PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis
AUTHORS	McLintock, Kate; Russell, Amy; Alderson, Sarah; West, Robert; House, AO; Westerman, Karen; Foy, Robbie

VERSION 1 - REVIEW

REVIEWER	Chris Burton University of Aberdeen, UK
REVIEW RETURNED	12-Apr-2014

GENERAL COMMENTS	I have two major concerns about the interpretation of the data which may require some further analysis. Both relate to the question about whether the increased diagnosis and treatment rates are related to screening directly (ie screened patients get diagnosed / treated) or indirectly (screening heightens awareness). This is an important question as current limited evidence suggests direct effects are small. The concerns are: 1. Apparent short term temporal dissociation of monthly screening and diagnosis rates. In the period 2007-2012 inspection of figure 1 suggests an annual cyclical pattern of screening. The prior probability of this being present would be high as practices are known to chase up missing QOF items before the end of the QOF year in March, so the data are likely to be correct. There is no appearance of periodicity in the diagnostic and prescribing rates. It should be possible to test for periodicity using time series analysis. If periodicity is present in the screening data and screening "works" then one should expect periodicity in the downstream events. 2. Reverse causality. The authors don't really pick up on the possibilities for reverse causality where patients (with one of the designated conditions) present with mood symptoms and either (a) the GP then codes them as "screened" because they have asked screening questions during the consultation or (b) the GP makes a diagnosis, completes a PHQ9 and then finds that their computer system automatically codes for screening (PHQ9 includes the 2 question screener). In our 2012 paper we conducted analysis which removed same-day screen and diagnosis / treat patients. This either needs to be addressed as a limitation or, preferably, addressed in further analysis. There are a couple of more minor criticisms 1. I don't think the 2006 blip is properly explained. What's going on here? is this multiple coding of the same patients? (it must be

	because subsequent rates are over 85%). Even if the authors don't know, some plausible explanation for the size of the effect would be useful 2. I think the discussion should include reference to some of Ziegelstein and Thoombs' more recent critiques of depression screening. This is important in setting the prior probability for large direct effects of screening to a low level. If taken at face value, this study seems to suggest quite a strong direct effect of screening which is at odds with other studies. I think the authors need to consider more carefully whether their observed findings indicate direct or indirect effects. That is not to say that indirect effects aren't either real or important - they are, but we need to know.

REVIEWER	Evangelos Kontopantelis University of Manchester, UK
REVIEW RETURNED	22-Apr-2014

GENERAL COMMENTS	Thank you for giving me the opportunity to review this interesting and generally well written paper. The authors set out to investigate depression case finding over time, for two groups of chronic condition patients. Case finding for the first group was incentivised through the Quality and Outcomes Framework from 2006-7 (QOF y3) until 2012-13 (QOF y9), while there was no incentive for the second group. Comparisons in case finding are made both within and between the two groups, over time. The paper is well written and interesting but I think clarifications are needed about the methods (the discussion I found excellent). Personally, I would not use a monthly time-window but would analyse using year aggregates and I will explain below why. Either way (monthly or yearly) methodological details will need to be provided. **Major points** 1) Interrupted Time Series Analysis, although the best quasi-experimental design in the absence of a control group, it does come with a lot of baggage. One such assumption is the linearity of the slopes pre- and post-intervention. Just looking at the graphs one can reason that this is not the case here, because of the expected QOF seasonality in coding. So if the authors did what I think they did (and like I said clarification on the methods is needed), adjusting for seasonality would be required to ensure linearity of the trends. Alternatively, the analysis could use yearly time windows to avoid this problem, and this would be my preferred choice. So: a. Seasonality is not mentioned at all (in Methods=>Data analysis) when the time series is defined – and hence not addressed b. Linearity assumption? (in Methods=>Data analysis)

c. Software used (+reference) and ITS model (+reference) need to be clarified (in Methods=>Data analysis).

d. As far as I know, when it comes to ITS, one intervention can be realistically modelled if one is to quantify the effect of an intervention (using components for: slope pre-, level change and slope change post-intervention). More than one intervention can be modelled but then interpretation of the slopes and level changes are problematic. It is not clear how the authors incorporate three interventions in a model (and how they model level and slope changes) and what their actual comparisons are e.g. is slope 2002-2004 always used as the baseline for comparison? The authors say “For each time period … the model has an overall constant and slope” but I did not find this description to be informative. The authors might need to describe the model in detail, using notation, in the appendix.

2) Graphs are nice but it can be difficult to obtain information on incidence and prevalence of depression. I would expect to see a table summarising these yearly, over time, and possibly by clinical computer system used in the practice.

