CLINICAL AND LABORATORY OBSERVATION

SpO_2 targeting and oximeter changes in the BOOST-II Australia and BOOST-II UK oxygen trials

Benjamin J. Stenson MD[1], Mark Donoghoe PhD[2], Peter Brocklehurst MB ChB[3], Peter G. Davis MD[4], Edmund Juszczak MSc[5], Ian C. Marschner PhD[2,6], John Simes MD[2], William O. Tarnow-Mordi MB ChB[2]

1, Neonatal Unit, Royal Infirmary of Edinburgh, Edinburgh, EH16 4SA, UK

2. NHMRC Clinical Trials Centre, University of Sydney, NSW, 2006, Australia.

3. Birmingham Clinical Trials Unit, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK

4. The Royal Women's Hospital; Department of Obstetrics and Gynecology, The University of Melbourne; Murdoch Childrens Research Institute Melbourne, Australia

5. National Perinatal Epidemiology Unit Clinical Trials Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

6. Department of Statistics, Macquarie University, NSW 2109, Australia

Corresponding author:

Benjamin Stenson, consultant neonatologist, Neonatal Unit, Royal Infirmary of Edinburgh, Little France Crescent, Edinburgh EH16 4SA, UK.

e-mail ben.stenson@luht.scot.nhs.uk

Telephone 0044 7768 426428

Abstract

Infants in the Australian and UK BOOST-II trials using revised oximeters spent more time within their planned SpO_2 target ranges than infants using the original oximeters (P<0.001). This may explain the larger mortality difference seen with revised oximeters. If so, average treatment effects from the NeoPrOM trials may be underestimates. (49 words)

Main text

The five Neonatal Oxygen Prospective Meta-analysis (NeoPrOM) trials evaluated the effects, in extremely preterm infants, of targeting lower (85-89%) versus higher (91-95%) pulse oximeter saturation (SpO_2) on death or disability.[1] Meta-analysis shows that for every 1,000 infants, targeting lower versus higher SpO_2 led to no difference in the primary composite outcome of death or major disability up to 18-24 months, and no difference in major disability, including blindness, but resulted in 28 more deaths, 22 more infants with NEC but 42 fewer infants receiving treatment for ROP.[2]

A challenge in analysing the trials was discovery in 2011 of an artefact in the study oximeters' calibration.[3] Downloaded SpO_2 values from Masimo oximeters in 176 preterm infants in 35 neonatal units had a non-physiological distribution with a large dip between 87 and 90% (figure 1, online). Masimo reported that this reflected their decision to adjust the

calibration of their oximeters so that at values above 87% the displayed SpO_2 was increased by 1-2% [3]. As well as fewer values than expected between 87-90%, this manufacturer-generated artefact returned more SpO_2 values than expected above 90%[3], thus affecting both target groups in the NeoPrOM trials. By elevating SpO_2 readings of 88 and 89% to higher displayed values the artefact would be expected to make the low target group range of 85-89% narrower and harder to target. By elevating SpO_2 values in the range 90-95% by 1-2% above the true value the artefact would mean that actual achieved SpO_2 values in the high target range with the original oximeters were lower than intended, narrowing the difference in SpO_2 between groups.

Although the original oximeters performed within required standards for accuracy, at the investigators' request Masimo provided revised software restoring the expected SpO_2 distribution[3,4] and the oximeters were changed to the revised software in three of the NeOProM trials [4,5]. According to Masimo, oximeters sold since then have incorporated revised software.

The effect of the oximeter revision on targeting accuracy and mortality is debated. In an interim safety analysis of the BOOST-II Australia and UK trials,[6] the relative risk of 36 week mortality was increased by 65% in low-target versus high-target infants using revised oximeters (21.8% vs. 13.3%; $P<0.001$), but was not significantly different in low-target versus high-target infants using original oximeters (test for interaction between results for revised and original oximeters $P = 0.006$). Each trial was stopped early to prevent avoidable deaths that might occur with continuing enrolment.[6]

Whyte et al[7] found no difference in targeting accuracy between original and revised oximeters in the BOOST-II trials[4,8,9], as measured by the difference in the median of the median SpO_2 distributions for the low and high-target groups. They concluded that decisions about optimal

SpO_2 targets should not be restricted to data obtained with revised oximeters.[7] However, an infant's median SpO_2 value does not describe the time they spent in their intended SpO_2 target range, which is a more meaningful measure of targeting accuracy. We now report an infant-specific analysis of time spent in the intended SpO_2 target ranges among infants managed using original and revised oximeters in the BOOST-II Australia and BOOST-II UK trials.

