PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cost-effectiveness of fecal calprotectin used in primary care in the diagnosis of inflammatory bowel disease
AUTHORS	Zhang, Wei; Wong, Chiew; Chavannes, Mallory; Mohammadi, Tima; Rosenfeld, Greg

VERSION 1 - REVIEW

REVIEWER	James Turvill York Teaching Hospital NHS Foundation Trust York UK
REVIEW RETURNED	13-Oct-2018

GENERAL COMMENTS	This is a well written and 'honest' paper. Its hypothetical design is clearly a limitation but the findings of the study are in line with a growing consensus which, I should say, I support. The discussion generally outlines the strengths and weaknesses of the paper well. This is a well written and 'honest' paper. Its hypothetical design is clearly a limitation but the findings of the study are in line with a growing consensus which, I should say, I support. The discussion generally outlines the strengths and weaknesses of the paper well. I would therefore support its publication. However I do have a number of comments that I think should be addressed head on. To my mind however my major criticism of the paper turns on the original premise of the argument: that if you have a raised CRP you refer the patient but if you do not you do not refer. This is clearly not the case in clinical practice. Many patients are referred despite normal baseline investigations and the utility of FC is particularly telling in this cohort of patients. So this paper is modelling the utility of FC versus the current best marker of inflammation (CRP (ESR)) not FC versus primary care practice should FC not be available. This I feel strongly needs to be enunciated and the implications that follow. The individual papers that allow the sensitivity and specificity of CRP to be determined (rather than the Jellema et al review that is quoted) should be identified. These (CRP/ESR) data are likely to derived from secondary care sources and are therefore open to criticism. This should be discussed. The authors take the trouble to

	demonstrate that the sensitivity and specificity of FC changes dependent upon clinical setting (primary or secondary care). The same will likely apply to CRP/ESR. I do not understand what is going on on page 9 lines 35 to 42 when compare with page 12 line 54 through to page 13 line 8. To my mind there is a contradiction here. At one stage 15% is stated at another 5%. Further I think 15% referral is not sufficiently conservative. Why not 30%? Lastly NICE who the authors quote have recently published a consensus document that I think should be referenced perhaps in relation to early diagnosis (page 15) or pathway design (page 17): Faecal Calprotectin in Primary Care as a Decision Diagnostic for Inflammatory Bowel Disease and Irritable Bowel Syndrome: https://www.nice.org.uk/.../endorsed-resource-consensus-paper-pdf-4595859614 There are two minor points: Page 8 line 36: CPR should read CRP Page 16 line 47: colposcopy should read colonoscopy

REVIEWER	Alison Smith Academic Unit of Health Economics (AUHE), University of Leeds, UK
REVIEW RETURNED	14-Nov-2018

GENERAL COMMENTS	General comments This is a well written article that provides a valuable addition to the literature on the cost-effectiveness of faecal calprotectin (FC) testing. In particular this analysis focuses on the primary care setting, in which few studies have previously been conducted. With some minor adjustments to more clearly signpost the limitations of the analysis and ongoing research in this area, this manuscript will be suitable for publication. Specific comments Abstract: In the results section I would say that FC testing 'is expected to cost more… but yield little higher QALYs..' rather than reporting the results as if they are a fact. Similarly in the conclusion section. There should be a recognition of the key limitation to this study – that the model only covers a short-term time horizon. For example, amend the conclusion to read 'Based on this analysis of short-term outcomes, screening adult patients in primary care using FC testing at a cut-off level of 100ug/g is expected to be cost-effective in Canada' (or add in a limitations statement). Strengths and Limitations of this study: The authors rightly highlight the key strengths of this study, being one of few cost-effectiveness analyses in this setting. However the authors do not discuss any of the limitations of the study – the main being the short-term time horizon of the analysis, which means that there is outstanding uncertainty over the long-term impact of FC testing in this setting.

	Introduction: First paragraph - Is the reported cost of IBD ($2.8 billion) an annual estimate? Assuming it is please report as such i.e. 'The corresponding annual economic costs…' Second paragraph - It would be helpful for readers less familiar with this clinical context to highlight up front that one of the main functional gut disorders that gets confused with IBD is Irritable Bowel Syndrome (IBS), and in addition why we should need to distinguish between IBD vs. IBS. A sentence or two about how these conditions are differentially managed (e.g. IBS by symptomatic management in primary care vs. IBD requiring specialist care management) and the different long term risks (e.g. common need for surgery in IBD patients & small mortality risk etc.) would provide better context for the need to accurately distinguish between these two patient groups. Methods: 1. Time horizon The authors' state that the one year time horizon used in the analysis is sufficient for patients to reach confirmed diagnosis of IBD vs. non-IBD. This may be so, but this is not a justification for adopting this time horizon, since diagnosis is only an intermediate outcome. Presumably earlier diagnoses expected to be achievable with FC may have long-lasting impacts e.g. if a bowel perforation or surgery can be avoided, for example; or, as the authors later highlight, there may be a long-term impact relating to avoided mortality resulting from reducing unnecessary colonoscopies for IBS patients. I think this is the primary limitation of the analysis as it stands, and this requires further discussion in the paper (See Discussion section comments). Can the authors provide a justification here as to why they did not construct a model with a longer time horizon? 2. Model structure/pathway The authors do not explicitly explain the rationale for choosing the 100 cut-off threshold for FC. To readers unfamiliar with the clinical context, it may seem odd to have used this cut-off when previous cited studies have used the 50 mcg/g cut-off. The Turvill 2014 reference (#16) explains this rational clearly, but it might be useful to have a comment on this in the main text (i.e. a higher threshold has been advocated in primary care to increase the positive predictive power of the test and counter the high false positive rate observed at the lower 50 mcg/g threshold). 3. Model parameters 3.1 Test accuracies The UK study (Walker 2018) used to inform the FC sensitivity and specificity and prevalence estimates included both alarm and non-alarm patients (i.e. both patients suspected of cancer and not suspected of cancer). In the UK patients with suspected cancer are immediately referred for emergency endoscopy, so these two cohorts are generally treated as separate populations. Were the authors here intending the model to apply to both suspected cancer and non-suspected cancer patients? Is there a distinction in how these patients are treated in the Canadian context? If so this may have implications for the model and applicability of these diagnostic accuracy estimates.

