PEER REVIEW HISTORY

This paper was submitted to the BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open where it was re-reviewed and accepted.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cost-effectiveness and quality of life in surgeon versus general practitioner organised colon cancer surveillance. A randomised controlled trial.
AUTHORS	Augestad, Knut Magne; Norum, Jan; Dehof, Stefan; Aspevik, Ranveig; Ringberg, Unni; Nestvold, Torunn; Vonen, Barthold; Skrøvseth, Stein; Lindsetmo, Rolv-Ole

VERSION 1 - REVIEW

REVIEWER	Bretthauer, Michael University of Oslo
REVIEW RETURNED	02-Oct-2012

GENERAL COMMENTS	The paper deals with a very important topic; surveillance programs of cancer survivors. The topic is also timely, given the current national cancerr survivor initiative in the UK (see ref in paper), and similar activities in other countries (including Norway). The method employed is a randomised trial, which should give high-quality data. The authors are to be congratulated for the enormous amount of work to put a trial like this together, involving many health care providers over a geographically large area and through a long period of time. However, unfortunately, there are a considerable number of major flaws in this study, which seriously diminish the validity of the work. Major issues 1. I don't understand the rationale for the choice of study endpoints; QOL is the primary endpoint, and the study was powered for this endpoint (see questions on power below). I question the importance of QOL as primary aim for cancer survivor surveillance programs. I think there is agreement on that the overarching goals are recurence-free survival and overall survival. Yes, QOL is important, but clearly not as important as prognosis (as defined by the two mentioned outcomes). By choosing QOL as the primary outcome (and powering the trial accordingly), the investigators missed the chance to get enough power for the really important outcomes. Further, there is also an inconsistency in the choice of the secondary outcomes; objective, confirmed recurrence is lumped together with suspicion for recurrence (such as diagnostic testing of various types, some within the surveillance proggram, some clearly not as they are not part of the program). This mixture of referrals for test, actual performed tests, and confirmed disease recurrence makes it very difficult to determine the real impact of the intervention. 2. Although the authors claim the opposite (page 8, results), There

	are obvious imbalances of the two comparison groups at baseline for variables which are important for the main outcome measures (such as age, tumor stage, and also hospital type, method of surgery). This is especially disturbing for tumor stage, which clearly has potential influence on QOL and recurrence risk, and according to the authors should have been stratified for at randomization (I don't understand how this could have happened if the randomization worked as planned). 3. Early stopping of the trial; this seems rather odd, and one may wonder if the real reason was slow recruitment, and not changes in the surveillance protocol on a national level (this could have been coped with, as the aim of the trial was to compare two surveillance settings not the surveillance scheme itself). As far as one can read from the manuscript, no predefined early stopping rule existed, no DSMB was employed, and the ethics committe was not consulted. This raises important ethical challenges, which are not addressed at all in the paper. 4. The application of detailed patient and GP information brochures (you call them "tools to support decision making" on page 10) for the GP group (the control group patients and surveillance surgeons did apparently not receive the same, newly developed information package as this is not mentioned anywhere in the paper) is clearly a benefit for the patients and the caregivers. However, it constitutes a big problem for this study, as it is also a new intervention, and as this was introduced together with the "real" trial intervention (the "real" intervention in this trial was GP surveillance), it is very difficult to determine which of the two were resonsible for the outcome. 5. The cost analysis is problematic, as it lumps together actual costs as observed and validated by the authors (who are experienced clinicans and health care workers, I don't doubt these numbers), and DRG based estimates. The problem is with the DRG's; the whole concept of DRG is that it is to be used for macroeconomic reimbursement policies in a health care system only. DRG's should not be used for cost analyses of single procedures or interventions. The basic principle of DRG's is that some procedures or services are underfinanced and some overfinanced with regard to the actual (real) costs. The use of DRG's in this study is particular problematic because the estimates of the different services provided in the two groups are not equally distributed between DRG and actual cost derivations (e.g. GP rates are not DRG based, hospital costs are). Therefore, althoguh the authors have done a lot of work to try to get this right, I am afraid that the obtained differences in costs for the two groups may not particularly valid. There may be no better way to do it, but that does not mean that this study is good enough to set the stage. Sometimes, it's better with no study than one with uncertain (and therefore probably wrong) results. 6. What was the reason for the application of three different QOL instruments? Are these overlapping in aim and focus? Why not just one, but three? In extention to this: with only about 100 patients and the large amount of questions and analyses; should a multiple testing correction have been performed? 7. Patients older than 75 years of age were not eligible for the study. I do understand the rationale for not surveilling them in the national guidelines (in which the main aims are to prevent recurrent disease and metachronous cancer over a period longer than 2 years, ref 7 of the manuscript), but in this study, QOL is the primary outcome and the evaluation is after only 2 years. Therefore, it may have been reasonable to include older patients in adequate health (with regard to co-morbitidy) in the study, This would have been advantageous also for the case load. With the design chosen, more than a third of