3) Methods are completely missing from the abstract. I appreciate that the authors are limited as to how much information they can report in an abstract but the BMJ group allows a reasonable word limit and the authors need to manage sections better, if this is a word limit issue. For example, odds ratios are reported in the abstract results, but there is no information on the statistical test/hypothesis they relate to.

4) On the incentive. I would expect the authors to mention the indicator, DEP1, and perhaps even provide its exact wording in the framework, as well as discuss it’s characteristics (lower, upper threshold, points worth etc). Perhaps in a table if they do not wish to discuss in the text. However, I would want to see the associated costs mentioned in the introduction and possibly the discussion section (using cross-sectional information, say for the last year 2011-12): overall cost estimate, mean cost per practice and, ideally, mean cost per surplus new case (above expectation, although this might require a new analysis). *Update*: I see now that the authors mention overall cost in the discussion but very briefly. Also need to be put in context of cost of 1bn per annum (i.e. whole of QOF), the overall benefits of which have are still debated.

5) Odds Ratios are reported in the abstract and in the results section but it is unclear what model was used for the analyses. They do mention binomial regression at one point in the results but this should have been discussed much earlier. Also the term ‘binomial’ is too generic and not very

informative e.g. what link function was used for the model? Logit, which is the most common, hence logistic regression or something else (e.g. probit). If you used logistic regression, please state as that, rather than binomial regression. Similarly tests for slopes are first mentioned at the end of the results section.

Minor points

1) Abstract, setting: I would prefer an inclusion of the geographical level. Probably: “General practices in Leeds Primary Care Trust (PCT; now Clinical Commissioning Group, CCG), UK”.
2) Sample representativeness is always a sore point. I do not like it when authors say with confidence that a sample is representative without formally comparing the distributions of a few key variables in the sample and the population (see for example this that discusses the subject and proposes a formal comparison http://www.jstatsoft.org/v55/c01), although I must admit that this is common practice and I have done it at times… So I would rephrase “representative sample” in the article summary section with a phrase that allows room for uncertainty e.g. “appears to be broadly representative on key parameters”.
3) The detailed table displaying characteristics (table 1) would benefit from inclusion of rurality (using census data) and computer system used. If the authors cannot find rurality data I’m happy to provide them (rural/urban classification and practice code, for all English practices). Clinical computer system use for all English practices has been discussed here: http://www.ncbi.nlm.nih.gov/pubmed/23913774
4) Table 1 should include another column with the characteristics of the practices that were not recruited as well (if not all info is available, report what is known e.g. list size, FTE, deprivation which can be very easily retrieved); I would expect them to be in more deprived areas and smaller.
5) Background, last sentence of first paragraph (“According to expected prevalence…”): this reference is now 19 years old and it is not entirely clear how this is still the case in general practice, after all the initiatives, potential changes in practice etc. Please rephrase to explain if that’s still the case and why.
6) Background, second paragraph (“…over 2006-13…”): I would add “(QOF years 3 -9)” in brackets.
7) Methods=>Practices and Participants=> “No distinction was made…” There is some evidence that clinical system is a predictor of practice QOF performance, certainly stronger than deprivation or other characteristics (Please see http://www.ncbi.nlm.nih.gov/pubmed/23913774). At least I would expect the authors to report the systems break-down

for their sample.

8) Methods=>Practices and Participants=> "Recorded depression in adults…" the authors seem to say that 2011-12 is the last year of incentivisation, when it actually is 2012-13? I suspect they mean that 2011-12 is the last year they have data for. This needs to be rephrased to explain the distinction better.
9) Methods=>Practices and Participants=> "Patients with conditions in both targeted and non-targeted groups…" I agree with the authors' approach, however, I do wonder if there is a risk that the targeted group will include patients with more comorbidities and hence will be more likely to be identified with depression. This should be mentioned as a potential limitation in the discussion section.
10) Methods=>Practices and Participants=> "…MIQUEST query" Is a reference needed here? Also do queries differ by clinical system? If yes please rephrase and mention clinical systems.
11) I must commend the authors on their practice of reporting all used Read and drug codes. If they have some time it might be worth uploading these to a new repository specifically aimed to address study transparency and reproducibility: www.clinicalcodes.org. Uploading the codes there will make them directly accessible to researchers and expose your work more. This is just a thought though: I do not expect you to use the website.
12) Methods=>Data analysis=> "We took the number of registered…" I would think this is an unnecessary assumption to make (i.e. using the correct denominator within each year shouldn't be that hard) but if they say that the sensitivity analysis gave the same results it means that they did use the more appropriate approach as well. So why are they not reporting that analysis as the main one to avoid the issue altogether? Even if they decide to stick to the current approach, more information will need to be provided on what exactly the sensitivity analysis entailed and comment on the differences (or just say none observed, more elaborately…) in the results section. Possibly also say that the results are available from the authors.
13) Methods=>Data analysis=> "A further discontinuity…" I think the authors need to explain why exception performance needed to be isolated and clarify how that was done through an interrupted time series analysis.
14) Results. Discuss characteristics of recruited, not recruited and all English practices (briefly).
15) Results. Section with 'raw' results required, linked to the requested table with annual incidence and prevalence rates. The authors do report month incidence rates when they discuss comparisons but I feel a dedicated short section is required to clarify the picture.
16) Results. The authors report on comparisons that were not clarified in the methods section and I was unsure of the

