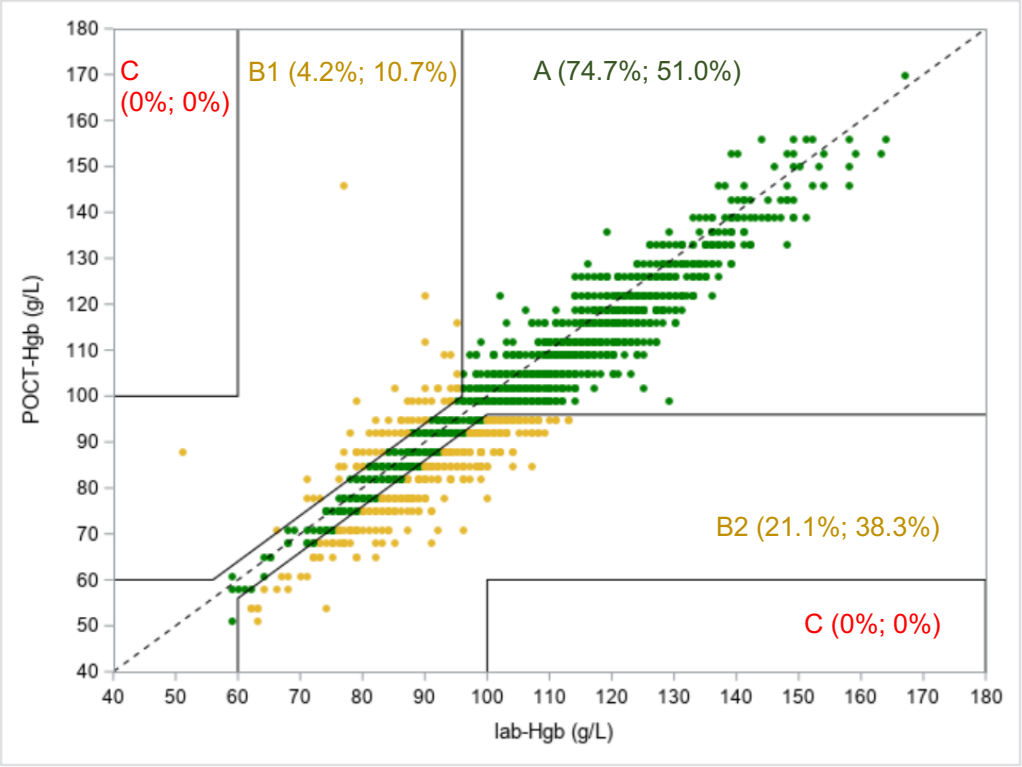



**Figure 1.** Limits of agreement plots for the full population and for the measurements with a higher potential for transfusions (lab-Hgb <100 g/L). Lab-Hgb, Laboratory-determined hemoglobin; LOA-L, Lower limit of agreement; LOA-U, Upper limit of agreement; MD, Mean difference; POCT-Hgb, Point-of-care testing of hemoglobin; RBC, Red blood cell.

