PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Diagnostic Reliability in Teledermatology: A Systematic Review and a Meta-Analysis
AUTHORS	Bourkas, Adrienn; Barone, Natasha; Bourkas, Matthew; Mannarino, Matthew; Fraser, Robert; Lorincz, Amy; Wang, Sheila; Ramirez-GarciaLuna, Jose

VERSION 1 – REVIEW

REVIEWER	Nicholas, Matilda Duke University School of Medicine, Dermatology
REVIEW RETURNED	16-Sep-2022

GENERAL COMMENTS	Worthwhile article and well done meta analysis that could benefit from some editing to decrease length.

REVIEWER	Choi, Ellie National University Hospital, Dermatology
REVIEW RETURNED	18-Sep-2022

GENERAL COMMENTS	This is an overall well written and conducted systematic review on the diagnostic concordance of teledermatology. It provides a good summary of a very relevant topic, which has not been done recently. A few major points for the authors consideration: 1. Teledermatology as authors described is divided into live video conferencing, store and forward and hybrid. However after the introduction/methods, this segretation is no longer mentioned. I dont see (on a quick screen) any live videoconferencing studies being included. Life videoconference and S&F would likely perform very differently and should be discussed and analysed separately (at least as a subgroup analysis). 2. There is a lot of data and information, and there are multiple subgroup analyses/long tables/compact figures. This can make the paper cognitively quite heavy to read, even though the actual word count may not be too great. Text wise, I feel that the subgroup analyses are important, so i would suggest for authors to considering trimming out the other less important descriptive statistics in the main text (e.g. page 9 lines 19-35 on study methodology and country can probably be excluded or trimmed to 1-2 lines, page 9 lines 40-19 can also be removed/significantly shortened)

	Might be good if authors can further rationalise the number of tables/figures in the main manuscript and shift out more to the supplementary files. I'm not sure all the main tables are really required, or if table 1A, 1B, 2 could perhaps be combined/compacted 3. I think the authors need to provide some element of the statistical analysis in the main text. I'm not familiar but is it appropriate to pool kappa values the same way as simple proportions? Minor comments page 9 lines 33-35. For this primary outcome, I would consider adding in the range for the % agreement and concordance (in addition to the CI). page 9 lines 45-48 "Diagnostic agreement rate between TDs, F2F dermatologists, and TD vs histopathology..." - this line is not clear to the reader and could read a few ways (although one would eventually understand after reading the whole paragraph). The comparison could easily read as the disagreement between TD and histopath, F2F and histopath, and again TD and histopath. - i feel that TD vs histopath involves a different nature of comparison (vs between TDs and between F2Fs). I think this point may be better in its own or the earlier paragraph, leaving this just for inter rater reliability within the same modality page 11 lines 6-8. would suggest to remove 'current gold standard' as it begets the question of why F2F evaluation by a dermatologist is considered gold standard. - would also suggest to shorten these 2 paragraphs on page 11. The main point on whether there is a significant difference and % agreement is diluted and lost int he sea of text. page 12. "diagnostic reliability.. by training involved and type of technology" - this paragraph seems to just be a subgroup by training involved, type of technology seems irrelevant? meanwhile, under "other sub-group analysis", "image capturing technologies" is unclear and one would need to refer to the supplementary figure to understanding. This can be elaborated in the text, e.g. no difference was seen by device used for photography (e.g. digital camera, phone....) I feel that a subgroup analysis between dermoscopic vs clinical images would be useful. I'm not sure if this was done but i could not find the efigure. The subgroup comparison for skin cancers vs non-skin cancers/inflammatory conditions could also be elaborated in text. I think that these two are important comparisons readers might want to know, and have bearings on implementation policies. The last half of the discussion is easy to read and informative. The first half, on page 13 and 14 on the other hand is harder to read. I think the points intended in each paragraph need to be stressed and emphasized instead of spending most of the paragraph repeating the results.

	I believe it should be confounders instead of 'confounds"? - page 4 "the potential for confounds across TD...." - page 5 "to control for potential confounds"

REVIEWER	Vestergaard, T. Odense University Hospital
REVIEW RETURNED	18-Oct-2022

GENERAL COMMENTS	Thank you for a thorough review comparing TD to F2F for dermatological conditions. I have the following comments: Please write more on data analysis and statistics. What analyses did you choose and why? Which programme did you use? Page 9, line 19: "44 papers that were included. Forty-one (93%) of the included studies were observational, of which 31 (76%) were prospective, nine (22%) were retrospective. One (2%) study was ambispective. Three studies were randomized controlled trials and one study was a quasi-randomized trial." 31+9+1+3+1=45? Page 11, line 8: "Twenty-seven studies (62%) included in this analysis were inclusive to all types of dermatoses, 13 (29%) studies looked specifically at suspicious lesions, and three (6.8%) studies excluded skin cancers completely." 27+13+3=43? I completely agree with you, that proper image acquisition and standardized referrals are essential for the success of TD. I would like to question, though, whether agreement on primary diagnosis is a relevant measure at all. I recognise, that many studies (including my own) have used this outcome, making it available to comparison and metaanalysis. But from a clinical perspective, I believe agreement on the management plan between TD and FTF is much more relevant. TD will often be used as a triage tool to differentiate mild/benign cases from severe/malignant/uncertain cases. If you agree, please elaborate a little bit on this in the discussion.

