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)	Incentivizing safe sex: A randomized trial of conditional cash transfers for HIV and sexually transmitted infection prevention in rural Tanzania
AUTHORS	Damien de Walque, William H. Dow, Rose Nathan, Ramadhani Abdul, Faraji Abilahi, Erick Gong, Zachary Isdahl, Julian Jamison, Boniphace Jullu, Suneeta Krishnan, Albert Majura, Edward Miguel, Jeanne Moncada, Sally Mtenga, Mathew Alexander Mwanyangala, Laura Packel, Julius Schachter, Kizito Shirima and Carol A. Medlin

This paper was submitted to the BMJ but declined for publication following peer review. The authors addressed the reviewer's comments and submitted the revised paper to BMJ Open where it was accepted for publication. These reviews were written for the BMJ.

VERSION 1 - REVIEW

REVIEWER	James Hargreaves Senior Lecturer, LSHTM
REVIEW RETURNED	06/06/2011

This is a potentially important paper - although I am not sure it is of sufficient scientific importance for publication for a general readership such as in the BMJ. It is not suitable for publication in the BMJ its current format – though this could probably be rectified.

I strongly encourage the authors to consider making the changes suggested here with regard the format of reporting - and resubmitting to the BMJ or to another public health / AIDS / STI journal. These are important, truly intersectoral studies – and I wish to encourage them. There is growing understanding of the importance of conditional cash transfers, and other social development and financial incentivisation programmes, for public health – and much interest in this within the public health sector. Much expertise in the design and delivery of these programmes lies outside the health sector as exemplified by this author group. There is much to learn in either direction between researchers in the economics and those in the public health field as this strand of research continues. One area of very clear difference is in normative approaches to reporting. I think this is an area where the public health / medical field has much to offer and I genuinely hope these authors will consider reporting their results in a format that facilitates publication in a high impact public health journal, and, even more importantly, facilitates greater understanding in the public health field and greater collaboration across disciplines.

Specific comments

Reporting

I refer the authors to the CONSORT statement and associated papers that outline approaches to reporting the results of clinical trials in medical / public health journals, and to literature on economic interventions reported in public health journals that adopt these standards – one example being the IMAGE study published by Pronyk, Hargreaves et al in the Lancet 2006. As examples:

- A participant flow diagram is required, and would probably remove the need for the recruitment and numbers analysed sections of the results. Some aspects of the section "participant flow" should be in the methods (eg that potential participants were randomly selected from Ifakara DHSS), while the numbers would probably also appear in the diagram.

- It is important to stipulate primary outcome measures from secondary ones and to identify when the specific analysis reported was planned (at study design; prior to a final data set being available; or during analysis of the final dataset)

- There should be coherence between the methods of analysis implied by the sample size calculation and those conducted (the sample size calc refers to a log rank test, but this analysis does not appear to have been conducted). When the sample size was calculated had the details of the combined endpoint been pre-specified?

- The results in tables 2/3 should include N,s and %s as well as measures of effect. Table 1 – for binary or categorical variables – should report these as n's / %s rather than "sample means" of binary variables.

- The BMJ has previously published good guidelines for structured discussions – that start with a clear statement of the findings as the authors see them in light of their a priori hypotheses. Its not quite clear to me what this statement should say – partly because as I have said above its not very clear what the primary outcome analysis was and when this was specified. It might say that there is from this trial some evidence that a high value CCT, but not a lower value CCT, was associated with decreased prevalence of STIs, but that its unclear what behavioural changes led to this shift and secondary outcome data do not necessarily support the suggestion that this was due to a shift towards sexual behaviour that would reduce risk of HIV in this population.

Other methodological comments

The rationale for a combined endpoint is unclear, and is perhaps difficult to justify – for reasons the authors allude to in their discussion. At the least more discussion of problems in interpretation should be discussed. The 4-bacterial STI endpoint appears to have been dominated by thrichomonas – though we do not currently have any baseline or follow up data on mycoplasma genitalium so its unclear whether this was common or how many people had co-infections. The infections each have different epidemiologies, and in turn this epidemiology is different to that of HIV – which is referred to in the title of the paper and appears to be the real motivation behind the study. There are, for example differences in the age-specific prevalence of HIV and the bacterial STIs used

which mean that great great caution should be taken in inferring that any effect on eg trichomonas or a bacterial STI composite could necessarily have any influence on HIV epidemiology. This is particularly important if there is any possibility that an influence of the intervention might have been to change the age of sexual partners – which is highly plausible in the case of this intervention. (I know of conference papers from Ross and White on this subject – I don't know if published). As the authors do point out – there is also the issue of treatment for bacterial STIs – which might have differed between arms here but would have had no influence on HIV epidemiology and might imply little or no change in sexual behaviour attributable to the intervention. Were sexual behaviour data not captured at all? There are well-described limitations to these data but they seem essential to at least try to get a handle on what it was that actually changed between the groups here.

