Manuscript committee comments:

1. We found this quite a technical paper and were not quite sure where it would fit in the journal. You haven't written this as a "Research Methods and Reporting" and we don't suggest you do. As written the main point is to demonstrate the use of a method and we don't usually publish this sort of paper. However, as editors we quite liked reading about this innovation and we think our readers will find this useful. Please can change the introduction and perhaps a little of the reporting so that you specify a "research question" and so that your manuscript addresses this question.

We have now added a research question in the abstract and introduction: How do clinicians vary in their response to new guidance? We have also altered the reporting to highlight how this question is addressed.

2. A couple of editors were not convinced that the examples you provide were good examples of where changes in clinical practice were necessary for all patients. There would seem to still be room for a good doctor to prescribe in Cerazette or trimethoprim in some circumstances, though we take your point that you would expect a large shift in average prescribing.

We agree that there will be some situations where it remains more appropriate to prescribe Cerazette as a branded product, or trimethoprim for a UTI. However this reflects the pragmatic reality of clinical practice: it would be difficult, if not impossible, to find any examples of changes and treatment guidance that are black and white, applying universally to all patients with no edge-cases. It is also the reason why detection of change is methodologically challenging. The two examples presented are important, affecting very large numbers of patients, where a substantial shift in practice should be seen for all general practices. This is reinforced by our detection of substantial changes in practice for the large majority of GPs.

3. 25% of practices excluded for desogestrel and 14% for trimethoprim/nitrofurantoin. It should be possible to reduce values by improving the algortihm. Perhaps practices with all values of 1 should not be excluded.

Since first submitting the paper, we have developed the analytic code to address this point. The core break detection algorithm already had the ability to include time-series with missing values, and we have now made use of this.

The algorithm can be applied to practices with missing observations by replacing the missing observation with an arbitrary value (such as 99) and subsequently removing the influence of this observation in the model by including a zero-one dummy variable for that exact observation. This allows the time series to be treated as complete, but with the missing observation still removed from analysis. This has enabled us to incorporate a large number of practices that were previously excluded.

The editors suggest that practices with all values of 1 should perhaps not be excluded. Practices that show zero variation naturally have no change-points, and it is therefore unnecessary to use the algorithm to determine whether they do. For this paper, we have also removed practices whose time series had more than half missing values, as it is unlikely that any method would be able to detect meaningful change in such data, and there is an increased likelihood of false positive detection in these practices. These remaining exclusions represent pre-filtering of the data to remove practices where it is highly unlikely that meaningful changes would be detected. We have noted this in the new "Redeploying this method elsewhere" section in the discussion.

4. Are examples in Figure 1 typical to illustrate a point, or specially selected to show a clear difference? The purple lines are dotted not dashed. Perhaps use las=1 in figures.

The examples are selected to illustrate some of the common types of change that are observed, and to show what the metrics shown in the paper relate to. We have corrected the figure text to say dotted. Axis tick labels are now horizontal as requested.

5. Figures 2 and 3 interesting, but still cross-sectional. Magnitude not particularly informative, as most shifting by same amount, particularly after run-in period.

As we noted in the discussion, we think that the magnitude being relatively uniform over time and having relatively little variation (as indicated by the standard deviation) is a valuable and informative observation: it indicates that once practices do make a change, most implement a relatively large change.

6. It would be interesting to see how timing and gradient correlate across practices – i.e. see scatterplot. This would distinguish between early/late and shallow/steep changers.

We have added these scatterplots as supplementary figures. They illustrate a similar point to the second panels in figures 2 and 3, but give a little more detail on range.

7. The mean magnitude of change would be informative. Around 0.7 for desogestrel, but < 0.4 for trimethoprim/nitrofurantoin, indicating the proportion of "uncomplicated" UTI cases.

We already report median and IQR in the manuscript. We have added mean as well, since this has been specifically requested in editorial feedback, but feel that only one should be reported; in our view the median is better since the data is somewhat skewed, hence our reporting IQR.

8. The discussion refers to 6000 practices, but the methods say 8078. Please resolve this difference.

This discrepancy was caused by the previous exclusion of practices with any missing values, as this has now been improved, we have changed the discussion to say "~8,000".