	all eligibile patients were excluded due to old age (see flow chart). 8. The description of the randomization process is lacking a number of important details. Please adhere to the CONSORT checklist. E.g., the randomization ratio is not mentioned (although it is 1:1, I guess); the method to allocate the sequence is not described, if there was any block randomization (for stratification, see above). The authors write that blinding was not possible because ”GP organised follow-up represented a new practice”. This is not a valid reason for non-blinding. It could be a reason for knowing if you are in a trial or not, but not for being in one or the other group (which blinding is about). But I agree with the authors, blinding would have been difficult. 9. The fact that patients received all questionnaires to be filled in over the 2-year course of the trial at baseline, constitutes a possible source of bias, as the authors don't know if the patrients adhered to the planned schedule (when did they fill in the forms?). 10. Were the research assistants blinded for the group allocation? 11. How many reminders did the patients get, and when? 12. How was the registration of SCE's secured from GP offices? Did you check the files here as well, as you did at the hospitals? 13. About the outcomes and the power analyses: a. Please state the time point for the primary outcome: after two years? And what was the comparison about? The difference of the change between baseline and two-year between the two groups; or the difference between the two groups at two-years? b. Is this study a superiority study or a non-inferiority study? In other words; did you expect superior performance of the GP group compared to the hospital group, or were you testing the hypothesis that the GP group was non-inferior to the hospital group? According to the power analysis, the former is the case. If that's correct; what was your ratioanle to choose a difference of 10 units on the EORTC global health score? 14. About the results a. You find a difference of 2.3 points on the EORTC global health score, which is far less than your predefined 10-point difference (which you wanted to detect). What is your interpretation of this result (in light of your design, which obviously is aiming at superiority, not non-inferiority)? In the discussion, you focus on the three other domains, where the GP group did better. But the main outcome is negative, and for insomnia and constipation, the hospital group did better. b. The result section lacks important information about the mean/median follow-up time, patient-years of follow-up, and cumulative data on the secondary endpoints (such as KM plots). c. You talk about ”time until diagnosis” in the result section. How is this time defined? From xxx to xxx? Minor issues 1. I suggest not to use the term ”salvage surgery”, it is unclear and unprecise. 2. It appears from the record in clinicaltrials.gov that the trial may have been registered some months after start of recruitment (but I am not certain, as the authors did not state the study period in the paper) 3. References; these references should be checked and, if indicated, revised: a. Background, first sentence; Ref 1 (Cancer in Norway) does not talk about the incidence/prevalence of colon cancer worldwide. As far as I can read, tt does not refer to surgery as ”only single curative treatment” either (I am also uncertain what you mean with ”only

	single curative treatment", could you rewrite?). b. Background, 2nd sentence: the authors cite ref. 2 for their statement "aorund one third of those resected will experience recurrent disease and most of them will survive less than 2 years". However, ref. 2 says: " Approximately two-thirds of patients will present with potentially curable disease (by surgery +/- adjuvant therapies). Of these 30-40% will relapse with metastatic disease (Rao 1981; Bohm 1993).", which is different from the manuscript. Also, as far as I can see, the Jeffery paper does not talk about the specific time in which recurrence and death occurs. c. According to the PDF document on the ngicg website, the year of publication of ref 7 is 2012, not 2010 as stated in the manuscript d. Check ref's 9 and 35; something odd here, typos. e. It is remarkable that ALL 550 GP's in the area agreed to particpate. Please confirm. f. The secondary outcomes include blood ins tool by "hemofec testing" (I believe the authors mean FOB testing; Hemofec is a brand name). However, this is not an examination involved in the surveillance scheme. Did the doctors order this for a reason, a clinical symptom or sign (and maybe some of these like weight loss or pain are already included as events)? Michael Bretthauer

REVIEWER	Blazeby, Jane University of Bristol, School of Social and Community Medicine
REVIEW RETURNED	03-Oct-2012

GENERAL COMMENTS	This trial is addressing an important question. It is important to the follow up of colon cancer and other cancer sites. There are a number of methodological issues that limit the interpretation of the results. 1. It is a single country study - limits generalisatbility 2. It is colon and not rectal cancer also - unclear why just colon 3. It is a small trial that seems to be underpowered 4. The Primary outcomes is not clear in abstract – QLQ-C30 does not produce a summary score – the abstract needs to be clear what the primary outcome was 5. The rationale and hypothesis for powering this study to expert a QOL benefit needs to be explained and the specific domain of QOL clarified - it appears to be the global QOL scale in the C30. 6. There are some 16 different QOL scales and items in the C30 and no key secondary outcome pre-specified. With this small sample size it is possible that the identified statistically significant end points are not true findings 7. it is unclear how missing QOL data were handled (both missing assessments and individual items) 8. It is possible that there is contamination between the trial arms (GPs may have received the intervention but be responsible for a

	patient randomised to hospital follow up) . can the degree of contamination be presented 9. it is unclear how the decision support pamphlets were designed and actually used in the trial 10. My overall view is that this trial is too small and has too many design weaknesses to make the results generalisable and it is not possible to be confident that the proposed QOL findings have not occurred by chance

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

The paper deals with a very important topic; surveillance programs of cancer survivors. The topic is also timely, given the current national cancerr survivor initiative in the UK (see ref in paper), and similar activities in other countries (including Norway). The method employed is a randomised trial, which should give high-quality data.The authors are to be congratulated for the enormous amount of work to put a trial like this together, involving many health care providers over a geographically large area and through a long period of time.

Reply: Thank you.

However, unfortunately, there are a considerable number of major flaws in this study, which seriously diminish the validity of the work.

Reply: Please see our answers below.

Major issues

1. I don't understand the rationale for the choice of study endpoints; QOL is the primary endpoint, and the study was powered for this endpoint (see questions on power below). I question the importance of QOL as primary aim for cancer survivor surveillance programs. I think there is agreement on that the overarching goals are recurence-free survival and overall survival. Yes, QOL is important, but clearly not as important as prognosis (as defined by the two mentioned outcomes). By choosing QOL as the primary outcome (and powering the trial accordingly), the investigators missed the chance to get enough power for the really important outcomes.