Methods:

For each of the 2096 infants in the BOOST II Australia and UK trials for whom oxygen saturation data were available, we calculated the percentage of all time on the oximeter that the infant spent at each SpO_2 value. We then determined the proportion of time that each infant spent in the low- and high-target ranges. Using these infant-specific proportions, within each randomised treatment group we calculated the difference between the original and revised oximeters in the mean proportion of time spent in each range, stratified by trial. For the low-target range, the net improvement in targeting was defined as the increase in the mean proportion of time spent in that range by infants in the low-target group minus any increase in the time spent in that range by infants in the high-target group. Similarly, the net improvement in targeting of the high-target range was defined as the difference between the increase in time for infants allocated to the high-target group and the increase for infants in the low-target group. The overall improvement in targeting was then defined as the sum of these two improvements.

Because the infant-specific proportions of time spent in each range must lie between 0 and 100%, and because the times in the low- and high-target ranges for a particular infant are not independent, the assumptions required by standard parametric statistical methods do not hold. Therefore, we used non-parametric methods to draw inferences from these data.

To estimate non-parametric 95% confidence intervals for the values of interest, we used bootstrapping[10]: randomly sampling infants from the original data with replacement mimics the process of sampling from the population, and we repeated this 10,000 times. By calculating the targeting improvements in each of these 10,000 resampled datasets, we obtained a distribution of 'bootstrapped' parameter estimates, from which we obtained a plausible range for the true parameter of interest.

Non-parametric two-sided p-values were calculated by conducting permutation tests[11,12]. This approach is based on the fact that we can produce a dataset in which we know the null hypothesis of no software effect is true by randomly shuffling the software labels between infants from the original dataset. We did this 10,000 times, allowing us to estimate the distribution of targeting improvements that could be observed if the null hypothesis were true. The p-value is then the proportion of times that the magnitude of the targeting improvements calculated from the permuted datasets exceeded that observed in the original data.

Results:

SpO_2 data were available for 1128/1135 infants in BOOST-II Australia and for 968/973 infants in BOOST-II UK. Low-target group infants in BOOST-II Australia using revised oximeters spent 5.8% more time in the low SpO_2 range of 85-89% than those using the original oximeters. Similarly, in BOOST-II UK low-target infants using revised oximeters spent 5.3% percent more time in the low SpO2 range of 85-89% than those using the original oximeters. High-target infants using revised oximeters spent 3.3% and 1.7% longer in the low SpO_2 range than high-target infants using original oximeters in the Australia and UK studies respectively, leading to a pooled net improvement of 2.9% (95% confidence interval: 1.4 –

4.4%; p<0.001) in targeting the low range using revised oximeters across the two trials (table 1 and figure 2).

In BOOST-II Australia, high-target infants using revised oximeters spent 3.2% more time in the high SpO_2 range of 91-95% than those using original oximeters, while low-target infants spent 0.1% longer in the high SpO_2 range, giving a net improvement in targeting the high SpO_2 range of 3.1%. In BOOST II UK, high-target infants using revised oximeters spent 4.5% less time in the high range than those using original oximeters, while low-target infants spent 2.2% less time in that range, giving a net deterioration in targeting accuracy of 2.2%. Pooling the data stratified by trial gave a net improvement in targeting the high target range of 1.0% (95% CI: 1.0% deterioration – 3.0% improvement; p = 0.322).

The combined net improvement of proportion of time spent in correct target range across both low- and high-target ranges associated with revised oximeters was 3.9% (95% CI: 1.6 – 6.2%; p < 0.001).

Further results from the individual trials are provided in tables 2 and 3 online. In relative terms, using revised versus original oximeters, low-target infants spent 30-40% more time in the low-target range in the 2 trials.