**3a HemoCue**




**3b i-STAT**



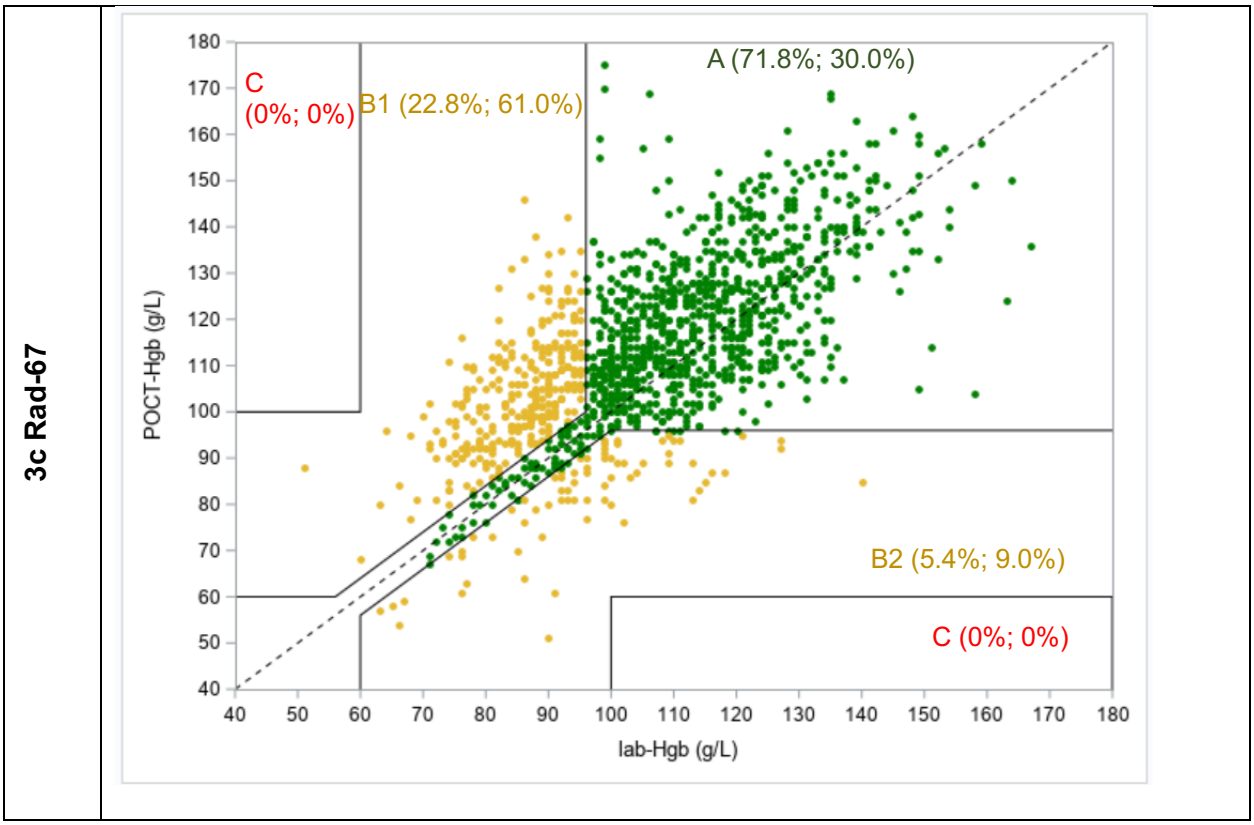



**Figure 2.** Error grid analyses. POCT-Hgb, Point-of-care testing of hemoglobin; lab-Hgb, laboratory-determined hemoglobin.

Zone A (green) represents low risk for potential transfusion errors, Zones B1 and B2 (yellow) represents moderate risk for potential transfusion errors, and Zone C (red) represents high risk for potential transfusion errors.

*The first and second percentages represent the proportion of measurements within a given zone for the full population and for measurements with a lab-Hgb<100 g/L, respectively.

## 6.12 APPENDIX 1. STANDARDS FOR REPORTING DIAGNOSTIC ACCURACY STUDIES (STARD) CHECKLIST

| Section & Topic | No | Item | Reported on page # |
|---|---|---|---|
| **TITLE OR ABSTRACT** | | | |
| | **1** | Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC) | NA |
| **ABSTRACT** | | | |
| | **2** | Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts) | 164, 165 |
| **INTRODUCTION** | | | |
| | **3** | Scientific and clinical background, including the intended use and clinical role of the index test | 166 |
| | **4** | Study objectives and hypotheses | 166, 167 |
| **METHODS** | | | |
| *Study design* | **5** | Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study) | 168 |
| *Participants* | **6** | Eligibility criteria | 168 |
| | **7** | On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry) | 168 |
| | **8** | Where and when potentially eligible participants were identified (setting, location and dates) | 168 |
| | **9** | Whether participants formed a consecutive, random or convenience series | 168 |
| *Test methods* | **10a** | Index test, in sufficient detail to allow replication | 168, 169 |
| | **10b** | Reference standard, in sufficient detail to allow replication | 169 |
| | **11** | Rationale for choosing the reference standard (if alternatives exist) | 169 |
| | **12a** | Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory | 169 |
| | **12b** | Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory | 169 |
| | **13a** | Whether clinical information and reference standard results were available to the performers/readers of the index test | 168, 169 |
| | **13b** | Whether clinical information and index test results were available to the assessors of the reference standard | 168, 169 |
| *Analysis* | **14** | Methods for estimating or comparing measures of diagnostic accuracy | 169-171 |
| | **15** | How indeterminate index test or reference standard results were handled | 171 |
| | **16** | How missing data on the index test and reference standard were handled | 171 |
| | **17** | Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory | 169 |
| | **18** | Intended sample size and how it was determined | 171 |
| **RESULTS** | | | |
| *Participants* | **19** | Flow of participants, using a diagram | Figure 1 |
| | **20** | Baseline demographic and clinical characteristics of participants | 172, Table 1 and 2 |
| | **21a** | Distribution of severity of disease in those with the target condition | NA |
| | **21b** | Distribution of alternative diagnoses in those without the target condition | NA |

| | | | |
|---|---|---|---|
| | **22** | Time interval and any clinical interventions between index test and reference standard | NA |
| *Test results* | **23** | Cross tabulation of the index test results (or their distribution) by the results of the reference standard | NA / Table 3 |
| | **24** | Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals) | 172, 173, Table 2 |
| | **25** | Any adverse events from performing the index test or the reference standard | 172 |
| **DISCUSSION** | | | |
| | **26** | Study limitations, including sources of potential bias, statistical uncertainty, and generalisability | 176, 177 |
| | **27** | Implications for practice, including the intended use and clinical role of the index test | 177 |
| **OTHER INFORMATION** | | | |
| | **28** | Registration number and name of registry | 178 |
| | **29** | Where the full study protocol can be accessed | 178 |
| | **30** | Sources of funding and other support; role of funders | 178 |

Abbreviation: NA, Not applicable.