Reply: An interesting point of view, we do however partly disagree. Overall and relapse-free survival is most important, but this has been addressed in other studies. At time of protocol writing (2007) the following issues were discussed:

1. **As three other large international trials (GILDA, COLOFOL, FACS) have survival as primary endpoint, we wanted to investigate other aspects of colon cancer surveillance (QoL and cost-effectiveness).**

2. **A follow-up trial powered for survival as primary endpoint would not be feasible in Norway, due to our population size. According to our pretrial power analyses several thousands had to be included in such a trail. Similarly, large international multicenter trials have slow inclusion, i.e. results from the GILDA trial have not been published yet.**
3. **Several cancer surveillance RCT trials have chosen to focus on other primary endpoints than survival. Please note:**

   *Wattchow DA, Weller DP, Esterman A, et al. General practice vs surgical-based follow-up for patients with colon cancer: randomised controlled trial. Br J Cancer 2006; **94**: 1116–21.*

   *Grunfeld E. Randomized Trial of Long-Term Follow-Up for Early-Stage Breast Cancer: A Comparison of Family Physician Versus Specialist Care. Journal of Clinical Oncology 2006; **24**: 848–55.*

   *Beaver K, Campbell M, Williamson S, et al. An exploratory randomized controlled trial comparing telephone and hospital follow-up after treatment for colorectal cancer. Colorect Dis 2012; **14**: 1201–9.*

   *Strand E, Nygren I, Bergkvist L, Smedh K. Nurse or surgeon follow-up after rectal cancer: a randomized trial. Colorect Dis 2010; **13**: 999–1003.*

   Grunfeld E. Routine follow up of breast cancer in primary care: randomised trial. BMJ 1996; : 1–5.

   *Kimman ML, Dirksen CD, Voogd AC, et al. Economic evaluation of four follow-up strategies after curative treatment for breast cancer: Results of an RCT. European Journal of Cancer 2011; **47**: 1175–85.*

4. **Finally: the hospital IRB board had no obligations to our protocol or choice of QoL as primary endpoint.**

Further, there is also an inconsistency in the choice of the secondary outcomes; objective, confirmed recurrence is lumped together with suspicion for recurrence (such as diagnostic testing of various types, some within the surveillance program, some clearly not as they are not part of the program).

This mixture of referrals for test, actual performed tests, and confirmed disease recurrence makes it very difficult to determine the real impact of the intervention.

Reply: Thank you for clarifying this to us. Please see our manuscript, where secondary endpoints (i.e. cost effectiveness and time to cancer diagnoses) are defined.

2. Although the authors claim the opposite (page 8, results), There are obvious imbalances of the two comparison groups at baseline for variables which are important for the main outcome measures (such as age, tumor stage, and also hospital type, method of surgery). This is especially disturbing for tumor stage, which clearly has potential influence on QOL and recurrence risk, and according to the authors should have been stratified for at randomization (I don't understand how this could have happened if the randomization worked as planned).

Reply: This is simply not correct. Please see calculated p values in table 3, showing no significant difference between control group and intervention group. The web based randomization service used is well known among Norwegian researchers and is used in

several ongoing randomized trials. We have no reason to mistrust the quality of this web service.

3. Early stopping of the trial; this seems rather odd, and one may wonder if the real reason was slow recruitment, and not changes in the surveillance protocol on a national level (this could have been coped with, as the aim of the trial was to compare two surveillance settings not the surveillance scheme itself).

Reply: No, the recruitments was as planned. Please see our next answer.

As far as one can read from the manuscript, no predefined early stopping rule existed, no DSMB was employed, and the ethics committe was not consulted. This raises important ethical challenges, which are not addressed at all in the paper.

Reply: We thank the peer reviewer for giving us the opportunity to discuss the ethics.

During the spring of 2012 the following situation occurred:

1) **The Norwegian Gastrointestinal Cancer Group introduced new follow-up guidelines in 2010. These guidelines were gradually implemented in our health care trust.**
2) **New guidelines recommended different radiological procedures and frequency of consultations and blood samples.**
3) **New national guidelines introduced caused confusion among participating surgeons and GPs of which guidelines to adhere to (old versus new).**
4) **There was a probability of data contamination of the ongoing trial by the new guidelines.**

These events triggered an interim analyses in July 2012. 1884 patient follow-up months were completed (i.e. 628 QoL questionnaires returned). At this point there was a 4 % probability of showing a statistically significant difference of 10 units of global QoL (EORTOC QLQ C30). A statistician (not originally involved in the trial) performed these analyses. On request conditional power calculations can be published. A preplanned stopping rule was not described. However, implementation of new national guidelines were not anticipated at time of protocol writing, and hence it was difficult to define this event in a preplanned stopping rule.

For clarification of futility based stopping please read:

Lachin JM. A review of methods for futility stopping based on conditional power. Statist Med 2005; ***24****: 2747–64.*

Jitlal M, Khan I, Lee SM, Hackshaw A. Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies. Br J Cancer 2012; ***107****: 910–7.*

Please see our manuscript (strength and limitations), for clarifications.