Discussion:

We have shown that the revision of the oximeter calibration software in the BOOST-II Australia and BOOST-II UK trials improved SpO_2 targeting. This was mainly because the low-target infants on revised oximeters spent longer in their intended SpO_2 range, increasing their exposure to lower SpO_2. This may, at least in part, explain the increased differences in mortality observed between randomized groups using revised oximeters[2] and suggests that

our current assessments of average treatment effects may underestimate the risk of targeting the lower SpO_2 range.

In pooled analyses of the BOOST-II Australia and BOOST-II UK trials, there was not a statistically significant difference in mortality at 2 years between randomisation groups with the original oximeters (p=0.49) [13] but there was a highly statistically significant difference in mortality between groups with the revised oximeters (p=0.001). The pooled mortality rate of the high-target groups decreased by 1.8% after the oximeter revision. The pooled mortality rate of the low-target groups increased by 7.5% after the oximeter revision. The changes in mortality observed following the oximeter revision were therefore largely explained by changes in mortality rates in the low target groups.

The small net improvement in the pooled results in targeting the high SpO_2 target range after the oximeter changes was not statistically significant. It is important to consider that displayed values in the higher target range after the oximeter revision represent higher true SpO_2 readings than similar readings obtained with the original oximeters. The different patterns between the two trials in the high target groups after the revision may be explained by chance

As previously suggested[7] our statistical method used "infant" as the unit of analysis by combining the data for each infant into a summary statistic followed by a comparison between target groups based on the between-infant variation in the summary statistic. Our use of percent time in the target range provides a more relevant metric of targeting accuracy than the median oxygen saturation for each infant previously presented.[7] It will be helpful if similar analyses are undertaken on the data from the other oxygen trials.

We analysed all time on the oximeter rather than time when breathing supplemental oxygen because the trials reported increased mortality with the lower SpO_2 target range. Lower SpO_2

values would be associated with relative hypoxaemia whether or not supplemental oxygen was in use and would not result in oxygen supplementation in infants in the low SpO_2 target groups.

In each trial using revised oximeters (BOOST-II UK[4,8,13], BOOST-II Australia[4,8,13] and the Canadian Oxygen Trial[5]) the observed risk of death was higher in low-target infants, and the pooled mortality rate was statistically significantly greater in low-target infants. A test for interaction showed strong evidence (P=0.009) that the pooled mortality results were different before and after the oximeter change (figure 3, online). The effect of additional centers joining the BOOST-II UK study is unlikely to explain this difference, as no new centers joined the Australian or Canadian trials after the oximeter revision. However, nurses in the three trials may have become better at SpO_2 targeting over time,[2,7] tending to increase the risk of mortality in lower-target infants. If so, this would strengthen the conclusion that spending more time in the lower saturation target increases mortality.

The NeOProM Trials show collectively, and particularly in the infants who were treated after the revision of the trial oximeters, that small differences in achieved SpO_2 distribution can have a significant effect on risk of mortality in extremely preterm infants. Clinicians should understand the performance of the oximeters that they use and this would be facilitated if the manufacturers of these devices published details of the calibration of their instruments.

In summary, changing the trial oximeters improved targeting, exposing low-target infants to more time with lower SpO_2 (P<0.001), which was associated with an increased risk of mortality. The improved targeting was captured by measuring time spent in the intended SpO_2 target ranges, but not by measuring the difference in median SpO_2.[7] Regardless of any effects of changing the oximeters on trial outcomes, the NeoPrOM trials show that aiming for the lower SpO_2 target has no significant effect on death or disability, or disability, or

blindness, but increases mortality[2]. Any proposed trade-off [2] between the benefits (less treatment for ROP, without differences in blindness or disability) and harms (more deaths and more NEC) of the low target should be made clear to parents. This information is already changing clinical practice.[14]

References:

1. Askie LM, Brocklehurst P, Darlow BA, Finer N, Schmidt B, Tarnow-Mordi W for the NeOProM Collaborative Group. NeOProM: Neonatal Oxygenation Prospective Meta-analysis Collaboration study protocol. BMC Pediatrics 2011, 11:6.