	In addition the UK study also focused on young adults (18-46 years old). Again this affects the applicability of this data to the model population and should be highlighted as a potential limitation. The authors state that for the blood tests, a meta-analysis was conducted to synthesize the logit-transformation of sensitivity and specificity. It's not completely clear to me exactly what the authors have done here – can they provide further details of this analysis? In particular has the correlation between the sensitivity and specificity values been accounted for? For the sensitivity and specificity values of FC, separate beta distributions have been applied. This approach fails to account for the correlation between these values. Why was an alternative distribution e.g. multi-variate normal not used? Results: The results should be average results hence reported as 'the FC testing strategy was about $21 more expensive on average than the standard practice…' etc. Scenario analyses: The description of what scenario analyses were conducted would be better placed in the Methods section rather than the Results. Under point (2), this should read 'FC testing accuracy was changed using an alternative data source'. I'm unsure of the validity of the scenario analysis using the Waugh study values for sensitivity and specificity. The Waugh study values relate to data mostly from the secondary (specialist) care setting, so the applicability of these values to the model population is poor. I understand the interest in wanting to look at the impact that increased or decreased sensitivity and/or specificity would have on the results, but what is the motivation for using the Waugh values in particular? Why weren't reduced FC test costs considered in the scenario analysis, as well as increased costs? It would be useful to see a scenario with reduced FC test costs as well, if possible. Can the authors report the cost-effectiveness plane scatter plot? If not in the main text then in supplementary material? It would be useful to see this figure if possible. Discussion: In relation to the limitations section discussion on secondary testing scenarios: This strategy has been recently assessed in the UK setting (see Turvill 2018, Evaluation of the clinical and cost-effectiveness of the York Faecal Calprotectin Care Pathway, http://dx.doi.org/10.1136/flgastro-2018-100962). Turvill's estimates of sensitivity and specificity for FC in the primary care setting, based on a cohort of 951 patients, were 0.94 and 0.92 respectively. These values may present a more applicable scenario analysis for consideration in this study, as opposed to the Waugh study values derived from primarily secondary care settings. Importantly, Turvill and colleagues found FC to be cost-saving, due to saving 100-150 unnecessary colonoscopies and

	140-190 gastroenterology outpatient appointments, with the trade-off being 4 incorrectly diagnosed IBD patients. The utility of the second FC test is that it can cut out a high proportion of false positive test results, resulting in overall cost-savings using this strategy. A discussion of the Turvill study would add value here. I suspect the future of FC testing will focus on these kinds of confirmatory testing strategies. As outlined in the Methods section comments, the limitation of the short time horizon adopted in this study should be highlighted in this section. Given that the sensitivity and specificity of the FC testing strategy are both expected to be higher than standard care testing, one would intuitively expect the short-term time horizon to be a conservative assumption for the FC strategy, since the potential long term benefits of early diagnosis and avoiding unnecessary endoscopy are not being captured. Hence, whilst I believe this is a limitation of this study, it may be reasonably argued that adopting a long-term horizon would produce more favourable results for FC, and hence the finding here that FC is cost-effective should hold in the long-run, if the stated assumptions regarding long term impacts hold. The Waugh study considered a longer time horizon: it would be useful to review their model and see what long-term impacts were included, to inform this discussion further. Another limitation of FC tests that it may be worthwhile highlighting is their low repeatability (i.e. within- and between-laboratory imprecision), and the fact that different FC tests produced by different manufacturers and using different platforms, can produce significantly different test results (i.e. between-method bias). This means that the sensitivity and specificity values adopted in this study model (based on a study using a specific ELISA test), may not hold for different laboratories with different pre-analytical and analytical operating procedures and/or using different test kits/methods. This is potentially a significant issue for home-testing kits also, since the benefits of increased uptake of testing may be negated by issues with test imprecision and bias.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: James Turvill

Institution and Country: York Teaching Hospital NHS Foundation Trust, York, UK

Please state any competing interests or state ‘None declared’: I have no competing interests

Please leave your comments for the authors below

This is a well written and 'honest' paper.

Its hypothetical design is clearly a limitation but the findings of the study are in line with a growing consensus which, I should say, I support.

The discussion generally outlines the strengths and weaknesses of the paper well.

I would therefore support its publication.

However I do have a number of comments that I think should be addressed head on.

Thank Dr. Turvill for your comments and supporting the publication.

