PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A randomized controlled pilot trial evaluating feasibility and acceptability of a computer-based tool to identify and reduce harmful and hazardous drinking among adolescents with alcohol-related presentations in Canadian pediatric emergency departments
AUTHORS	Newton, Amanda; Dow, Nadia; Dong, Kathryn; Fitzpatrick, Eleanor; Wild, Cameron; Johnson, David; Ali, Samina; Colman, Ian; Rosychuk, RJ

VERSION 1 - REVIEW

REVIEWER	Benedict Weobong London School of Hygiene and Tropical Medicine
REVIEW RETURNED	12-Jan-2017

GENERAL COMMENTS	**General comments: 1** The paper may be considered for publication as it addresses and contributes to tackling a mental health condition of public health importance, using innovative and potentially scalable strategies. Most of my comments are minor as this is a straightforward report on a pilot study. **Abstract: Minor Essential Revisions** • Objectives: looks to me as the aim as that's what pilot/feasibility studies are designed for. Authors should state the specific objectives instead • Results: include numbers randomised, analysed in each group. **Methods:** General comments: Major Compulsory Revisions • need to state/describe what the design is: was this individual or cluster randomised? • Data collection: it appears the authors missed out on the main measure of 'feasibility' • Outcomes: suggest to use the full AUDIT in the future definitive RCT, and this could be discussed as part of 'lessons' learnt in this report. Using the AUDIT-C is likely to lead to a skewed spread of AUDIT scores as there will be many individuals with scores clustered just above the cut-off. The response options for the three questions are poorly

	graduated, such that a participant's drinking can decrease modestly without being reflected in a reduced score. This could bias the effect estimate toward the null. The use of the AUDIT-C for screening in this context is however a good idea. • Sample size: Not practice to use pilot studies to estimate sample sizes for definitive RCTs for the simple reason that there's an inherent imprecision in the effect size estimates because of the small sample size, and also because they are not designed as hypotheses-testing studies. I am curious why the authors did this and if this should not be discussed as a limitation of this report.

REVIEWER	Tapan Rai University of Technology Sydney
REVIEW RETURNED	07-Feb-2017

GENERAL COMMENTS	Overall, this is a well-written description of a well-designed pilot RCT. For the most part, it needs only minor revisions. However, there is a MAJOR issue with the statistics that needs to be properly addressed. Addressing this could result in a change in the findings and discussion. Therefore, I am treating this as being in need to a MAJOR REVISION. The issues are as follows: MAJOR ISSUE: In the Sub-section on Sample Size (Page 11, lines 41-51), you state that you want the definitive RCT to be able to detect an effect size of f=0.1. However, your calculations on page 16 (line 10) suggest that this would require a sample size of 1571 per group to achieve power =0.8 at the 5% significance level. This is not correct for an effect size of f=0.1. The required sample size of 1571 per group would be correct, if you were planning an independent samples t-test with effect size, d=0.1. However, the effect size index, f (which is used for F-tests such as ANOVA) is not equivalent to the effect size index, d (which is used for independent sample t-tests. In fact, an effect size of d=0.1 is much smaller (approximately 1/2) than an effect size of f=0.1, which requires a total sample size of approximately 786 to achieve 80% power at the 5% significance level. I would also like to note that these sample sizes are not determined by your trial, but are in fact fixed, once you have fixed the power, significance level and the effect size. In fact, it is not clear to me why you fixed the sample size at f=0.1 (is it just because some other trial did? This is not an acceptable reason for you to use the same sample size.) The issues you need to consider are: do you want to show that your intervention has the same effect as the other one? Perhaps, it has a larger effect... that would be more interesting! Or perhaps it has a smaller effect, but is more cost-effective or more feasible for other reasons or more likely to be accepted by the population concerned; the could be interesting too! This is something that needs to be included in your discussion.