question they were addressing (discussed in detail as a main point) e.g. comparison of rates between periods 2002-4 and 2007-11 (I imagine a typo and you mean 2007-12), or 2004-6 to 2007-12.

17) Results “that is the rates can be, and were, taken as constant during these periods” probably rephrase to something using “can be assumed to be constant”.
18) Results “…with fitted constants and slopes” Unclear what the authors mean.
19) Results “During the period after QOF was introduced but before incentives…” My guess would be that the authors got the years wrong there, should be April 2004 to March 2006?
20) Discussion. “Rates of new prescriptions for antidepressants exceeded…” This statement, although verified by the graphs, is not supported by the evidence presented in the results section. It needs to be.
21) Discussion=>Limitations. The authors are right about self-selected practices and the risk of selection bias. However, the relative benefits of this study would be obvious if the response rate was higher i.e. there is quite a lot of room for selection bias in a response rate of 50%. Usage of a single computer system is potentially a bigger headache than self-selection (see paper I mentioned before). Also I understand what you mean about “ceilings on performance” (but I am not sure everyone would) and for example CPRD practices are better QOF performers (by approximately 1%) but the “ceiling” is inherently a QOF problem (i.e. that 1% does not cause the issue). In reference 17 we used a CPRD sub-sample specifically selected to be as representative as possible i.e. we did not use the whole of the CPRD (and they were not higher performers). To summarise, agree there are potential issues but probably rephrase: lose ceiling, add single computer system?
22) Table 1 (revisited…): year of comparison needs to be placed on title rather than in footnotes. England average list size looks odd. I’m pretty sure it’s over 6,500 but it might have been that in 2002? In that case might be worth adding 2 more columns and list characteristics at the beginning and (2002) and the end of the study (2012).

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

I have two major concerns about the interpretation of the data which may require some further analysis. Both relate to the question about whether the increased diagnosis and treatment rates are related to screening directly (ie screened patients get diagnosed / treated) or indirectly (screening heightens awareness). This is an important question as current limited evidence suggests direct effects are small. The concerns are:

1. Apparent short term temporal dissociation of monthly screening and diagnosis rates. In the period 2007-2012 inspection of figure 1 suggests an annual cyclical pattern of screening. The prior probability of this being present would be high as practices are known to chase up missing QOF items before the end of the QOF year in March, so the data are likely to be correct. There is no appearance of periodicity in the diagnostic and prescribing rates. It should be possible to test for periodicity using time series analysis. If periodicity is present in the screening data and screening "works" then one should expect periodicity in the downstream events.

We acknowledge that annual periodicity, seasonality, is clearly seen in the group of patients with targeted conditions following the introduction of incentives for depression case finding. There may also be some seasonality for the other conditions and for other periods, and the profile of the seasonality may vary. Explicit seasonality was explored and it did improve the model fit but at the expense of model complexity with tens of additional terms. We sought to reduce complexity where it was not necessary. Since the periods within the model are complete years, there is no change to the overall level or to the slope related to these periods. This is because the terms are orthogonal to the seasonality terms. We have added text to the manuscript in order to:
(1) Acknowledge seasonality, and that it is especially apparent after case finding incentives have become established (after 2006/7)
(2) Note that our modelling is robust so that levels and slope are unaffected by seasonality which is incorporated in the error term.

2. Reverse causality. The authors don't really pick up on the possibilities for reverse causality where patients (with one of the designated conditions) present with mood symptoms and either (a) the GP then codes them as "screened" because they have asked screening questions during the consultation or (b) the GP makes a diagnosis, completes a PHQ9 and then finds that their computer system automatically codes for screening (PHQ9 includes the 2 question screener). In our 2012 paper we conducted analysis which removed same-day screen and diagnosis / treat patients. This either needs to be addressed as a limitation or, preferably, addressed in further analysis.