1. To my mind however my major criticism of the paper turns on the original premise of the argument: that if you have a raised CRP you refer the patient but if you do not you do not refer. This is clearly not the case in clinical practice. Many patients are referred despite normal baseline investigations and the utility of FC is particularly telling in this cohort of patients. So this paper is modelling the utility of FC versus the current best marker of inflammation (CRP (ESR)) not FC versus primary care practice should FC not be available. This I feel strongly needs to be enunciated and the implications that follow.

We had considered what should be the comparison group. We agree that in reality, physicians examine patients and might refer patients with normal CRP/ESR or might not refer patients with abnormal CRP/ESR. However, we did not have good data sources for the sensitivity and specificity of the primary care practice. Thus, we chose CRP/ESR as the comparison group, which is also consistent with previous CEAs by Whitehead and Hutton (2010) and Turvill et al. (2018). In our model, if a patient has ongoing symptoms despite a negative CRP/ESR, they would subsequently be referred. It is only the initial referral where patients with a normal CRP/ESR are not referred. This more closely reflects clinical practice although some patients with severe symptoms would be referred irrespective of CRP/ESR value. The rational and implication of choosing CRP/ESR as the comparison group can be found in lines 22-38 on page 7. To address this limitation, consistent with Turvill et al. (2018), in one of our scenario analyses (lines 22-27 on page 13 and Table 2), we have applied the sensitivity (=1) and specificity (=0.788) of primary care practice without FC testing reported in Waugh et al. although Turvill et al. (2018) criticized the high specificity and sensitivity used in Waugh et al..

2. The individual papers that allow the sensitivity and specificity of CRP to be determined (rather than the Jellema et al review that is quoted) should be identified. These (CRP/ESR) data are likely to derived from secondary care sources and are therefore open to criticism. This should be discussed. The authors take the trouble to demonstrate that the sensitivity and specificity of FC changes dependent upon clinical setting (primary or secondary care). The same will likely apply to CRP/ESR.

Thank you for the suggestions. Reviewer 2 also had the same comment. We have now identified the individual papers on sensitivity and specificity of CRP/ESR (line 8 on page 10 and Table 1) and provided detailed method how we synthesized the sensitivity and specificity (Supplementary file). Also, as suggested, we have discussed the sensitivity/specificity of CRP/ESR might be different depending on the clinical setting (primary or secondary care) (lines 3-6 on page 17).

3. I do not understand what is going on on page 9 lines 35 to 42 when compare with page 12 line 54 through to page 13 line 8. To my mind there is a contradiction here. At one stage 15% is stated at another 5%. Further I think 15% referral is not sufficiently conservative. Why not 30%?

To clarify, this is the proportion of non-IBD patients with negative test results who had persistent symptoms and were later referred to secondary care and colonoscopy (shown in Figure 3). Based on our expert opinion, we used 15% (now lines 40-43 on page 10) in our base-case analysis. We applied 5%, 10%, 20% and 25% in our scenario analyses. The 5% was obtained from two previous UK studies. Now as suggested, we also used 30% as one scenario analysis (lines 38-40 on page 13) and the results can be found in Table 2.

4. Lastly NICE who the authors quote have recently published a consensus document that I think should be referenced perhaps in relation to early diagnosis (page 15) or pathway design (page 17):

Faecal Calprotectin in Primary Care as a Decision Diagnostic for Inflammatory Bowel Disease and Irritable Bowel Syndrome: https://www.nice.org.uk/.../endorsed-resource-consensus-paper-pdf-4595859614

We have now referenced the published consensus document on pages 6-8 and reference #11.

5. There are two minor points:

Page 8 line 36: CPR should read CRP

Done (line 42 on page 9).

6. Page 16 line 47: colposcopy should read colonoscopy

Done (line 22 on page 18).

Reviewer: 2

Reviewer Name: Alison Smith

Institution and Country: Academic Unit of Health Economics (AUHE), University of Leeds, UK

Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

General comments

This is a well written article that provides a valuable addition to the literature on the cost-effectiveness of faecal calprotectin (FC) testing. In particular this analysis focuses on the primary care setting, in which few studies have previously been conducted. With some minor adjustments to more clearly signpost the limitations of the analysis and ongoing research in this area, this manuscript will be suitable for publication.

Thank Dr. Smith for your comments and supporting the publication.

Specific comments

Abstract:

1) In the results section I would say that FC testing 'is expected to cost more… but yield little higher QALYs..' rather than reporting the results as if they are a fact. Similarly in the conclusion section.

We have changed the sentence as suggested (line 45 on page 2).

2) There should be a recognition of the key limitation to this study – that the model only covers a short-term time horizon. For example, amend the conclusion to read 'Based on this analysis of short-term outcomes, screening adult patients in primary care using FC testing at a cut-off level of 100ug/g is expected to be cost-effective in Canada' (or add in a limitations statement).

We have changed the conclusion as suggested (lines 8-10 on page 3).

Strengths and Limitations of this study:

3) The authors rightly highlight the key strengths of this study, being one of few cost-effectiveness analyses in this setting. However the authors do not discuss any of the limitations of the study – the

main being the short-term time horizon of the analysis, which means that there is outstanding uncertainty over the long-term impact of FC testing in this setting.

We have added our main limitation, i.e., short-term time horizon of the analysis (lines 29-32 on page 4).