4. The application of detailed patient and GP information brochures (you call them ”tools to support decision making” on page 10) for the GP group (the control group patients and surveillance surgeons did apparently not receive the same, newly developed information package as this is not mentioned anywhere in the paper) is clearly a benefit for the patients and the caregivers. However, it constitutes

a big problem for this study, as it is also a new intervention, and as this was introduced together with the ”real” trial intervention (the ”real” intervention in this trial was GP surveillance), it is very difficult to determine which of the two were resonsible for the outcome.

Reply: The intervention is defined as a complex intervention (consist various interconnecting parts), thoroughly discussed and defined in several BMJ papers. Please read:

Campbell NC. Designing and evaluating complex interventions to improve health care. BMJ 2007; : 1–5.

Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008; ***337****: a1655–5.*

The decision support tool was introduced to ensure high adherence to the follow-up program. Randomised trials with decision support tools are well known:

Please read:

Augestad KM, Berntsen G, Lassen K, et al. Standards for reporting randomized controlled trials in medical informatics: a systematic review of CONSORT adherence in RCTs on clinical decision support. J Am Med Inform Assoc (JAMIA) 2012.

5. The cost analysis is problematic, as it lumps together actual costs as observed and validated by the authors (who are experienced clinicans and health care workers, I don't doubt these numbers), and DRG based estimates.

The problem is with the DRG's; the whole concept of DRG is that it is to be used for macroeconomic reimbursement policies in a health care system only. DRG's should not be used for cost analyses of single procedures or interventions. The basic principle of DRG's is that some procedures or services are underfinanced and some overfinanced with regard to the actual (real) costs. The use of DRG's in this study is particular problematic because the estimates of the different services provided in the two groups are not equally distributed between DRG and actual cost derivations (e.g. GP rates are not DRG based, hospital costs are).

Reply: We thank the peer- reviewer for this interesting comment with regard to DRG. However, we do disagree, as the DRG value is the closest we can come to quantify the societal cost of a surgical procedure. The DRG weights and consequently costs are based on specific costs calculated and analysed in a selected number of Norwegian hospitals. When comparisons have been performed, northern Norwegian hospitals have generally a somewhat higher cost than the Norwegian figure. This is due to the fact that northern Norway has many hospitals (eleven) serving a population of only 500.000 inhabitants. This is due to the significant distances in the region (Northern Norway covers a land area of about two-thirds of that of UK). Correcting for this, the difference between groups would have been even greater. However,

employing the national DRG-figures make the results more comparable to other more dense populated areas. Furthermore, the GP rates and the DRG values are set by the same national health authority (Department of Health). A national perspective are employed in both settings and based on calculated costs. We therefore argue that they are reliable.

Therefore, althoguh the authors have done a lot of work to try to get this right, I am afraid that the obtained differences in costs for the two groups may not particularly valid. There may be no better way to do it, but that does not mean that this study is good enough to set the stage. Sometimes, it's better withno study than one with uncertain (and therefore probably wrong) results.

Reply: We believe our data on health care recourse utilization are solid. Please see methods of data collection. Furthermore, we have used a well-known method of analyzing cost-effectiveness alongside a randomized trial, i.e. BMJ guidelines.

This provides the first estimate of cost-effectiveness of a decentralized colon cancer follow-up program which we believe is of interest for decision makers organizing colon cancer follow-up.

Uncertainty in cost-effectiveness analyses is not a new phenomena and sensitivity analyses are commonly used to address this. Please see figure 4.

For clarification of methods used please read:

Drummond M. Guidelines for authors and peer reviewers of economic submissions to the BMJ. BMJ 1996; : 1–9.

6. What was the reason for the application of three different QOL instruments? Are these overlapping in aim and focus? Why not just one, but three?

Reply: These QOL instruments should be well known to health services researchers. We chose one generic QoL instrument (EQ-5D-3L), one cancer specific (EORTC QLQ-C30) and a modified version of the Outpatient Experiences Questionnaire. By our point of view, these QoL instruments are overlapping in a very limited extent. Please see our answer to reviewer 3 regarding the latter QoL instrument.

In extention to this: with only about 100 patients and the large amount of questions and analyses; should a multiple testing correction have been performed?

Reply: Please note that 110 patients were included in the trial. Patients answered the QOL questionnaires every third month (after each follow-up consultation). In total we received 628 questionnaires during the trial period (1884 months) with a response rate of 95%. A Bonferroni correction was discussed with the statistician. However, due to the high response rate and high number of questionnaires analysed, multiple testing corrections were not performed.

7. Patients older than 75 years of age were not eligible for the study. I do understand the rationale for not surveilling them in the national guidelines (in which the main aims are to prevent recurrent disease and metachronous cancer over a period longer than 2 years, ref 7 of the manuscript), but in this study, QOL is the primary outcome and the evaluation is after only 2 years. Therefore, it may have been reasonable to include older patients in adequate health (with regard to co-morbitidy) in the study, This would have been advantageous also for the case load. With the design chosen, more than a third of all eligibile patients were excluded due to old age (see flow chart).

Reply: The intervention had to adhere to the national follow-up guidelines, i.e. patients >75 years are not included in the national follow-up program. Including older patients would obviously increased the difference in cost as more of them would have needed support from spouses when travelling to hospital.

8. The description of the randomization process is lacking a number of important details. Please adhere to the CONSORT checklist. E.g., the randomization ratio is not mentioned (although it is 1:1, I guess); the method to allocate the sequence is not described, if there was any block randomization (for stratification, see above). The authors write that blinding was not possible because ”GP organised follow-up represented a new practice”. This is not a valid reason for non-blinding. It could be a reason for knowing if you are in a trial or not, but not for being in one or the other group (which blinding is about). But I agree with the authors, blinding would have been difficult.