Our study was mainly designed to evaluate the effects of incentivised case-finding on targeted patient populations with diabetes and coronary heart disease. We could not directly answer the question as to how incentivised case-finding exerted its effects, although we can offer insights from our analysis of non-targeted patient populations and our accompanying ethnography. From these we judge that incentivised case-finding exerted its effects in at least three ways: case-finding applied with fidelity to the recommended questions; case-finding with individual and often loose adaptations of the recommended case finding questions; and through awareness-raising. There may have been other mechanisms, as Guthrie and Morales recently highlighted (BMJ 2014;348:g1413). We now amplify this point for readers in the text.
On a minor point, we are uncertain from our own enquiries as to whether coding a PHQ9 did automatically trigger a DEP1 code.

Minor criticisms
1. I don't think the 2006 blip is properly explained. What's going on here? is this multiple coding of the same patients? (it must be because subsequent rates are over 85%). Even if the authors don't know, some plausible explanation for the size of the effect would be useful

The results section has been expanded to include further comment on the spikes in coding activity seen in 2006. We do not believe this is multiple coding of the same patients as only the first screening, diagnosis or antidepressant prescribing code recorded in the patient’s record were included in the analysis.

2. I think the discussion should include reference to some of Ziegelstein and Thoombs' more recent

critiques of depression screening. This is important in setting the prior probability for large direct effects of screening to a low level.

Thank you for highlighting this body of work by Thombs and Ziegelstein which we now cite (Thombs B, Ziegelstein R, Roseman M, et al. There are no randomized controlled trials that support the United States Preventive Services Task Force guideline on screening for depression in primary care: a systematic review. BMC Medicine 2014;12(1):13). Corresponding changes have been made to text in the Background.

If taken at face value, this study seems to suggest quite a strong direct effect of screening which is at odds with other studies. I think the authors need to consider more carefully whether their observed findings indicate direct or indirect effects. That is not to say that indirect effects aren't either real or important - they are, but we need to know.

We were initially surprised by apparent strong effects. Comparison with the modest effects found elsewhere, and acknowledging rates of new depression-related diagnoses rose in non-targeted long-term conditions coincident with only a modest rise in recorded case finding in these patients in this work, suggests a sizeable indirect effect. However, we are unable to indicate whether the observed findings were due to direct or indirect effects as our study was designed to assess population effects. This issue has been highlighted in the discussion.

Reviewer 2
Thank you for giving me the opportunity to review this interesting and generally well written paper. The authors set out to investigate depression case finding over time, for two groups of chronic condition patients. Case finding for the first group was incentivised through the Quality and Outcomes Framework from 2006-7 (QOF y3) until 2012-13 (QOF y9), while there was no incentive for the second group. Comparisons in case finding are made both within and between the two groups, over time.
The paper is well written and interesting but I think clarifications are needed about the methods (the discussion I found excellent). Personally, I would not use a monthly time-window but would analyse using year aggregates and I will explain below why. Either way (monthly or yearly) methodological details will need to be provided.

Major points
1) Interrupted Time Series Analysis, although the best quasi-experimental design in the absence of a control group, it does come with a lot of baggage. One such assumption is the linearity of the slopes pre- and post-intervention. Just looking at the graphs one can reason that this is not the case here, because of the expected QOF seasonality in coding. So if the authors did what I think they did (and like I said clarification on the methods is needed), adjusting for seasonality would be required to ensure linearity of the trends. Alternatively, the analysis could use yearly time windows to avoid this problem, and this would be my preferred choice. So:
a. Seasonality is not mentioned at all (in Methods=>Data analysis) when the time series is defined – and hence not addressed
b. Linearity assumption? (in Methods=>Data analysis)

We refer to our response on similar points raised by Reviewer 1.

c. Software used (+reference) and ITS model (+reference) need to be clarified (in Methods=>Data analysis).

We have added details of the software used, including citations as requested by the software authors.

d. As far as I know, when it comes to ITS, one intervention can be realistically modelled if one is to quantify the effect of an intervention (using components for: slope pre-, level change and slope change post-intervention). More than one intervention can be modelled but then interpretation of the slopes and level changes are problematic. It is not clear how the authors incorporate three interventions in a model (and how they model level and slope changes) and what their actual comparisons are e.g. is slope 2002-2004 always used as the baseline for comparison? The authors say “For each time period … the model has an overall constant and slope” but I did not find this description to be informative. The authors might need to describe the model in detail, using notation, in the appendix.