Introduction:

4) First paragraph - Is the reported cost of IBD ($2.8 billion) an annual estimate? Assuming it is please report as such i.e. 'The corresponding annual economic costs…'

Done (line 22 on page 5).

5) Second paragraph - It would be helpful for readers less familiar with this clinical context to highlight up front that one of the main functional gut disorders that gets confused with IBD is Irritable Bowel Syndrome (IBS), and in addition why we should need to distinguish between IBD vs. IBS. A sentence or two about how these conditions are differentially managed (e.g. IBS by symptomatic management in primary care vs. IBD requiring specialist care management) and the different long term risks (e.g. common need for surgery in IBD patients & small mortality risk etc.) would provide better context for the need to accurately distinguish between these two patient groups.

As suggested, we have added clinical context to show the need to accurately distinguish between IBD and functional gut disorders such as IBS (lines 26-38 on page 5).

Methods:

6) 1. Time horizon

The authors' state that the one year time horizon used in the analysis is sufficient for patients to reach confirmed diagnosis of IBD vs. non-IBD. This may be so, but this is not a justification for adopting this time horizon, since diagnosis is only an intermediate outcome. Presumably earlier diagnoses expected to be achievable with FC may have long-lasting impacts e.g. if a bowel perforation or surgery can be avoided, for example; or, as the authors later highlight, there may be a long-term impact relating to avoided mortality resulting from reducing unnecessary colonoscopies for IBS patients. I think this is the primary limitation of the analysis as it stands, and this requires further discussion in the paper (See Discussion section comments). Can the authors provide a justification here as to why they did not construct a model with a longer time horizon?

We agree that the short term time horizon (one year) is one of our main study limitations. We have explained why we did not consider a longer time horizon (lines 17-24 on page 8) and discussed this limitation in Discussion section (lines 29-41 on page 18). We did not consider a longer time horizon mainly due to the limited direct data/evidence to enable us to estimate the long term impact and the possibility of adding more uncertainties and assumptions in terms of management/treatment pathway for IBD and non-IBD. However, in long term, because of the early diagnosis, we expect FC to generate more benefits, e.g., avoiding mortality/risk resulting from reduced unnecessary colonoscopies or bowel perforations/surgeries. Therefore, our study provides a relatively conservative cost-effectiveness results. Adopting a long-term horizon would produce more favourable results for FC and hence our finding that FC is cost-effective should hold in the long-run.

7) 2. Model structure/pathway

The authors do not explicitly explain the rationale for choosing the 100 cut-off threshold for FC. To readers unfamiliar with the clinical context, it may seem odd to have used this cut-off when previous cited studies have used the 50 mcg/g cut-off. The Turvill 2014 reference (#16) explains this rational clearly, but it might be useful to have a comment on this in the main text (i.e. a higher threshold has

been advocated in primary care to increase the positive predictive power of the test and counter the high false positive rate observed at the lower 50 mcg/g threshold).

As suggested, we have added the rationale for choosing the 100 cut off for FC (lines 15-22 on page 7).

3. Model parameters

8) 3.1 Test accuracies

The UK study (Walker 2018) used to inform the FC sensitivity and specificity and prevalence estimates included both alarm and non-alarm patients (i.e. both patients suspected of cancer and not suspected of cancer). In the UK patients with suspected cancer are immediately referred for emergency endoscopy, so these two cohorts are generally treated as separate populations. Were the authors here intending the model to apply to both suspected cancer and non-suspected cancer patients? Is there a distinction in how these patients are treated in the Canadian context? If so this may have implications for the model and applicability of these diagnostic accuracy estimates.

The UK study by Walker et al. was the only reliable data source we can found to inform the important parameters for our model. Walker et al., excluded patients with suspected colorectal cancer (one of their exclusion criteria) from their study and commented that their data “are only representative of patients deemed unsuitable for urgent cancer referral by their GP”. Therefore, consistent with patient population of Walker et al., we are looking at patients who are suspected of having IBD but not suspected of having cancer that needs for urgent referral. We have clarified this in lines 49-54 on page 7. Accordingly, our model applies the prevalence and test accuracy among patients with and without gastrointestinal alarm symptoms. We have now added a scenario analysis, which applies the prevalence of IBD among patients without alarm symptoms only and the corresponding FC test sensitivity and specificity reported in Walker et al. (lines 49-52 on page 13). The results are quite consistent with our base case and other senator analyses (Table 2).

9) In addition the UK study also focused on young adults (18-46 years old). Again this affects the applicability of this data to the model population and should be highlighted as a potential limitation.

We have highlighted this as a potential limitation (lines 38-47 on page 15 and lines 12-20 on page 20).

10) The authors state that for the blood tests, a meta-analysis was conducted to synthesize the logit-transformation of sensitivity and specificity. It’s not completely clear to me exactly what the authors have done here – can they provide further details of this analysis? In particular has the correlation between the sensitivity and specificity values been accounted for?

For the sensitivity and specificity values of FC, separate beta distributions have been applied. This approach fails to account for the correlation between these values. Why was an alternative distribution e.g. multi-variate normal not used?

We have added details on the meta-analysis in a supplementary file. We had tried a bivariate meta-analysis model (Model specification: Appendix 1 of Reitsma, p988), but the estimated covariance matrix was not full rank and might be unreliable, possibly due to a small sample size (only 3 studies). Therefore, we estimated the sensitivity and specificity independently (i.e., without accounting for the correlation between the sensitivity and specificity).

