PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Current State of Science in Machine Learning Methods for Automatic Infant Pain Evaluation using Facial Expression Information: Study Protocol of A Systematic Review and Meta-analysis
AUTHORS	Cheng, Dan; Liu, Dianbo; Philpotts, Lisa Liang; Turner, Dana P; Houle, Timothy T.; Chen, Lucy; Zhang, Miaomiao; Yang, Jianjun; Zhang, Wei; Deng, Hao

VERSION 1 – REVIEW

REVIEWER	Laure Perrier University of Toronto, Canada
REVIEW RETURNED	08-Apr-2019

GENERAL COMMENTS	The authors have identified an important topic on infant pain and the challenge of achieving good pain control through objective assessment. Overall, the authors have presented a well-designed protocol that is written clearly. However, a few items that would strengthen the manuscript have been identified. Several items outlined below involve clarifications: Eligibility criteria • Provide a definition of infant (i.e., there would likely be an age range for this group) – Since you are searching the term neonate, should this be defined and clarified in your criteria? – Cut-off ages should be noted in your protocol for inclusion/exclusion criteria • Are premature infants included or excluded? Clarify or provide explanations if not including • Identify the study types that will be included in the review Search strategy • Provide rationale for the date limits of 2008-2018 • Table 2: For an exhaustive search, – PubMed: MeSH (Medical Subject Headings) need to be added to the search, e.g., Infant, Newborn; Pain; Machine Learning; Support Vector Machine; Limit of Detection; Algorithms; Expert Systems; Deep Learning; etc. – Why has the term algorithm been left out of IEEE and Web of Science search? – Acronyms (such as SVM) should also be searched in their full length, e.g., support vector machine(s) – Consider other search terms for infant such as newborn – The search should include wildcards in order to include variations of words, e.g., infant* = infant, infants; detect* = detection, detecting, detects; etc.

	– The section called ‘Machine learning methods’ lists several models besides SVM (such as RVM, AAM, K-NN) – why was SVM the only one that was included in the search strategy? – The term artificial intelligence is mentioned in the Discussion – should this term be included in the searches? – Web of Science: Why is pain searched in the title only? Study selection • Clarify and describe each stage of screening, i.e., titles and abstracts screening; full-text screening Primary outcome • Page 13, Line 32: Provide references for the ‘previous experience’ mentioned Risk of bias • Since the study types have not been declared it is unclear how to determine if developing a new risk of bias tool is necessary • Once the study types that will be included are declared it may be possible to identify and use risk of bias tools currently available – if this is not feasible, rationale must be provided for why these pre-established tools are not appropriate

REVIEWER	Erik Loeffen University of Groningen, University Medical Center Groningen, Beatrix Children's Hospital, Groningen, the Netherlands
REVIEW RETURNED	15-Apr-2019

GENERAL COMMENTS	Review manuscript bmjopen-2019-030482 I think the authors have asked an important question and I applaud research into pain measurement in children. Very interesting topic this machine learning is! I should note that I am no expert on machine learning so please take this into account when interpreting my comments (I do however know a fair bit about systematic reviews and pain measurement in children). The manuscript is interesting but does need some work, especially the search strategy needs substantial improvement, before I can recommend it for acceptance in BMJ Open. During reviewing I have written comments down, I’ve grouped these in major and minor remarks. MAJOR REMARKS - English is suboptimal in various places throughout the manuscript, please have a native speaker revise the manuscript. For example: “By far, there is no research has quantitatively and systematically summarized and compared the performance of these ML methods.”. ‘By far’ is a strange way to start this sentence (I think authors mean ‘To date’), and between ‘research’ and ‘has’ a word is missing (that?). - Should there be a place for CINAHL in the databases? A lot of measurement studies are indexed there. - Throughout the manuscript authors only mention that ‘pain’ is measured. However, in infants it is not possible to distinguish between pain and stress during procedures. This should be mentioned somewhere and discussed upon. - The search strategy needs work: 1) it is not comprehensive, e.g. there are missing terms in especially the infant string (e.g.

	newborn, baby, etc) and pain string (e.g. agony, hurt, etc), 2) asterisks should be used, i.e. studies that have “infants” in the abstract will now be missed as the strategy says ‘infant’ instead of ‘infant*’, 3) why only search in tiab? Keywords/mesh terms are then missed. - I would encourage authors to provide an example risk of bias extraction sheet, so we can see how the studies will be scored, although the approach sounds OK, I am having difficulties visualizing when a study will be scored low, moderate, or high risk. (also, can existing bias judgements be partially used? E.g. COSMIN criteria, Cochrane ROB? MINOR REMARKS INTRODUCTION - “relief treatments are rarely provided to infants and neonates undergoing painful procedures” – That statement is too bold, especially since it has no reference. Sucrose is for example often provided to neonates during painful procedures. METHODS - Population not described in sufficient detail: all types of pain? Procedural? Post OR? Nociceptive? Etc - Idem dito for control: which assessment instruments would authors consider a sufficiently solid tool? Or any tool? - “To account for publication bias, we will include qualitative review articles, systematic review and meta-analysis articles for missing unpublished literatures through their reference lists.” – How does this account for publication bias? Do authors mean ‘to identify missing studies?’. Also, will authors perform forwards and backwards citation screening? - Which societies (conference meeting abstracts)? Be precise. - Will the review/screening process be piloted? - Subgroup analysis: which medical procedure types will be distinguihshed? - Do authors also take things as costs / resources into account when interpreting data? I can imagine these are costly machines / software programs. Perhaps can authors expand upon this in the discussion?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Laure Perrier

Institution and Country: University of Toronto, Canada

Please state any competing interests or state ‘None declared’: None declared

The authors have identified an important topic on infant pain and the challenge of achieving good pain control through objective assessment. Overall, the authors have presented a well-designed protocol that is written clearly. However, a few items that would strengthen the manuscript have been identified. Several items outlined below involve clarifications:

Comment 1: Eligibility criteria

• Provide a definition of infant (i.e., there would likely be an age range for this group)

– Since you are searching the term neonate, should this be defined and clarified in your criteria?