## 6.13 APPENDIX 2. SUPPLEMENTARY TABLES AND FIGURES

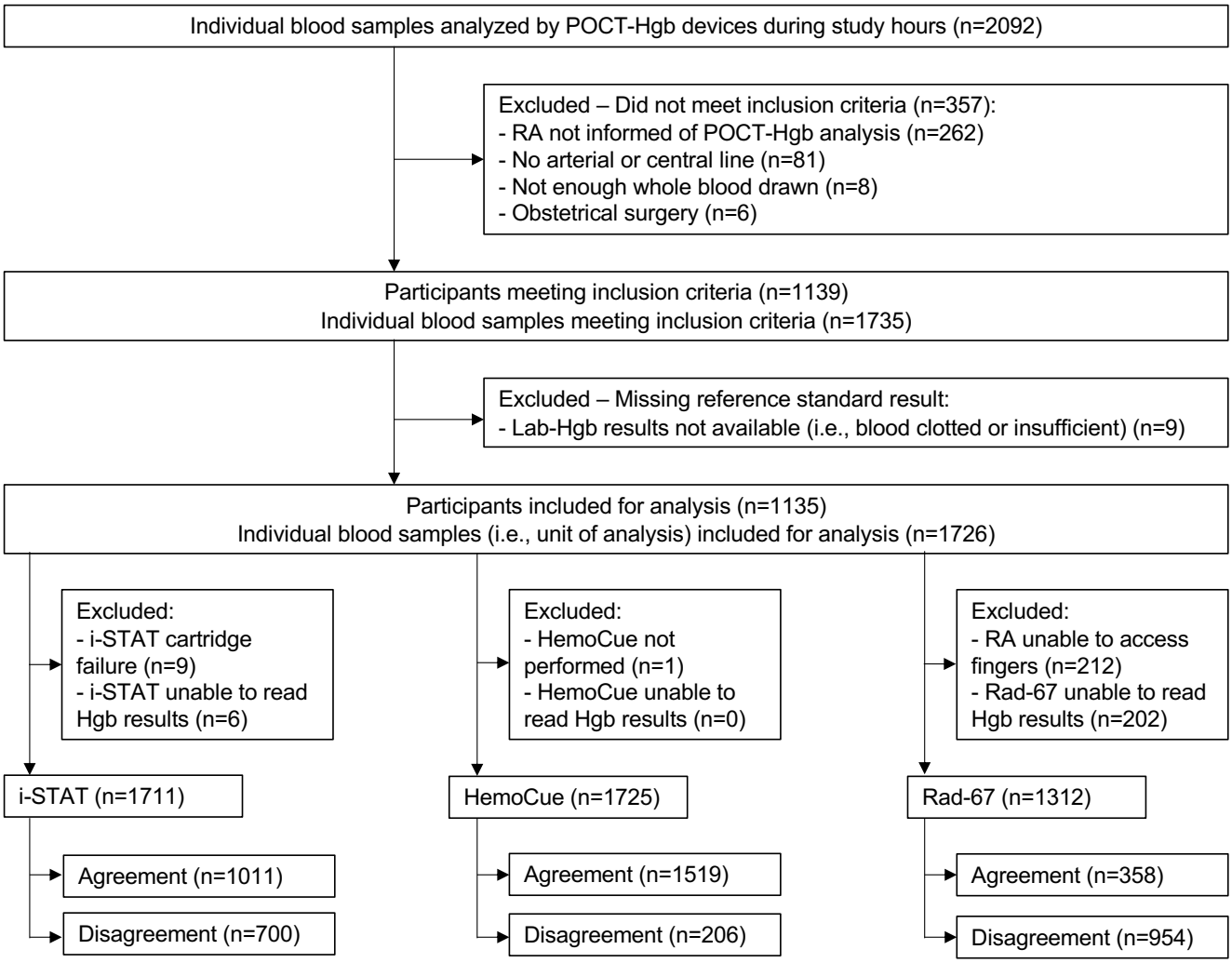



**Figure 3.** Flow of participants. Hgb, Hemoglobin; Lab-Hgb, Laboratory-determined hemoglobin; POCT-Hgb, Point-of-care testing of hemoglobin; RA, Research assistant.
Agreement and disagreement incidence for each POCT-Hgb device are based on the Institute for Quality Management in Healthcare (IQMH) definition.