Results:

11) The results should be average results hence reported as ‘the FC testing strategy was about $21 more expensive on average than the standard practice…’ etc.

Done (line 20 on page 14).

Scenario analyses:

12) The description of what scenario analyses were conducted would be better placed in the Methods section rather than the Results.

We have moved the description of scenario analyses under “Analyses” of Methods section (page 13).

13) Under point (2), this should read ‘FC testing accuracy was changed using an alternative data source’.

I’m unsure of the validity of the scenario analysis using the Waugh study values for sensitivity and specificity. The Waugh study values relate to data mostly from the secondary (specialist) care setting, so the applicability of these values to the model population is poor. I understand the interest in wanting to look at the impact that increased or decreased sensitivity and/or specificity would have on the results, but what is the motivation for using the Waugh values in particular?

We agree with the reviewer that the Waugh study values for sensitivity and specificity are not applicable to the population in primary care setting and thus we removed this scenario. As suggested in the comment below, we have applied the sensitivity and specificity used in Turvill 2018 study (point 2) on page 13 and Table 2).

14) Why weren’t reduced FC test costs considered in the scenario analysis, as well as increased costs? It would be useful to see a scenario with reduced FC test costs as well, if possible.

As suggested, we have added scenarios with reduced FC test costs, i.e., $20 and $30 (Table 2).

15) Can the authors report the cost-effectiveness plane scatter plot? If not in the main text then in supplementary material? It would be useful to see this figure if possible.

We have included the cost-effectiveness plane scatter plot for base case in the Supplementary file.

Discussion:

16) In relation to the limitations section discussion on secondary testing scenarios:

This strategy has been recently assessed in the UK setting (see Turvill 2018, Evaluation of the clinical and cost-effectiveness of the York Faecal Calprotectin Care Pathway, http://dx.doi.org/10.1136/flgastro-2018-100962). Turvill’s estimates of sensitivity and specificity for FC in the primary care setting, based on a cohort of 951 patients, were 0.94 and 0.92 respectively. These values may present a more applicable scenario analysis for consideration in this study, as opposed to the Waugh study values derived from primarily secondary care settings. Importantly, Turvill and colleagues found FC to be cost-saving, due to saving 100-150 unnecessary colonoscopies and 140-190 gastroenterology outpatient appointments, with the trade-off being 4 incorrectly diagnosed IBD patients. The utility of the second FC test is that it can cut out a high proportion of false positive test results, resulting in overall cost-savings using this strategy. A discussion of the Turvill study would add value here. I suspect the future of FC testing will focus on these kinds of confirmatory testing strategies.

As suggested, we have used the sensitivity and specificity used in Turvill 2018 study as one scenario although their sensitivity and specificity values were based on repeating FC testing (point 2) on page 13). In addition, we have added this study in the Introduction section (lines 38-43 on page 6) and discussed the study under Discussion section (lines 22-38 on page 19).

17) As outlined in the Methods section comments, the limitation of the short time horizon adopted in this study should be highlighted in this section. Given that the sensitivity and specificity of the FC testing strategy are both expected to be higher than standard care testing, one would intuitively expect the short-term time horizon to be a conservative assumption for the FC strategy, since the potential long term benefits of early diagnosis and avoiding unnecessary endoscopy are not being captured. Hence, whilst I believe this is a limitation of this study, it may be reasonably argued that adopting a long-term horizon would produce more favourable results for FC, and hence the finding here that FC is cost-effective should hold in the long-run, if the stated assumptions regarding long term impacts hold. The Waugh study considered a longer time horizon: it would be useful to review their model and see what long-term impacts were included, to inform this discussion further.

Please see our response to the comment #6) above, which is the same as this comment.

18) Another limitation of FC tests that it may be worthwhile highlighting is their low repeatability (i.e. within- and between-laboratory imprecision), and the fact that different FC tests produced by different manufacturers and using different platforms, can produce significantly different test results (i.e. between-method bias). This means that the sensitivity and specificity values adopted in this study model (based on a study using a specific ELISA test), may not hold for different laboratories with different pre-analytical and analytical operating procedures and/or using different test kits/methods. This is potentially a significant issue for home-testing kits also, since the benefits of increased uptake of testing may be negated by issues with test imprecision and bias.

We have emphasized the low repeatability of FC tests under our Discussion section (lines 12-36 on page 20).

VERSION 2 – REVIEW

REVIEWER	James Turvill York Teaching Hospital NHS Foundation Trust York UK
REVIEW RETURNED	29-Dec-2018

GENERAL COMMENTS	I would recommend acceptance of this paper for publication. There are just a few typographical errors: 1 the grammar lines 28/29 and 33/34 on page 5 (introduction second paragraph need correcting) 2 in the Analyses section page 13, paragraph 2, number 4) is missing. The authors go from 3) to 5). All else I think is good.

REVIEWER	Alison Smith University of Leeds, UK
REVIEW RETURNED	02-Mar-2019

GENERAL COMMENTS	In general the authors have adequately responded to the initial review comments. There are several typos within the manuscript, and areas where the quality of writing could be improved. Given these issues, the manuscript would benefit from a thorough check for any further errors which may have been missed and where the readability could be improved. Once these issues have been addressed I believe this manuscript would be appropriate for publication.

