PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Comparison of handheld rebound tonometry with Goldmann applanation tonometry in children with glaucoma: a cohort study
AUTHORS	Dahlmann-Noor, Annegret ; Bunce, Catey; Puertas, Renata; Tabasa-Lim, Shenille; El-Karmouty, Ahmed; Kadhim, Mustafa; Wride, Nicholas; Lewis, Amanda; Grosvenor, Dawn; Rai, Poornima; Papadopoulos, Maria; brookes, john; Khaw, Peng

VERSION 1 - REVIEW

REVIEWER	Sharon F. Freedman MD Professor of Ophthalmology and Pediatrics Duke Eye Center
REVIEW RETURNED	30-Aug-2012

GENERAL COMMENTS	The authors present the results of a prospective study comparing Icare rebound and Goldmann applanation tonometry in 102 pediatric glaucoma patients, and conclude that Icare does not correlate well with Goldmann IOP measurement, but that “normal” Icare readings are likely to be accurate and may spare the child an EUA. Icare IOP was found to frequently and significantly measure higher IOP than Goldmann applanation. 1. Introduction - the authors state that the gold standard IOP technique under anesthesia is Perkins tonometry, but this reviewer would submit that in the US, tonopen is more often used as the standard technique. 2. Introduction, page 4, 1st three lines – this sentence does not make grammatical sense as written. Similarly, the secondary objective would seem to be the influence of CCT ON IOP, but not the influence of IOP AND CCT? 3. Protocol – when using the Icare, did the investigators include readings with poor reliability (e.g., flashing P and TOP line), or only those of good reliability (Psolid, Pflashing with Bottom line, Pflashing with Middle line and below 21 mmHg)? Was there any provision for when the two sequential Icare readings were not within a certain number of mm Hg of one another? 4. Results – what does “the variability reading on the RBT was high” mean? What was the threshold for exclusion? Similarly, what does “11 children with high SD scores” mean? How is this different from the “variability reading on the RBT was high”? 5. Results – page six, line 11 – what does “LOA” mean? Need to write out “limits of agreement” the first time this is used, and in the table. Please also explain the terminology used in the following sentence (the *3.19 and the *7.94)? 6. Results – Pachymetry Data- since so few children had CCT above the “normal range”, it would be nice to know if these children also had nystagmus, corneal abnormalities, aphakia, Haab striae, or any other features which might confound the association of higher CCT with greater disagreement between RBT and GAT? How was this

	disagreement measured? 7. Figure 1D – why does the y-axis say “density” rather than Number of eyes? Why did the authors choose to divide the age groups into two for this graph, and what is the significance of that data? 8. Discussion – first line – did the authors seek to examine any other possible determinants of poor correlation between Icare and GAT IOP, such as nystagmus, aphakia, etc.? 9. Discussion – second paragraph – the authors suggest that regarding with caution those RBT readings taken on an anxious or squeezing child, or repeating RBT until readings with low variability are obtained – might not be the usual practice in a busy clinic. But the reviewer feels that this is vital for IOP measurement with any technique, and RBT is no exception. Surely this is necessary with GAT and also Tonopen, which are the standard methods in children when RBT is not available or used. 10. Table 1 is somewhat interesting but represents only a small number of practitioners, and could be omitted from the manuscript if space is at a premium.

REVIEWER	Richard A. Russell Post-doctoral Research Fellow Department of Optometry and Visual Science, City University London, UK.
REVIEW RETURNED	26-Oct-2012

GENERAL COMMENTS	According to the Methods on Page 4, the sample size was determined by recommendations for Bland-Altman analysis in reference 18. Having read this paper, I cannot see any reference to sample size required for Bland-Altman analysis. Could the authors provide a different reference – I know Bland typically recommends 100-200 on his homepage (http://www-users.york.ac.uk/~mb55/meas/sizemeth.htm). For the primary analysis, 74 children were included, which appears adequate in this example. The inclusion criteria were children with an established diagnosis of glaucoma, yet the age range was up to 19 years old. I would suggest the authors only include ages up to and not including 16 or 18 years old if they are to refer to their patients as “children”. Furthermore, on Page 5 the authors state that (under the header “Analysis of GAT data”) two children were excluded as glaucoma suspects yet these should have been excluded from the start… Could the others clarify please? In the abstract the authors should define “EUA”. It is not uncommon for researchers to examine agreement between two devices where one device (tonometer) is tested against another but the observers are different for each device. I wonder if the authors could comment on this in their Discussion since there is a potential for systematic bias between Observer 1 (who took the RBT measurement) and Observer 3 (who took the GAT measurement)… I.e., there is the possibility (although I do not think this is the case) that inflated IOP readings from RBT are due to human error rather than instrument error. It may be beneficial to discuss this as a limitation in the Conclusion.