**Table 4.** Limits of agreement estimates when outliers are removed from analysis.

| | n outliers removed | Mean bias (g/L) | Lower limit (g/L) | 95% CI of the lower limit | Upper limit (g/L) | 95% CI of the lower limit |
|---|---|---|---|---|---|---|
| **Full population** | | | | | | |
| i-STAT (n=1708) | 3 | -3.2 | **-14.6** | -15.1, -14.0 | **8.1** | 7.6, 8.6 |
| HemoCue (n=1721) | 4 | -1.0 | **-8.5** | -8.8, -8.1 | **6.4** | 6.0, 6.7 |
| Rad-67 (n=1308) | 4 | 7.6 | **-20.2** | -21.7, -18.7 | **35.4** | 33.9, 36.8 |
| **Higher potential for transfusions (lab-Hgb <100 g/L)** | | | | | | |
| i-STAT (n=668) | 3 | -2.6 | **-14.7** | -15.6, -13.8 | **9.6** | 8.7, 10.5 |
| HemoCue (n=675) | 5 | -0.7 | **-7.5** | -8.0, -7.0 | **6.2** | 5.7, 6.7 |
| Rad-67 (n=486) | 4 | 12.8 | **-13.1** | -15.3, -10.9 | **38.8** | 36.6, 40.9 |

Abbreviations: CI, confidence intervals; Lab-Hgb, Laboratory-determined hemoglobin; RBC, Red blood cells.

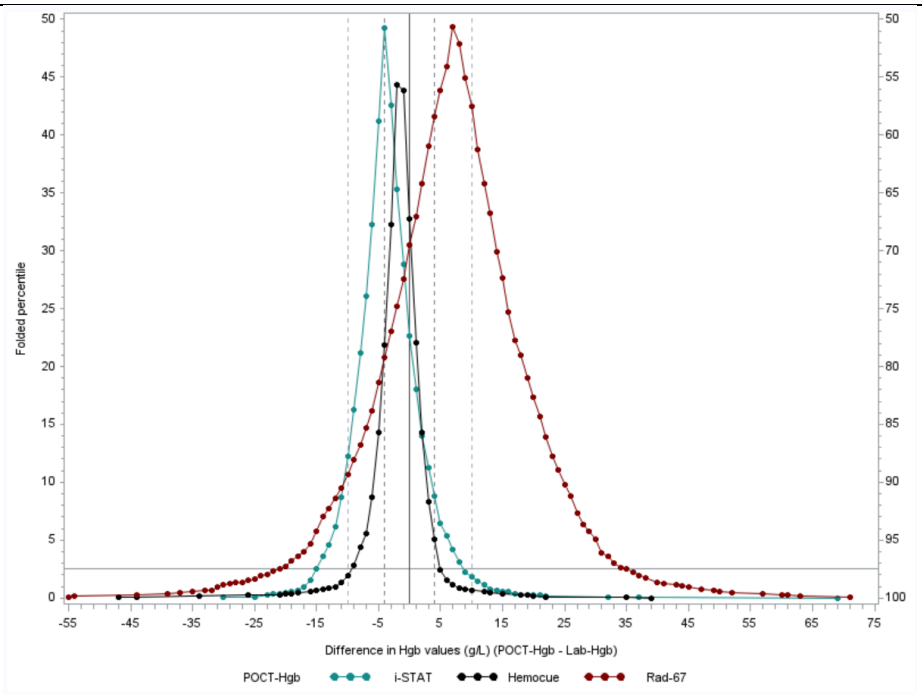
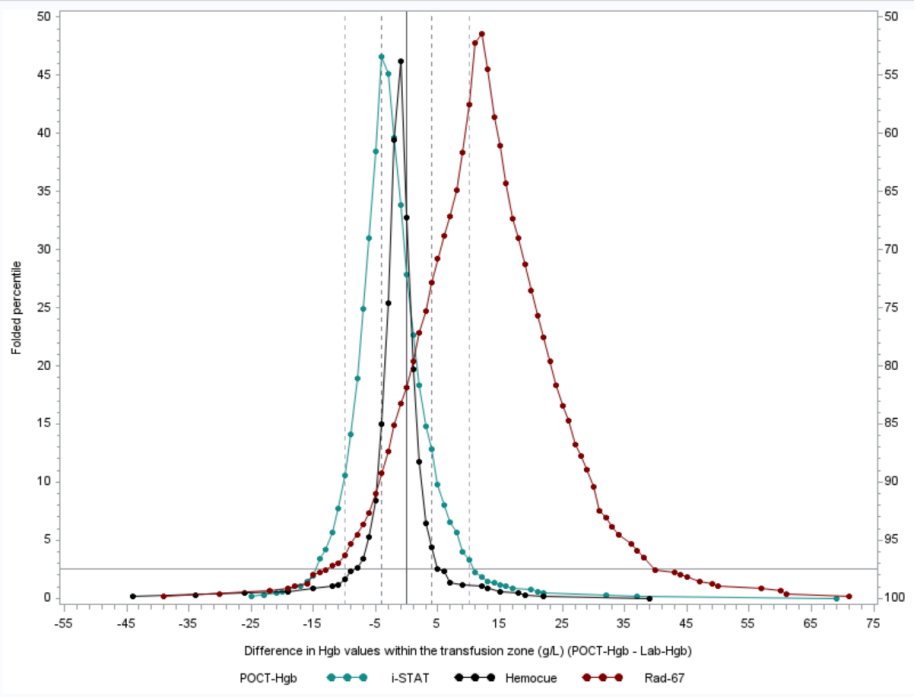
| | |
|---|---|
| **3a Full population** |  |
| **3b Lab-Hgb <100 g/L** |  |