9. Please revise your paper to respond to all of the comments by the reviewers. Their reports are available at the end of this letter, below.

In your response please provide, point by point, your replies to the comments made by the reviewers and the editors, explaining how you have dealt with them in the paper.

We have responded below to all reviewer comments below, and made changes to the manuscript as needed.

Comments from Reviewers

Reviewer: 1

Comments:
Dr. Walker and colleagues present an automated statistical detection approach to detect changes in prescribing behavior in order to quantify variation in speed of adoption and magnitude of warranted changes at healthcare institutions. To evaluate the performance of their approach, they used two example time-series in a large study with English primary care prescribing data: the prescribing of generic desogestrel around the expiry of the Cerazette patent; and the prescribing of nitrofurantoin over trimethoprim around a change in prescribing guidelines. They found that the method was able to automatically and robustly detect changes in prescribing behaviors in both examples. Great variation exists in speed of implementation for these warranted changes. I commend the authors' efforts in creating this computational approach to automatically detect changes in clinical practice, which addresses an important question with rigorous and efficient methodology. Overall, the manuscript is well-written and can be informative to researchers interested in studying diffusion of change in medical practice. I have some major and minor comments that hopefully will help strengthen the manuscript.

We would like to thank the reviewer for their comments, which we found constructive and useful.

1. Methods, page 5 line 43 – One important feature of the approach is the choice of the level of significant for breaks to control the false-positive rate. The authors used p=0.000001 in the current study. Could you please include some discussion on how to select the level of significance in a given study/sample? It would be helpful for researchers who are interested in applying the method in their studies. Furthermore, how big an impact did this choice of level of significance have on the results in the current study?

*The p-value controls the false-positive rate of detection of the algorithm. As the sample size increases (approaches infinity) the false-positive rate converges to the chosen p-value, which in turn allows us to control the false rate of detection. For example, in a sample of T=1000 observations, choosing a p-value of 0.001 results in an expected false-discovery rate of p*T=1 breakpoint. In small samples the false-positive rate of trend indicator saturation can lie above the chosen p-value (see Figure 2 in Pretis, Reade, Sucarrat, 2018, Journal of Statistical Software), thus to ensure a low false-positive rate and increase confidence in detected breaks actually reflecting underlying changes we chose a very conservative rate. We have added discussion on p-value choice in our new "Redeploying this method elsewhere" in the discussion.*

2. Methods, page 5 line 14-15 – the investigation excluded "practices with incomplete time series, or those that did not vary during the time series". Please clarify the latter part. It appears that latter part refers to absolute change in value rather than trend in the time series based on the description of results in Data section on page 7 line 13-14.

Practices that 'did not vary' refers to practices where the proportion of prescriptions stayed absolutely constant throughout the sample (ie. the variance is zero). These practices therefore do not exhibit any changepoints. We have clarified this in the text (pg 5 para 1), and noted in the text that only one practice met this criterion (pg 7 para 1).

3. Results, page 7 – This point is related to the second comment. As about 1/4 of practices were excluded from the Cerazette analyses, it would be informative to know the representativeness of the analytical sample to interpret the results. Please describe how these excluded practices compare to those included in the analysis.

We have now amended the algorithm so that is is able to handle time-series with missing values. Consequently, only 3% and 5% of practices are now excluded from the Cerazette and trimethoprim analyses respectively. The number of excluded practices is now extremely small. Those that have missing values are generally smaller practices with a lower prescribing volume of the drugs in question (causing the missing values by having a denominator of 0). We have added figures showing the much smaller mean patient list size of excluded practices (pg 7 para 1)

4. Results, page 7 – The authors noted that the method could become hypersensitive to change and result in inappropriate detection when the initial variance of the time series was very low. They overcame this problem by "tweaking the maximum size of the block-partitioning". Please elaborate on how this parameter should to be set to avoid this problem of hypersensitivity. Moreover, would this method be able to differentiate larger structural changes from smaller changes such as seasonal variations if both were present?