– Cut-off ages should be noted in your protocol for inclusion/exclusion criteria

Reply:

Thank you. We agree with the reviewer that a clear definition of infant is necessary. We followed your advice and clarified the definition of infants as young children no more than 12 months, including newborns or neonates, full term, premature, and post-mature infants. We added the definition to our manuscript and clarified that neonates would be included in the inclusion criteria. Details are presented below.

BEFORE:

We did not provide a definition of infants and did not note cut-off ages in inclusion/exclusion.

AFTER:

 # added definition

The population of our study will be infants undergoing pain. Infants will be defined as young children no more than 12 months, including newborn or neonate, full term, premature, and post-mature infants. Infant pain can be heel stick, arterial puncture, intravenous cannula, finger stick, nasal aspiration, or post operation pain, etc. (P9 L7-11)

 # added exclusion criteria

(6) Children more than 12 months and adults. (P9 L14-15)

• Are premature infants included or excluded? Clarify or provide explanations if not including

Reply:

Thank you. We agree with the reviewer that this information needs to be clarified. We will include premature infants in our study.

We decide to include this population because several studies show that premature infants are more easily exposed to clinical painful procedures and also experience more pain. [Johnston CC, Fernandes AM, Campbell-Yeo M. Pain in neonates is different. Pain. 2011;152:S65–73.][Hall RW, Anand KJ. Pain management in newborns. Clinics in perinatology. 2014;41(4):895-924.][Porter FL, Wolf CM, Miller JP. Procedural pain in newborn infants: the influence of intensity and development. Pediatrics. 1999;104(1):e13-.]. In addition, Carbajal et al found that preterm neonates experienced more than 10 painful procedures pre day, the majority of which (80%) were not preceded by specific analgesia. [Carbajal R, Rousset A, Danan C, et al. Epidemiology and treatment of painful procedures in neonates in intensive care units. JAMA. 2008; 300:60–70.] Zhi et al found that gestational age was one of the most influencing factors for infant pain assessment, and it was necessary to construct specific models depending on gestational age for infants with low gestational age that had limited ability for behavioral communication. [Zhi R, Zamzmi G, Goldgof D, Ashmeade T, Sun Y. Automatic Infants' Pain Assessment by Dynamic Facial Representation: Effects of Profile View, Gestational Age, Gender, and Race. Journal of clinical medicine. 2018;7(7):173.] Therefore, we believe this inclusion will provide necessary and clinical important information for our audiences.

BEFORE:

We did not note if premature infants were included or excluded.

AFTER :

The population of our study will be infants undergoing pain. Infants will be defined as young children no more than 12 months, including newborn or neonate, full term, premature, and post-mature infants. (P9 L7-9)

• Identify the study types that will be included in the review

Reply:

Thank you for your comments. We now identify the study types as quantitative prediction model studies.

BEFORE:

(3) Not a quantitative study concerning ML methods;

AFTER :

(3) Not a quantitative prediction model study concerning ML methods; (P9 L12-13)

Comment 2: Search strategy

• Provide rationale for the date limits of 2008-2018

Reply:

Thank you for your comments. We will search the literature until present day in order to review the most current literature. We choose to search from 2008 onwards since our study topic is focused on "current state of science in machine learning methods for automatic infant pain evaluation using facial expression information". Advances in machine learning methods for pain assessment, especially deep learning methods, have received increasing attention in this recent five to six years.[Valstar MF, Almaev T, Girard JM, McKeown G, Mehu M, Yin L, Pantic M, Cohn JF. Fera 2015-second facial expression recognition and analysis challenge. In2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 2015 May 4 (Vol. 6, pp. 1-8). IEEE.] [Chen Z, Ansari R, Wilkie D. Automated Pain Detection from Facial Expressions using FACS: A Review. arXiv preprint arXiv:1811.07988. 2018 Nov 13.] In order to include the majority of studies about this topic, we will search studies starting from January 2008. To the best of our knowledge, our study is the first systematic review and meta-analysis about machine learning methods for automatic infant pain evaluation using facial expression information, and we believe that the 10-year range is appropriate.

• Table 2: For an exhaustive search,

– PubMed: MeSH (Medical Subject Headings) need to be added to the search, e.g., Infant, Newborn; Pain; Machine Learning; Support Vector Machine; Limit of Detection; Algorithms; Expert Systems; Deep Learning; etc.

Reply:

Thank you for your suggestion. We followed the reviewer's advice and invited Lisa Philpotts, BSN, MSLS from the Treadwell Virtual Library at Massachusetts General Hospital to help us develop an updated search strategy. Ms. Philpotts is a professional medical librarian who has experience creating search strategies for systematic reviews and meta-analysis.

The search now includes subject headings in any databases that have controlled vocabulary in addition to the keywords that were listed in the prior protocol.

BEFORE:

The search strategies did not include subject headings. (P10-12 Table 2)

AFTER :

The PubMed strategy includes MeSH terms, and the Embase strategy includes Emtree subject headings.

– Why has the term algorithm been left out of IEEE and Web of Science search?

Reply:

Thank you for your advice. We now added “algorithm” and other keywords (e.g. newborn, expert systems, limit of detection, Support Vector Machine, etc) to the Web of Science search. We decided to remove the machine learning concept from the IEEE search in order to maximize retrieval. Unlike Web of Science, PubMed, and Embase, which have a wide variety of literature necessitating the inclusion of the machine learning concept, the IEEE database is limited to the engineering and computer science literature. The number of citations on infant pain in the IEEE database is relatively small compared to the other databases.