**Figure 4.** Mountain plots for the full population and for the transfusion zone (lab-Hgb <100 g/L). Lab-Hgb, Laboratory-determined hemoglobin; POCT-Hgb, Point-of-care testing of hemoglobin.

**Table 5. Cohen's κ statistic results**

| Agreement threshold | Cohen's κ[a] (95% CI) (n=1725) | Cohen's κ (95% CI) (n=1711) | Cohen's κ (95% CI) (n=1312) |
|---|---|---|---|
| | HemoCue | i-STAT | Rad-67 |
| <100 g/L | 0.88 (0.86, 0.90) | 0.74 (0.71, 0.78) | 0.47 (0.42, 0.52) |
| <90 g/L | 0.85 (0.82, 0.88) | 0.72 (0.68, 0.76) | 0.34 (0.28, 0.40) |
| <85 g/L | 0.84 (0.80, 0.87) | 0.72 (0.67, 0.77) | 0.34 (0.26, 0.42) |
| <80 g/L | 0.83 (0.78, 0.88) | 0.65 (0.59, 0.71) | 0.34 (0.23, 0.45) |
| <75 g/L | 0.77 (0.68, 0.86) | 0.63 (0.54, 0.72) | 0.33 (0.18, 0.48) |
| <70 g/L | 0.84 (0.73, 0.94) | 0.53 (0.41, 0.66) | 0.38 (0.14, 0.62) |

Abbreviations: CI, confidence interval.
[a]Kappa statistic <0.00 represents poor agreement, within a range of 0.00 to 2.00 is slight agreement, 0.21 to 0.40 is fair agreement, 0.41 to 0.60 is moderate agreement, 0.61 to 0.80 is substantial agreement, and 0.81 to 1.00 is almost perfect agreement.[57,58]

# CHAPTER 7. INTEGRATED DISCUSSION

This thesis was primarily focussed on identifying and understanding strategies to inform the analysis of method comparison studies that include repeated measurements. The following chapter presents an integrated discussion encompassing the major themes explored in the context of this thesis along, with the main findings arising from the articles composing this thesis.[1] For this, I will start by summarizing the main findings of each article presented in this thesis, followed by a critical review of the themes raised by all manuscripts. Finally, I will outline the strengths and limitations of this thesis, along with the implications for practice and future directions arising from this work.

The overarching objective of this thesis was to identify best practices related to the analysis of method comparison studies that included repeated measurements to inform the statistical analysis of the PREMISE study prior to initiating its analysis. Concurrently, some objectives of this thesis were motivated by enhancing my learning experience in epidemiological methods and increasing my exposure to various research and statistical methods.

Chapters 2 and 3 were dedicated to the scoping review, the first study used for the elaboration of this thesis. The protocol for the scoping review is presented in Chapter 2, and the manuscript reporting the findings of the scoping review is presented in Chapter 3. In the context of this scoping review, the available literature was systematically searched to identified statistical methods that may be used to inform the analysis of method comparison studies as alternatives to the limits of agreement (LOA) method adjusted for repeated measurements[2]. The identified statistical methods were then summarized and compared to the LOA method, both quantitatively and qualitatively. The identified statistical methods were further grouped in a way that is logical and that will help researchers identify robust statistical methods for the analysis of their own method comparison studies.

Chapters 4 to 6 were dedicated to the PREMISE study. In Chapter 4, the protocol for the PREMISE study is presented to describe the background and methods associated with this study. In Chapter 5, the statistical analysis plan (SAP) is presented, which complements the PREMISE protocol. Finally, the findings from the analysis of the outcomes pertaining to agreement between POCT-Hgb and lab-Hgb are presented in Chapter 6. Findings from this study provide useful information on the accuracy of POCT-Hgb devices, with the conclusion that none of the devices should be used interchangeably with lab-Hgb based on current definitions for agreement.[3] Furthermore, it was found that the different devices under study each have widely different levels of agreement with the lab-Hgb, which has important clinical implications.