	The way you have structured your paper, your trial does not have any effect on the sample size required. As I mentioned earlier, that is fixed by the effect size, power and significance level. Most pilot studies would be used to determine effect size and from the effect size, then calculate the sample size required. The effect size should depend on the size of a clinically meaningful effect and the variability that you observe. However, you have made no effort to address what effect (difference on the AUDIT-C scale) is clinically meaningful. While you have assessed variability from your trial, you have then gone on to calculate the size of the difference (0.153 points on the AUDIT-C scale) that you would be able to detect with a sample size of 1571 per group. This does not appear to be the correct approach to take here. I think you should be actually calculating sample size required based on your trial (and not on some arbitrary effect size that is not justified). However, if you do choose to persist with your approach (this would take some effort to justify), some of the things that you need to address are: 1. Is the effect of 0.153 points clinically meaningful? What does it mean in practical terms? 2. Why would you do a trial that detects such a small effect size, when your pilot study suggests that your effect size is much larger (d > 1, I believe) 3. Is it feasible to recruit 1571 patients per group for a more definitive trial? Given your recruitment/retention rate, this may take several years even if you recruited from all appropriate EDs in Canada. Such a large trial with so many EDs involved would be a nightmare to manage, and the statistical analysis would need to adjust for clustering within each site, so the effective sample size that you need would in fact be greater than what you would require for a t-test. If you have the answers to these issues, you should address them in your Discussion section. This is the one major issue, due to which I recommend a major revision. There are also a few MINOR issues, some of which I will list below. MINOR ISSUES: 1. On page 6, Lines 46-51, you state: "Removal of this criterion reduced the concern that we were not able to accurately identify or exclude those youth who had used other substances just prior to ED visit." While this may be true, it appears to put too much of a positive spin on a minor mis-step in the trial. It needs to be stated in a more negative way and included in the limitations. Report how many who self-identified as having used another drug were excluded before the criterion changed (if this answer is none, please state this. 2. You have a large number of outcome measures for an RCT. Please classify your outcomes as primary/secondary/tertiary, based on the primary goals of the pilot study. If this paper focusses on one of the secondary/tertiary, please mention this. Also, make it a point to mention what would be the primary outcome for the main trial. Is it change in AUDIT-C score at 1 month or at 3 months? I would have expected the primary endpoint to be at 1 month (especially since you can expect loss to follow up at the 3-month mark). However, you have done the main calculations for the paper at the 3-month mark, suggesting that this is your primary end-point for the definitive trial. If this is the case, what is the purpose of the 1-month data collection point? With sample size calculation as a major focus of your paper, you would need to address the total number that you would need to recruit (based on your retention rate) to achieve the required sample

	size at the primary endpoint, and the power that it would provide at the secondary endpoint 3. Please explain, more carefully, how you determined that a sample size of 44 would be sufficient for the pilot trial. Lancaster et al (whom you cite) suggest a ballpark figure of 30 per group. While 22 per group is not unreasonable for a pilot study, your write-up seems to suggest that you did something deeper than that. If this is the case, please elaborate. If it was just a matter of feasibility with Lancaster et al as a guide, please state this. This is something that would provide a guideline to other researchers. 4. Please explain more carefully how you use GPower – what are the inputs? What test is your estimation based on? Again, this would provide guidance to other researchers. 5. I appreciate the fact that you have addressed the reason for the imbalance in allocation, but it does come across as a poor data management system and poor practice for a clinical trial – just a comment. I am very pleased to see that you have addressed this as a lesson learned in your discussion. I hope that a more definitive trial that you conduct has a more professional data management system (there are several professional packages for management of clinical trial data, some of which have free versions) and a more professional approach to data look up. 6. I like the fact that you have tested the blinding (Perception of Group Allocation, p.13). However, hypothesis tests for categorical data (even Fisher's exact test) are not very sensitive. I would recommend that you report the raw counts or percents of correct/incorrect guesses of allocations. 7. Please check the reference to Lancaster et al [37]. I believe it was published in 2004, not 2004, as you state. Overall, this is a well-designed trial, which can serve to guide other researchers in conducting pilot clinical trials. Therefore, I would like to see you publish the best possible paper base in the work that you have done, and I hope that these comments guide you in achieving that in the next iteration of this paper.