I can perhaps see a rationale for the combined endpoint if this directly links to the "condition" attached to the cash transfer – but overall I think my preference in design would have been not to use a composite marker. I would certainly encourage the authors to report in a lot more detail what happened over time to each of the specific STIs. Graphs showing the unadjusted prevalence results (and confidence intervals) over time for each of the STIs in I and C groups – and perhaps also shown stratified by sex – would be particularly useful as an adjunct to the reporting of hypothesis tests on the composite outcome.

Process data

For a complex intervention such as that reported on here it seems essential to report data on "process" as is recommended now for trials of complex interventions in public health (Anne Oakleys group have been particularly active and have published on this in the BMJ). I could imagine a number of forms such data might have taken – but the overall aim of such data would to convey to the critical reader things like: was the intervention delivered as intended; was it acceptable and accessed by participants; did intermediary markers (such as sexual behaviour) change in the direction hypothesised etc. There are obvious issues of space limitation – but some data of this type seem essential here.

Detailed comments

The term "group randomised trial" has a specific meaning – the randomisation of groups – and is confusing here. The trial appears to be an unblinded, individually randomised and controlled trial.

The authors use the term "risk ratio" throughout – I think what they present are Odds Ratios from logistic regression models. Risk Ratios in epidemiology refer to ratios of cumulative incidence proportions – here we are talking about relative odds of prevalence. This is an important distinction - particularly so as the outcome is not particularly rare.

I'm also not sure about the emphasis on "proof of concept" – if this phrase is to be used the concept in question should be much more clearly articulated – and the discussion of limitations should highlight how proof of this concept relates to other relevant concepts. For example, I don't think that this study in anyway proves the concept that cash transfers can influence sexual behaviour – no data on this

were collected and the overall balance from outcome results suggests that at least those behaviours relevant to shifts in HIV were not achieved. The abstract must stipulate something about what the "condition" was and how this was monitored.

The additional complexity of a sub-village level randomisation intended to allow the study of "potential peer-effects" is interesting but complex and requires a bit more detail in the limitations section. The inclusion of a fixed term for sub-village in the analysis appears to be intended to "re-correct" the results for this aspect of the design I think (we don't see here any of these analysis on peer-effects – so in the context of this paper this design aspect feels like a limitation, though there may have been interesting reasons for this). However, there are greater difference that one would normally expect in a randomised trial between the unadjusted and adjusted results, and I wonder whether this is partly due to this group level random selection step (given that there were only 10 sub villages, and this step therefore may well have introduced imbalance). However, it might also be related to the problem of non-collapsibility of Odds Ratios from cohorts. It may be useful to provide some more detailed baseline data on the differences between villages at baseline – and how the stratified sampling might have led to some of the baseline differences described. More on steps taken to ensure randomisation was truly blind would have been useful. As it reads it has the potential to be very easily influenced at field level. At least this should be reflected in the limitations.

REVIEWER	Dr Surinder Singh Senior Lecturer in General Practice University College London
REVIEW RETURNED	17/06/2011

Firstly, a word of caution. There are some quite intricate statistics used in this paper and I am no expert (though, importantly I managed to acquire some help with this in that I consulted one of my senior 'stats' colleagues in the dept). If it is likely that this paper is published it ought to be looked at independently from a statistical point of view.

I liked this paper - not simply because of the subject material but because it highlights an important, perhaps all-to-simple, way of reducing high-risk behaviour in a certain population. The results are fascinating.

The introduction is fine along with the methods and the trail design, including a description of the participants and interventions. One question:I wondered why one of the key STIs being incentivised was mycoplasma genitalium (MG) when it seemed to be an unfamiliar infection to the clinicians (page 8). I understand that HIV was also not included for genuine reasons. The role of MG seemed to be ambiguous since there was no testing at baseline.

I also have a question about sample size (page 10); the description is that a log rank test is being used, however the results/tables show logistic regression (which is the entirely appropriate tool to use) - does the initial description need minor modification?

A couple of questions/comments about the tables:

Table 1: Do we need columns 3 & 5? My understanding that as these are baseline descriptive results 'P' values are not necessary. This may be a debatable point? Also in columns 1 & 2 the figures are presented as fractions of 1 (0.499 & 0.511); would percentages be clearer (perhaps with standard deviations)?