BEFORE:

IEEE: 2008 to 2018

(infant OR neonate OR neonatal) AND (pain OR painful) AND (face OR facial OR expression OR recognition OR detection OR automatic OR machine learning OR deep learning OR neural network OR SVM OR computer vision)

Web of Science: 2008 to 2018

TS=(infant OR neonate OR neonatal) AND (pain OR painful) AND (face OR facial OR expression OR recognition OR detection OR automatic OR machine learning OR deep learning OR neural network OR SVM OR computer vision) AND TI=pain

AFTER :

We have added algorithm to the search strategies in all of the databases. (P10-12 Table 2)

– Acronyms (such as SVM) should also be searched in their full length, e.g., support vector machine(s)

Reply:

Thank you. We agree with the reviewer’s suggestion and have included the spelled out acronyms as keywords in the searches.

BEFORE:

The item Support Vector Machine was not included in our search strategy.

AFTER :

The item Support Vector Machine was included in our search strategy. (P10-12 Table 2)

– Consider other search terms for infant such as newborn

Reply:

Thank you for your comment. We agree with the reviewer's advice. Now we added other terms for infant such as newborn, baby, babies, and neonat*.

BEFORE:

The items, such as newborn, baby, babies, and neonat* were not included in our search strategy.

AFTER :

The items, such as newborn, baby, babies, and neonat* were included in our search strategy. (P10-12 Table 2)

– The search should include wildcards in order to include variations of words, e.g., infant* = infant, infants; detect* = detection, detecting, detects; etc.

Reply:

Thank you for your comments. We agree with the reviewer's suggestion. Wildcards have been incorporated into the search to include word variants in the search strategy.

BEFORE:

Wildcards were not included in the search.

AFTER :

Keywords were truncated with wildcards when appropriate: infant*, neonat*, newborn*, pain*, hurt*, suffer*, distress* (P10-12 Table 2)

– The section called 'Machine learning methods' lists several models besides SVM (such as RVM, AAM, K-NN) – why was SVM the only one that was included in the search strategy?

Reply:

Thank you for your comments. We agree with your suggestion. Now we added RVM, AAM, PNN, and K-NN to the search strategy. We also spelled out these acronyms (ex: active appearance model) as suggested.

BEFORE:

The items, such as RVM, AAM and K-NN, were not included in the search strategy.

AFTER :

The items, such as such as RVM, AAM, PNN and KNN, were included in our search strategy. (P10-12 Table 2)

– The term artificial intelligence is mentioned in the Discussion – should this term be included in the searches?

Reply:

Thank you for your suggestion. We now added artificial intelligence into the search strategy.

BEFORE:

The items, artificial intelligence, was not included in the search strategy.

AFTER :

The items, artificial intelligence, was included in the search strategy. (P10-12 Table 2)

– Web of Science: Why is pain searched in the title only?

Reply:

Thank you for the insightful comments. We have incorporated this feedback into the search strategy and search for pain in the title, abstract, and keyword fields of Web of Science, similar to the search strategies for the other databases.

BEFORE:

Pain was searched in the title field in Web of Science.

AFTER:

Pain* is searched in the TS (title, abstract, keyword) field in Web of Science.

Study selection

• Clarify and describe each stage of screening, i.e., titles and abstracts screening; full-text screening

Reply:

Thank you for your comments. We followed the reviewer's suggestion, and now we described the screening process more clearly.

BEFORE:

Two authors (D.L. and D.C.) will independently review and screen searched article records to identify eligible studies according to our inclusion and exclusion criteria using the Abstrackr platform.[29] A third

investigator (H.D. or W.Z.) will resolve the disagreements between these two evaluators. Excluded studies will be listed in the PRISMA flowchart specifying reasons for exclusion in **Figure 1**.

AFTER :

Two authors (D.L. and D.C.) will independently review and screen searched article records to identify eligible studies according to our inclusion and exclusion criteria using the *Covidence* digital platform. A third investigator (H.D. or W.Z.) will resolve the disagreements between these two evaluators. In this step, two authors screen titles and abstracts of searched articles for primary exclusion on the *Covidence* platform. We will pilot the process by screening the first ten studies under the primary investigator's supervision. Once all the articles screened, we will download the full-text of included studies for further identification. Once eligible studies are identified, we will extract information for qualitative synthesis. Excluded studies will be listed in the PRISMA flowchart specifying reasons for exclusion in **Figure 1**. (P12 L3-11 and Appendix 1)

Primary outcome

• Page 13, Line 32: Provide references for the 'previous experience' mentioned

Reply:

Thank you for your suggestion. Now we take the reviewer's advice and provide references for the 'previous experience'. References are listed as below.

1. Liu D, Cheng D, Houle TT, Chen L, Zhang W, Deng H. Machine learning methods for automatic pain assessment using facial expression information: Protocol for a systematic review and meta-analysis.

2. Zamzmi G, Kasturi R, Goldgof D, Zhi R, Ashmeade T, Sun Y. A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases. IEEE reviews in biomedical engineering. 2017 Nov 27;11:77-96.

BEFORE:

Based on our previous experience

AFTER :

Based on our previous experience,[23,34] (P15 L17)

Comment 3: Risk of bias

• Since the study types have not been declared it is unclear how to determine if developing a new risk of bias tool is necessary

Reply:

Thank you for your comments. We agree with the reviewer's suggestions. We followed the advice of reviewer 2 and checked OSMIN criteria and Cochrane ROB. The OSMIN checklist can be applied for Systematic reviews of measurement properties, Measurement instrument selection, Identification of the need for further research measurement properties, Designing a study on measurement properties, Reporting a study on measurement properties, and Reviewing the quality of a submitted manuscript on measurement properties. Cochrane ROB is suited for randomized studies.

In addition, we checked another existing guideline in the medical field, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement. We conclude that TRIPOD is the most suitable guideline for this case because the majority of the items in the checklist are suited for machine learning prediction methods, and the TRIPOD statement is widely used in the medical field. Therefore, we will apply TRIPOD for assessing risk of bias.

To account for highly specialized machine learning aspects, we added specific items (designed by our data scientist Dianbo Liu) including input data selection, model performance, and outcome reporting.