REVIEWER	Noertjojo, Kukuh WorkSafe BC, Clinical Services
REVIEW RETURNED	31-Jan-2023

GENERAL COMMENTS	I appreciate very much your attempt to investigate this relevant topic particularly when we move more toward telehealth practices. My biggest concern with your paper is the fact of very high and statistically significant heterogeneity in almost everyone of the meta-analysis you did on which the results affecting the conclusion you made. I am conservative and will not advocate for meta-analysis in such an event (last week I attended a Cochrane Method group training on meta-analysis in case of very few study. One of the conclusion that the trainer told us was that in the event of large heterogeneity especially if it was unexplained one NOT to do meta-analysis and just report it qualitatively. I tend to hold this view as well hence I am suggesting to the Editor to get opinion of meta-analysis methodology expert.

	Further, I'd like to suggest that you provide the Quadas-2 (and perhaps also Quadas-C) summary sheet in the paper.

REVIEWER	Ong, Mei-Sing Harvard University
REVIEW RETURNED	12-Apr-2023

GENERAL COMMENTS	1. Literature search strategy was not specified in the manuscript. The authors stated the terms used for searching the International Prospective Register of Systematic Reviews and OSF. Were the same terms used to screen for articles included in this study? How were the terms combined? Please be specific about the methods used for literature search. There should be enough details so that the study can be replicated by others. 2. There were a number of exclusion criteria, described in eTable1. These criteria should be described in the main text and appropriate justification should be provided for these criteria. E.g. Why exclude studies where patients captured their own photographs, when this is likely the most common practice? 3. Studies included in the review spanned many countries, with vastly different demographics and access to advanced teledermatology technology. How generalizable is this review? Could this have explained the heterogeneity in the results of the studies? 4. Recent advances in artificial intelligence has had a huge impact on teledermatology. Can the authors comment on this? 5. Please clarify the table header on pages 61-67. 6. Use of non-standard abbreviations is distracting. E.g. MA for meta-analysis, TD for teledermatology. I suggest spelling out the terms instead. Many acronyms were also used without specifying what they were. E.g. TD1, TD2, TD3, Derm, Histo, publ.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Matilda Nicholas, Duke University School of Medicine

Comments to the Author:
Worthwhile article and well done meta analysis that could benefit from some editing to decrease length.

- Did the authors assess whether TD by dermatologists improves primary care physician accuracy? This is because that's the main reason for having that kind of service available. If not (I'm aware the number of studies looking at the performance of primary care physicians was small), that should be acknowledged in the discussion.

- Response: Thank you for reviewing our manuscript and for your insightful comments. While the included studies did not directly assess the impact of TD on primary care physician accuracy, we agree that this is an important consideration for the effectiveness of TD services. One article by Costello et al 2019, reported high levels of satisfaction among PCPs with the teleconsultation process, with a 96% agreement rate that they learned about dermatologic diagnosis through the teleconsultation and a 100% agreement rate that the teleconsultation helped patient care. We have added a discussion of these points to our manuscript. Please see lines 268-273 for the new changes.

Editor(s)' Comments to Author:
- Please minimize the use of abbreviations in your Abstract and ensure all abbreviations are defined at first use.

- Response: We have made the appropriate changes to address the Reviewer's concerns. Please see line 25 for the new changes.

- We noted that you state that your systematic review has been performed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Please be aware that the PRISMA guidelines relate to the reporting of systematic reviews, NOT to their design. Please revise this sentence accordingly. Please also note that PRISMA-P is for the reporting of systematic review protocols, not the results paper.

- Response: We clarified the guidelines we followed in our study. In the Method section, it now reads: "This study was performed in accordance with the Preferred Reported Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.
Protocol Registration
Prior to the conduct of this review, a protocol which adhered to the PRISMA-protocols (i.e., PRISMA-P) guidelines was developed and then registered on Open Science Framework. Access: https://osf.io/fjdvg."

- Please include, as a supplementary file, the precise, full search strategy (or strategies) for all databases, registers and websites, including any filters and limits used.

- Response: The search strategies have been included in the supplementary materials. Please see lines 24-229 in the supplementary appendix for the new changes.

- Please ensure that you have fully discussed the methodological limitations of the study in the Discussion section of the main text.