Point 1: In large samples the block-partitioning will not affect outcomes, however, in small samples, the block partitioning can affect the detection of individual changepoints. In particular, with the present dataset (in which some practices exhibit zero variation over some time periods) block partitioning can result (by chance) in time series blocks that exhibit no variation. In simulations to produce the generalised software library for all use-cases, the rule that was found to perform well is a block size of at most 30 break variables in each block; however we halved this to 15 in the present context to account for practices that show long periods of little variation. We added a note to the manuscript that the block-partitioning was changed from the default (which is described in the reference provided).

Point 2: the method is able to detect variation due to seasonal (or other factors) by including these as additional regressors in the model. We have added a note on this in the new "Redeploying this method elsewhere" section.

5. Discussion, page 13 line 52-57 – While I understand the advantage of using the proportion of "undesirable" prescribing (over all prescribing) compared to its absolute volume, the meaning of the term "confounding by indication" is not very clear. As the term could be easily confused with its typical use, (i.e. referring to one type of confounding in studies of treatment-outcome relationship), I would suggest describing or replacing the term "confounding by indication" here.

We have changed this to read "variation in the prevalence of underlying conditions".

6. Discussion, page 14 line 10 – there is a typo: the last word in "focused our analyses on practises".

Many thanks, this has been corrected.

7. The authors described their method and results well, but the manuscript can benefit from elaborating the discussion to include practical guidance for researchers in practice, for example, the requirements of this method on the size of an analysis unit (here, practice) and quality of data.

This is a good suggestion, following this we have added a section on this in the discussion - "Redeploying this method elsewhere".

Reviewer: 2

Comments:
Thank you for the opportunity to review this manuscript. This was a very interesting topic - as someone who regularly uses interrupted time series to investigate the impact of medicine policies, I can see it being very useful for many applications. The manuscript was very well written and easy to follow, although I'd appreciate a few more details about certain things as described below:

1. From the perspective of a potential user of this method who isn't familiar with trend-indicator saturation, I would be interested in a few more details about the underlying regression model. Presumably it can account for autocorrelation, seasonality and cyclic or secular trends in the time series?

The method can easily account for autocorrelation, seasonality, cyclic or secular trends, or other explanatory factors by including these as regressors in the model that are not selected over, while simultaneously selecting over trend breaks (as currently done). We have added this to our "Redeploying this method elsewhere" section in the discussion.

2. Additionally, I am curious if this method is broadly applicable, or are there specific requirements and/or assumptions of the time series data that must be met? (for example, a minimum number of time points).

The method is broadly applicable and has no assumptions on a minimum sample size beyond standard time series regressions (that is, much of the theoretical analysis is conducted using asymptotics, i.e. large sample theory, but simulations have been used to study small sample properties). The method relies on independent, identically distributed error terms, this can be ensured by including autoregressive lags and is automatically tested during identification of the breakpoints.

Small sample simulation results are shown in Pretis, Reade, and Sucarrat (2018, Journal of Statistical Software) showing that the false-positive rate is slightly higher than the chosen p-value in small samples, but approaches the chosen p-value in samples approximately larger than 200 observations.

We have described this in our new "Redeploying this method elsewhere" section.

3. I note that many practices did not observe a significant shift until well after the intervention(s). Obviously the further in time from the intervention the change occurs, the less sure you can be it is due to the intervention itself. Now, for the examples in this paper it may not matter so much, where the focus is more on improvements in prescribing and there are few other alternative explanations - but for other scenarios, it should be noted that an automated approach may identify changes potentially unrelated to the intervention and strategies would be needed to exclude these.

The method described in our paper can turn a very large volume of complex time series data into a smaller number of coefficients describing the timing, slope, and magnitude of change; and we have used this to describe the variation in changes in clinical practice across a very large number of organisations. If the suggestion is that changes may be detected, by those re-using our technique, and that these changes would then need thoughtful interpretation to judge whether they are related to a given cause, then we agree: thoughtful interpretation is needed by all users of any analytic method, on any given dataset.

4. What about practices that experienced a change prior to the intervention date? Are these included in the summary measures in Table 1? It may be worth mentioning in how many cases the algorithm identified changes clearly unrelated to the interventions (i.e. beforehand). Also, how many practices experienced no change?