2. Askie LM, Darlow BA, Davis PG, Finer N, Stenson B, Vento M et al. Effects of targeting lower versus higher arterial oxygen saturations on death or disability in preterm infants. Cochrane Database Syst Rev 2017;4:CD011190.

3 Johnston ED, Boyle B, Juszczak E, King A, Brocklehurst P, Stenson BJ. Oxygen targeting in preterm infants using the Masimo SET Radical pulse oximeter. Arch Dis Child Fetal Neonatal Ed 2011;96:F429-33.

4. Stenson BJ, Tarnow-Mordi WO, Darlow BA, et al. for the Boost II United Kingdom, Australian and New Zealand Collaborative Groups. Oxygen saturation and outcomes in preterm infants. N Engl J Med 2013;368:2094-104.

5. Schmidt B, Whyte RK, Asztalos EV, et al. Effects of targeting higher vs lower arterial oxygen saturations on death or disability in extremely preterm infants: a randomized clinical trial. JAMA 2013;309:2111-20.

6. Stenson B, Brocklehurst P, Tarnow-Mordi W. Increased 36-week survival with high oxygen saturation target in extremely preterm infants. N Engl J Med 2011;364:1680-2.

7. Whyte RK, Nelson H, Roberts RS, Schmidt B. Benefits of Oxygen Saturation Targeting Trials: Oximeter Calibration Software Revision and Infant Saturations. J Pediatr 2017;182:382-4.

8. Tarnow-Mordi W, Stenson B, Kirby A. Oxygen-Saturation Targets in Preterm Infants. N Engl J Med 2016;375:187-8.

9. Darlow BA, Marschner SL, Donoghoe M, et al. Randomized controlled trial of oxygen saturation targets in very preterm infants: two year outcomes. J Pediatr 2014;165:30-5 e2.

10. Efron B. Bootstrap methods: another look at the jackknife. Annals of Statistics 1979;7:1-26.

11. Good, PI. Resampling Methods: A Practical Guide to Data Analysis. Boston: Birkhäuser, 2006: Chapter 3.

12. Still, AW, White AP. The approximate randomization test as an alternative to the F test in analysis of variance. British Journal of Mathematical and Statistical Psychology 1981;34:243-252.

13. BOOST-II Australia and United Kingdom Collaborative Groups, Tarnow-Mordi W, Stenson B, Kirby A, Juszczak E, Donoghoe M, Deshpande S et al. Outcomes of two trials of oxygen-saturation targets in preterm infants. N Engl J Med. 2016;374:749–760.

14. Huizing MJ, Villamor-Martínez E, Vento M, Villamor E.Pulse oximeter saturation target limits for preterm infants: a survey among European neonatal intensive care units. Eur J Pediatr. 2017 Jan;176(1):51-56.

Figure legends

Figure 1, online

Average frequency distributions of the proportion of time spent by individual infants at each pulse oximeter saturation value while receiving supplemental oxygen in 176 infants monitored using un-modified Masimo SET Radical pulse oximeters.

From: Arch Dis Child Fetal Neonatal Ed. Johnston ED, Boyle B, Juszczak E, King A, Brocklehurst P, Stenson BJ. Oxygen targeting in preterm infants using the Masimo SET Radical pulse oximeter, Volume 96 Page F430. Copyright © (2011) BMJ Publishing. Reprinted with permission.

Figure 2:

Distribution of patients by proportion of time spent in target range.

Black vertical bars signify the mean proportion of time spent in each target range.

Figure 3, online:

Meta-analysis of Mortality at 18-24 months in the BOOST-II UK and Australian Trials and the Canadian Oxygen Trial, sub-grouped according to whether infants were treated with the original or the revised oximeters, restricted to the UK, Australian and Canadian trials. Analysed using RevMan version 5.3, Cochrane Collaboration.