- Response: We updated the limitations in our discussion section to include 1) We mainly included observational studies, so no real conclusions of whether TD is superior to F2F can be drawn. 2) The data is highly heterogeneous which lowers the generalizability of our findings. However, this can also be a strength as it is representative of most reported studies, a wide population, and a wide range of lesions. 3) Database search was limited, and we did not include any Grey Literature sources. Thus, there is a risk that we may be missing reports mainly from the developing world. Please see lines 326-354 for the new changes in the main text.
- Following Reviewer 2 and 3's comments, we acknowledge the absence of a clear gold standard. While the agreement rate for the treatment plan would be an ideal standard, it was seldom reported in the papers included in the systematic review, and thus, not discussed in our paper. Furthermore, it falls outside the scope of our current findings.

Reviewer: 2
Dr. Ellie Choi, National University Hospital

Comments to the Author:
This is an overall well written and conducted systematic review on the diagnostic concordance of teledermatology. It provides a good summary of a very relevant topic, which has not been done recently.

A few major points for the authors consideration:

1. Teledermatology as authors described is divided into live video conferencing, store and forward and hybrid. However after the introduction/methods, this segretation is no longer mentioned. I dont see (on a quick screen) any live videoconferencing studies being included. Life videoconference and S&F would likely perform very differently and should be discussed and analysed separately (at least as a subgroup analysis).
- Response: We thank the reviewer for her thorough feedback. Regarding the suggestion to analyze and discuss live video conferencing and store-and-forward separately, we have included a discussion of the potential differences between the two methods, taking into account the available evidence from the three studies that included live video conferencing. We have also acknowledged the limitations of the available studies and proposed avenues for future research. Please see lines 339-346 for the new changes.

2. There is a lot of data and information, and there are multiple subgroup analyses/long tables/compact figures. This can make the paper cognitively quite heavy to read, even though the actual word count may not be too great.
- To address this concern, we have implemented several changes:
- Firstly, we have updated the results section by reducing the text for descriptive statistics and referring readers to eTable 1 (we moved Table 1 to the supplementary appendix to adhere to the journal formatting guidelines and renamed it as eTable 1) for the relevant information.
- Additionally, we have merged Tables 1 and 2, eliminating redundancy (new title is eTable 1).
- We also condensed the information in eTable 4 in the supplementary appendix.
- Lastly, we have transferred one significant subgroup analysis on technology devices from the supplementary material to the main text to ensure its prominence. These modifications aim to improve the focus of the paper while still highlighting the significant results. Please see lines 237-248 in the main text for the new changes.
- We also moved subgroup analyses on TD vs TD, F2F vs F2F and TD vs histo to the supplementary and moved subgroup analysis on imaging device to main text following one of the reviewer's comments. Please see lines 270-284 in the supplementary appendix for the new changes.

Text wise, I feel that the subgroup analyses are important, so i would suggest for authors to considering trimming out the other less important descriptive statistics in the main text (e.g. page 9 lines 19-35 on study methodology and country can probably be excluded or trimmed to 1-2 lines, page 9 lines 40-19 can also be removed/significantly shortened)
- Response: We removed lines with descriptive statistics on pages 9 and 10.
- Response: We also moved the technology device subgroup analysis from supplementary to the results (main text) because of the statistically significant findings. Please see lines 237-243 in the main text for the new changes.

Might be good if authors can further rationalise the number of tables/figures in the main manuscript and shift out more to the supplementary files. I'm not sure all the main tables are really required, or if table 1A, 1B, 2 could perhaps be combined/compacted
- We have addressed the issue by merging Tables 1 and 2 to eliminate redundancy. We moved the merged table to the supplementary appendix to adhere to the journal formatting guidelines and renamed it as eTable 1.
- Additionally, we have removed sections of the paper that duplicated information already present in eTable 4. These changes streamline the presentation of data.

3. I think the authors need to provide some element of the statistical analysis in the main text. I'm not familiar but is it appropriate to pool kappa values the same way as simple proportions?
- Response: We acknowledge that pooling kappa values in the same manner as simple proportions is not appropriate. Unlike proportions, which range from 0 to 1, kappa values are a special type of

correlation that range from -1 to 1. However, negative values of kappa are generally considered unlikely to occur (Biochem Med (Zagreb). 2012 Oct; 22(3): 276–282).
- Therefore, when pooling kappa values, it is essential to consider the co-variation among them. In our study, we addressed this issue by utilizing point-biserial correlations to assess the effect of the continuous kappa variable in relation to the dichotomous explanatory variables (British Journal of Mathematical and Statistical Psychology 202;73:113–44).
- We have revised and incorporated these points in the Statistical Methods section accordingly, and moved key points from the supplementary to the main text. Please see lines 176-188 for the new additions in the main manuscript.