For the examples presented in this paper, we did include changes that occurred before the "intervention dates". For both examples, even before the first intervention dates, practices could have made legitimate decisions to change practice, and these are of active, positive clinical interest. For example, a practice may have changed antibiotic prescribing behaviour in response to a preponderance of lab reports describing trimethoprim resistance in mid-stream urines that clinicians themselves had sent to microbiology in the course of their own clinical practice, as resistance to trimethoprim grew across the country, prior to PHE making a change in national policy in response to these reports. Similarly it is a positive finding of our paper that some practices made early changes to prescribing behaviour of a long-term medicine (desogestrel) in anticipation of the subsequent change in cost due to patent expiry. However, the analytic code already identifies all changes within a time-series, meaning that it is simple to filter out changes that occurred before any specified date, if needed, for a specific analysis.

5. It wasn't entirely clear to me if the algorithm could potentially identify multiple changes within a practice, and if so how it deals with them. Looking at the graphs in the supplementary material there seems to always be one main change identified.

Yes, the algorithm can identify multiple changes. As per the manuscript, in the specific applications presented we set the code to detect the single largest behaviour change in each individual practice. In cases where we find multiple breaks, we focus on the single largest break that explains at least 50 percent of the total observed change. We have further clarified this in the methods and also noted it in the discussion.

6. Can you clarify the Cerazette intervention? Prior to patent expiry and availability of the generic options, wouldn't only Cerazette be available and thus Cerazette prescribing be 100%?

It is correct that Cerazette was the only available option before patent expiry, and thus 100% of prescribed desogestrel would be dispensed as Cerazette. However it is always possible for a GP to prescribe generically, even if there is no generic available. We have clarified this in the text (pg. 5 para. 2).

Reviewer: 3

Comments:
Thank you for the opportunity to review this paper. I believe this work adds substantially to the published literature, describing a novel, original method which can be applied across various domains of healthcare to capture behaviour change in response to new evidence or new developments in a field.

Although this paper is somewhat technical in its description of this novel method, behaviour change and implementation are relevant across any field of medicine or healthcare. In an era of increasing availability of routine health data, this approach has multiple applications which are well described in the discussion section (including differentiating between warranted and unwarranted variation in healthcare, identifying best practices in implementation, and driving quality improvement). I do believe it is important to multiple audiences and is best suited to a general medical journal, and has the potential to direct general practitioners to the team's web platform to examine these measures for their own practices.

Many thanks for this positive feedback.

The research question and study design are described appropriately. I have some minor suggestions of some elements of the methods could be clarified.
Page 5/19, Line 16 - It would be helpful to clarify here whether 'incomplete time series' complete data for all months during the study period, and perhaps include the reasons for missing values here, rather than at the beginning of the results.

Many thanks, we have now reduced the number of practices excluded because they had missing values, and now only exclude those where over half the observations are missing. We have moved the explanation of why values are missing to the methods section.

Page 5/19, Line 55 - Although clear further in the paper, I would suggest mentioning here that each of these graphs related to an individual practice.

Done, many thanks.

Page 6/19, Line 20 - The final line describing the slope measure ("until the mean of the time series at the end of the time series") is somewhat unclear.

We have removed the second half of this sentence as it was confusing, and the first half of the sentence should be sufficient to explain how the slope is derived.

Page 6/19, Line 26 - Refers to "mean proportion at the end of the time period", can the authors perhaps clarify if this is study time period rather than behaviour change time period? Also, should this be the proportion at the end of the study time period, or mean proportion over some period of time? For the final part of this sentence "at the time of the largest detected change", I would suggest amending to "at the start time of the largest detected change" or similar.

Thank you for drawing attention to this definition. It was not quite accurate in a few ways, we have corrected and clarified this as suggested.

Overall the results do address the research question and I found them clear. One suggestion in relation to Table 1, would be to clarify whether the timing measure for "UTI antibiotics" relates to timing after the guidance change. It could also be interesting to add the equivalent metric for timing after the Quality Premium incentive.

We have noted in the text that in Tab 1 "intervention" means the first guideline change. We are not sure that adding change time in relation to the Quality Premium would add any more information for the reader (as it would be the same value minus the 30 months between interventions). There are three lines on the time series figure 3 for nitrofurantoin which will allow readers to understand the relationship between changes and various national interventions.