We will define three rates, 0 as not mentioned or unclear (high risk), 1 as mentioned but not with details (moderate risk), 2 as mentioned with details (low risk). An example of design details is presented at the end of this file, Appendix 2 and Appendix 3.

BEFORE:

There was no established criteria or tool to assess the risk of bias for ML prediction algorithm research. Therefore, we will develop and utilize a customized risk of bias assessment tool. They will be evaluated in three main aspects including input data selection, model performance, and outcome reporting. Factors influencing input data selection include database sponsorship (e.g., organization or single study data) and image/video quality (e.g., Dpi of video, camera setting). Factors influencing model performance include research team (i.e. whether there is a professional computer scientist or clinician), innate prior of ML algorithm, algorithm training process, and optimization and evaluation method. Factors introducing reporting bias include incomplete reporting, selective reporting, non-standard reporting (e.g. only report point estimate without standard errors or confidence intervals). Based on these aspects, risk of bias judgment of included studies will be ranked as low risk, moderate risk, high risk or unclear. In order to better assess the quality of included studies, we plan to compare reported items in these studies with established and recommended reported items according to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement in medical field.[30]

AFTER:

There is no established criteria or tool to assess the risk of bias for ML prediction algorithm research. In order to better assess the quality of included studies, we plan to compare reported items in these studies with established and recommended reported items according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement for the medical field.[35] To account for machine learning aspects, we will develop and include a project-specific risk of bias appendix as a complementary assessment tool. Three main aspects, including input data selection, model performance, and outcome reporting, will be evaluated. Factors influencing input data selection include database sponsorship (e.g., organization or single study data) and image/video quality (e.g., Dpi of video, camera setting). Factors influencing model performance include research team (e.g., whether there is a professional computer scientist, clinician or nurse), innate prior of ML algorithm, algorithm training process, and optimization and evaluation method. Factors influencing reporting bias include incomplete reporting, selective reporting, and non-standard reporting (e.g. only report point estimate without standard errors or confidence intervals). Based on these aspects, risk of bias judgment of included studies will be ranked as low risk, moderate risk, high risk or unclear. (P16 L19-30 and P17 L1-5)

• Once the study types that will be included are declared it may be possible to identify and use risk of bias tools currently available

– if this is not feasible, rationale must be provided for why these pre-established tools are not appropriate

Reply:

Thank you for your comments. We agree with the reviewer's suggestions. We followed the advice of reviewer 2 and checked OSMIN criteria and Cochrane ROB. The OSMIN checklist can be applied for Systematic reviews of measurement properties, Measurement instrument selection, Identification of the need for further research measurement properties, Designing a study on measurement properties, Reporting a study on measurement properties, and Reviewing the quality of a submitted manuscript on measurement properties. Cochrane ROB is suited for randomized studies.

In addition, we checked another existing guideline in the medical field, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement. TRIPOD is the most suitable guideline for this case because the majority of the items in the checklist are suited for machine learning prediction methods, and the TRIPOD statement is widely used in the medical field. Therefore, we will apply TRIPOD for assessing risk of bias.

To account for machine learning aspects, we added specific items including input data selection, model performance, and outcome reporting.

We will define three rates, 0 as not mentioned or unclear (high risk), 1 as mentioned but not with details (moderate risk), 2 as mentioned with details (low risk). An example of design details is presented at the end of this file, Appendix 2 and Appendix 3.

BEFORE:

There was no established criteria or tool to assess the risk of bias for ML prediction algorithm research. Therefore, we will develop and utilize a customized risk of bias assessment tool. They will be evaluated in three main aspects including input data selection, model performance, and outcome reporting. Factors influencing input data selection include database sponsorship (e.g., organization or single study data) and image/video quality (e.g., Dpi of video, camera setting). Factors influencing model performance include research team (i.e. whether there is a professional computer scientist or clinician), innate prior of ML algorithm, algorithm training process, and optimization and evaluation method. Factors introducing reporting bias include incomplete reporting, selective reporting, non-standard reporting (e.g. only report point estimate without standard errors or confidence intervals). Based on these aspects, risk of bias judgment of included studies will be ranked as low risk, moderate risk, high risk or unclear. In order to better assess the quality of included studies, we plan to compare reported items in these studies with established and recommended reported items according to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement in medical field.[30]

AFTER:

There is no established criteria or tool to assess the risk of bias for ML prediction algorithm research. In order to better assess the quality of included studies, we plan to compare reported items in these studies with established and recommended reported items according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement for the medical field.[35] To account for machine learning aspects, we will develop and include a project-specific risk of bias appendix as a complementary assessment tool. Three main aspects, including input data selection, model performance, and outcome reporting, will be evaluated. Factors influencing input data selection include database sponsorship (e.g., organization or single study data) and image/video quality (e.g., Dpi of video, camera setting). Factors influencing model performance include research team (e.g., whether there is a professional computer scientist, clinician or nurse), innate prior of ML algorithm, algorithm training process, and optimization and evaluation method. Factors influencing reporting bias include incomplete reporting, selective reporting, and non-standard reporting (e.g. only report point estimate without standard errors or confidence intervals). Based on these aspects, risk of bias judgment of included studies will be ranked as low risk, moderate risk, high risk or unclear. (P16 L19-30 and P17 L1-5)

Reviewer: 2

Reviewer Name: Erik Loeffen

Institution and Country: University of Groningen, University Medical Center Groningen, Beatrix Children's Hospital, Groningen, the Netherlands

Please state any competing interests or state 'None declared': None declared.

I think the authors have asked an important question and I applaud research into pain measurement in children. Very interesting topic this machine learning is! I should note that I am no expert on machine learning so please take this into account when interpreting my comments (I do however know a fair bit about systematic reviews and pain measurement in children).

The manuscript is interesting but does need some work, especially the search strategy needs substantial improvement, before I can recommend it for acceptance in BMJ Open.