Minor comments

-page 9 lines 33-35. For this primary outcome, I would consider adding in the range for the % agreement and concordance (in addition to the CI).
- Response: We have added the range to all appropriate statistics. Please see line 215 in the main text for the new changes.

-page 9 lines 45-48 "Diagnostic agreement rate between TDs, F2F dermatologists, and TD vs histopathology..."
- this line is not clear to the reader and could read a few ways (although one would eventually understand after reading the whole paragraph). The comparison could easily read as the disagreement between TD and histopath, F2F and histopath, and again TD and histopath.
- Response: We have made the appropriate changes to address these concerns. Please see lines 270-271 in the supplementary appendix for the new changes.

- i feel that TD vs histopath involves a different nature of comparison (vs between TDs and between F2Fs). I think this point may be better in its own or the earlier paragraph, leaving this just for inter rater reliability within the same modality
- We have taken your feedback into consideration and made revisions to the paragraph structure to better align the comparison between TDs and histopathology with the comparisons between TDs and between F2F dermatologists and PCPs. We moved this information to the supplementary section along with the earlier paragraph on TD vs TD, F2F vs F2F and TD vs Histo for better flow. Please see lines 270-284 in the supplementary appendix for the new changes.

-page 11 lines 6-8. would suggest to remove 'current gold standard' as it begets the question of why F2F evaluation by a dermatologist is considered gold standard.
- Response: We have proceeded with appropriate changes to address concerns. Please see line 259 in the supplementary appendix for the new change.

- would also suggest to shorten these 2 paragraphs on page 11. The main point on whether there is a significant difference and % agreement is diluted and lost in the sea of text.
- Response: We have proceeded with appropriate changes to address concerns. We moved this section to the supplementary and refer to lines 260-268 in the supplementary appendix for the new changes.

-page 12. "diagnostic reliability.. by training involved and type of technology"
- this paragraph seems to just be a subgroup by training involved, type of technology seems irrelevant?
- Response: We have proceeded with appropriate changes to address concerns. Please see line 228 in the main text for the new change.

-meanwhile, under "other sub-group analysis", "image capturing technologies" is unclear and one would need to refer to the supplementary figure to understanding. This can be elaborated in the text, e.g. no difference was seen by device used for photography (e.g. digital camera, phone....)
- Response: We proceeded with appropriate changes to address concerns. Please see new paragraph (lines 228-235) in the main text for the new changes. We also moved the figure form the supplementary to the main text (Figure 3).

-I feel that a subgroup analysis between dermoscopic vs clinical images would be useful. I'm not sure if this was done but i could not find the efigure. The subgroup comparison for skin cancers vs non-skin cancers/inflammatory conditions could also be elaborated in text. I think that these two are important comparisons readers might want to know, and have bearings on implementation policies.
- We apologize for not clearly mentioning the subgroup analysis comparing dermoscopic and clinical images in the main text. We understand the concern about length limitations, which is why this analysis was included in the supplementary materials. We revised the main text (under the heading “Other sub-group analyses”) to at least acknowledge the existence of this analysis and provide a reference to the corresponding supplementary material. This will ensure that readers are aware of its inclusion and can access the detailed information if needed. Please see lines 250-254 in the main text for the new changes.

-The last half of the discussion is easy to read and informative. The first half, on page 13 and 14 on the other hand is harder to read. I think the points intended in each paragraph need to be stressed and emphasized instead of spending most of the paragraph repeating the results.
- Response: We thank the reviewer for this helpful feedback, and we proceeded with appropriate changes to address concerns. The first half of discussion was reworked to highlight the main findings. Please see lines 264-283 in the main text for the new changes.

-I believe it should be confounders instead of 'confounds"?
- page 4 "the potential for confounds across TD...."
- page 5 "to control for potential confounds"
- Response: We have proceeded with appropriate changes to address concerns. Please see lines 79, 95 in the main text for the new changes.

Reviewer: 3
T. Vestergaard, Odense University Hospital

Comments to the Author:
Dear authors,

Thank you for a thorough review comparing TD to F2F for dermatological conditions.
I have the following comments:

Please write more on data analysis and statistics. What analyses did you choose and why?
- Response: Thank you for reviewing our manuscript and for your feedback. In the supplementary methods section, we have included the following information on statistics: For the percentage of agreement, meta-analyses were performed using the aggregated data. Proportions were calculated as the ratio of the number of reported agreement to the total sample size. Proportions are provided with the corresponding 95 percent confidence intervals. For the kappa values, point-biserial correlations were calculated and aggregated. In both cases, combined with the inspection of forest plots, the I2 index and the τ2 statistic were used to investigate statistical heterogeneity. Given the high degree of heterogeneity noted, the authors proceeded with a random-effects model for overall

complications with a logit transformation. Publication bias was not statistically pursued given the substantial heterogeneity in addition to the authors choosing to pursue a meta-analysis of proportions.
- We also expanded on how agreement rates and Cohen's kappa concordances were treated as individual and independent values, while proportions were aggregated and analyzed using meta-analyses with corresponding confidence intervals. We discuss utilization of point-biserial correlations to obtain kappa values, considering the continuous nature of the kappa variable and the dichotomous explanatory variables.
- We moved key points from the supplementary on the statistics section to the main text. Please see lines 171-183 in the main text, and lines 241-256 in the supplementary appendix for the new changes.