REVIEWER	Nanhua Zhang Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio USA
REVIEW RETURNED	08-Feb-2017

GENERAL COMMENTS	This article examines the feasibility of a computer-based tool to identify and reduce harmful and hazardous drinking among adolescents with alcohol-related presentation in Canadian pediatric emergency departments, using a multi-site randomized control pilot study. The authors conclude that a definitive RCT is feasible. From design and statistical analysis perspective, I have the following comments: 1. It is not clear to me how the randomization was done, which resulted in imbalance in the number of subjects in the intervention and the control group. The research team accessing the research website during the trial shouldn't disrupt the randomization since they were also randomized to one of the two groups. 2. Another drawback in the study was the changing in exclusion criterion regarding other drug use during the study could result in incomparable samples. 3. Another big concern is that the reduction of AUDIT-C score in the control group is bigger than that in the intervention group, implying

	no efficacy of the intervention. With a reverse effective, the power analysis for the definitive trial is problematic. 4. The study retention is poor and the authors should assess differential attrition by comparing the subjects who stayed and those dropped out. 5. On page 12 line 30: Fisher's exact test was used which is inappropriate because the true and perceived allocation were correlated. The correct procedure to use McNenar's test. 6. Table 1 should also report the summary statistics by treatment group, and use proper tests to assess group differences. Same for Table 3 (maybe also table 4).

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
1.Objectives: looks to me as the aim as that's what pilot/feasibility studies are designed for. Authors should state the specific objectives instead.
Author Response: We have revised the wording in our abstract.

2. Results: include numbers randomised, analysed in each group.
Author Response: This information from Figure 1 has been added to the text as requested.

3. Need to state/describe what the design is: was this individual or cluster randomised?
Author Response: This information is stated under Randomization: "Participants were assigned in a 1 to 1 allocation to either intervention or control group"

4. Data collection: it appears the authors missed out on the main measure of 'feasibility'
Author Response: Clarifications have been made in the Methods section.

5. Outcomes: suggest to use the full AUDIT in the future definitive RCT, and this could be discussed as part of 'lessons' learnt in this report. Using the AUDIT-C is likely to lead to a skewed spread of AUDIT scores as there will be many individuals with scores clustered just above the cut-off. The response options for the three questions are poorly graduated, such that a participant's drinking can decrease modestly without being reflected in a reduced score. This could bias the effect estimate toward the null. The use of the AUDIT-C for screening in this context is however a good idea.
Author Response: Thank you for this suggestion! We have left this point out of our discussion as it did not fall under our study objectives, but we have brought this issue forward to the team as we plan for the definitive trial.

6. Sample size: Not practice to use pilot studies to estimate sample sizes for definitive RCTs for the simple reason that there's an inherent imprecision in the effect size estimates because of the small sample size, and also because they are not designed as hypotheses-testing studies. I am curious why the authors did this and if this should not be discussed as a limitation of this report. As stated in Lancaster et al 2004 "A major reason for conducting a pilot study is to determine initial data for the primary outcome measure, in order to perform a sample size calculation for a larger trial (Ross-McGillet al. 2000; Stevinson & Ernst 2000)."
Author Response: We agree that pilot studies are not designed as hypothesis-testing studies. We also agree that there is imprecision in the effect size estimates and our sample size calculation is based on a fixed effect size d that is not determined from the pilot data. The pilot data helped inform our sample size by indicating the mean difference between groups given our estimates of standard deviation from the pilot study.

Reviewer 2

1. In the Sub-section on Sample Size (Page 11, lines 41-51), you state that you want the definitive RCT to be able to detect an effect size of f=0.1. However, your calculations on page 16 (line 10) suggest that this would require a sample size of 1571 per group to achieve power =0.8 at the 5% significance level. This is not correct for an effect size of f=0.1. The required sample size of 1571 per group would be correct, if you were planning an independent samples t-test with effect size, d=0.1. However, the effect size index, f (which is used for F-tests such as ANOVA) is not equivalent to the effect size index, d (which is used for independent sample t-tests. In fact, an effect size of d=0.1 is much smaller (approximately 1/2) than an effect size of f=0.1, which requires a total sample size of approximately 786 to achieve 80% power at the 5% significance level. I would also like to note that these sample sizes are not determined by your trial, but are in fact fixed, once you have fixed the power, significance level and the effect size.
Author Response: We regret the use of notation has caused confusion. The notation f was unintentionally used as we did conduct the calculations with a two-tailed t-test of independent means, effect size d=0.1, Type I Error=0.05, and Power=0.80. Yes, we agree that the sample size is not determined by the pilot and what we were trying to accomplish was inform what an effect size of d=0.1 would mean in terms of mean difference between groups given our estimates of standard deviation from the pilot.