The first main theme of this thesis was to enhance my learning experience and deepen my understanding of epidemiologic and statistical methods by maximizing my exposure to various research projects and aims. All articles written in the context of this thesis and presented within the previous chapters have undoubtedly contributed to reaching this objective. The scoping review exposed me to all steps involved in reviewing the literature systematically. Indeed, I lead all aspects of the scoping review, including the designing, implementing, and conducting of the study, as well as analyzing, summarizing, and reporting the findings from the review. Moreover, the findings from this review have further complemented my learning by exposing me to various statistical methods that have been proposed as alternate techniques to inform the analysis of method comparison studies with repeated measurements, which may be used in place of the LOA method. This scoping review has also improved my critical-thinking skills while I was analyzing our findings as I was able to find a logical way of presenting our findings, based on the techniques that have been used to deal with repeated measurements. The review further increased my knowledge of the various LOA methods as many authors of the identified publications discussed the strengths and limitations of the LOA methods, as well as many different statistical technicalities to take into consideration when using this method for the analysis of method comparison studies. In terms of the PREMISE study, the three related manuscripts combined have enhanced my

learning experience and exposure to research methods. In fact, I was exposed to the design of a large and complex observational prospective study when writing the protocol, along with the challenges associated with the submission of manuscripts for publication in peer-reviewed medical journals. During the creation of the SAP, I was exposed to the challenges associated with producing a rigorous document informed by complex statistical methods. Furthermore, conducting the analysis of the PREMISE study has highly increased my understanding of the advantages and limitations by following an SAP that was designed a priori, improved my knowledge of the LOA method for repeated measurements, and evolved my skills in performing advanced statistical methods. Finally, the summarizing and reporting of my findings during the analysis of the PREMISE study has helped me learn to prioritize the information that I want to provide and how to write a compelling report of the analysis of a large study that generated an abundance of data.

The second main theme involved in this thesis was to understand and further inform the analysis of method comparison studies that include repeated measurements. When designing the analytic component of the PREMISE protocol and SAP, we were aware that the LOA method adapted for repeated measurements was likely going to be an appropriate analytic technique, but we had found little information on alternate statistical methods that could be used to analyze the study data. Thus, a knowledge gap was identified in the current literature in terms of guidance documents outlining existing statistical methods to inform such analysis, as well as outlining the various ways of handling repeated measurements in the context of method comparison studies. In fact, only one publication had been identified as a potential guidance document, referred to as a “tutorial for practitioners”, to inform the analysis of method comparison studies, using five different statistical methods including the LOA method.[4] However, a variety of other statistical methods had been promoted in the literature within independent publications and it was felt that it might be helpful to gather them all within one publication. Nonetheless, the tutorial for practitioners may be helpful in further understanding those five statistical methods more

specifically and helping researchers choose the most appropriate method for the analysis of their data when hoping to use those methods specifically. The lack of a complete guidance document led to the idea of the scoping review to systematically search the literature for proposed statistical methods to inform method comparison studies that include repeated measurements. The main objective of the scoping review was to summarize and gather all statistical methods that could be compared to the LOA method and generate a guidance document as a reference tool to help support future studies generating similar data. The findings of this review also helped inform the analysis of the PREMISE study. Furthermore, the analysis of the PREMISE study is a concrete example of the analysis of a large method comparison study that include repeated measurements, which demonstrated various statistical methods to inform the analysis of such studies.

The third and final main theme related to this thesis was to fill an important knowledge gap in terms of transfusion medicine within the operating room setting. This was mainly done via the analysis of the PREMISE study. This study was designed to facilitate a critical review of the accuracy of commonly used POCT-Hgb devices in the operating room. This study was of high importance as POCT-Hgb devices are often relied upon for transfusion-decision making in this setting, as red blood cell (RBC) transfusions are often guided by hemoglobin (Hgb) measurements, among with other clinical and physiological parameters.[5–10] However, current evidence lacks adequate validation of the accuracy of POCT-Hgb devices in measuring Hgb during major surgery, where the Hgb values have the potential to vary drastically and rapidly.[11,12] Furthermore, the evidence on the accuracy of such devices is very limited within the range of Hgb values with a higher risk for transfusions, being between 60 and 100 g/L.[13–15] To our knowledge, the PREMISE study is the largest method comparison study examining the accuracy of commonly used point-of-care testing devices for hemoglobin (POCT-Hgb) in the operating room, with the most data on the accuracy of those devices when lab-Hgb measurements are within 60-100 g/L. Thus, with its important accrual of data and findings on the accuracy of POCT-Hgb devices, the

PREMISE study has the potential to contribute largely to advancing the knowledge in the field of transfusion medicine in the operating room.