	Important comment: on Page 6, the authors describe their methods for analysing the agreement between GAT and RBT – they correctly identify evidence for a relationship between the magnitude of difference between methods and the magnitude of the measurement. They state that due to this “limits of agreement cannot be simply produced” so they stratify the data: above and below 21 mmHg. I do not see how this stratification helps since the data both above and below this threshold still look to have a relationship between the magnitude of difference between methods and the magnitude of the measurement. Additionally, visual inspection of Figure 1C suggests that the sample size of the two groups is about 54 for data with an IOP below 21 mmHg, and only 20 for the group with IOP above 21 mmHg. The authors should attempt to log the data and back-transform as a function of the mean in order to remove the relationship (or an alternative transformation). I certainly think that their current stratification has not helped, and has significantly reduced the sample size for the higher IOP group. The limits of agreement should be plotted on the Bland-Altman figures.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

1. Introduction - the authors state that the gold standard IOP technique under anesthesia is Perkins tonometry, but this reviewer would submit that in the US, tonopen is more often used as the standard technique.

We thank the reviewer for this comment. Whilst the Reichert Tono-PenR may be more commonly used in the US, the current reference standard worldwide for tonometry is the Goldmann applanation tonometer (ISO 8612) mounted on a slit lamp. The Perkins tonometer, as the hand-held version of GAT, is a suitable reference standard for IOP measurements for patients in the supine position, and hence for children during an examination under anaesthesia.

We have changed the manuscript accordingly.

2. Introduction, page 4, 1st three lines - this sentence does not make grammatical sense as written. Similarly, the secondary objective would seem to be the influence of CCT ON IOP, but not the influence of IOP AND CCT?

Thank you for pointing this out; we have corrected the manuscript.

3. Protocol - when using the Icare, did the investigators include readings with poor reliability (e.g., flashing P and TOP line), or only those of good reliability (Psolid, Pflashing with Bottom line, Pflashing with Middle line and below 21 mmHg)? Was there any provision for when the two sequential Icare readings were not within a certain number of mm Hg of one another?

We may have to clarify the terminology here: flashing P indicates a standard deviation greater than normal, a line in the up position indicates that the standard deviation of the measurements is high and a new measurement is recommended (manufacturer’s manual). We would call these, indices of “poor reliability”, or “high variability”.

We recorded whether a line was absent (solid P), or in High, Middle, or Low Position. We excluded measurements from 16 participants in whom the line was in the High position (see “Results/Analysis”.

We did not make a provision for sequential iCare measurements of certain difference.

4. Results - what does "the variability reading on the RBT was high" mean? What was the threshold for exclusion? Similarly, what does "11 children with high SD scores" mean? How is this different from the "variability reading on the RBT was high"?

We apologise for the lack of clarity. By “high variability reading” we mean a flashing P or TOP line in

the “up” position. This was the case in 16 participants (page 5, line 10), of which 11 occurred in readings by observer 1 (page 5, line 15), and in 6 children in readings by both observer 1 and 2 (page 5, line 18/19). We have clarified the terminology in the manuscript.

5. Results - page six, line 11 - what does "LOA" mean? Need to write out "limits of agreement" the first time this is used, and in the table. Please also explain the terminology used in the following sentence (the *3.19 and the *7.94)?
Thank you, we have made the suggested changes. We have also provided figures for our limits of agreement rather than arithmetic computations and apologise for this oversight.

6. Results - Pachymetry Data- since so few children had CCT above the "normal range", it would be nice to know if these children also had nystagmus, corneal abnormalities, aphakia, Haab striae, or any other features which might confound the association of higher CCT with greater disagreement between RBT and GAT? How was this disagreement measured?
Of the thirteen children with CCT > 650um, one had corneal clouding/oedema and enlarged cornea, and in two children “multiple previous glaucoma interventions” were recorded. None of these children had nystagmus, aphakic glaucoma, aniridia or anterior segment dysgenesis syndromes, or other confounding corneal abnormalities. By "disagreement" we mean the absolute difference between IOPs obtained using GAT and RBT, we have modified our paper to make this more explicit.

7. Figure 1D - why does the y-axis say "density" rather than Number of eyes? Why did the authors choose to divide the age groups into two for this graph, and what is the significance of that data?
Figure 1D is a histogram, which plots density rather than numbers. In statistical terminology, the height of a rectangle on the histogram is equal to the “frequency density of the interval”, i.e., the frequency divided by the width of the interval.
We divided the age groups into 5-9 years and 10 years and over based on normative data published by Hussein et al (Corneal thickness in children, AJO, 138(5):744-8; reference now inserted), who reported a mean corneal thickness of 566 um (SD 48) in the younger and 554 um (SD 35) in the older age group.
The significance of this data is that only few children in our cohort had CCT outside the “normal” range, and that whilst disagreement between RBT and GAT increased with higher CCT measurements, CCT was not the only factor contributing to disagreement between methods.
8. Discussion - first line - did the authors seek to examine any other possible determinants of poor correlation between Icare and GAT IOP, such as nystagmus, aphakia, etc.?
Only three children in this cohort had aphakic glaucoma, of whom one also had nystagmus. One additional child had nystagmus. We did not attempt to analyse these cases separately.