2. In fact, it is not clear to me why you fixed the sample size at f=0.1 (is it just because some other trial did? This is not an acceptable reason for you to use the same sample size.) The issues you need to consider are: do you want to show that your intervention has the same effect as the other one? Perhaps, it has a larger effect... that would be more interesting! Or perhaps it has a smaller effect, but is more cost-effective or more feasible for other reasons or more likely to be accepted by the population concerned; this could be interesting too! This is something that needs to be included in your discussion.
Author Response: To clarify, this decision was not based on another trial. Brief interventions such as the one we are studying (and the one in the other trial that was cited) are not expected to have a large effect on alcohol-related outcomes. They are a time-limited intervention (5-10 minutes). It is more likely for a brief intervention to have a small effect size and we need to plan the definitive trial to detect a small effect. We have clarified our position in the Sample Size section of the manuscript. We appreciate you pointing out the need to improve our reporting.

3. The way you have structured your paper, your trial does not have any effect on the sample size required. As I mentioned earlier, that is fixed by the effect size, power and significance level. Most pilot studies would be used to determine effect size and from the effect size, then calculate the sample size required. The effect size should depend on the size of a clinically meaningful effect and the variability that you observe. However, you have made no effort to address what effect (difference on the AUDIT-C scale) is clinically meaningful. While you have assessed variability from your trial, you have then gone on to calculate the size of the difference (0.153 points on the AUDIT-C scale) that you would be able to detect with a sample size of 1571 per group. This does not appear to be the correct approach to take here. I think you should be actually calculating sample size required based on your trial (and not on some arbitrary effect size that is not justified). However, if you do choose to persist with your approach (this would take some effort to justify), some of the things that you need to address are:
1. Is the effect of 0.153 points clinically meaningful? What does it mean in practical terms?
2. Why would you do a trial that detects such a small effect size, when your pilot study suggests that your effect size is much larger (d > 1, I believe)
3. Is it feasible to recruit 1571 patients per group for a more definitive trial? Given your recruitment/retention rate, this may take several years even if you recruited from all appropriate EDs in Canada. Such a large trial with so many EDs involved would be a nightmare to manage, and the statistical analysis would need to adjust for clustering within each site, so the effective sample size that you need would in fact be greater than what you would require for a t-test.

If you have the answers to these issues, you should address them in your Discussion section.
Author Response: The belief is that a brief intervention would likely provide a small effect, and amongst adolescents this small effect could have an important impact on health outcomes. We have added some details to our Introduction and Methods on the expected impact of brief intervention. Thank you for pointing out this important gap in our manuscript. The pilot trial provides estimates of the standard deviations that help us translate the mean difference that corresponds to the small effect size. We note that pilot studies are not hypothesis testing and "a pilot study does not provide a meaningful effect size estimate for planning subsequent studies due to the imprecision inherent in data from small samples." (Leon, Davis, & Kraemer, 2011) Thabane et al (2010) echo the same comment that "it can be dangerous to use pilot studies to estimate treatment effects." We agree that the definitive trial would be large. We note how we are considering various scenarios for planning this trial vis-à-vis the results of the pilot trial.

4. On page 6, Lines 46-51, you state: "Removal of this criterion reduced the concern that we were not able to accurately identify or exclude those youth who had used other substances just prior to ED visit." While this may be true, it appears to put too much of a positive spin on a minor mis-step in the trial. It needs to be stated in a more negative way and included in the limitations. Report how many who self-identified as having used another drug were excluded before the criterion changed (if this answer is none, please state this.
Author Response: We have added information to the Methods and commented on this protocol change in the discussion.