Tables 2/3: Fairly clear - though why is mycoplasma genitalium included; see my comments above about MG.

Discussion - I think this is fine and is a well-written account of the what the trial has shown, including within that a robust examination of the limitations. I agree that some of the findings are perverse (positive results only seen at specific points) - but further study may shed light on this.

Generalizability: again I agree and this is well-written. For those unfamiliar with "proof of concept" - might this warrant a definition or description?

Abstract: I think this is fine and describes well what follows in the trial.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1 – James Hargreaves

Thank you for your comments about the interest of the study and the need for interdisciplinary exchanges. We have followed your recommendations and those of the editors to improve the reporting of our results.

We have included a participant flow diagram that is included as figure 1 (separate file) and further described under the participant flow sub-section in the results section. The selection from the Ifakara Health and Demographic Surveillance System is described under the participants sub-section in the methods section.

In the outcomes sub-section of the methods section, we define our primary outcome measure: “The primary outcome measure, as defined in the study protocol, is the round-specific combined point prevalence of the four sexually transmitted infections that were regularly tested – *Chlamydia trachomatis, Neisseria gonorrhea, Trichomonas vaginalis,* and *Mycoplasma genitalium* – at months 4, 8, and 12. “This primary outcome measure was planned at the study design in order to have sufficient power. We now say it explicitly in the text.

The sample size sub-section has now been revised to reflect the power calculations for the analysis reported. The previous power calculations were those used when the project was initially conceived,

but these were updated to the currently reported power calculations during the final stages of project design and recruitment planning.

We have made the suggested changes in all tables.

As suggested, we have added a statement at the beginning of the discussion section describing the main findings of the study.

The measure of combined point prevalence was constructed at study design to ensure sufficient power to detect differences in the control and treatment groups in response to the intervention (the conditional cash transfer). While it would have been interesting to study specific trends in prevalence rates of the various STIs tested over time, we were not powered to do so, and in any case, these specific trends would yield more insight into the specific transmission patterns of each STI and the susceptibility of the research population to infection than into sexual practices, per se. Our objective was in some ways much narrower than this, as we were seeking sufficient power to detect differences that would indicate changes in behavior as a result of the intervention, rather than transmission patterns of different STIs within this population. The impact of financial incentives on behavior relating to sexual health was in fact confirmed by our study, although the biological outcomes cannot be used to infer the relative importance of STI treatment seeking behavior versus a reduction in risky sexual activity (e.g. increased condom use, number of partners). It is true as the reviewer states that this result does not confirm a reduced risk of infection for HIV in this population, it does point to the importance of a behavioral pathway (via treatment seeking behavior or changing sexual practices).

In the text (outcome sub-section of the methods section), we have strengthened the rationale for reliance on the composite marker.

As requested, we are also providing as a supplemental table including the prevalence by study arm for each STI at each round. We provide it as additional information for the reviewers and editors, but we would think that because our study was not powered to detect reduction in individual STI prevalence and because of space limitation that table should not be included in the published paper. Of course, if the referees and editors think otherwise, we would be happy to include in this or another format.

As far as sexual behavior data are concerned, we have collected quantitative and qualitative data. We are developing separate manuscript for its analysis and it has been analyzed in the following PhD dissertation:

We collected a significant amount of process data in conducting this study. As the reviewer indicates, space limitation makes it difficult to provide the details that would be of interest to many readers. However, we include some information about whether the intervention was acceptable and accessed by participants and how it was perceived.

We refer the reviewer to the participant flow diagram, and have added a process sub-section in the results section. We are copying here this sub-section:

We have made the suggested modifications in the abstract and the trial design sub-section of the methods section.

What we are presenting are relative risks, i.e. the probability of being STI positive in the intervention group, divided by the probability of being STI positive in the treatment group. The term relative risk ratio can be confusing, thus we now instead use the clearer term “relative risk” in the statistical methods.

We have removed the reference to “proof of concept” in the abstract and the body of the text.

First, we note that there were 10 villages and 50 sub-villages. It is true that we introduced indicator variables for each sub-village in the adjusted models to account for the different selection probability (and potentially associated peer-effects) at the sub-village level. We have added one sentence in the limitation section to account for this.

“In order to study potential peer-effects, in randomly selected sub-villages, the probably of selection in the intervention arm was 75% and in the other sub-villages, it was 25%. This might have led to baseline imbalances. For this reason, we included sub-village indicator variables in the adjusted models. This might explain some of the differences between the results from the unadjusted and the adjusted models”.