The interventions we consider are (1) the introduction of depression case finding incentives and (2) the introduction of QOF two years prior to this. Consequently our interventions are not concurrent. The reviewer might be concerned that migration of practices from EMIS to SystmOne could be viewed as a further intervention. This would introduce identification issues into the model. We note though that migration occurred throughout the study period and so any differences in electronic record systems would arise as changes in slopes rather than step changes in level. We have not added the further complexity of modelling record systems as it is just one of several potential influences on the observed rates. We report descriptively and do not aim to claim causality. We did however observe substantial step changes following the introduction of case finding incentives which differ between the group of targeted condition and the group of other long-term conditions.

2) Graphs are nice but it can be difficult to obtain information on incidence and prevalence of depression. I would expect to see a table summarising these yearly, over time, and possibly by clinical computer system used in the practice.

We have data only on incidence and so only report incidence. A table summarising annual counts and rates per 100,000 patients has been added to results. For comments on the potential differences between clinical systems, please see our response to 2.1. We have added text to the limitations section to acknowledge this possibility.

3) Methods are completely missing from the abstract. I appreciate that the authors are limited as to how much information they can report in an abstract but the BMJ group allows a reasonable word limit and the authors need to manage sections better, if this is a word limit issue. For example, odds ratios are reported in the abstract results, but there is no information on the statistical test/hypothesis they relate to.

We have supplied more information on statistical methods as the reviewer requested.

4) On the incentive. I would expect the authors to mention the indicator, DEP1, and perhaps even provide its exact wording in the framework, as well as discuss it’s characteristics (lower, upper threshold, points worth etc). Perhaps in a table if they do not wish to discuss in the text. However, I would want to see the associated costs mentioned in the introduction and possibly the discussion section (using cross-sectional information, say for the last year 2011-12): overall cost estimate, mean cost per practice and, ideally, mean cost per surplus new case (above expectation, although this might require a new analysis). Update: I see now that the authors mention overall cost in the discussion but very briefly. Also need to be put in context of cost of 1bn per annum (i.e. whole of QOF), the overall benefits of which have are still debated.

Text has been added to the background, expanding information on the indicator, its characteristics and updating the estimate of costs. We acknowledge that we were unable to provide a comprehensive cost analysis but place the 2012-13 estimates in the context of the cost of QOF as a whole.

5) Odds Ratios are reported in the abstract and in the results section but it is unclear what model was used for the analyses. They do mention binomial regression at one point in the results but this should have been discussed much earlier. Also the term 'binomial' is too generic and not very informative e.g. what link function was used for the model? Logit, which is the most common, hence logistic regression or something else (e.g. probit). If you used logistic regression, please state as that, rather than binomial regression. Similarly tests for slopes are first mentioned at the end of the results section.

We have supplied more information on statistical methods as the reviewer requested.

Minor points
1) Abstract, setting: I would prefer an inclusion of the geographical level. Probably: "General practices in Leeds Primary Care Trust (PCT; now Clinical Commissioning Group, CCG), UK".

Unfortunately the word limit did not permit this change to be made following the addition of methods to the abstract, as suggested in major comment (3).

2) Sample representativeness is always a sore point. I do not like it when authors say with confidence that a sample is representative without formally comparing the distributions of a few key variables in the sample and the population (see for example this that discusses the subject and proposes a formal comparison http://www.jstatsoft.org/v55/c01), although I must admit that this is common practice and I have done it at times… So I would rephrase "representative sample" in the article summary section with a phrase that allows room for uncertainty e.g. "appears to be broadly representative on key parameters".

Wording has been changed following the reviewer's advice.

3) The detailed table displaying characteristics (table 1) would benefit from inclusion of rurality (using census data) and computer system used. If the authors cannot find rurality data I'm happy to provide them(rural/urban classification and practice code, for all English practices). Clinical computer system use for all English practices has been discussed here: http://www.ncbi.nlm.nih.gov/pubmed/23913774

The table of practice characteristics has been updated to include rural/urban classification as proportion of practices in urban areas. These data are based on Office for National Statistics figures of Lower Super Output Area from the 2001 census and the data on practices are from the Health and Social Care Information Centre Indicator Portal (https://indicators.ic.nhs.uk/) published in 2011.

A breakdown of clinical computing systems used by participating practices at the time of data collection (in 2011) is also added. Data are presented by three categories of clinical system; TPP SystmOne, EMIS (combined LV, PCS and Web) and other (iSoft Premiere, iSoft Synergy, InPS Vision, Healthysoft). We were aware that some EMIS and 'other' products were used by single practices and that presenting this more detailed data, already identified as originating from Leeds, might make individual practices identifiable.