9. Discussion - second paragraph - the authors suggest that regarding with caution those RBT readings taken on an anxious or squeezing child, or repeating RBT until readings with low variability are obtained - might not be the usual practice in a busy clinic. But the reviewer feels that this is vital for IOP measurement with any technique, and RBT is no exception. Surely this is necessary with GAT and also Tonopen, which are the standard methods in children when RBT is not available or used.
We agree that anxiety and lid squeezing affect all tonometry methods, and have removed the reference to busy clinics.

10. Table 1 is somewhat interesting but represents only a small number of practitioners, and could be omitted from the manuscript if space is at a premium.
Thank you for this comment; we would like to leave this decision to the Editor.

Reviewer 2
According to the Methods on Page 4, the sample size was determined by recommendations for

Bland-Altman analysis in reference 18. Having read this paper, I cannot see any reference to sample size required for Bland-Altman analysis. Could the authors provide a different reference - I know Bland typically recommends 100-200 on his homepage (http://www-users.york.ac.uk/~mb55/meas/sizemeth.htm). For the primary analysis, 74 children were included, which appears adequate in this example.
Thank you, we have provided a different reference.

The inclusion criteria were children with an established diagnosis of glaucoma, yet the age range was up to 19 years old. I would suggest the authors only include ages up to and not including 16 or 18 years old if they are to refer to their patients as "children". Furthermore, on Page 5 the authors state that (under the header "Analysis of GAT data") two children were excluded as glaucoma suspects yet these should have been excluded from the start... Could the others clarify please?
We have modified our inclusion criteria. We actually wished to study all attendees of a paediatric glaucoma clinic which included some subjects over 17 and glaucoma suspects.

In the abstract the authors should define "EUA".
Thank you, we have inserted examination under anaesthesia.

It is not uncommon for researchers to examine agreement between two devices where one device (tonometer) is tested against another but the observers are different for each device. I wonder if the authors could comment on this in their Discussion since there is a potential for systematic bias between Observer 1 (who took the RBT measurement) and Observer 3 (who took the GAT measurement)... I.e., there is the possibility (although I do not think this is the case) that inflated IOP readings from RBT are due to human error rather than instrument error. It may be beneficial to discuss this as a limitation in the Conclusion.
We realise that when comparing methods of measurement, disagreement can be introduced by the method itself and by the observer (if the measurement is subjective). We did consider minimising this by using the same observer to use both methods however we felt that this would give unrealistic measurements due to recall bias. We felt that the comparison between observer 1 and observer 2 showed that inflated readings from RBT were consistent between observers and not simply due to one inaccurate observer and that since all observers were trained IOP assessors, the possibility of one rogue assessor leading to the findings was as the reviewer has concluded unlikely.

Important comment: on Page 6, the authors describe their methods for analysing the agreement between GAT and RBT - they correctly identify evidence for a relationship between the magnitude of difference between methods and the magnitude of the measurement. They state that due to this "limits of agreement cannot be simply produced" so they stratify the data: above and below 21 mmHg. I do not see how this stratification helps since the data both above and below this threshold still look to have a relationship between the magnitude of difference between methods and the magnitude of the measurement. Additionally, visual inspection of Figure 1C suggests that the sample size of the two groups is about 54 for data with an IOP below 21 mmHg, and only 20 for the group with IOP above 21 mmHg. The authors should attempt to log the data and back-transform as a function of the mean in order to remove the relationship (or an alternative transformation). I certainly think that their current stratification has not helped, and has significantly reduced the sample size for the higher IOP group.
Stratification removed statistical evidence of a linear relationship between the difference and magnitude of measurements although the reviewer is quite correct to comment on the fact that such a stratification resulted in smaller numbers within each strata. Unfortunately logging the data did not remedy the non normality which is why this approach was not taken. The main rationale behind stratifying was to demonstrate how different limits of agreement were when a cut-off was used. We believe we should urge caution in applying these based on the fact that the numbers are small.

The limits of agreement should be plotted on the Bland-Altman figures.
The Bland Altman diagram has been computed in order to assess the assumptions behind the limits of agreement. We agree with the reviewer that some caution needs to be applied in using these figures which is why we did not include them on the plots.

VERSION 2 – REVIEW

REVIEWER	Sharon F. Freedman MD Professor of Ophthalmology and Pediatrics Duke Eye Center
REVIEW RETURNED	25-Nov-2012

GENERAL COMMENTS	The revision seems adequate. My remaining question has to do with the use of the term "density" on the y-axis of Figure 1, panel D. This still does not make sense to me.

REVIEWER	Richard A. Russell Post-doctoral Research Fellow Department of Optometry and Visual Science, City University London, UK. I have no competing interests.
REVIEW RETURNED	12-Nov-2012

GENERAL COMMENTS	The authors have addressed all my concerns in the revised manuscript.