Figure 1, online

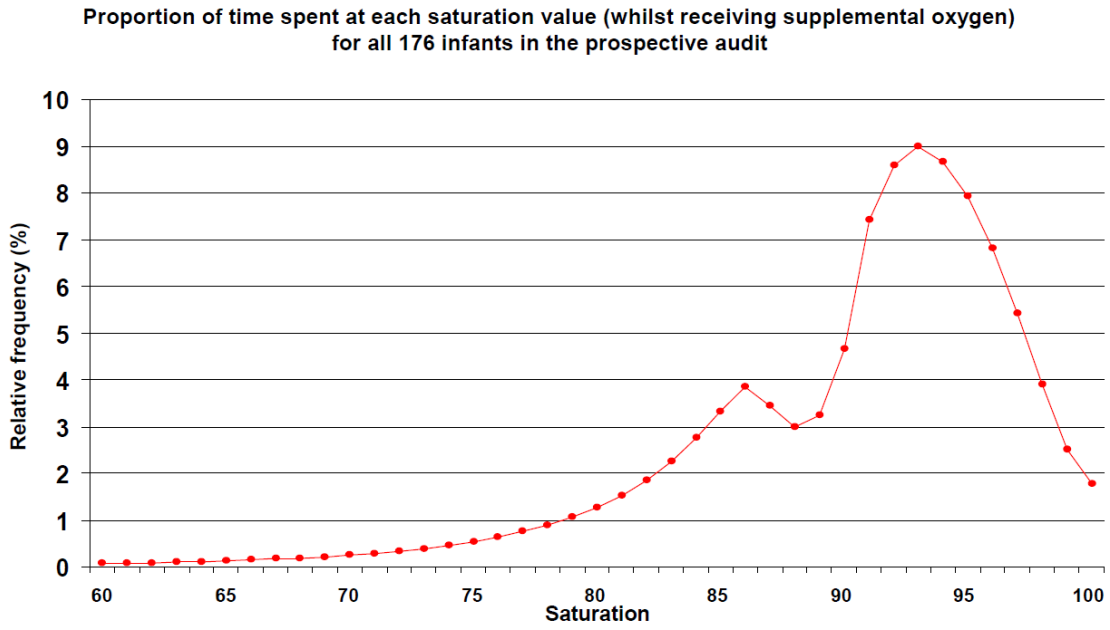

Figure 2

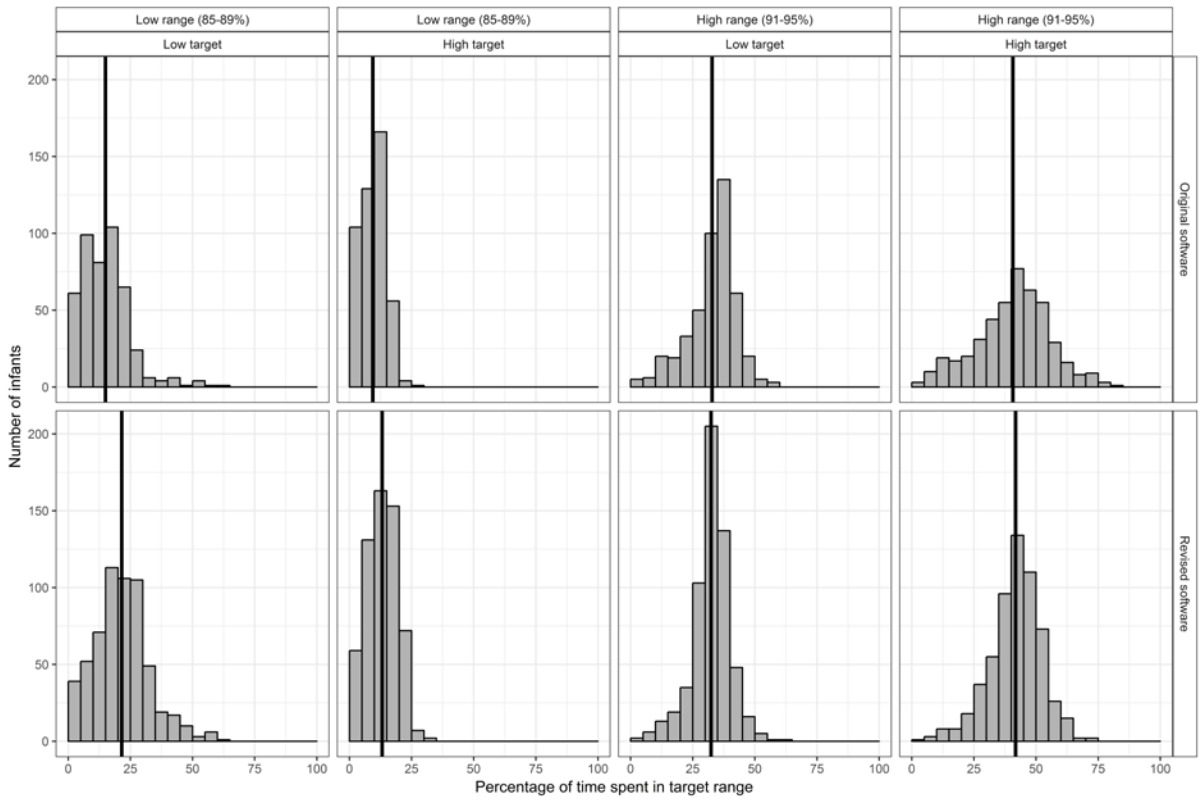

Figure 3

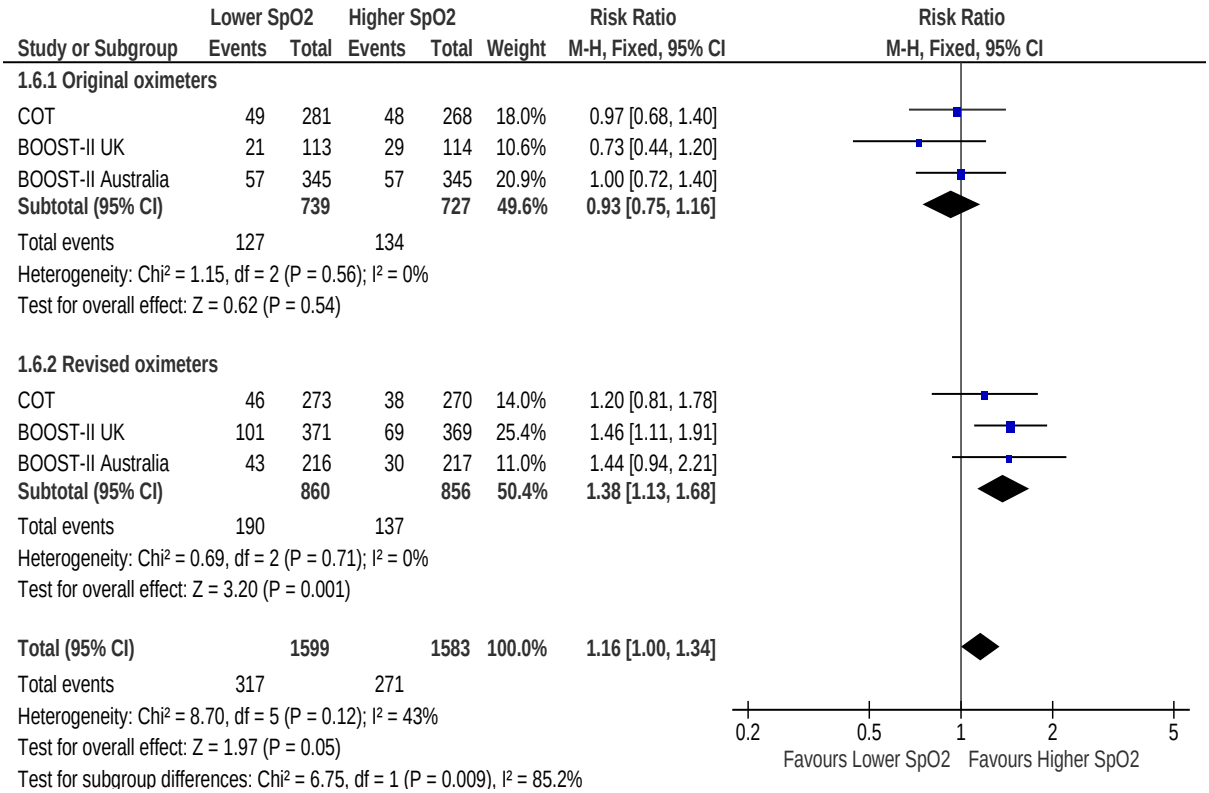

Table 1: Combined results. Mean distributions of the proportions of time spent by each infant in low (85-89%) and high (91-95%) ranges of actual SpO_2 using revised versus original oximeters. Pooled net improvements stratified by trial.

