20th November 2019

Dear editors/reviewers:

RE: World Journal of Gastrointestinal Endoscopy Manuscript NO: 52262

Thank you for taking the time to review our submission and for your helpful feedback. Our responses to your specific comments and queries are indicated in red below your original comments.

Reviewer 4:

This article tried to evaluate the long-term outcome of SBT-based SPRINT course. In the end, the authors found that the course may contribute to a shorter period to achieve JAG certification, with more unsedated procedures. I have several concerns regarding this study.

1. Formatting: Too many words for the statistical analysis part. Many of the methodological descriptions should be documented in trial design, intervention or outcome part.

Response: Thank you. The statistical methods section has been revised and considerably shortened, with some aspects relocated under study design and study outcome.

2.Methods: (a) Although this is a pilot study, proper sample size should be calculated to explain some negative results. But the author neither described the calculation process, nor explained why they did not do so.

Response: The study was based on a convenience sample of all trainees attending a single training session that met the inclusion criteria, along with corresponding control trainees. As such, the sample size was fixed, hence no formal a priori power calculation was performed. We have now mentioned this under Limitations.

We had considered post-hoc power calculations, however, performing these for the generalized estimating equation analyses of long-term trainee outcomes would require a large number of assumptions to be made, in order to simplify the complex analysis to the point that such a calculation was feasible. In doing this, the resulting power calculation may not be an accurate reflection of the analysis that was actually performed, hence would be of dubious reliability. In addition, as reported by Hoenig et al. (DOI: 10.1198/000313001300339897), the use of post-hoc

power calculations to assist in the interpretation of negative results is logically flawed, and could result in misleading conclusions. As such, post-hoc power calculations were not reported.

(b) One of the endpoint event was achieving gastroscopy certification. I am curious about the detailed criteria of the certification. Why did not the author try to use these criteria as endpoints？

Response: Thank you. The endpoint of JAG gastroscopy certification has now been referenced and further elaborated upon under Methods, Study Outcomes. Whilst JAG certification is a composite endpoint, this is primarily dependent on minimum procedural numbers (200+), and performance in formative and summative assessment through the use of the direct observation of procedural skills (DOPS). This was the basis our choice of "time to 200 procedures" as an outcome. In terms of formative assessment, we found that there were no significant differences in the number of DOPS performed by cases and controls. However, due to the differences in intervals at which DOPS were performed (in relation to lifetime procedure count) and the small study sample, this was not analysed.

3.Results: (a) The results showed that the baseline characteristics of both groups were comparable, regarding number of procedures. 60% of the trainees were naïve to endoscopic procedure, which, in other words, means that 6 trainees in case group and 10 trainees in control group had experiences. However, the average number of procedures were 10 and 3 in case and control groups, respectively. Thus, it gives me an impression, that non-naïve trainees in case group had a mean of 25 procedures and that those trainees in control group had a mean of 7 procedures. Could this possibly contribute to the earlier acquisition of certification in case group? Perhaps a subgroup analysis might be helpful.

Response: Thank you. The number of trainees with experience was N=6 in the case group (40% of 15 trainees) and N=9 in the control group (37% of 24 trainees). In the subgroup of trainees with experience, the mean number of procedures prior to the training course was 24 vs. 8 for cases vs. controls, similar to the numbers that you quoted. However, the distribution of experience was different in these two subgroups, with the experienced trainees in the control group tending to have completed a similar small number of procedures (range: 1-18), whilst the cases were made up of three with minimal experience (1, 1 and 7 procedures), and three with more experience (43, 44 and 48 procedures).

We performed a subgroup analysis to include only trainees with <10 lifetime procedure counts, which led to similar Kaplan-Meier plots and statistically significant results.

(b) With the doubling of procedures, the trainees had similar degree of improvements in D2 intubation. Did that mean we can obtain elevation of experience through actual practice instead of SBT?

Response: Thank you. Yes, our learning curve plots shows the relationship between competency acquisition (as measured by unassisted D2 intubation rates) with lifetime patient-based OGD count. We did not include SBT procedures under the lifetime procedure count as it is clear that SBT should not be a substitute for patient-based training procedures and for this to count towards JAG certification. It is unclear where the differences in gradients between cases and controls (OR 1.99 vs OR 1.74; P=0.205) could be a Type 2 error due to the small sample sizes.

(c) Although the Kaplan-Meier curves of 200th procedure accomplishment were comparable, the achievement rate from the 6th month to the 10th month seemed higher in case group. Could there be some reasons for this?

Response: This is an interesting observation. From further observation of the Kaplan-Meier curves, the larger difference between groups observed between 6-10 months is not driven by the case group, for which the proportion of trainees with 200+ procedures increases at a relatively consistent rate over the whole follow up period. Instead, it is caused by the control group, which has a low proportion of trainees achieving the outcome initially, followed by a preponderance of trainees reaching 200 procedures between 10-12 months. This causes the two groups to become more similar in the later period of follow-up.

During UK higher specialty training, trainees tend to undertake rotations in different hospitals and change on a year basis. We suspect that trainees will push to achieve certification within their first year of training and before they rotate, otherwise they will require assessment of OGD skills in a new hospital by trainers who are not familiar with their abilities and competencies. Because achieving 200 procedures is one of the criteria for certification, this may explain the rise in the proportion of trainees in the control group who reach the 200-procedure mark closer to the 12-month time point. The fact that this preponderance is only visible in the control group may be since those trainees in the case group who were aiming to achieve this target did so earlier, in light of their higher overall

rate of procedures per month (median: 16.2 vs. 13.8), hence achieved this target as part of their routine training, rather than having to make a particular effort to increase procedure numbers when approaching the one year milestone.

In light of your comment, we also performed an additional analysis, which truncated the follow up period at 10 months. This found the difference between the groups to narrowly reach statistical significance (p=0.040). However, the effect size had a wide confidence interval (hazard ratio: 5.4, 95% CI: 1.1-26.7), which crossed that of the primary analysis (HR: 1.7, 95% CI: 0.8-3.5). In addition, choosing to truncate the follow up at the point with the largest difference between groups would constitute "cherry-picking", and would increase the risk of a false-positive. In light of these limitations, this additional analysis has not been reported in the paper.

4. Discussion: The authors stated that several limitations existed, like inadequate sample size, difference of training intensity among different regions and selection bias. So, with these limitations, could the final result be credible? We all think that SBT prior to actual practice is helpful. But with such concerns illustrated above, the long-term outcome of this course may be questioned. Certain explanations should be addressed to make the result more credible.

Response: Thank you. We agree that interpretation is difficult due to small sample sizes. This reflects the complete lack of similar studies offering long term trainee outcomes in the existing literature (as mentioned within our Discussion). Another limitation was that, because trainee performance was studied (in order to compared performance with experts), this had to be initially performed without coaching/feedback, which are seminal for improving performance post-SBT. This may have somewhat compromised the effectiveness of the hands-on technical elements of the course. It is hoped that our pilot data will provide the impetus for a larger and more robust future study with objective assessments at regular intervals which will better determine their impact on trainee and patient outcomes. This has been added at the end of the final discussion paragraph.

In addition to revisions in response to your suggestions, we have also made refinements to enhance the manuscript. We would like to take the opportunity to thank you for your comments and for considering our revised submission.

Yours faithfully,

Keith Siau, James Hodson and Neil Hawkes (on behalf of co-authors)