Which programme did you use?
- Response: The supplementary methods section also includes the following additional information on statistics: Statistical analysis was performed using the dmetar package in R v.4.0.1 (R Foundation for Statistical Computing, 2022). Please see lines 172-173 in the main text for the new changes.

Page 9, line 19: "44 papers that were included. Forty-one (93%) of the included studies were observational, of which 31 (76%) were prospective, nine (22%) were retrospective. One (2%) study was ambispective. Three studies were randomized controlled trials and one study was a quasi-randomized trial." 31+9+1+3+1=45?
- Response: We apologize for the mistake in our original statement. Upon review, we have found that there were actually 40 observational studies, not 41. We apologize for any confusion caused. Please see revised statement lines 197-200 in the main text, Due to limitations in the available space and to better adhere to journal formatting guidelines, we moved the merged Table 1+2 to the supplementary appendix and renamed it as eTable 1.

Page 11, line 8: "Twenty-seven studies (62%) included in this analysis were inclusive to all types of dermatoses, 13 (29%) studies looked specifically at suspicious lesions, and three (6.8%) studies excluded skin cancers completely." 27+13+3=43?
- Response: We apologize again for the mistake in our original statement. Upon review, we have found that there were actually 28 studies that were inclusive to all types of dermatoses, not 27. We apologize for any confusion caused. We have made the necessary corrections in our revised manuscript, please see lines 207-209 for new changes.

I completely agree with you, that proper image acquisition and standardized referrals are essential for the success of TD.

I would like to question, though, whether agreement on primary diagnosis is a relevant measure at all. I recognise, that many studies (including my own) have used this outcome, making it available to comparison and metaanalysis. But from a clinical perspective, I believe agreement on the management plan between TD and FTF is much more relevant. TD will often be used as a triage tool to differentiate mild/benign cases from severe/malignant/uncertain cases. If you agree, please elaborate a little bit on this in the discussion.
- Response: We thank the reviewer for this helpful feedback and we proceeded with appropriate changes to address concerns. We included a brief paragraph in the discussion. Please see lines 281-286 in the main text for the new changes.

Best regards,
Tine Vestergaard

Reviewer: 4
Dr. Kukuh Noertjojo, WorkSafe BC

Comments to the Author:
I appreciate very much your attempt to investigate this relevant topic particularly when we move more toward telehealth practices.
My biggest concern with your paper is the fact of very high and statistically significant heterogeneity in almost everyone of the meta-analysis you did on which the results affecting the conclusion you made. I am conservative and will not advocate for meta-analysis in such an event (last week I attended a Cochrane Method group training on meta-analysis in case of very few study. One of the conclusion that the trainer told us was that in the event of large heterogeneity especially if it was unexplained one NOT to do meta-analysis and just report it qualitatively. I tend to hold this view as well hence I am suggesting to the Editor to get opinion of meta-analysis methodology expert.
***Comment from the Editor: Although there's some heterogeneity (and the concordance rates vary tremendously between studies, which suggest that), the authors already readily acknowledge that in the discussion, so we support keeping the meta-analyses.
- Response: Thank you for your feedback and for the comment regarding the heterogeneity in our study. Following the Editor's and rest of the Reviewers endorsement, we are proceeding with the analysis of pooled data. Please see lines 323-335 in the discussion section and lines 61-62 under "Article Summary" regarding heterogeneity.

Further, I'd like to suggest that you provide the Quadas-2 (and perhaps also Quadas-C) summary sheet in the paper.
- Response: Thank you for your feedback. We would like to clarify that we have already utilized the Quadas-2 tool for our ROB analysis in the initial version of our paper. However, following your suggestion, we have made an additional update. To address your input, we have now included a summary table (eTable 6A) in the supplementary material that highlights our tailored approach to the ROB analysis, incorporating the Quadas-2 tool. We hope that this will provide readers with a clear and concise overview of our methodology and the assessment of potential original studies biases.
- After careful consideration, we have decided not to incorporate QUADAS-C into our study. While we recognize the potential value of QUADAS-C in certain contexts, we believe that our current methodology, which utilizes QUADAS-2, adequately addresses the risk of bias assessment for our research question. As such, we have chosen to maintain consistency within our study by solely relying on QUADAS-2.