Reply: Thank you for agreeing with us that blinding would have been difficult. Suggested corrections are made to increase CONSORT adherence. Please see methods section.

9. The fact that patients received all questionnaires to be filled in over the 2-year course of the trial at baseline, constitutes a possible source of bias, as the authors don’t know if the patrients adhered to the planned schedule (when did they fill in the forms?).

Reply: Patients set the present date on their returned questionnaires. Thus we could follow their progression month by month, and subsequently adherence to their planned follow-up schedule.

10. Were the research assistants blinded for the group allocation?

Reply: No they were not blinded for the group allocation.

11. How many reminders did the patients get, and when?

Reply: Approximately 50 reminders were sent to the patients. This was done when patients were 3 months overdue with their questionnaires. This is clarified in the text.

12. How was the registration of SCE's secured from GP offices? Did you check the files here as well, as you did at the hospitals?

Reply: The SCE was defined after review of the hospital electronic medical record (i.e. information from general practitioner referral notes and surgeon notes). Due to legal constraints, we had no possibility to check the GPs EMR.

13. About the outcomes and the power analyses:

a. Please state the time point for the primary outcome: after two years? And what was the comparison about? The difference of the change between baseline and two-year between the two groups; or the difference between the two groups at two-years?

Reply: Neither alternatives listed by the peer-reviewer were used. A general linear model was employed, were time (1-24 months) and intervention group (GPs versus surgeon) were predictors in analyses of variance (between groups ANOVA). I.e. we compared the mean difference (global QoL) of follow-up consultation at 1,3,6,9,12,15,18, 21,24 months.

b. Is this study a superiority study or a non-inferiority study? In other words; did you expect superior performance of the GP group compared to the hospital group, or were you testing the hypothesis that the GP group was non-inferior to the hospital group? According to the power analysis, the former is the case. If that's correct; what was your ratioanle to choose a difference of 10 units on the EORTC global health score?

Reply: Superiority study. Please read Kings paper where a difference of 10 on global QoL is defined as a small to moderate difference in QoL global health:

*King M. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Quality of Life Research 2004; **5**: 555–67.*

14. About the results

a. You find a difference of 2.3 points on the EORTC global health score, which is far less than your predefined 10-point difference (which you wanted to detect). What is your interpretation of this result (in light of your design, which obviously is aiming at superiority, not non-inferiority)?

Reply: In the retrospect we should have aimed for a non-inferiority study. However, as no such study were published at time of protocol writing (2007), we believed at time of study start that our intervention would cause a small to moderate improvement in QoL, hence the trial was planned accordingly. This dilemma of choosing a inferiority versus a non-inferiority trial is well known, please read:

*Gayet-Ageron A, Agoritsas T, Combescure C, Bagamery K, Courvoisier DS, Perneger TV. What differences are detected by superiority trials or ruled out by noninferiority trials? A cross-sectional study on a random sample of two-hundred two-arms parallel group randomized clinical trials. BMC Med Res Methodol 2010; **10**: 93.*

In the discussion, you focus on the three other domains, where the GP group did better. But the main outcome is negative, and for insomnia and constipation, the hospital group did better.

Reply: No, this is simply not correct. The GP group showed no statistical significance on the main outcome measure (global QoL), but better performance on several subscales. Please remember that for EORTC QLQ C30 the higher score the better on the five functional scales (physical, role, cognitive, emotional and social) and global health. On the contrary, the lower score the better on the six single-item scales (dyspnoea, insomnia, appetite loss, constipation, diarrhoea and financial difficulties) (table 4).

b. The result section lacks important information about the mean/median follow-up time, patient-years of follow-up, and cumulative data on the secondary endpoints (such as KM plots).

Reply: Thank you for pointing this out for us. Please see the results section, which are updated with these data. As only 14 patients experienced cancer recurrence, we believe that a KM plot will not add any new information (please see table 7).

c. You talk about ”time until diagnosis” in the result section. How is this time defined? From xxx to xxx?

Reply: From: the start date of a serious clinical event (date given either the GP referral or in the hospital EMR) to: date of diagnosis of either a) a true cancer recurrence or b) a false positive event (defined as normal repeated radiological tests or normal histology from the biopsy taken at screening colonoscopy). Please see the text for clarification.

Minor issues

1. I suggest not to use the term ”salvage surgery”, it is unclear and unprecise.

Reply: We agree, changed to metastases surgery.

2. It appears from the record in clinicaltrials.gov that the trial may have been registered some months after start of recruitment (but I am not certain, as the authors did not state the study period in the paper)

Reply: The first patient was recruited in June 2007. The study was registered in clinicaltrials.gov in December 2007. As the research phases of a complex intervention were followed, the first months were defined as a pilot to test the feasibility of the survey.

3. References; these references should be checked and, if indicated, revised:
a. Background, first sentence; Ref 1 (Cancer in Norway) does not talk about the incidence/prevalence of colon cancer worldwide. As far as I can read, tt does not refer to surgery as ”only single curative treatment” either (I am also uncertain what you mean with ”only single curative treatment”, could you rewrite?).