5. You have a large number of outcome measures for an RCT. Please classify your outcomes as primary/secondary/tertiary, based on the primary goals of the pilot study. If this paper focusses on one of the secondary/tertiary, please mention this. Also, make it a point to mention what would be the primary outcome for the main trial. Is it change in AUDIT-C score at 1 month or at 3 months? I would have expected the primary endpoint to be at 1 month (especially since you can expect loss to follow up at the 3-month mark). However, you have done the main calculations for the paper at the 3-month mark, suggesting that this is your primary end-point for the definitive trial. If this is the case, what is the purpose of the 1-month data collection point? With sample size calculation as a major focus of your paper, you would need to address the total number that you would need to recruit (based on your retention rate) to achieve the required sample size at the primary endpoint, and the power that it would provide at the secondary endpoint.
Author Response: Changes made as requested. The choice to make the primary end-point 3-months was based on team discussions. This information is now communicated in the Analysis and Results sections. Thank you for pointing out this omission.

6. Please explain, more carefully, how you determined that a sample size of 44 would be sufficient for the pilot trial. Lancaster et al (whom you cite) suggest a ballpark figure of 30 per group. While 22 per group is not unreasonable for a pilot study, your write-up seems to suggest that you did something deeper than that. If this is the case, please elaborate. If it was just a matter of feasibility with Lancaster et al as a guide, please state this. This is something that would provide a guideline to other researchers.
Author Response: We did indeed do several calculations to arrive at the choice of sample size for the pilot RCT. We had intended to cite Lancaster et al for the general points rather than the rule of thumb figure stated in Browne 1995. We have clarified our calculations for the sample size for the pilot study and indicated the upper confidence interval values and confidence interval widths that guided our choice.

7. Please explain more carefully how you use GPower – what are the inputs? What test is your estimation based on? Again, this would provide guidance to other researchers.
Author Response: We used GPower with a two-tailed t-test of independent means, effect size d=0.1,

Type I Error=0.05, and Power=0.80. We have added these settings to the Methods section.

8. I appreciate the fact that you have addressed the reason for the imbalance in allocation, but it does come across as a poor data management system and poor practice for a clinical trial – just a comment. I am very pleased to see that you have addressed this as a lesson learned in your discussion. I hope that a more definitive trial that you conduct has a more professional data management system (there are several professional packages for management of clinical trial data, some of which have free versions) and a more professional approach to data look up.
Author Response: It was an unfortunate error on our part, and one that we wish hadn't occurred. Several changes are being made to trial management to ensure this does not happen in the definitive trial.

9. I like the fact that you have tested the blinding (Perception of Group Allocation, p.13). However, hypothesis tests for categorical data (even Fisher's exact test) are not very sensitive. I would recommend that you report the raw counts or percents of correct/incorrect guesses of allocations.
Author Response: Thank you for the suggestion. This table has been added to the Results (Table 2).

10. Please check the reference to Lancaster et al [37]. I believe it was published in 2004, not 2004, as you state.
Author Response: We stated that Lancaster et al was published in 2002, and you are correct that it is published in 2004.

Reviewer 3
1. It is not clear to me how the randomization was done, which resulted in imbalance in the number of subjects in the intervention and the control group. The research team accessing the research website during the trial shouldn't disrupt the randomization since they were also randomized to one of the two groups.
Author Response: Additional description has been provided in the results section. In brief, team members continued to access the website until they could view intervention content (computer-based SBIRT). This occurred over time and while participants were being enrolled, and unfortunately meant that more participants were allocated to the control arm while the team was ensuring they could see intervention content.

2. Another drawback in the study was the changing in exclusion criterion regarding other drug use during the study could result in incomparable samples.
Author Response: Indeed, had we not made this change early on, there may have been an impact of who was recruited. Comment has been added to the Methods and Discussion sections.

3. Another big concern is that the reduction of AUDIT-C score in the control group is bigger than that in the intervention group, implying no efficacy of the intervention. With a reverse effective, the power analysis for the definitive trial is problematic.
Author Response: We note that pilot studies are not hypothesis testing. The effect observed in the pilot has imprecision because of the small sample size. The direction of the difference in the pilot cannot imply efficacy or lack thereof.

4. The study retention is poor and the authors should assess differential attrition by comparing the subjects who stayed and those dropped out.
Author Response: This was not an objective of the pilot study, but will be an objective for the definitive trial we hope to conduct.