Reviewer: 5
Dr. Mei-Sing Ong, Harvard University

Comments to the Author:
1. Literature search strategy was not specified in the manuscript. The authors stated the terms used for searching the International Prospective Register of Systematic Reviews and OSF. Were the same terms used to screen for articles included in this study? How were the terms combined? Please be specific about the methods used for literature search. There should be enough details so that the study can be replicated by others.
Response: We thank the reviewer for the helpful feedback. We addressed the search strategy feedback earlier. In the revised manuscript, we included full search strategies for all databases in the supplementary, INCLUSION and EXCLUSION criteria are also discussed in the supplementary. Search strategies can be found in the supplementary appendix, lines 72-277. The INLCUSION / EXCLUSION criteria are available in eTable 2.

2. There were a number of exclusion criteria, described in eTable1. These criteria should be described in the main text and appropriate justification should be provided for these criteria. E.g. Why

exclude studies where patients captured their own photographs, when this is likely the most common practice?
- Response: We have now included the key exclusion criteria in the main text. We acknowledge the concern regarding the exclusion of studies where patients captured their own photographs, as it is a common practice. However, we chose to exclude such studies to ensure consistent and controlled image quality. The reliability and quality of patient-captured images may vary significantly due to differences in lighting, focus, and other factors. By excluding these studies, we aimed to allow for a more accurate comparison of diagnostic reliability between TD and F2F methods. We hope this clarification addresses the reviewer's concern, and we have incorporated this justification in the revised paragraph.
- Please see lines 129-135 for the new changes in the main text.
- eTable 1 is now eTable 2 due to other formatting changes.

3. Studies included in the review spanned many countries, with vastly different demographics and access to advanced teledermatology technology. How generalizable is this review? Could this have explained the heterogeneity in the results of the studies?
- Response: We agree that the studies in our review encompass diverse demographics and teledermatology technologies. However, rather than viewing this as a limitation, we see it as a strength that enhances the generalizability of our findings. This diversity reflects real-world scenarios and underscores the adaptability of teledermatology across different contexts.
- We've included detailed patient and study characteristics in eTable 1 for context, despite some suggestions to remove certain information. We acknowledge that not all studies reported comprehensive information on their resources, limiting our analysis. However, we tried to control for these factors through sub-group analysis. We also address the generalizability and potential limitations of our findings in an expanded discussion section. Please see lines 326-337 for the new changes in the main text.
- We also wanted to note that studies did not consistently report information on the tools and resources available to them, which may limit the comprehensiveness of our analysis. However, attempts to control for these confounding factors were done through sub-group analysis of the TD technique.
4. Recent advances in artificial intelligence has had a huge impact on teledermatology. Can the authors comment on this?
- Response: We appreciate the reviewer's interest in the impact of recent advances in AI on teledermatology. Indeed, AI’s application in teledermatology has been promising, particularly in the areas of image recognition and diagnosis. However, our article primarily focuses on the diagnostic reliability of telemedicine in dermatology, specifically comparing the outcomes of store-and-forward teledermatology and video-based consultations to traditional face-to-face diagnoses. Therefore, we believe that adding an additional layer of complexity in our study by including AI studies is beyond our current scope. We believe that discussing AI in teledermatology warrants a separate, dedicated investigation that thoroughly explores its capabilities, limitations, and implications for the future of dermatological care.
- However, we have addressed this concern in the Discussion section as a perspective for future studies. Please see lines 361-365 in the main text for the new changes.

5. Please clarify the table header on pages 61-67.
- Response: We have proceeded with appropriate changes in the supplementary appendix to address these concerns. Please see new header for eTable 4 of the supplementary appendix.

6. Use of non-standard abbreviations is distracting. E.g. MA for meta-analysis, TD for teledermatology. I suggest spelling out the terms instead. Many acronyms were also used without specifying what they were. E.g. TD1, TD2, TD3, Derm, Histo, publ.

- Response: We revised abbreviations and included definitions where appropriate. We have made several updates to improve the readability of our figures and table legends. We have removed the abbreviations "MA" and "publ." Furthermore, we have incorporated abbreviation descriptions within the legends to ensure better understanding. In order to accommodate space constraints and maintain a visually appealing format, we have retained the abbreviation "TD" (referring to teledermatology or teledermatologist) in both figures and tables. Additionally, we have revised the main text to explicitly spell out "teledermatologist" or “teledermatology” instead of using the TD abbreviation for greater precision.