	General comments Introduction: p5 typos: “One of the most common functional gut disorders that is difficult to distinguish…” “causes no serious consequences or permanent damage” “However, IBD can have serious complications” P5. I would be wary about saying that IBS causes no serious consequences or permanent damage – I suspect patients with IBS would not agree with this statement. I suggest amending – e.g.: “Whilst IBS can be safely managed within primary care services, the risk of serious complications associated with IBD (such as …) necessitates specialist care management”. P5. Do you have a reference for the statement that 11% of the Canadian population are affected by IBS? P6. The paragraph beginning “Recently, the detection of fecal…” could be improved. It is currently chronologically incorrect: the Waugh study is the first in the serious of studies discussed, and was based predominantly on secondary care data using the standard cut-off of 50 mcg/g. This study is what informed the original NICE positive guidance in 2013. The Walker and Turvill studies are more recent additions to the literature, both based in the primary care setting and adopting a higher cut-off of 100 mc/g to defend against the high false positive rate when using the standard cut-off in the primary care setting. In addition, the Turvill study included a repeat FC test for patients with an initially raised result. Based on the Turvill study, NICE have released a new consensus statement advocating the two-testing algorithm using the higher cut-off. As a minimum the two additional statements added need to be switched around, e.g.: “More recently Turvill et al. have demonstrated that repeating FC testing among those with a first FC test ≥100µg/g in primary care is cost-saving compared with CRP/ESR testing or single FC testing using the standard cut-off of 50µg/g. NICE have subsequently endorsed this repeated testing algorithm, using the higher 100µg/g cut-off, within a recent consensus document”. This section could be improved further by making it chronologically correct and more concise throughout. Materials and Methods P7. I found the added text under ‘Comparison groups’ confusing to read and suggest amending this. E.g.: “A higher 100 µg/g cut-off in primary care has been recently advocated and demonstrated to increase the positive predictive power of the test and counter the high false positive rate observed at the lower standard 50 µg/g cut-off in the primary care setting.9,11,12,14 Therefore, we chose one-off FC testing using the 100 µg/g cut-off for FC testing in primary care setting as the intervention for our analysis. Referrals based on standard care ESR/CRP testing in primary care were used as the comparator. This assumes that patients with a normal CRP/ESR would not be referred initially (but may be subsequently referred if symptoms persist). if they have ongoing symptoms, they would subsequently be referred. This is a simplification of real-world practice – clinicians are known, for example, to refer patients with normal ESR/CRP to secondary care. Nevertheless, there is currently a lack of reliable data on the accuracy of real-world primary care referral practices, hence we based the comparator on ESR/CRP testing, in line with previous analyses.15,12. An alternative estimate of primary care referral accuracy was based on the Waugh study, which estimated a high sensitivity (=1) and specificity (=0.788); since the reliability of these estimates has

	been previously questioned12, this was used as a scenario analysis only. ” Ideally, we would have the current primary care practice as our control group. However, there was not good data sources for the sensitivity and specificity of the primary care practice. Waugh et al.6 used a very high sensitivity (=1) and specificity (=0.788) for the primary care practice. Turvill et al. considered it unlikely that general practitioners (GPs) were more accurate at referring patients based on symptomatology than based on ESR/CRP testing alone.12 Thus, we chose CRP/ESR as the comparison group, which is consistent with previous CEAs by Whitehead and Hutton15 and Turvill et al.12. I am not personally aware of the literature on real world referral practices and/or what other specific information clinicians may use alongside CRP/ESR testing and if it is true that there is no reliable data on this in the literature (particularly in Canada). The authors may be able to clarify this point. Under ‘decision model’, to improve readability I suggest changing the new text “but are not suspected of having cancer that needs for urgent referral” to “but are not suspected of having cancer (which requires urgent specialist referral)”. P8. I would remove the added text explaining why a longer term horizon was not conducted. The argument provided is not convincing, and better left unsaid in my opinion. This point has now been highlighted as a limitation in the Discussion section, which is sufficient. P8. A couple of typos. “Under the proposed strategy of adding FC testing (Figure 2), patients with positive FC results of FC test will be referred to specialist care…” P10. Suggest amending to read “Previous studies estimated a 50% or 60% probability of non-IBD patients still having persistent symptoms after the initial management by GPs, estimates were based on expert opinion” P13. Remove the first sentence under ‘Analyses’ – the price year is already reported under the costs section. P15. “When the prevalence was increased to 20%, the probability of the FC testing strategy being cost-effective would increased to 96.7%” and “The price threshold at which FC testing strategy became..” P16. Change “the prevalence would be likely to be higher” to “the prevalence would likely be higher” P19. The added text here is a bit strong. Suggest the following amendments: “Secondly, we onlydid not considered a short longer time horizon. In the long term, because of the earlier diagnoses, we expect FC to generate more benefits, e.g., by avoiding mortality/risk resulting from reduced unnecessary colonoscopies or bowel perforations/surgeries. Therefore, we expect our study to provide a relatively conservative cost-effectiveness estimate for FC. Nevertheless, further research on the long-term impact of early diagnosis of IBD and IBS is needed to validate this claim. Adopting a long-term horizon would likely produce more favourable results for FC and hence our finding that FC is cost-effective should hold in the long-run.” P19. The text in the paragraph beginning “Thirdly..” could be significantly cut down. E.g. you could cut out the middle section of this paragraph, which adds little: “Thirdly, the model assumes that FC would be used as a single test applying a fixed threshold of 100 ug/g. Alternative two-stage testing strategies may also be used. Turvill et al., for example, recently evaluated such a retesting FC strategy, using an FC cut-off threshold of ≥100μg/g and conducting a repeat FC test for patients with an initial test