Randomization was not blind. Participants were not blinded to arm assignment since awareness of their eligibility for the conditional cash transfer was a critical component of the intervention. We believe the reviewer refers to the possibility of manipulation of the randomization at the field level. We think our very transparent procedures eliminated this risk. We added in the randomization sub-section of the methods section that the randomization step took place in public view, minimizing the potential for manipulation.

Reviewer 2 – Surinder Singh

We have modified the tables to include percentages, but we have kept the p-values in table 1 to underline the few imbalances at baseline.

As indicated in the response to the editor’s comments, the sample sub-section has now been revised to reflect the power calculations for the analysis reported. The previous power calculations were those used when the project was initially conceived, but these were updated to the currently reported power calculations during the final stages of project design and recruitment planning. As per the reviewer’s intuition, we too had initially considered the log rank test to be most appropriate. However, fieldwork in preparation for project launch suggested the strong possibility of time-varying effect sizes, which led us to instead prefer an approach that would estimate separate models at each post-treatment time point.

Mycoplasma Genitalium was included in the primary study endpoint calculation to increase power. However, we did not tie the CCT payments to participants to a negative test result for m Gen. While m Gen has been shown to be incontrovertibly linked to risky sexual activity, there is some uncertainty around transmission pathways. Rather than risk penalizing participants from testing positive if it was unrelated to risky behavior, we chose to use the aggregate results in the composite measure to increase power for the study.

We have modified the sentence about the lack of familiarity of participants and clinicians with m Gen, and replaced it with an explanation of why m Gen can be used to increase power but is less appropriate for conditionality.

VERSION 2 – REVIEW

REVIEWER	Jon Deeks Professor of Biostatistics University of Birmingham
REVIEW RETURNED	20/09/2011

This manuscript reports results of a randomized trial of financial incentives to increase safer sexual behaviours. The study did not proceed as planned, and the results do not provide convincing evidence that the intervention works. The authors find a significant effect at one time point using an adjusted analysis, and interpret their findings as being more conclusive than probably can be justified.

1. There appears to be a degree of post hoc rationalization of sample size calculations which is not made entirely clear to the reader, with a discrepancy between the sample size calculation reported in the study protocol and the calculation reported in the paper. The authors do acknowledge at the end of the sample size calculation section that they had initially intended to recruit more participants but they do not explicitly state that the calculation which is reported was undertaken post hoc and based upon the incidence observed in the study. The protocol reports a calculations based on detecting (relative) differences of a 30% magnitude or greater in the treatment arm compared to the control arm with incidence rates varying between 15%-20% across research sites and drop-out rates as high as 20% per year giving a total sample size of 3000 individuals. The paper states that the power calculation was based on an incidence rate of 12%, controlling for one baseline measure for a one-third reduction requiring 2400 individuals with no mention of any drop-out.

2. Calculation of P-values between randomized groups at baseline is illogical – the P-values indicate the probability that differences have occurred by chance – as all differences are created by randomization, they must have occurred by chance, so why calculate a probability? What is important is the magnitude of the differences, not the P-values. Please remove the P-values from Table 1 and the baseline data section of the results. The use of 2 decimal places on the percentages in this table is also not justified – it implies excessive precision.

3. There is no mention in the statistical methods whether an intention-to-treat process was followed for the analysis and how missing data were handled (although there is a section describing how much missing data existed).

4. I would have expected the results section to have reported on the incidence of the outcome measure – this is not mentioned at all in the text and is somewhat cryptically reported in Table 2 (wrongly labeled “sample mean”) and more appropriately in Table 3. However, the actual numbers positive are never reported by randomized group, which is highly desirable (and I believe required by the CONSORT guidelines).

5. Please note that the logistic regression model will have estimated odds ratios and not relative risks. With an incidence rate of 12% the figures will be close to relative risks but they should be described properly.

6. The results section on "outcomes and estimation" focused on the statistical significance of the comparisons, with very little "estimation" of treatment effects for the main comparisons. It would be helpful to give the estimates (with 95% CIs) in this section – for example you can state that the unadjusted analysis estimated a reduction in the odds of STI of 20% (95% CI: 6% increase to 46% reduction) at 12 months, whereas the adjusted model estimated a reduction of 27% (95% CI: 1% to 53% reduction)

7. Tables 2 and 3 do not state what the comparator group is for computation of the relative risks (odds ratios). Inclusion of standard errors in these tables probably isn't helpful – they are figures on the log odds ratio scale. The confidence intervals are more useful, and should a reader require a standard error they could be computed from these values.