During reviewing I have written comments down, I've grouped these in major and minor remarks.

MAJOR REMARKS

Comment 1

- English is suboptimal in various places throughout the manuscript, please have a native speaker revise the manuscript. For example: "By far, there is no research has quantitatively and systematically summarized and compared the performance of these ML methods.". 'By far' is a strange way to start this sentence (I think authors mean 'To date'), and between 'research' and 'has' a word is missing (that?).

Reply:

Thank you for your comments. We appreciate the reviewer's recommendation and have asked a native English-speaking colleague Dr. Dana T Turner to assist us to improve the quality of the English throughout our manuscript.

Comment 2

- Should there be a place for CINAHL in the databases? A lot of measurement studies are indexed there.

Reply:

Thank you for your advice. We ran a preliminary search in CINAHL, and screening did not reveal unique citations meeting our inclusion criteria. We invited a medical librarian, Lisa Philpotts, BSN, MSLS, from the Treadwell Virtual Library at Massachusetts General Hospital to help us develop an updated search strategy. Ms. Philpotts is a professional medical librarian who has experience creating search strategies for systematic reviews and meta-analysis. She agreed that the search could benefit from the addition of another database. At Ms. Philpotts's suggestion, we have added Embase, another major biomedical database, to the list of databases to be searched.

BEFORE:

 # Structured Abstract

We will search three major public electronic medical and computer science databases including Web of Science, PubMed, IEEE Xplore Digital Library from 2008 January to 2018 December.

 # Search strategy

We will conduct a systematic global search strategy on three major public electronic medical and computer science databases including Web of Science, PubMed, IEEE Xplore Digital Library from 2008 January to 2018 December (10 years).

AFTER:

 # Structured Abstract

We will search four major public electronic medical and computer science databases including Web of Science, PubMed, Embase, and IEEE Xplore Digital Library from 2008 January to present. (P3 L14-16)

 # Search strategy

An experienced medical librarian (L.P.) with systematic review expertise will conduct searches in four major public electronic medical and computer science databases including Web of Science, PubMed, Embase, and IEEE Xplore Digital Library from 2008 January to present. (P9 L18-20)

Comment 3

- Throughout the manuscript authors only mention that 'pain' is measured. However, in infants it is not possible to distinguish between pain and stress during procedures. This should be mentioned somewhere and discussed upon.

Reply:

Thank you for suggestion. We agree with the reviewer that it is difficult to distinguish pain and stress during procedures in infants in clinical practice. Infants' responses to pain and stress are quite nonspecific to human eyes and can easily be misinterpreted. There were several instruments available specifically designed for assessing stress. For instance, the Newborn Individualized Developmental Care and Assessment Program (NIDCAP) is a widely used instrument for infant stress assessment. [Als H. A synactive model of neonatal behavioral organization: framework for the assessment of neurobehavioral development in the premature infant and for support of infants and parents in the neonatal intensive care environment. Physical & Occupational Therapy in Pediatrics. 1986 Jan 1;6(3-4):3-53.] In addition, Holsti et al attempted to use multidimensional assessments, including the full NIDCAP to distinguish pain and stress. [Holsti L, Grunau RE, Oberlander TF, Whitfield MF, Weinberg J. Body movements: an important additional factor in discriminating pain from stress in preterm infants. The Clinical journal of pain. 2005;21(6):491.] These instruments demonstrated the possibilities to distinguish infant pain and stress but are both labor intensive and expensive to implement, which made understanding how pain and stress are currently assessed is a highly specialized practice area. [Holsti L, Grunau RE. Extremity movements help occupational therapists identify stress responses in preterm infants in the neonatal intensive care unit: A systematic review. Canadian Journal of Occupational Therapy. 2007 Jun;74(3):183-94.] Recent ML research showed that machine algorithms might be the future to address this issue. Mansor et al provided a Probalistic Neural network (PNN) classifier to distinguish pain from other non-pain tasks (rest/cry, air puff, friction) in Classification of Pain Expressions (COPE) database and showed remarkable results with accuracy higher than 90%.[Mansor MN, Junoh AK, Ahmed A, Osman MK. Single Scale Self Quotient Image and PNN for infant pain detection. In2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014) 2014 Nov 28 (pp. 553-555). IEEE.] It indicated that machine learning methods have the potential to distinguish pain and stress combined with other clinical factors and this is a central component of our research. For these reasons, we believe we will

not include the term of “stress” in our searching strategy as we are interested in research predicting infant pain within a very specifically defined scale, which is achievable concept in ML research.

BEFORE:

We did not discuss infant pain and stress.

AFTER:

Infants’ response to pain and stress are nonspecific and can be misinterpreted. The Newborn Individualized Developmental Care and Assessment Program (NIDCAP®) is widely used for infant stress assessment.[39] Holsti et al attempted to use multidimensional assessments, including the full NIDCAP to distinguish pain and stress.[40] However, these instruments are both labor intensive and expensive to implement. In addition, understanding how pain and stress can affect infant development and how they are currently assessed is a highly specialized practice area.[41] Mansor et al provided a PNN classifier to distinguish pain from other non-pain tasks (rest/cry, air puff, friction) in the Classification of Pain Expressions (COPE) database and showed a remarkable results with accuracy higher than 90%.[26] Machine learning methods have the potential to distinguish pain and stress combined with other clinical directors. (P18 L25-30 and P19 L1-4)

Comment 4

- The search strategy needs work: 1) it is not comprehensive, e.g. there are missing terms in especially the infant string (e.g. newborn, baby, etc) and pain string (e.g. agony, hurt, etc), 2) asterisks should be used, i.e. studies that have “infants” in the abstract will now be missed as the strategy says ‘infant’ instead of ‘infant*’, 3) why only search in tiab? Keywords/mesh terms are then missed.

Reply:

Thank you for your advice. We agree with the reviewer’s points, and reviewer one had the same feedback. We consulted with a medical librarian to update the search, and multiple additional synonyms have been added, truncation using wildcards has been incorporated, and controlled vocabulary fields have been searched in addition to the free text title and abstract fields. The machine learning concept in particular has been greatly expanded.