	above this cut-off.12 They found this FC strategy to be cost-saving in a UK primary care setting, due to saving 100-150 unnecessary colonoscopies and 140-190 gastroenterology outpatient appointments in IBS patients with an initially raised FC result, with the trade-off being an additional 4 incorrectly diagnosed IBD patients compared to ESR/CRP testing alone [[?assumed]]. The utility of the second FC test is that it can cut out a high proportion of false positive test results, resulting in overall cost-savings" In addition, the authors do not discuss the results of the added scenario analysis using the Turvill sensitivity and specificity estimates, which indicate a higher cost-effectiveness of FC using this strategy vs. the one-test base case strategy. This should be added into the above section also, and is a key finding from the scenario analyses that deserves greater attention. The scenario based on the Waugh primary care accuracy values has also not been discussed, however given the lack of face validity of these estimates I do not see a problem with this, given that they are reported in the results table. P21 typo: "It is worth noting that…"

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: James Turvill

Institution and Country: York Teaching Hospital NHS Foundation Trust

York

UK

Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

I would recommend acceptance of this paper for publication.

There are just a few typographical errors:

1 the grammar lines 28/29 and 33/34 on page 5 (introduction second paragraph need correcting)

Done as suggested by Reviewer 2 below (page 5).

2 in the Analyses section page 13, paragraph 2, number 4) is missing. The authors go from 3) to 5).

Done. The numbers have been changed (page 13).

All else I think is good.

Reviewer: 2

Reviewer Name: Alison Smith

Institution and Country: University of Leeds, UK

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

In general the authors have adequately responded to the initial review comments. There are several typos within the manuscript, and areas where the quality of writing could be improved. Given these issues, the manuscript would benefit from a thorough check for any further errors which may have been missed and where the readability could be improved. Once these issues have been addressed I believe this manuscript would be appropriate for publication.

General comments

Introduction:

p5 typos:

“One of the most common functional gut disorders that is difficult to distinguish…”

“causes no serious consequences or permanent damage”

Done (page 5).

“However, IBD can have serious complications”

P5. I would be wary about saying that IBS causes no serious consequences or permanent damage – I suspect patients with IBS would not agree with this statement. I suggest amending – e.g.: “Whilst IBS can be safely managed within primary care services, the risk of serious complications associated with IBD (such as …) necessitates specialist care management”.

As suggested, we have changed the sentence into “While IBS can be safely managed within primary care services, the risk of serious complications associated with IBD (such as bowel obstruction and toxic megacolon) necessitates specialist care management.” (page 5)

P5. Do you have a reference for the statement that 11% of the Canadian population are affected by IBS?

Reference has been added (page 5).

P6. The paragraph beginning “Recently, the detection of fecal…” could be improved. It is currently chronologically incorrect: the Waugh study is the first in the serious of studies discussed, and was based predominantly on secondary care data using the standard cut-off of 50 mcg/g. This study is what informed the original NICE positive guidance in 2013. The Walker and Turvill studies are more recent additions to the literature, both based in the primary care setting and adopting a higher cutoff of 100 mc/g to defend against the high false positive rate when using the standard cut-off in the primary care setting. In addition, the Turvill study included a repeat FC test for patients with an initially raised result. Based on the Turvill study, NICE have released a new consensus statement advocating the two-testing algorithm using the higher cut-off. As a minimum the two additional statements added need to be switched around, e.g.: “More recently Turvill et al. have demonstrated that repeating FC testing among those with a first FC test ≥100µg/g in primary care is cost-saving compared with CRP/ESR testing or single FC testing using the standard cut-off of 50µg/g. NICE have subsequently endorsed this repeated testing algorithm, using the higher 100µg/g cut-off, within a recent consensus document”. This section could be improved further by making it chronologically correct and more concise throughout.

Thank you for your helpful suggestion. These studies have been rearranged in a chronological order as suggested (page 6).

Materials and Methods

P7. I found the added text under ‘Comparison groups’ confusing to read and suggest amending this.

E.g.: “A higher 100 µg/g cut-off in primary care has been recently advocated and demonstrated to increase the positive predictive power of the test and counter the high false positive rate observed at the lower standard 50 µg/g cut-off in the primary care setting.9,11,12,14 Therefore, we chose one-off FC testing using the 100 µg/g cut-off for FC testing in primary care setting as the intervention for our analysis. Referrals based on standard care ESR/CRP testing in primary care were used as the comparator. This assumes that patients with a normal CRP/ESR would not be referred initially (but may be subsequently referred if symptoms persist). if they have ongoing symptoms, they would subsequently be referred. This is a simplification of real-world practice – clinicians are known, for example, to refer patients with normal ESR/CRP to secondary care. Nevertheless, there is currently a lack of reliable data on the accuracy of real-world primary care referral practices, hence we based the comparator on ESR/CRP testing, in line with previous analyses.15,12. An alternative estimate of primary care referral accuracy was based on the Waugh study, which estimated a high sensitivity (=1) and specificity (=0.788); since the reliability of these estimates has been previously questioned12, this was used as a scenario analysis only. ” Ideally, we would have the current primary care practice as our control group. However, there was not good data sources for the sensitivity and specificity of the primary care practice. Waugh et al.6 used a very high sensitivity (=1) and specificity (=0.788) for the primary care practice. Turvill et al. considered it unlikely that general practitioners (GPs) were more accurate at referring patients based on symptomatology than based on ESR/CRP testing alone.12 Thus, we chose CRP/ESR as the comparison group, which is consistent with previous CEAs by Whitehead and Hutton15 and Turvill et al.12.