4) Table 1 should include another column with the characteristics of the practices that were not recruited as well (if not all info is available, report what is known e.g. list size, FTE, deprivation which can be very easily retrieved); I would expect them to be in more deprived areas and smaller.

The table of practice characteristics has been extended to include data on non-participating practices. The first column of the figures is for 'All England', while the next two columns are recruited and non-recruited practices. The column "p" is from the comparison of those recruited and not recruited. There is no comparison to ‘All England’ as the local practices are in the ‘All England’ group as well, and therefore cannot be compared to a group containing themselves. The practices recruited were larger but with no difference in Indices of Multiple Deprivation (IMD). The IMD values in this table are changed from the previous draft due to an update in the way the data was calculated since the table was originally compiled. Previously, the IMD was the deprivation score of the practice postcode (as with urban classification), and now it is the average IMD of individual patients across the practice, providing a more representative figure.

5) Background, last sentence of first paragraph (“According to expected prevalence…”): this reference is now 19 years old and it is not entirely clear how this is still the case in general practice, after all the initiatives, potential changes in practice etc. Please rephrase to explain if that’s still the case and why.

This reference has been deleted and the text updated.

6) Background, second paragraph (“…over 2006-13…” PubMed ;): I would add “(QOF years 3 -9)” in brackets.

This annotation has been inserted.

7) Methods=>Practices and Participants=> “No distinction was made…” There is some evidence that clinical system is a predictor of practice QOF performance, certainly stronger than deprivation or other characteristics (Please see http://www.ncbi.nlm.nih.gov/pubmed/23913774). At least I would expect the authors to report the systems break-down for their sample.

Thank you for alerting us to this omission. A breakdown has been provided, along with insertion of this reference on the bearing of practice choice of clinical computing system in the discussion. Comment on the lack of data on clinical computing systems has been added to discussion of limitations of this study.

8) Methods=>Practices and Participants=> “Recorded depression in adults…” the authors seem to say that 2011-12 is the last year of incentivisation, when it actually is 2012-13? I suspect they mean that 2011-12 is the last year they have data for. This needs to be rephrased to explain the distinction better.

Thank you for highlighting this ambiguity. The section has been rephrased to ensure the distinction between the end of data collection and the last year of incentivisation is clear.

9) Methods=>Practices and Participants=> “Patients with conditions in both targeted and non-targeted groups…” I agree with the authors’ approach, however, I do wonder if there is a risk that the targeted group will include patients with more comorbidities and hence will be more likely to be identified with depression. This should be mentioned as a potential limitation in the discussion section.

Thank you for drawing our attention to this possibility. We now cite work which suggests that patients in the targeted group may indeed have more comorbidities. (K Barnett, S W Mercer, M Norbury, G Watt, S Wyke & B Guthrie. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet 2012;380, 37-43.) The implications have been highlighted in discussion of limitations.

10) Methods=>Practices and Participants=> “…MIQUEST query” Is a reference needed here? Also do queries differ by clinical system? If yes please rephrase and mention clinical systems.

Additional text explaining MIQUEST has been added, along with a reference to use of the software.

11) I must commend the authors on their practice of reporting all used Read and drug codes. If they have some time it might be worth uploading these to a new repository specifically aimed to address study transparency and reproducibility: www.clinicalcodes.org. Uploading the codes there will make them directly accessible to researchers and expose your work more. This is just a thought though: I do not expect you to use the website.

Thank you for inviting us to contribute to this clinical codes repository. We plan to upload our lists of codes in the near future.

12) Methods=>Data analysis=> "We took the number of registered…" I would think this is an unnecessary assumption to make (i.e. using the correct denominator within each year shouldn't be that hard) but if they say that the sensitivity analysis gave the same results it means that they did use the more appropriate approach as well. So why are they not reporting that analysis as the main one to avoid the issue altogether? Even if they decide to stick to the current approach, more information will need to be provided on what exactly the sensitivity analysis entailed and comment on the differences (or just say none observed, more elaborately…) in the results section. Possibly also say that the results are available from the authors.

We acknowledge that assuming a constant annual denominator as a limitation in the revised manuscript. The denominator for the binomial regression varies monthly as patients are subjected to case finding and thereby become ineligible for the incentive for a period of 15 months, or as new long-term diagnoses are identified or resolved. There are also variations due to patients dying or leaving the practice. We used annual QOF reports for the denominator values and took them to be constant for that year. Since the denominator is large compared to the numerator, the error of the model will be small.

13) Methods=>Data analysis=> "A further discontinuity…" I think the authors need to explain why exception performance needed to be isolated and clarify how that was done through an interrupted time series analysis.