There were some limitations related to this thesis. For the scoping review, the LOA method was used as an anchor to identify publications that discussed this specific statistical method along with alternate methods to inform the analysis of method comparison studies with repeated measurements, to allow direct comparison between the methods. This aspect of the inclusion criteria was chosen in part to examine how other statistical methods perform when compared with the LOA method, but also as a way of limiting the scope of the review in order to be feasible within the confines of a Master's thesis. Nonetheless, this scoping review was vast in scope with screening over 5000 different publications and reports and including 39 publications in the review. In terms of the PREMISE analysis, only the planned analyses pertaining to agreement between POCT-Hgb devices and lab-Hgb measurements were performed as it was felt that this fell within the scope of a Master's thesis. However, the predictive modelling, as well as transfusion evaluation, were planned secondary analyses. These were not presented herein as they are outside of the scope of this thesis.

There are several implications related to both main studies presented in this thesis. The scoping review has identified several statistical methods as alternate approaches to the LOA method, that may be used to inform the analysis of method comparison studies with repeated measurements, along with providing a complete and comprehensive overview of such statistical methods. Findings from this review will potentially help other researchers inform the analysis of future method comparison studies that include repeated measurements. The reporting of our findings from this scoping review also provides an overview of various existing methods for the handling of repeated data in method comparison studies. Moreover, informed statistical analyses may lead to improved reporting of method comparison studies, which has the potential to improve patient care by informing healthcare providers concerning the accuracy of various devices used to measure physiological parameters measured in a continuous fashion. The findings from this

study may potentially help other researchers choose robust statistical methods for the analysis of future method comparison studies that include repeated measurements. Future research opportunities were identified following this scoping review, such as the conduct of another scoping review that aims to identify all possible statistical methods that are appropriate for the analysis of method comparison studies with repeated measurements, without using the LOA method as an anchor limiting the findings.

Results of the PREMISE study will have significant implications for the care of surgical patients in the operating room. Indeed, the agreement statistics reported in the PREMISE study manuscript (Chapter 6) may be used directly by clinicians providing care to surgical patients to inform transfusion-decision making in this population, specifically when using POCT-Hgb devices for such clinical decisions. While those devices can provide useful and rapid information on hemoglobin (Hgb) concentrations intra-operatively, our findings will potentially encourage clinicians to use prudence when using those devices to inform transfusion decision-making in the operative setting. Moreover, the PREMISE study generated a wealth of data and information on surgical patients, including data on peri-operative blood transfusions, surgical and anesthetic parameters, and post-operative complications. Further analyses are planned for this study, and will be performed and reported in the future. Indeed, predictive models will be built to predict the context in which those devices provide inaccurate measurements. Further analyses will be conducted to examine the appropriateness of intra-operative transfusion events as well as transfusion outcomes. Finally, the protocol for the PREMISE study was designed to study the POCT-Hgb devices readily available for use in the operating room at our institution (HemoCue, i-STAT, and Rad-67). Blood gas analyzers were not examined in PREMISE, as they are not frequently used in the operating rooms in North America and were not readily available. This has been identified as a limitation of the PREMISE study. Future research opportunities would be to repeat this study and include a blood gas analyzer for examination of accuracy, along with other

POCT-Hgb devices. Furthermore, it may be of interest to repeat this study in other institutions, where other types of POCT-Hgb devices are in used.

In conclusion, the analysis of method comparison studies with repeated measurements comes with many challenges due to the correlated nature of repeated measurements. This thesis has outlined the challenges related to the analysis of method comparison studies that include repeated measurements and identified different approaches to inform their analysis. An example of some of the statistical methods that may be used for the analysis of this study design was provided using the PREMISE study dataset, a large prospective method comparison study that included repeated measurements. The findings from the analysis of PREMISE help fill an important knowledge gap related to transfusion practices in the operating room, by reporting on the accuracy of commonly used point-of-care devices to measure hemoglobin in the operating room.

**REFERENCES**


1. Lewis KB, Graham ID, Boland L, Stacey D. Writing a compelling integrated discussion: a guide for integrated discussions in article-based theses and dissertations. *Int J Nurs Educ Scholarsh*. 2021;18(1):20200057. doi:10.1515/ijnes-2020-0057

2. Bland JM, Altman DG. Agreement Between Methods of Measurement with Multiple Observations Per Individual. *J Biopharm Stat*. 2007;17(4):571-582. doi:10.1080/10543400701329422

3. Johnston A, Bourner G, Martin T, et al. Guidance for quality control practices and precision goals for CBCs based on IQMH patterns-of-practice survey. *Int J Lab Hem*. 2019;41:15-22. doi:10.1111/ijlh.12915

4. Parker RA, Scott C, Inácio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. *BMC Med Res Methodol*. 2020;20:1-14. doi:10.1186/s12874-020-01022-x