I am not personally aware of the literature on real world referral practices and/or what other specific information clinicians may use alongside CRP/ESR testing and if it is true that there is no reliable data on this in the literature (particularly in Canada). The authors may be able to clarify this point.

This paragraph has been reorganized and changed as suggested (page 7).

Under ‘decision model’, to improve readability I suggest changing the new text “but are not suspected of having cancer that needs for urgent referral” to “but are not suspected of having cancer (which requires urgent specialist referral)”.

Done as suggested (page 8).

P8. I would remove the added text explaining why a longer term horizon was not conducted. The argument provided is not convincing, and better left unsaid in my opinion. This point has now been highlighted as a limitation in the Discussion section, which is sufficient.

The added text has been removed (page 8).

P8. A couple of typos. “Under the proposed strategy of adding FC testing (Figure 2), patients with positive FC results of FC test will be referred to specialist care…”

It has been changed into “Under the FC testing strategy (Figure 2), patients with positive FC test results will be referred to specialist care …” (page 9).

P10. Suggest amending to read “Previous studies estimated a 50% or 60% probability of non-IBD patients still having persistent symptoms after the initial management by GPs, estimates were based on expert opinion”

It has been changed into “Based on expert opinions, previous studies estimated that the probability of non-IBD patients still having persistent symptoms after the initial management by GPs was 50% or 60%.” (page 10).

P13. Remove the first sentence under ‘Analyses’ – the price year is already reported under the costs section.

Done (page 12).

P15. “When the prevalence was increased to 20%, the probability of the FC testing strategy being cost- effective would increased to 96.7%” and “The price threshold at which FC testing strategy became..”

These sentences have been changed into “When the prevalence was increased to 20%, the probability of the FC testing strategy being cost-effective would increase to 96.7% at the threshold of $50,000/QALY. The maximum price at which the FC testing strategy would still be cost-effective was about $70.” (page 14).

P16. Change “the prevalence would be likely to be higher” to “the prevalence would likely be higher”

Done as suggested (page 15).

P19. The added text here is a bit strong. Suggest the following amendments: “Secondly, we only did not considered a short longer time horizon. In the long term, because of the earlier diagnoses, we expect FC to generate more benefits, e.g., by avoiding mortality/risk resulting from reduced unnecessary colonoscopies or bowel perforations/surgeries. Therefore, we expect our study to provide a relatively conservative cost-effectiveness estimate for FC. Nevertheless, further research on the long-term impact of early diagnosis of IBD and IBS is needed to validate this claim. Adopting a long-term horizon would likely produce more favourable results for FC and hence our finding that FC is cost-effective should hold in the long-run.”

Done as suggested (page 18).

P19. The text in the paragraph beginning “Thirdly..” could be significantly cut down. E.g. you could cut out the middle section of this paragraph, which adds little: “Thirdly, the model assumes that FC would be used as a single test applying a fixed threshold of 100 ug/g. Alternative two-stage testing strategies may also be used. Turvill et al., for example, recently evaluated such a retesting FC strategy, using an FC cut-off threshold of ≥100μg/g and conducting a repeat FC test for patients with an initial test above this cut-off.12 They found this FC strategy to be cost-saving in a UK primary care setting, due to saving 100-150 unnecessary colonoscopies and 140-190 gastroenterology outpatient appointments in IBS patients with an initially raised FC result, with the trade-off being an additional 4 incorrectly diagnosed IBD patients compared to ESR/CRP testing alone [[?assumed]]. The utility of the second FC test is that it can cut out a high proportion of false positive test results, resulting in overall cost-savings”

As suggested, we have cut out the middle section of this paragraph and made the changes as suggested (pages 18-19).

In addition, the authors do not discuss the results of the added scenario analysis using the Turvill sensitivity and specificity estimates, which indicate a higher cost-effectiveness of FC using this strategy vs. the one-test base case strategy. This should be added into the above section also, and is a key finding from the scenario analyses that deserves greater attention. The scenario based on the Waugh primary care accuracy values has also not been discussed, however given the lack of face validity of these estimates I do not see a problem with this, given that they are reported in the results table.

We have also discussed the results from the added scenario analysis using the Turvill et al. sensitivity and specificity estimates and added the discussion in the same paragraph above as suggested (page 19).

P21 typo: “It is worth noting that…”

Done (page 19).

VERSION 3 - REVIEW

REVIEWER	Alison Smith University of Leeds, UK
REVIEW RETURNED	09-Mar-2019

GENERAL COMMENTS	I believe the authors have now addressed all the previously highlighted issues and this manuscript is ready for submission. One editorial comment: p19, sentence beginning 'The future research..', should be 'Future research...'.