BEFORE:

The search strategy only included title and abstract terms, did not utilize wildcards, and had insufficient synonyms.

AFTER:

The search strategies for all of the databases have been edited significantly in order to run a more comprehensive search. (P10-12 Table 2)

Comment 5

- I would encourage authors to provide an example risk of bias extraction sheet, so we can see how the studies will be scored, although the approach sounds OK, I am having difficulties visualizing when a study will be scored low, moderate, or high risk. (also, can existing bias judgements be partially used? E.g. COSMIN criteria, Cochrane ROB?

Reply:

Thank you for your comments. We followed the reviewer’s advice and checked OSMIN criteria and Cochrane ROB. The OSMIN checklist can be applied for Systematic reviews of measurement

properties, Measurement instrument selection, Identification of the need for further research measurement properties, Designing a study on measurement properties, Reporting a study on measurement properties, and Reviewing the quality of a submitted manuscript on measurement properties. Cochrane ROB is suited for randomized studies.

In addition, we checked another existing guideline in the medical field, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement. TRIPOD is the most suitable guideline for this case because the majority of the items in the checklist are suited for machine learning prediction methods, and the TRIPOD statement is widely used in medical fields. Therefore, we will apply TRIPOD for assessing risk of bias.

To account for machine learning aspects, we added specific items including input data selection, model performance, and outcome reporting.

We will define three rates, 0 as not mentioned or unclear (high risk), 1 as mentioned but not with details (moderate risk), 2 as mentioned with details (low risk). An example of design details is presented at the end of this file, Appendix 2 and Appendix 3.

BEFORE:

There was no established criteria or tool to assess the risk of bias for ML prediction algorithm research. Therefore, we will develop and utilize a customized risk of bias assessment tool. They will be evaluated in three main aspects including input data selection, model performance, and outcome reporting. Factors influencing input data selection include database sponsorship (e.g., organization or single study data), and image/video quality (e.g., Dpi of video, camera setting). Factors influencing model performance include research team (i.e. whether there is a professional computer scientist or clinician), innate prior of ML algorithm, algorithm training process, and optimization and evaluation method. Factors introducing reporting bias include incomplete reporting, selective reporting, non-standard reporting (e.g. only report point estimate without standard errors or confidence intervals). Based on these aspects, risk of bias judgment of included studies will be ranked as low risk, moderate risk, high risk or unclear. In order to better assess the quality of included studies, we plan to compare reported items in these studies with established and recommended reported items according to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement in medical field.[30]

AFTER:

There is no established criteria or tool to assess the risk of bias for ML prediction algorithm research. In order to better assess the quality of included studies, we plan to compare reported items in these studies with established and recommended reported items according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement for the medical field.[35] To account for machine learning aspects, we will develop and include a project-specific risk of bias appendix as a complementary assessment tool. Three main aspects, including input data selection, model performance, and outcome reporting, will be evaluated. Factors influencing input data selection include database sponsorship (e.g., organization or single study data) and image/video quality (e.g., Dpi of video, camera setting). Factors influencing model performance include research team (e.g., whether there is a professional computer scientist, clinician or nurse), innate prior of ML algorithm, algorithm training process, and optimization and evaluation method. Factors influencing reporting bias include incomplete reporting, selective reporting, and non-standard reporting (e.g. only report point estimate without standard errors or confidence intervals). Based on these aspects, risk of bias judgment of included studies will be ranked as low risk, moderate risk, high risk or unclear. (P16 L19-30 and P17 L1-5)

MINOR REMARKS

INTRODUCTION

Comment 6

- “relief treatments are rarely provided to infants and neonates undergoing painful procedures” – That statement is too bold, especially since it has no reference. Sucrose is for example often provided to neonates during painful procedures.

Reply:

Thank you for the suggestion. We agree with your points. Now we softened our language about this statement, and we referenced related studies to support our statement. References are listed as below.

1. Carbajal R, Rousset A, Danan C, et al. Epidemiology and treatment of painful procedures in neonates in intensive care units. JAMA. 2008; 300:60–70.

2. Johnston CC, Fernandes AM, Campbell-Yeo M. Pain in neonates is different. Pain. 2011;152:S65–73.

3. Byrd PJ, Gonzales I, Parsons V. Exploring barriers to pain management in newborn intensive care units: a pilot survey of NICU nurses. Advances in Neonatal Care. 2009; 9:299–306.

4. Cong X, Delaney C, Vazquez V. Neonatal nurses' perceptions of pain assessment and management in NICUs: a national survey. Advances in Neonatal Care. 2013; 13:353–60.

5. Pillai Riddell RR, Stevens BJ, McKeever P, et al. Chronic pain in hospitalized infants: health professionals' perspectives. Journal of Pain. 2009; 10:1217–25.

BEFORE:

Relief treatments are rarely provided to infants and neonates undergoing painful procedures despite these treatments are provided to elder children and adults undergoing similar painful procedures every day.

AFTER:

Relief treatments are inadequately provided to infants and neonates undergoing painful procedures despite the fact that these treatments are provided to older children and adults undergoing similar painful procedures every day.[3-7] (P5 L6-8)

METHODS

Comment 7

- Population not described in sufficient detail: all types of pain? Procedural? Post OR? Nociceptive? Etc

Reply:

Thank you for your comments. We took the reviewer’s advice and described in sufficient detail. The population in our study is defined as infants undergoing all types of painful procedures. Pain can be heel stick, arterial puncture, intravenous cannula, finger stick, nasal aspiration, and post operation pain, etc.

BEFORE:

We did not describe the population in sufficient detail.

AFTER:

The population of our study will be infants undergoing pain. Infants will be defined as young children no more than 12 months, including newborn or neonate, full term, premature, and post-mature infants. (P9 L7-9)

Comment 8

- Idem dito for control: which assessment instruments would authors consider a sufficiently solid tool? Or any tool?