There is a striking effect in 2006/7 as is shown in the graphs we provide. The effect in this year is obvious without statistical modelling. Our focus and interest was on the long-term sustained effect seen after the introduction of case finding incentives rather than the immediate change. To avoid bias from this first year (2006/7) rates were permitted to be different in that year, so isolating it from the sustained effect we sought to assess.

14) Results. Discuss characteristics of recruited, not recruited and all English practices (briefly).

This information has been added to the manuscript.

15) Results. Section with 'raw' results required, linked to the requested table with annual incidence and prevalence rates. The authors do report month incidence rates when they discuss comparisons but I feel a dedicated short section is required to clarify the picture.

The Results section has been expanded and a table added.

16) Results. The authors report on comparisons that were not clarified in the methods section and I was unsure of the question they were addressing (discussed in detail as a main point) e.g. comparison of rates between periods 2002-4 and 2007-11 (I imagine a typo and you mean 2007-12), or 2004-6 to 2007-12.

Additional information has been added to this section of the method to aid interpretation of results by the reader.

Thank you for identifying the typographical error. This has been corrected.

17) Results “that is the rates can be, and were, taken as constant during these periods” probably rephrase to something using “can be assumed to be constant”.

This sentence has been rephrased, as suggested by the reviewer.

18) Results “…with fitted constants and slopes” Unclear what the authors mean.

Description has been added to the text.

19) Results “During the period after QOF was introduced but before incentives…” My guess would be that the authors got the years wrong there, should be April 2004 to March 2006?

Thank you for highlighting this error, which has been corrected.

20) Discussion. “Rates of new prescriptions for antidepressants exceeded…” This statement, although verified by the graphs, is not supported by the evidence presented in the results section. It needs to be.

A table has been added to the results to illustrate this point.

21) Discussion=>Limitations. The authors are right about self-selected practices and the risk of selection bias. However, the relative benefits of this study would be obvious if the response rate was higher i.e. there is quite a lot of room for selection bias in a response rate of 50%. Usage of a single computer system is potentially a bigger headache than self-selection (see paper I mentioned before). Also I understand what you mean about “ceilings on performance” (but I am not sure everyone would) and for example CPRD practices are better QOF performers (by approximately 1%) but the “ceiling” is inherently a QOF problem (i.e. that 1% does not cause the issue). In reference 17 we used a CPRD sub-sample specifically selected to be as representative as possible i.e. we did not use the whole of the CPRD (and they were not higher performers). To summarise, agree there are potential issues but probably rephrase: lose ceiling, add single computer system?

We have reworked and clarified this section of this discussion. We acknowledge our oversight concerning the method of reference 17 and have removed the citation from this paragraph. As noted in minor point 7, discussion of study limitations has ben expanded to include comment on clinical computing systems.

22) Table 1 (revisited…): year of comparison needs to be placed on title rather than in footnotes. England average list size looks odd. I’m pretty sure it’s over 6,500 but it might have been that in 2002? In that case might be worth adding 2 more columns and list characteristics at the beginning and (2002) and the end of the study (2012).

The data presented are taken from multiple sources, each recorded at different times e.g. census data and QOF data. It certainly is not a perfect snap-shot of practice demographics but gives information for comparison. The year referred to in the majority of characteristics has been placed in the title as requested, with exceptions annotated in footnotes.

Data to compare practice characteristics in 2002 and 2012 were not available in the public domain.

The majority of papers and NHS figures refer to an average practice size of over 6,500. The table provided uses the median practice size which is 5,987 patients. The mean practice size from our data is 6,835, but this is misleading as the distribution of practices is positively skewed. As such the majority (57 per cent) of practices in England have fewer than the mean number patients on their list. The most frequently seen (modal) practice size in England is between 2,000 and 3,000 patients. When discussing typicality of a practice it may not be useful to think that a practice with between 6,000 and 7,000 patients is normal, as less than 9 per cent of practices in England fall into this category.

VERSION 2 – REVIEW

REVIEWER	Christopher Burton University of Aberdeen UK
REVIEW RETURNED	28-Jun-2014

GENERAL COMMENTS	I commend the authors for addressing the points raised in the first review. I am happy with most of them but I do think they have side-stepped the issue I raised about the seasonality of screening compared to the apparent lack of it for diagnosis or treatment. While stating explicitly that case-finding may work directly or indirectly they don't examine this dissociation (or same-day screen and treat cases) as ways of unpicking that. I'm going on about these again because I think they matter in terms of how we use this evidence in relation to policy: if case-finding predominantly works by increasing awareness / lowering treatment thresholds, then we need to make sure that those factors are explicitly considered in any future programmes. I suspect the authors recognise that this indirect effect is likely, but they don't really examine it; my concern is that their data will be picked up by people with a more literal approach to interpretation.