VERSION 2 – REVIEW

REVIEWER	Choi, Ellie National University Hospital, Dermatology
REVIEW RETURNED	02-Jul-2023

GENERAL COMMENTS	Thank you to the authors for their revision, most of the comments have been appropriately addressed. I have a few additional minor comments 1. “Our sub-group analyses revealed that agreement rates between teledermatology consultations and F2F physicians were significantly higher when dermatologists conducted in-person assessments compared to non-specialists.” - I think these results need to be mentioned in the main text if it is going to be mentioned in the discussion. - Also, was there a test to see if the pooled kappa of 0.69 (CI 0.6-0.75) is significantly higher than 0.52 (CI 0.26-0.71)? Or does this only refer to the pooled % accuracy and not the agreement. - Can I also confirm that the F2F doctor in this context refers to the patient facing physician at the point of referral? I can imagine this causing confusion as F2F physician can also mean the doctor providing the gold standard diagnosis. Consider using the term referring physician instead. 2. High heterogeneity in outcome measures - I imagine the studies would vary quite significantly in how they define a diagnosis being in agreement. Some may perhaps require the exact same diagnosis (seb k, bcc, tinea pedis), other studies may group diagnoses into looser categories (e.g. benign lesions, skin cancer, fungal infections). Do authors feel that this is a potential reason for the widely varying accuracy reported? (As low as 13% as high as 98%) 3. Supplementary file “Publication bias was not statistically pursued due to the substantial heterogeneity observed, in addition to the authors' decision to pursue a meta-analysis of proportions.” - I get that heterogeneity might limit assess of publication bias, but I’m not sure if that’s a sufficient reason to not assess for it at all (what about using other methods). - Also I’m not sure how not pursuing publication bias is contingent on the authors decision to do a meta-analysis of proportions

REVIEWER	Vestergaard, T. Odense University Hospital

REVIEW RETURNED	03-Jul-2023

GENERAL COMMENTS	Thank you for this revised manuscript. I have a few comments: Lines 197-199: "Forty of the included studies were observational, of which 31 were prospective, nine were retrospective. One study was ambispective." 31+9+1=41? Figure 2 is missing, it is a duplicate of figure 3. eFigure 4: percentage agreement is missing, only Kappa values presented. eTable 4: Some of the vertical headings cannot be read as half of the letters are missing.

REVIEWER	Noertjojo, Kukuh WorkSafe BC, Clinical Services
REVIEW RETURNED	07-Jul-2023

GENERAL COMMENTS	I am happy that you addressed almost all of reviewers concerns/suggestions almost completely

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2 - Dr. Ellie Choi, National University Hospital

Comments to the Author:

Thank you to the authors for their revision, most of the comments have been appropriately addressed. I have a few additional minor comments

1. “Our sub-group analyses revealed that agreement rates between teledermatology consultations and F2F physicians were significantly higher when dermatologists conducted in-person assessments compared to non-specialists.”

- I think these results need to be mentioned in the main text if it is going to be mentioned in the discussion.

o We appreciate your suggestion and have incorporated a brief summary of the sub-group analysis results into the main body of the text. Due to figure number limitations, we opted to keep the corresponding figures and additional text in the supplementary appendix.

 New paragraph in main manuscript: “Teledermatologists' 70.96% agreement rate with F2F dermatologists significantly exceeded the 44.1% rate from non-specialists ($p < 0.001$). Non-specialists consistently showed lower diagnostic concordance across studies; see supplementary appendix and eFigure 3 for further details.”

- Also, was there a test to see if the pooled kappa of 0.69 (CI 0.6-0.75) is significantly higher than 0.52 (CI 0.26-0.71)? Or does this only refer to the pooled % accuracy and not the agreement.

o We did conduct meta-regressions to assess group differences in all sub-group analyses, following the methodology outlined by Morton SC et al. Thus, we have performed formal tests to determine if the pooled kappa of 0.69 (CI 0.6-0.75) is significantly higher than 0.52 (CI 0.26-0.71), and the corresponding p-values have been provided.

 To improve clarity, we added the following to the Data Analysis and Synthesis section in the supplementary appendix: “To evaluate the statistical significance of differences between kappa values, we performed meta-regressions and derived corresponding p-values.

- Can I also confirm that the F2F doctor in this context refers to the patient facing physician at the point of referral? I can imagine this causing confusion as F2F physician can also mean the doctor providing the gold standard diagnosis. Consider using the term referring physician instead.

o We understand the potential confusion surrounding the term "F2F physician". However, in our study, the term 'Face-to-Face (F2F) physician' refers to healthcare professionals in the comparison group, who conducted in-person assessments only. We would like to note that these assessments could occur concurrently or in a sequential manner, depending on the case. This term does not exclusively refer to the referring physicians.

 To clarify to our readers the definition of F2F control group, we included the following paragraph in the ‘Eligibility criteria’ section under ‘Methods’: “In this context, an 'F2F physician' refers to healthcare professionals, such as dermatologists, general practitioners, or emergency department physicians, who conducted in-person assessments only. This term is used to represent the comparison group in our analyses, and these assessments may occur concurrently or sequentially with teledermatology consultations, depending on the case.”