Reply:

Thank you for your comments. We agree with the reviewer's points and clarified assessment instruments. There are several tools that can be used for infant acute pain and procedural pain assessment, including PIPP, N- PASS, NIPS, the CRIES scale, and EVENDOL scale, etc. Nurses and pediatricians may choose different tools in clinical settings. In our review, we will not limit the included studies to one single tool. Any instruments used for infant pain assessment are allowed in the review.

Comment 9

- "To account for publication bias, we will include qualitative review articles, systematic review and meta-analysis articles for missing unpublished literatures through their reference lists." – How does this account for publication bias? Do authors mean 'to identify missing studies?'. Also, will authors perform forwards and backwards citation screening?

Reply:

Thank you for your comments. Yes, we meant that we will include qualitative review articles, systematic review and meta-analysis articles to identify missing studies. We will perform forwards and backwards citation screening through citations and reference lists of systematic reviews to find more relevant papers. We will also search the arXiv database to identify preprint studies and unpublished literature.

Comment 10

- Which societies (conference meeting abstracts)? Be precise.

Reply:

Thank you for your comments. We followed the reviewer's suggestion, and now we described the societies more clearly.

BEFORE:

We will also search related professional meeting abstracts and preprints (e.g. IEEE conferences, pain conferences, arXiv.org)

AFTER:

We will also search related professional meeting abstracts and preprints (e.g. IEEE conferences, Conference on Computer Vision and Pattern Recognition, Conference on Neural Information Processing Systems, Topics and Advances in Pediatrics, Florida Academy of Pain Medicine, 2018 and 2019 Annual Scientific Meeting, arXiv.org). (P9 L22-26)

Comment 11

- Will the review/screening process be piloted?

Reply:

Thank you for your comment. Yes, we will pilot the searching process by screening the first ten studies under the primary investigator's supervision. We will discuss any problems or questions that arise and develop solutions before screening all studies. If the reviewer were talking about the preliminary identified research results as a pilot, we are glad that we already found around ten studies eligible for our analyses from a very rough preliminary search (PubMed, google scholar search), previous experience (our adult pain meta-analysis exclusions) and our readings in this topic (e.g. review article).

BEFORE:

We did not describe about piloted screen

AFTER:

We will pilot the process by screening the first ten studies under the primary investigator's supervision. (P12 L7-8)

Comment 12

- Subgroup analysis: which medical procedure types will be distinguished?

Reply:

Thank you for your comments. We accepted the reviewer's suggestion and described the subgroup analysis of medical procedures types that may be involved. The medical procedure of infant pain can be acute procedural pain and postoperative pain.

BEFORE:

We did not describe the procedure types in detail.

AFTER:

medical procedure type (e.g. acute procedural pain vs postoperative pain). (P17 L15)

Comment 13

- Do authors also take things as costs / resources into account when interpreting data? I can imagine these are costly machines / software programs. Perhaps can authors expand upon this in the discussion?

Reply:

Thank you for your suggestion. We agree with the reviewer's point that taking things such as costs/resources into account is useful. We will include computational efficiency, which is one aspect of resources, as one of the secondary outcomes. To the best of our knowledge, computer scientists tend not to disclose the costs of their studies.[Sikka K, Sharma G, Bartlett M. Lomo: Latent ordinal model for facial analysis in videos. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016 (pp. 5580-5589).] [Chen Z, Ansari R, Wilkie D. Automated Pain Detection

from Facial Expressions using FACS: A Review. arXiv preprint arXiv:1811.07988. 2018 Nov 13.] Therefore, it might be difficult for us to assess costs in machine learning studies. However, we hope that, in the future, computer scientists will be able to disclose the costs of studies to provide references for the health-care budget. We consulted machine learning scientists about the costs of machines and software programs and learned that it is free to use machine learning platforms such as tensorflow or pytorch, while the prices are costly for the computer hardware such as GPU. We took the reviewer's advice and expanded upon costs and resources in the **Discussion** section.

BEFORE:

We did not discuss costs and resources of machine learning studies.

AFTER:

To the best of our knowledge, it's not common for computer scientists to clearly disclose their costs and courses in their studies.[42,43] Therefore, it is difficult for us to evaluate costs in machine learning studies. However, we hope that, in the future, computer scientists will be able to disclose the costs of studies in order to provide references for the health-care budget. (P19 L17-20)

VERSION 2 – REVIEW

REVIEWER	Laure Perrier University of Toronto, Canada
REVIEW RETURNED	20-Aug-2019

GENERAL COMMENTS	Current State of Science in Machine Learning Methods for Automatic Infant Pain Evaluation using Facial Expression Information: Study Protocol for A Systematic Review and Meta-analysis The authors have done excellent work to provide clarity on a number of issues previously identified. There are a few items that remain outstanding that would strengthen the manuscript and these are outlined below: Eligibility criteria • Clearly identify the study types to be included in the review. They are listed in your PROSPERO registration but do not appear in your protocol. Search strategy • Provide rationale for the date limits of 2008-present Study selection • Clarify full-text screening – will this be done by two reviewers independently? • Page 13, Line 25-27: States that information will be extracted "for qualitative synthesis". The Objectives and Data Extraction section indicate data extraction will include data that will be used for quantitative synthesis – consider harmonizing this information.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Laure Perrier

Institution and Country: University of Toronto, Canada
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
Current State of Science in Machine Learning Methods for Automatic Infant Pain Evaluation using Facial Expression Information: Study Protocol for A Systematic Review and Meta-analysis

The authors have done excellent work to provide clarity on a number of issues previously identified. There are a few items that remain outstanding that would strengthen the manuscript and these are outlined below:

Eligibility criteria
• Clearly identify the study types to be included in the review. They are listed in your PROSPERO registration but do not appear in your protocol.