Reply: ok, corrected in text

b. Background, 2nd sentence: the authors cite ref. 2 for their statement ”aorund one third of those resected will experience recurrent disease and most of them will survive less than 2 years”. However, ref. 2 says: ” Approximately two-thirds of patients will present with potentially curable disease (by surgery +/- adjuvant therapies). Of these 30-40% will relapse with metastatic disease (Rao 1981; Bohm 1993).”, which is different from the manuscript. Also, as far as I can see, the Jeffery paper does not talk about the specific time in which recurrence and death occurs.

Reply: ok, corrected in text

c. According to the PDF document on the ngicg website, the year of publication of ref 7 is 2012, not 2010 as stated in the manuscript.

Answer: ok, corrected.

d. Check ref’s 9 and 35; something odd here, typos.

Reply: ok, corrected.

e. It is remarkable that ALL 550 GP’s in the area agreed to particpate. Please confirm.

Reply: Yes, this is truly remarkable. We were surprised by the positive response from the GPs. We wrote to all 550 GPs prior to the trial start, giving them detailed information about the trial. We asked to notify us by mail if they had any obligations towards trial participation. No feedback was defined as a confirmation of trial participation. 15 GPs wrote to us telling us that they were looking forward to trial participation. Two GPs notified us about trial withdrawal prior to the survey, i.e. 448 GP agreed to participate and 2 withdrew. The text is updated wit this information.

f. The secondary outcomes include blood ins tool by ”hemofec testing” (I believe the authors mean FOB testing; Hemofec is a brand name). However, this is not an examination involved in the surveillance scheme. Did the doctors order this for a reason, a clinical symptom or sign (and maybe some of these like weight loss or pain are already included as events)?

Reply: That is corrected, **text updated. Most FOB test were initiated by a SCE.**

Reviewer: 2

Recommendation:

Comments:
This trial is addressing an important question. It is important to the follow up of colon cancer and other cancer sites. There are a number of methodological issues that limit the interpretation of the results.

1. It is a single country study - limits generalisatbility

Reply: We agree with the peer-reviewer that it might limit generalizability that this is a single country survey. However: Similar surveys have been performed in many different countries (USA, Canada, UK, Australia, Netherlands). These surveys are commonly cited and thus accepted as generalizable. Please see our references given to peer-reviewer 1.

Furthermore: Many countries have a comparable health care organization as Scandinavia, with the GPs having a strong role as a gatekeeper to access use of hospital services, i.e. Sweden, Denmark, UK, Australia, Netherlands and Canada. To further enhance generalizability we have:

1) **Adhered to ISPORS (International Society of Pharmacoeconomics and Outcomes Research)recommendations of transferability between jurisdictions. Please read:**

   Drummond M, Barbieri M, Cook J, et al. Transferability of Economic Evaluations Across Jurisdictions: ISPOR Good Research Practices Task Force Report. Value in Health 2009; ***12****: 409–18.*

2) **In the sensitivity analyses we have increased cost elements by 30-40% to reflect unit cost reported in similar UK trials. Please see Figure 4.**

2. It is colon and not rectal cancer also - unclear why just colon

Reply: In Norway colon and rectum have different follow-up guidelines (tests, frequency of tests and consultations). Therefore we only chose to include colon in the trial, as this made it possible to intervene with a similar set of follow-up guidelines for all patients.

3. It is a small trial that seems to be underpowered

Reply: We agree that this trial is small compared to trials like FACS, GILDA and COLOFOL. However, our primary endpoint were QoL and not survival and the trial was powered according to this. The trial was stopped at 80% of preplanned inclusion and 628 follow-up cycles (i.e. 1884 patient follow-up months) completed due to futility, i.e. no effect of the intervention on QoL. Please see our answer to peer-reviewer 1.

4. The Primary outcomes is not clear in abstract – QLQ-C30 does not produce a summary score – the abstract needs to be clear what the primary outcome was

Reply: Thanks for informing us. The abstract is clarified, global health score was primary outcome.

5. The rationale and hypothesis for powering this study to expert a QOL benefit needs to be explained and the specific domain of QOL clarified - it appears to be the global QOL scale in the C30.

Reply: That is correct, clarified in the text. Please see our answer to peer-reviewer 1.

6. There are some 16 different QOL scales and items in the C30 and no key secondary outcome pre-specified. With this small sample size it is possible that the identified statistically significant end points are not true findings

Reply: We have taken this view on board, global health score was the primary outcome measure. Due to the amount of QoL questionnaires returned, over a long period of time, with a high response rate, and a low conditional power (4%), we do believe our findings (no decline in patient related QoL due to GP organized follow-up) represent significant findings. Several previous published trials support this point of view, i.e. no decline in QoL by GP organized follow-up. Please see citations above.

Our key secondary endpoint is cost-effectiveness. This is clarified in the text.

7. it is unclear how missing QOL data were handled (both missing assessments and individual items)

Reply: Thank you for informing us about this. We have applied methods described in:

Quality of life outcomes in clinical trials and health care evaluation by Stephen J Walters (Wiley 2009).

i.e.

Missing items within a form: We have treated the score for that scale as missing.

Missing forms: Missing data imputation by the last observation carried forward (LOCF).

This is clarified in the manuscript (methods).

8. It is possible that there is contamination between the trial arms (GPs may have received the intervention but be responsible for a patient randomised to hospital follow up) . can the degree of contamination be presented

Reply: We believe this potential contamination is practically non-existing. According to our database and electronic medical record review, patients received follow-up in the correct arm i.e. follow-up organized either by GPs or surgeons. As shown in table 5, patients are utilizing resources in both primary and secondary care, however we believe this represent a natural drift between primary and secondary care. I.e. patients with multiple conditions needs assistance from both primary and secondary care.