5. On page 12 line 30: Fisher's exact test was used which is inappropriate because the true and perceived allocation were correlated. The correct procedure to use McNenar's test.

Author Response: Chi-square tests are an acceptable way of assessing blinding. Here we have a 2x3 table with small cell counts and have chosen to use a Fisher's Exact test on account of the relatively small cell sizes.

6. Table 1 should also report the summary statistics by treatment group, and use proper tests to assess group differences. Same for Table 3 (maybe also table 4).
Author Response: We chose not to present summary statistics by treatment allocation because in the pilot study we were not powered to detect differences. We agree that data should be presented by group for the definitive RCT we hope to conduct.

VERSION 2 – REVIEW

REVIEWER	Tapan Rai University of Technology Sydney
REVIEW RETURNED	09-Apr-2017

GENERAL COMMENTS	The authors have satisfactorily addressed most of the minor issues raised in previous reviews. However, their insistence on presenting statistical calculations continues to pose a problem. First, they state (in response to previous reviews) that pilot studies such as this one should not be used for efficacy hypothesis testing or to determine effect size that is used to inform a larger definitive trial. They provide references to support this statement. However, they proceed to do precisely that. In fact, the main aim stated in the abstract of the current version is "change in alcohol consumption from baseline to 1- and 3-months post intervention". The then go on to to determine standard deviations for the control and intervention group and use it to determine the size of the effect (in clinical terms) that they would be able to detect in a more definitive study. They base this on a small standardised effect size that they have picked rather arbitrarily as what they would expect to detect in the more definitive study they propose. There are several problems with this. These include: 1. If the sample size of a (pilot) study is too small to determine a mean, logic dictates that it is too small to establish the standard deviation for that mean. 2. If the standard deviation is not accurate, it should not be used to inform effect size. 3. The authors state that (in response to previous reviews) that the small effect size they chose is not based on any previous trial. Yet in the paper they state that it is based on extant research based on a comparable intervention. 4. The use of a standardised effect size to estimate sample size for a large definitive trial is problematic. The standardised effect size of d=0.1 will always produce a minimum estimated sample size of 1571 at 5% significance and 80% power. This is not related to the pilot study. However, the authors use this information together with the admittedly inaccurate results of the trial to "determine" the size of the effect (on the AUDIT-C scale) that they would observe in a trial with 1571 participants. They do not discuss whether this difference that they would observe is clinically meaningful. This is not an appropriate method to calculate the sample size for a clinical trial. An analogy in a pharmaceutical trial could be something like this: a drug company would like to test the efficacy of a drug in reducing blood pressure; the drug is made for all natural ingredients, so the effect is likely to be small. They arbitrarily choose d=0.1 as the

	expected standardised effect size, and calculate that a definitive clinical trial would require 1571 participants. However, they are not able to explain what d=0.1 means to the public/funding body/ethics committee. So they conduct a small pilot study in which they determine standard deviations for a control and intervention group. Based on these standard deviations, they calculate that d=0.1 corresponds to a difference of 1mmHg between a placebo and intervention group. The effect is clearly not meaningful, but they don't bother discussing this anywhere. Instead they try to publish the paper and seek funding for a larger trial in the hope that their marketing team could convince the public to buy a drug that is clinically proven to have a definite (but clinically meaningless) effect in lowering blood pressure. Should such a paper be published? Should the larger trial be funded? In the case of the pharmaceutical company the answer is clearly No. In the case of the current authors' trial, they have not provided enough information to convince me that this is worth pursuing. There are several other issues as well. 1. Recruitment feasibility is not adequately assessed. The trial has a 60% dropout rate at the primary endpoint of 3 months. If this were to persist (although they suggest measures to mitigate the issue), they would need to recruit approximately 4000 participants to be able to analyse 1571 patients at the 3-month followup. The recruitment rate of 37% is fairly good; however, combined with their dropout rate, this would mean that they would need to screen something in the order of 10,000 participants in order to be able to analyse 1571 participants at 3 months. This does not appear to be feasible even with 14 EDs. 2. The authors state that their pilot study showed feasibility in terms of blinding. However, while they were successful in maintaining blinding in the current study, this is insufficient evidence to suggest that they would be successful in maintaining blinding in a larger study that recruits 400 participants in 14 EDs across Canada. Given these issues, my first instinct is to suggest rejection of the paper. However, I believe that the authors have done considerable work in getting to this stage. There should be a publishable paper in this, simply based on feasibility, retention and other artefacts of a complex trial. The authors are more likely to achieve this paper by sticking to the basics and focussing simply on the feasibility, lessons learned and descriptive statistics rather than making questionable and unjustified claims about their sample size for a larger trial.