REVIEWER	Evangelos Kontopantelis University of Manchester England
REVIEW RETURNED	20-Jun-2014

GENERAL COMMENTS	I am happy with the revision and the authors have addressed the vast majority of my previous comments satisfactorily. There are a few minor points that are up to the authors whether they want to act on them or not. 1) Thought out text: I would replace “binomial” with “logistic”, as explained before. 2) In the added modelling information I wanted to see how the slopes were modelled, rather than the standard logistic regression model. The only relevant info is “slope terms were added where appropriate”. The key advantage of an interrupted time-series design is its ability to account for pre-intervention trends rather than just levels and without the modelling details it is unclear how this is an ITS. (this relates to my 'the study cannot be replicated' choice: it cannot without that info from the authors. However, I appreciate

	that's a bit harsh for complex analyses like this where access to the full code files and the original data is needed, and a description in the methods section is hardly ever enough). 3) I didn't comment before, but the 'first year of coding' issue is quite common across almost all QOF conditions. Less 'severe' cases end up in the registers and GPs are (I believe falsely) accused of aggressive case finding to increase denominators. Even more so for conditions which are a bit 'controversial'. The most extreme example is CKD for which the prevalence jump after incentivisation (in 2006) is massive. It might be worth adding a sentence in the discussion.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

I am happy with the revision and the authors have addressed the vast majority of my previous comments satisfactorily. There are a few minor points that are up to the authors whether they want to act on them or not.

1) Thought out text: I would replace "binomial" with "logistic", as explained before.

Our statistician offered the following explanation. The outcome is binomial with the numerator being the number of screens and the denominator specifying the number eligible for screening for the targeted or non-targeted group. This is then regressed upon factors specifying the time-series model. Thus it is a binomial regression.

2) In the added modelling information I wanted to see how the slopes were modelled, rather than the standard logistic regression model. The only relevant info is "slope terms were added where appropriate". The key advantage of an interrupted time-series design is its ability to account for pre-intervention trends rather than just levels and without the modelling details it is unclear how this is an ITS. (this relates to my 'the study cannot be replicated' choice: it cannot without that info from the authors. However, I appreciate that's a bit harsh for complex analyses like this where access to the full code files and the original data is needed, and a description in the methods section is hardly ever enough).

Our statistician offered the following explanation. In all circumstances encountered in this study, a slope was initially modelled and a Wald test undertaken to establish the statistical significance of that slope. Often the slope was very close to zero so that it was of not of significance statistically or clinically. In those cases only a change in level was ultimately modelled. Description was restricted for the sake of simplicity but the zero slopes are clearly seen in the figures provided.

3) I didn't comment before, but the 'first year of coding' issue is quite common across almost all QOF conditions. Less 'severe' cases end up in the registers and GPs are (I believe falsely) accused of aggressive case finding to increase denominators. Even more so for conditions which are a bit 'controversial'. The most extreme example is CKD for which the prevalence jump after incentivisation (in 2006) is massive. It might be worth adding a sentence in the discussion.

Thank you for highlighting this phenomenon. The first and third paragraphs of the discussion have been amended to augment this point:

"The spike in diagnoses immediately following incentivisation probably reflects coding patterns before general practitioners began to realise they would trigger alerts for further assessments required by QOF when recording depression related diagnoses. Similar phenomena have been observed in first years of new QOF indicators."

Reviewer: 1

I commend the authors for addressing the points raised in the first review. I am happy with most of them but I do think they have side-stepped the issue I raised about the seasonality of screening compared to the apparent lack of it for diagnosis or treatment. While stating explicitly that case-finding may work directly or indirectly they don't examine this dissociation (or same-day screen and treat cases) as ways of unpicking that.

I'm going on about these again because I think they matter in terms of how we use this evidence in relation to policy: if case-finding predominantly works by increasing awareness / lowering treatment thresholds, then we need to make sure that those factors are explicitly considered in any future programmes. I suspect the authors recognise that this indirect effect is likely, but they don't really examine it; my concern is that their data will be picked up by people with a more literal approach to interpretation.

We thank Reviewer 1 for emphasizing this point, which we now address more explicitly in the third paragraph of the Discussion:

"A combination of these explanations seems likely for two reasons. First, we found strong evidence of seasonality for coded case-finding but not for new diagnoses or prescribing. Second, our parallel ethnographic study of general practices demonstrated the absence of a systematic approach to following up and managing screen-positive cases."