9. it is unclear how the decision support pamphlets were designed and actually used in the trial

Reply: The decision support pamphlets were received by a) patients allocated to GP follow-up and b) GPs organizing the follow-up. The decision support pamphlet was introduced as part of the intervention, to ensure high follow-up program adherence by participating patients. Please see the manuscript for details.

10. My overall view is that this trial is too small and has too many design weaknesses to make the results generalisable and it is not possible to be confident that the proposed QOL findings have not occurred by chance

Reply: please see our answers above.

Additional Questions:
Please enter your name: Jane M Blazeby

Job Title: Professor of Surgery

Institution: University of Bristol

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: No

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests (either as indicated above or any other financial or non-financial interests) please declare them here:

VERSION 2 – REVIEW

REVIEWER	Dimitri Aristotle Raptis, MD, MSc Surgeon Department of Surgery University Hospital Zurich Switzerland I have no competing interests to declare.
REVIEW RETURNED	13-Dec-2012

THE STUDY	This paper would benefit from some editing by the production editors as there are still several grammatical errors.
GENERAL COMMENTS	The authors should be congratulated for the efforts and the excellent study they conducted.

REVIEWER	Dr. Bellinda King-Kallimanis Research Fellow Department of Medical Gerontology, Trinity College Dublin
REVIEW RETURNED	18-Dec-2012

THE STUDY	In Figure 2, the authors state that 474 patients were excluded and 321 did not meet the inclusion criteria, in the CONSORT check list, this says criteria is stated on page 11, however I could not find clear inclusion/exclusion criteria, can the authors please elaborate on why 321 patients were excluded. CONSORT checklist has items under the results section marked OK, please refer to where in the manuscript this information is located.
RESULTS & CONCLUSIONS	See general comments with respect to issues around clarifying missing data, this is why I am concerned about appropriate conclusions.

	In addition only mean costs are presented, there is no information given to know how skewed these costs were and whether the mean is an appropriate way to report these costs, could authors please provide some additional information with respect to this.
REPORTING & ETHICS	The authors do not mention anything about ethical approval, can a sentence please be included.
GENERAL COMMENTS	This is a 1:1 randomized trial of standard care post curative colon cancer resection (care lead by surgeon) compared to an alternative mode for delivery of care (GP lead). The primary outcomes were health-related quality-of-life and cost. Secondary to this the author examined differences in rate of serious clinical events, false positive tests, time to recurrences and frequency of metastases surgery. This paper addresses an important question in a timely way that has the potential to guide care guidelines for this patient group. The paper is clearly written, however, I had trouble following the missing data patterns and would like the authors to clarify this. There were 110 patients enrolled in the study and 600 completed follow-up questionnaires, there is no mention of how many forms were missing and what percent were due to drop out and what percent were related to a missed follow-up. Also, the authors report that data from these forms were imputed by using the last observation carried forward method. Did the authors look at the profiles of the patients who dropped out to try and understand whether these observations were missing completely at random or missing at random? While the sample size is small, limiting more complex analyses, I would recommend the authors looking at Fairclough, Thijs, Huang et al, (2008). Handling missing quality of life data in HIV clinical trials: what is practical. Quality of Life Research, 17; 61-73 and Graham. (2009). Missing data analysis: making it work in the real world. Annual Reviews Psychology, 60; 549–76. The authors also state that missing items were treated as missing, what was the final N for the p-values reported in Table 4 from the adjusted general linear model? In the results section the authors report that 11 people withdrew, with no wish for follow up, and that 84 participants were followed for 12 months and 58 were followed for 24 months, this comes to 142 completers and 153 at baseline, but it appears data was collected on 110 patients. Perhaps I am missing something, but could the authors please clarify how these numbers came about. Also in the flow chart in Figure 2 suggests that 32 dropped out, were these the people who's data was LOCF? I was also wondering what percentage of participants in the GP arm also had a surgical consultation as there were 218 consults (phone or in person), if there were a group who did not have surgical contact, I wonder if their QoL was different, did the authors look at this? While I realize it might not have been part of the SAP to conduct a subgroup analysis, I think it would be helpful to know when thinking about implementing this intervention

REVIEWER	Kjetil Søreide MD PhD Professor of Surgery

	Department of Surgical Sciences University of Bergen, Norway Department of Surgery Stavanger University Hospital Armauer Hansensvei 20 POB 8100, N-4068 Stavanger, Norway
REVIEW RETURNED	02-Jan-2013