REVIEWER	Nanhua Zhang Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
REVIEW RETURNED	02-Apr-2017

GENERAL COMMENTS	The authors have addressed my comments in this revision. I have no additional comments.

VERSION 2 – AUTHOR RESPONSE

Reviewer 2 Comment 1: The authors have satisfactorily addressed most of the minor issues raised in previous reviews. However, their insistence on presenting statistical calculations continues to pose a problem.First, they state (in response to previous reviews) that pilot studies such as this one should

not be used for efficacy hypothesis testing or to determine effect size that is used to inform a larger definitive trial. They provide references to support this statement. However, they proceed to do precisely that. In fact, the main aim stated in the abstract of the current version is "change in alcohol consumption from baseline to 1- and 3-months post intervention". The then go on to determine standard deviations for the control and intervention group and use it to determine the size of the effect (in clinical terms) that they would be able to detect in a more definitive study. They base this on a small standardised effect size that they have picked rather arbitrarily as what they would expect to detect in the more definitive study they propose. There are several problems with this. These include:
1. If the sample size of a (pilot) study is too small to determine a mean, logic dictates that it is too small to establish the standard deviation for that mean.
2. If the standard deviation is not accurate, it should not be used to inform effect size.
3. The authors state that (in response to previous reviews) that the small effect size they chose is not based on any previous trial. Yet in the paper they state that it is based on extant research based on a comparable intervention.

Author Response to Comment 1: Since the submission of the manuscript you reviewed, we have confirmed 10 pediatric emergency departments from across Canada to participate in recruiting potential participants. To definitive trial will be conducted as a cluster RCT. This approach reduces the risk of contamination bias that could be introduced by clinicians with patient level randomization. Month is the appropriate unit of randomization for this study as other potential choices (ED, week) do not allow for us to also optimally study implementation of the SBIRT tool in a range of EDs over a sufficient time-period. We still plan on basing the definitive trial sample size on an expected effect size of d=0.1 for the primary outcome, which will be alcohol-related consequences rather than alcohol consumption. During recent consultation with clinicians across the ED sites, we determined that while both are important outcomes, clinicians would consider the SBIRT intervention effective and important to provide during routine ED care if there was a greater change in alcohol-related consequences among those adolescents who receive SBIRT. These consequences are often what bring adolescents to the ED for alcohol-related care. Two recent ED trials of computer-based SBIRT by Cunningham and Walton et al. have defined alcohol-related consequences as a primary outcome of interest. In the Cunningham trial, the effect size reported for this outcome at 3 months was 0.11 (as an aside, the effect size reported for alcohol consumption at 3 months was 0.13). Walton et al. did not report effect sizes.

In response to your points 1 and 2: We are not using any of the alcohol-related outcome data in the pilot trial to inform a sample calculation given our change in primary outcome and trial design. In response to point 3: To clarify, we what meant by the response to reviews was that the use of f=0.1 was not based on another trial. The decision of d=0.1 was based on extant literature. When we designed the pilot trial, there was one study published reporting effect sizes for alcohol consumption (reference 45 cited in the manuscript). Since this time, another trial has published (Cunningham et al. in Pediatrics) similar effect sizes for both alcohol consumption and alcohol-related consequences.

We have made extensive changes to the manuscript that we hope eliminate the outstanding concerns you voiced about the study.

Reviewer 2 Comment 2: Recruitment feasibility is not adequately assessed.