	Low-target range (SpO_2 85-89%)		High-target range (SpO_2 91-95%)	
	Aus	**UK**	**Aus**	**UK**
Low-target infants				
Using original oximeters (n=457)	14.3%	17.2%	32.2%	34.7%
Using revised oximeters (n=591)	20.0%	22.4%	32.3%	32.5%
Change using revised vs original oximeters	+5.8%	+5.3%	+0.1%	-2.2%
High-target infants				
Using original oximeters (n=460)	8.5%	12.3%	38.9%	46.0%
Using revised oximeters (n=588)	11.8%	14.0%	42.1%	41.6%
Change using revised vs original oximeters	+3.3%	+1.7%	+3.2%	-4.5%

Net improvement in targeting	+2.5%	+3.6%	+3.1%	-2.2%
Pooled net improvement	+2.9%		+1.0%	
(95% CI)	(+1.4%, +4.4%)		(-1.0%, +3.0%)	
p-value	<0.001		0.322	
Overall improvement for both ranges	+3.9%			
(95% CI)	(+1.6%, +6.2%)			
Overall p-value	<0.001			

Table 2, online

Distributions of the proportions of time spent by each infant in low (85-89%), high (91-95%) and other ranges of actual SpO_2 using revised versus original oximeters. BOOST-II Australia.

		SpO_2 range				
	N	**< 85%**	**85-89%**	**90%**	**91-95%**	**> 95%**
Low-target infants						
Using original oximeters	344	13.0%	14.3%	3.8%	32.2%	36.8%
Using revised oximeters	219	13.0%	20.0%	5.7%	32.3%	29.0%
Change using revised vs original oximeters		+0.0%	+5.8%	+1.9%	+0.1%	-7.8%
High-target infants						
Using original oximeters	345	7.4%	8.5%	2.8%	38.9%	42.4%
Using revised oximeters	220	6.8%	11.8%	4.6%	42.1%	34.7%
Change using revised vs original oximeters		-0.6%	+3.3%	+1.8%	+3.2%	-7.7%
Net improvement in targeting			+2.5%		+3.1%	

(95% CI)			(+0.4%, +4.6%)		(+0.4%, +5.8%)	
p-value			0.017		0.032	
Overall improvement for both ranges			+5.6%			
(95% CI)			(+2.1%, +8.7%)			
Overall p-value			<0.001			

Table 3, online

Mean distributions of the proportions of time spent by each infant in low (85-89%), high (91-95%) and other ranges of actual SpO_2 using revised versus original oximeters. BOOST-II UK.

		SpO_2 range				
	N	**< 85%**	**85-89%**	**90%**	**91-95%**	**> 95%**
Low-target infants						
Using original oximeters	113	19.0%	17.2%	4.2%	34.7%	24.9%
Using revised oximeters	372	16.1%	22.4%	6.3%	32.5%	22.8%
Change using revised vs original oximeters		-3.0%	+5.3%	+2.1%	-2.2%	-2.1%
High-target infants						
Using original oximeters	115	11.4%	12.3%	3.9%	46.0%	26.3%
Using revised oximeters	368	9.5%	14.0%	5.0%	41.6%	29.9%
Change using revised vs original oximeters		-2.0%	+1.7%	+1.1%	-4.5%	+3.6%
Net improvement in targeting			+3.6%		-2.2%	
(95% CI)			(+1.5%, +5.7%)		(-5.3%, +0.7%)	
p-value			0.003		0.100	

Overall improvement for both ranges			+1.3%	
(95% CI)			(-2.3%, +4.8%)	
Overall p-value			0.465	