2. High heterogeneity in outcome measures

- I imagine the studies would vary quite significantly in how they define a diagnosis being in agreement. Some may perhaps require the exact same diagnosis (seb k, bcc, tinea pedis), other studies may group diagnoses into looser categories (e.g. benign lesions, skin cancer, fungal infections). Do authors feel that this is a potential reason for the widely varying accuracy reported? (As low as 13% as high as 98%)

o Thank you for your insightful observation. We agree that varying definitions of diagnostic agreement across studies could contribute to the wide accuracy range. We sought to mitigate this by selecting studies that reported on primary diagnoses rather than grouping conditions into broad categories. Despite this, variability in study designs and clinician expertise likely contributed to the heterogeneity.

o Furthermore, while we conducted subgroup analysis for disease categories like inflammatory skin conditions, skin cancers, and other dermatoses, no statistically significant trends emerged. This highlights that factors beyond disease categorization may affect reported accuracies.

3. Supplementary file

“Publication bias was not statistically pursued due to the substantial heterogeneity observed, in addition to the authors' decision to pursue a meta-analysis of proportions.”

- I get that heterogeneity might limit assess of publication bias, but I'm not sure if that's a sufficient reason to not assess for it at all (what about using other methods). Also I'm not sure how not pursuing publication bias is contingent on the authors decision to do a meta-analysis of proportions

o We appreciate your query on publication bias. Given the inherent nature of proportional data, standard assessments, such as the I^2 statistic, are often employed to estimate heterogeneity in meta-analyses of proportions, despite being originally designed for comparative data. In these analyses, high I^2 values are an expected outcome due to the independent variance observed in proportional data, unrelated to sample sizes. In addition, variations in the geographical and temporal dimensions of the studies contribute to the heterogeneity seen in prevalence and incidence estimates. Thus, a high I^2 should not be construed as indicative of inconsistency in this context. Regarding publication bias, tests such as Egger's test and funnel plots, developed for comparative data, may not be apt due to their reliance on the assumption that studies yielding 'positive' results are published more frequently. In a meta-analysis of proportions, this assumption becomes less clear due to the absence of a universally accepted definition of a 'positive' result. Moreover, there is no conclusive evidence suggesting that proportional data can be effectively adjusted or suited for these tests. Considering these factors, our study opted for a qualitative assessment of publication bias within the framework of a meta-analysis of proportions. This approach was guided by the specific attributes of proportional data and our commitment to delivering the most accurate and trustworthy results.

 To further clarify to our readers, we included the following paragraph in the 'Data Analysis and Synthesis' section in the supplementary: "Given the unique properties of proportional data and the considerable heterogeneity observed, conventional publication bias tests, specifically designed for comparative data, were not considered applicable. As such, statistical pursuit of publication bias was not undertaken. Instead, a methodologically appropriate qualitative assessment of publication bias was implemented for this type of analysis. This approach was deemed to provide the most accurate and robust outcome."

o We hope this response and the additional text in the supplementary addresses your concern and clarifies our methodology.

Reviewer: 3 - T. Vestergaard, Odense University Hospital

Comments to the Author:

Dear authors,

Thank you for this revised manuscript. I have a few comments:

1. Lines 197-199: "Forty of the included studies were observational, of which 31 were prospective, nine were retrospective. One study was ambispective." 31+9+1=41?

o We apologize for the confusion in our previous text. We corrected the text and now it reads: "eTable 1 summarizes the study and participant characteristics for the 44 included papers. Forty-one of the included studies were observational, of which 32 were prospective, eight were retrospective. One study was ambispective."

2. Figure 2 is missing, it is a duplicate of figure 3.

o Thank you for bringing this to our attention, we apologize for the mistake. Figure 2 was inadvertently missing, and Figure 3 was duplicated in its place. We will promptly upload the correct Figure 2 to the portal.

3. eFigure 4: percentage agreement is missing, only Kappa values presented.

o Thank you for bringing this to our attention, we apologize, part A of the figure was accidentally deleted after our earlier revisions. We added the missing figure back.

4. eTable 4: Some of the vertical headings cannot be read as half of the letters are missing.

o Thank you for bringing this to our attention, we reformatted the headings of the table to correct this issue.

Reviewer: 4 - Dr. Kukuh Noertjojo, WorkSafe BC

Comments to the Author: I am happy that you addressed almost all of reviewers concerns/suggestions almost completely.

VERSION 3 – REVIEW

REVIEWER	Choi, Ellie National University Hospital, Dermatology
REVIEW RETURNED	15-Jul-2023

GENERAL COMMENTS	Thank you to the authors for the clear and detailed explanation, I have no further to add.