Author Response to Comment 2: As noted in our response to Comment 1, we have confirmed 10 pediatric EDs from across Canada to participate in recruiting potential participants for the full-scale trial. In 2016, these EDs treated 976 adolescents with visits for alcohol-related concerns. Using baseline data from our pilot trial, we expect that 89% of these adolescents will screen positive for harmful and hazardous alcohol use. We will aim to enrol these adolescents into the definitive trial. In the same year, these EDs also treated 853 mental health visits that involved harmful and hazardous

alcohol use. Because ED care for these adolescents should also include SBIRT, we will also aim to enrol these adolescents into the definitive trial. From this annual estimate of 1,722 visits per year, we expect 64% of adolescents will consent to participate. This estimate takes into account our pilot trial results (38%) and 13 other prospective ED studies recruiting adolescents with alcohol- and mental health-based visits (recruitment rates ranged from 30% to 90%). We have reviewed the recruitment processes of these studies, and for the definitive trial, propose more comprehensive RA coverage to ensure that a recruitment rate of 64% is realistic (from 10 hours/day on weekends in the pilot study to 10 hours/day, 7 days/week in the cluster trial). This potentially translates into 1,102 enrolled adolescents over a 12-month period. We will recruit over a 3-year period.

The drop-out rate in the pilot trial is a concern. As noted in our manuscript, we propose several new approaches to reduce drop-out in the definitive trial. These include financial incentive and text/web-based follow-up rather than telephone follow-up. These approaches were successfully used in Cunningham trial that recruited adolescents with harmful and hazardous alcohol use. The retention rate for this trial was 86% at 3-month post ED discharge. While we are in the process of confirming the sample size required for the definitive RCT, given that there are over 3,000 potentially eligible adolescents that may consent to participation, if all these adolescents enrolled, we would have >2,800 at 3-month follow-up.

Reviewer 2 Comment 3: The authors state that their pilot study showed feasibility in terms of blinding. However, while they were successful in maintaining blinding in the current study, this is insufficient evidence to suggest that they would be successful in maintaining blinding in a larger study that recruits 400 participants in 14 EDs across Canada.

Author Response to Comment 3: Yes, we agree that there is no guarantee that maintaining successful blinding in the pilot trial will translate to the same outcome in the definitive trial. However, it is still prudent to review approaches to blinding in a pilot study.

In the cluster RCT, we propose that clinicians deliver the intervention in conjunction with routine care so these individuals will not be blinded in the cluster RCT. Any research staff working in the ED will not be blinded because of the randomization schedule. ED clinicians will need to know when a month is ‘experimental care’ versus ‘regular care’ so they know what care to provide. Posters will be placed up in the ED and research staff will remind clinicians what care they are to deliver each month. Because we wish to study the effectiveness of SBIRT in relation to routine care alone, the comparison intervention will not include a sham intervention like the pilot trial.

We have outlined these changes in our manuscript.

Reviewer 2 Comment 4: Given these issues, my first instinct is to suggest rejection of the paper. However, I believe that the authors have done considerable work in getting to this stage. There should be a publishable paper in this, simply based on feasibility, retention and other artefacts of a complex trial. The authors are more likely to achieve this paper by sticking to the basics and focussing simply on the feasibility, lessons learned and descriptive statistics rather than making questionable and unjustified claims about their sample size for a larger trial.

Author Response to Comment 4: We have made extensive changes throughout to carefully weave in the proposal that we are currently developing for the definitive trial. We have used findings from the pilot trial including feasibility, outcome measurement, and recruitment/retention rates to inform our decisions, and justify our new decisions with supporting literature.

It has been a worthwhile experience to write the proposal for the definitive trial while having this manuscript under review. Many of the decisions that we proposed based on the pilot trial results have

now been made and we can comment on them in the manuscript.

Additional Revisions Made to the Manuscript: Please note that we have revised Table 1 based on earlier feedback from a reviewer to present demographics according to allocation. We have also revised Figure 1 for clarity.

VERSION 3 – REVIEW

REVIEWER	Tapan Rai University of Technology Sydney, Australia
REVIEW RETURNED	03-Jul-2017

GENERAL COMMENTS	I am happy with the effort that the authors have put into addressing my previous concerns. I believe that this final version is far more focussed and more readable than the version I first read, and it should be published in its current form.