5. Rekman J, Workneh A, Rao P, et al. Point-of-care hemoglobin testing and liver surgery. *HPB*. 2018;20:S490. doi:10.1016/j.hpb.2018.06.2966

6. Giraud B, Frasca D, Debaene B, Mimoz O. Comparison of haemoglobin measurement methods in the operating theatre. *Br J Anaesth*. 2013;111(6):946-954. doi:10.1093/bja/aet252

7. Baker L, Park L, Gilbert R, et al. Intraoperative Red Blood Cell Transfusion Decision-making: A Systematic Review of Guidelines. *Ann Surg*. 2021;274(1):86-96. doi:10.1097/SLA.0000000000004710

8. Practice Guidelines for Perioperative Blood Management: An Updated Report by the American Society of Anesthesiologists Task Force on Perioperative Blood Management. *Anesthesiology*. 2015;122(2):241-275. doi:10.1097/ALN.0000000000000463

9. Bennett S, Ayoub A, Tran A, et al. Current practices in perioperative blood management for patients undergoing liver resection: a survey of surgeons and anesthesiologists. *Transfusion*. 2018;58:781-787. doi:10.1111/trf.14465

10. Jouffroy R, Baugnon T, Carli P, Orliaguet G. A survey of blood transfusion practice in French-speaking pediatric anesthesiologists. *Pediatr Anesth*. 2011;21:385-393. doi:10.1111/j.1460-9592.2011.03531.x

11. Wu PBS, Morey TE, Harris NS, Gravenstein N, Rice MJ. Intravenous Fluids Cause Systemic Bias in a Conductivity-Based Point-of-Care Hematocrit Meter. *Anesth Analg*. 2012;114(2):314-321. doi:10.1213/ANE.0b013e31823fecbd

12. Hiscock R, Kumar D, Simmons SW. Systematic review and meta-analysis of method comparison studies of Masimo pulse co-oximeters (Radical-7™ or Pronto-7™) and HemoCue® absorption spectrometers (B-Hemoglobin or 201+) with laboratory haemoglobin estimation. *Anaesth Intensive Care*. 2015;43(3):341-350. doi:10.1177/0310057x1504300310

13. Johnson M, Marwick PC, Coetzee JF. Comparison of Hemoglobin Measurements by 3 Point-of-Care Devices With Standard Laboratory Values and Reliability Regarding Decisions for Blood Transfusion. *Anesth Analg*. 2020;131(2):640-649. doi:10.1213/ANE.0000000000004533

14. Rice MJ, Gravenstein N, Morey TE. Noninvasive Hemoglobin Monitoring: How Accurate Is Enough? *Anesth Analg*. 2013;117(4):902-907. doi:10.1213/ANE.0b013e31829483fb

15. Morey TE, Gravenstein N, Rice MJ. Let's Think Clinically Instead of Mathematically About Device Accuracy. *Anesth Analg*. 2011;113(1):89-91. doi:10.1213/ANE.0b013e318219a290

# CHAPTER 8. ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis co-supervisor, Dr. Guillaume Martel. Guillaume, you provided me with my first opportunity and exposure to the realm of clinical research, a field for which I have grown to be passionate about. Your thoroughness, attention to detail, excitement for new discoveries, and passion for research are truly inspiring and I aspire to continue learning from you in the future.

I would also like to thank my thesis co-supervisor, Dr. Dean Fergusson. Your knowledge in the field of clinical research is unparalleled. I am extremely fortunate to have had the opportunity to learn from you in the past two years and I will be forever grateful for this. Both you and Dr. Martel have undoubtedly contributed to my growth and my learning journey throughout this program by being generous with your time, insight, and guidance.

I would like to extend my gratitude to the members of my thesis advisory committee, Drs. Daniel McIsaac and Tim Ramsay. Your combined set of skills in the fields of statistics and research have been valuable and stimulating. The time and support you have both provided for my thesis in the past two years have been extremely valuable and appreciated.

To Leah and Tori, I am forever indebted to you for your help and support during my thesis. Leah, I am lucky to have you as a colleague, and also as a friend I can always count on. Your capacity to support me, make me laugh, and keep me grounded have been incredibly helpful. Tori, I truly appreciate you taking some of your valuable time during your residency to collaborate on the scoping review. Your efficacy and determination are inspirational.

Finally, to my partner Martin, my family, and my friends who have supported me unconditionally throughout my education adventures. From cheering for me, to cooking countless meals, you have kept me going through this whole process and I will cherish your love and support forever.

# CHAPTER 9. FUNDING

This thesis was partly funded by an admission scholarship provided by the University of Ottawa and a stipend from the Liver and Pancreas Unit research group which was funded by a Canadian Blood Services Intramural Research Grant.