Reply:

Thank you for your comments. Now we clarify and list the study types eligible for inclusion in our revised manuscript. We intend to include computer science algorithms paper (methodology and performance evaluation), clinical research (application studies), systematic reviews and meta-analysis in this topic. Regular reviews and qualitative studies will be excluded. SR and meta-analysis will be used for extraction of citations for reducing publication bias.

BEFORE:

#inclusion:

We did not clearly identify and list the study types eligible for inclusion in our protocol.

#exclusion:

(3) not a quantitative prediction model study concerning ML methods;

AFTER:

#inclusion:

We intend to include computer science algorithms paper (methodology and performance evaluation), clinical research (application studies), systematic reviews and meta-analysis in this topic. Regular reviews and qualitative studies will be excluded. SR and meta-analysis will be used for the extraction of citations for reducing publication bias. (P9 L11-14)

#exclusion:

(3) not computer science algorithms paper (methodology and performance evaluation), clinical research (application studies), or systematic reviews and meta-analysis concerning ML methods; (P9 L16-18)

Search strategy
• Provide rationale for the date limits of 2008-present

Reply:

Thank you for your comments. We have now included the rationale for the date limits of 2008 to present in our revised manuscript. We choose to search from 2008 onward since our study topic is primarily focused on the methodological aspects of automatic infant pain prediction algorithms. To the best of our knowledge, major advances in machine learning techniques, especially for deep learning methods (e.g. CNN), started to be widely applied and rapidly evolved within the recent five to six years because of the unprecedented functional improvement of hardware (e.g. GPU for computing) and parallel-computing capacity (e.g. Hadoop).[Valstar MF, Almaev T, Girard JM, McKeown G, Mehu M, Yin L, Pantic M, Cohn JF. Fera 2015-second facial expression recognition and analysis challenge. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 2015 May 4 (Vol. 6, pp. 1-8). IEEE.] [Chen Z, Ansari R, Wilkie D. Automated Pain Detection from Facial Expressions using FACS: A Review. arXiv preprint arXiv:1811.07988. 2018 Nov 13.] Thus some outdated algorithms developed ten or twenty years ago may not be applicable or clinically usable, because any modern algorithm may outperform them due to aforementioned technological advances. In our study, we intend to include a balanced collection of both classic and modern algorithms, so we decide to search studies starting from January 2008 to assure an extended ten-year long search period to assure good coverage of studies.

BEFORE:

We did not provide rationale for the date limits of 2008-present in the manuscript.

AFTER:

We choose to search from 2008 onward since our study topic is primarily focused on the methodological aspects of automatic infant pain prediction algorithms. To the best of our knowledge, major advances in machine learning techniques, especially for deep learning methods (e.g. CNN), started to be widely applied and rapidly evolved within the recent five to six years because of the unprecedented functional improvement of hardware (e.g., GPU for computing) and parallel-computing capacity (e.g. Hadoop).[35,36] Thus some outdated algorithms developed ten or twenty years ago may not be applicable or clinically usable, because any modern algorithm may outperform them due to aforementioned technological advances. In our study, we intend to include a balanced collection of both classic and modern algorithms, so we decide to search studies starting from January 2008 to assure an extended ten-year long search period to assure good coverage of studies. (P9 L25-30 and P10 L1-6)

Study selection

• Clarify full-text screening – will this be done by two reviewers independently?

Reply:

Thank you for your comments. Since our study topic is uniquely focused on computer science algorithms on infant pain prediction, it will require expertise on two fields including computer science and medicine. We believe that it will be better and time-efficient to have both the computer scientist and the physician work together to review the full-text to avoid selection bias (i.e. only read the sections they are familiar about) in the full-text screening process. It will be ideal if we have two independent groups of specialists (two computer scientist-physician pairs as review teams), but our research team only have one data scientist (D.L.). Therefore, in our study, two authors D.L. (computer scientist) and D.C.(physician) will screen the full-text together to decide the eligibility of included studies after title and abstract screening round. And the PI (H.D.) will resolve the conflicts or answer questions when needed. Moreover, because of the poor reporting among health-related computer science papers based on our previous experience, it is difficult for single field specialist (either computer scientist or physician) to understand and extract the necessary information alone. We will further discuss this issue in the discussion section of our formal paper.

BEFORE:

Once all the articles screened, we will download the full-text of included studies for further identification.

AFTER:

Since our study topic is uniquely focused on computer science algorithms on infant pain prediction, it will require expertise on two fields including computer science and medicine. We believe that it will be better and time-efficient to have both the computer scientist and the physician work together to review the full-text to avoid selection bias (i.e. only read the sections they are familiar about) in the full-text screening process. It will be ideal if we have two independent groups of specialists (two computer scientist-physician pairs as review teams), but our research team only have one data scientist (D.L.). Therefore, in our study, two authors D.L. (computer scientist) and D.C.(physician) will screen the full-text together to decide the eligibility of included studies after title and abstract screening round. And the PI (H.D.) will resolve the conflicts or answer questions when needed. Moreover, because of the poor reporting among health-related computer science papers based on our previous experience, it is difficult for single field specialist (either computer scientist or physician) to understand and extract the necessary information alone. We will further discuss this issue in the discussion section of our formal paper. (P12 L8-10 and P13 L1-11)

• Page 13, Line 25-27: States that information will be extracted “for qualitative synthesis”. The Objectives and Data Extraction section indicate data extraction will include data that will be used for quantitative synthesis – consider harmonizing this information.

Reply:

Thank you for your comments. Now we make the states consistent in the revised manuscript that we will extract information for data synthesis. The measures in the identified studies contain both quantitative and qualitative data, thus we will perform quantitative data synthesis or qualitative summary reports respectively

BEFORE:

Once eligible studies are identified, we will extract information for qualitative synthesis.

AFTER:

Once eligible studies are identified, we will extract information for data synthesis. The measures in the identified studies contain both quantitative and qualitative data, thus we will perform quantitative data synthesis or qualitative summary reports respectively. (P13 L12-14)