8. There is no explanation of how the adjustment variables were chosen, whether they were prespecified (the protocol has no statistical methods section so I would presume that they were not prespecified) and the manner in which they were categorized or used as continuous measures. The comparison of the effect in males and females must have been undertaken using a test of interaction but this is not currently mentioned.

9. From the results which have been obtained I am not clear that the authors can conclude with any reasonable degree of certainty that there is a benefit of this intervention. However, the headline for the discussion (and the summary points) is that there was a significant reduction for the higher $20 payments (opening sentence of discussion). However, this reduction was only observed as being statistically significant when adjustments were made, only at one of the three time points, and only when the serum tests were not considered. It therefore seems to be an overstatement of the findings.

10. One conclusion from the trial (mentioned in the abstract) is that a further study needs to be done to clarify the magnitude of the benefit. This statement is presumptive about there being a proven benefit, and one wonders how well the case can be made for a larger trial which would be needed (probably requiring over 10,000 participants).

VERSION 2 – AUTHOR RESPONSE

We have been more explicit on the smaller sample size than originally anticipated in the "Sample size" section (this was the only key aspect in which the study did not proceed as planned). We have qualified our results in the abstract, the first paragraph of the discussion section and the summary points box.

In the revised "Sample size" section we have now explicitly indicated that the power calculation presented is an ex-post calculation based on the observed properties of the actual recruited sample and infection rates. In the original manuscript submission we had presented the power calculations from the study protocol, but were asked to remove these. We agree with the earlier reviewers that the power calculations corresponding to our actual data analysis are most directly relevant for readers. We have removed the p-values from table 1 and from the baseline data section of the results. We have also removed the second decimal from table 2.

We have added the following sentences to the beginnng of the statistical methods section to clarify these two points:
"Each individual was coded as per their initial randomized assignment as per an intent-to-treat design. However, individuals who were not present at any given round were treated as missing and dropped from the analysis for that round due to lack of outcome data."

Our tables do in fact report relative risks, not odds ratios. We understand that there is some difference in use of the term relative risk, but we are using the term in the same sense as indicated in the BMJ Clinical Evidence glossary (http://clinicalevidence.bmj.com/ceweb/resources/glossary.jsp). Although logistic regression will yield estimated odds ratios, we have transformed the effects into relative risks using the "margins" and "nlcom" commands in *Stata 12* (using the method as recommended e.g. in: Kleinman LC, Norton EC. What's the risk? a simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. Health Services Research 2009; 44: 288-302). This is now explicitly stated in the "Statistical methods" section of the text.

We prefer to report the relative risks for two reasons. First, this was an interdisciplinary project, and while we believe that publication in BMJ is highly appropriate, we also want the magnitudes to be readily interpretable by other audiences as well such as economists who do not typically use odds ratios. Second is the related point referred to by the reviewer, that with an incidence rate of .12, the odds ratio will show a larger reduction than the relative risk. For example, where we report a relative risk of 0.73, the corresponding odds ratio would have been reported as 0.69. To avoid the problem of readers having to do the calculation themselves to understand how different the odds ratio is from the more interpretable relative risk, we prefer to directly report the relative risk.

We have now provided this statement of estimates of the effects for the main results (high value CCT group at month 12) for the unadjusted and adjusted model in the outcome and estimation section. Expanded notes under tables 2 and 3 now indicate that the reference group for the computation of the relative risks is the control group. We have removed the standard errors from tables 2 and 3, retaining the confidence intervals.

While the adjustment variables were not explicitly pre-specified in the protocol, they are standard socio-demographic variables. We have now indicated in the statistical methods that age and income are continuous variable and that the other adjustment variables are categorical. In the outcomes and estimation sub-section of the results section, we now make clear that we ran a test of interaction for the difference between males and females and that the interaction term for female was not significant. We have revised the headline of the discussion to make clear that the statistically significant results were only obtained in the adjusted model and only at month 12 and not at earlier time points, for the high value cash payments and for the serum tests. We would be even more tentative if the results were only statistically significant during an early time point rather than the final time point; however the pattern of increasing effect over time is consistent with our a priori hypothesis and indeed is the reason why we structured the intervention to have multiple incentivized testing rounds. As noted in the "Interpretation" section, "the impact of the conditional cash transfer may take time to materialize, perhaps because it is not easy to extricate oneself from complicated sexual relationships, or perhaps because participants needed time to become accustomed to (and trust) the incentive mechanism."

In the abstract and the summary box, we now say that the intervention is "potentially promising" rather than promising and we add that a larger study would be useful to clarify the effect size, to calibrate the size of the incentive, and to determine whether the intervention can be delivered cost effectively.