RESULTS & CONCLUSIONS	Limitations concerning closure of trial, study number and consequence for conclusions drawn.
GENERAL COMMENTS	The randomised controlled trial by Augestad and colleagues investigates how a GP led vs surgeon led follow up of patients operated on for colon cancer with curative intent would interfere with patient reported quality of life on a global health perspective as well as implications for overall cost in following the programme. The programme is based on the national recommendations by the Norwegian Gastrointestinal Cancer Group (NGICG), as defined at the time. The trial was stopped prematurely before reaching the endpoint at 24 months. The authors conclude that there was no difference in follow-up as measured by the outcomes of QoL and costs. I congratulate the authors for having designed and performed this trial. As they correctly state, very little is known about GP led follow up and, as such, these results are a welcomed remedy to the present knowledge gap in this respect. However, I have some concerns and remarks that the authors may wish to address to further improve the presentation and to, specifically, acknowledge some limitations: Major comments: - I lament the fact that the trial was stopped early, as it hampers drawing firm conclusions based on the stated endpoints in the study trial design. While this cannot be rectified at this time, I think the authors should take a more modest view on the interpretation of the results and their conclusions based on this limitation. Intro Based on this Norwegian cohort, I do think it would be in place to discuss previous findings on surveillance and outcome – even though based on smaller cohort studies – that would warrant the assumed conditions that the authors have of the effectiveness and compliance to this. Please see results obtained from: - Körner H, Söreide K, Stokkeland PJ, Söreide JA. Systematic follow-up after curative surgery for colorectal cancer in Norway: a population-based audit of effectiveness, costs, and compliance. J Gastrointest Surg. 2005 Mar;9(3):320-8 This study found no difference in survival between followed and non-followed patients (although not an RCT, it reflects real-life). Please also note the considerable actual difference in compliance to the tools, and the respective costs associated with each asymptoimatic, curable recurrence per test.

	- Körner H, Söreide K, Stokkeland PJ, Söreide JA. Diagnostic accuracy of serum-carcinoembryonic antigen in recurrent colorectal cancer: a receiver operating characteristic curve analysis. Ann Surg Oncol. 2007 Feb;14(2):417-23. Exemplifies the poor diagnostic ability of CEA as a surveillance tool. - Søreide K, Træland JH, Stokkeland PJ, Glomsaker T, Søreide JA, Kørner H. Adherence to national guidelines for surveillance after curative resection of nonmetastatic colon and rectum cancer: a survey among Norwegian gastrointestinal surgeons. Colorectal Dis. 2012 Mar;14(3):320-4. Points to the actual differences and change in practice outside guidelines, which actually affects adherence and thus also the assumed “intention to treat” approach in your outcomes analyses. Should at least be discussed in the Discussion. Methods Please give some more info on the non-included cohort as this was a rather large proportion of the entire eligible patients Please remember that surveillance is based on the assumption that this approach will detect asymptomatic, yet curable recurrence and this is the main target approach. The cost analyses should take into consideration the uncertainty with “intention to treat”, f.ex. some tests wil have lower compliance than others. Furthermore, if CEUS is performed (and I assume only some hopsitals offered this investigation during the study period), the only benefit from having GPs doing the surveillance is the clinical examination and blood-tests, as the other diagnostics would be performed in-hospital anyway (and as such should incur on the hospital costs, not the GP costs). It may not be feasible to do this calculation, but then again it should at least be dealt with in the discussion, as it infers with the actual cost-analyses for GP vrs hospital led surveillance. Results You mention a lower than expected recurrence rate during fo-up (about 15%). This is no surprise as you have a median of 1,5 years (=18 mo) of fo-up, and to catch the majority of recurrences you would need at least 3 years (about 95% of all recurrences occur within 3 years of surgery). So this is simply reflecting the short fo-up time rather than improvements in surgery. The fo-up time severely hampers drawing conclusions for the 24-months period, as less than 50% had reached this fo-up time. Please, again, remember that surveillance is based on the assumption that this approach will detect asymptomatic, yet curable

	recurrence and this is the main target outcome – 3 of 6 in the GP group, and 5 of 8 in the hospital group. Recurrence per se is not the main target, and as such not the outcome of interest. Are salvage surgery thus considered only for asymptomatic, surveillance detected recurrence, or for all recurrences? The same goes for costs, please specify. Conclusions Should be modified based on the comments above. Tables Table 7 contains info based on a case series approach for each group. This is best left out, or, as a compromise could be included as supplementary info, but has very little to do with the trial as designed and the endpoints investigated. Figures None of the figures are numbered, at least in my version for review. This needs to be amended. - The figure number 1 could go as supplementary info. - Figure 5 is not comprehensible as it now stands in my view, I cannot see any relation between the x-axis and the figures given for each column. Other minor details: - Several text passages are marked in red text, I assume either as a sign of author revisions before submission or tracked changes. Primary submitted manuscript files should be devoid of authors' marked text.

REVIEWER	Dr Andrew Hinde Division of Social Statistics and Demography and Southampton Statistical Sciences Research Institute University of Southampton Southampton SO17 1BJ United Kingdom
REVIEW RETURNED	03-Jan-2013

THE STUDY	On p. 10 there are some typographical errors. In the section headed 'Statistics' , l. 6, 'moths' should be 'months'; l. 8, 'were' should be 'where'; l. 10 'imputated' should be 'imputed'. The only comment I have about the description of the statistical methods relates to the treatment of missing data. What proportion of the forms were missing and were therefore imputed using the last observation carried forward?
GENERAL COMMENTS	The cost savings of the GP follow-up compared with the surgeon follow-up are entirely due to the lower costs of travel and sick-leave (i.e. work loss) among the GP group. These more than outweigh the higher costs of serious clinical events and follow-up tests. At least this is my understanding of the (rather complex) Table 5. This raises two questions in my mind. (1) How much would these costs vary if you were to study a less dispersed and remote population than that

	of northern Norway? (2) In the UK for example, the costs of travel would be borne by the patient, the costs of work loss mainly by the employer, and the costs of health care contacts and follow-up tests by the National Health Service. In other words, while assessing the global cost difference is interesting, it may not resonate with any of the individual actors involved in the drama (the patient, the health care provider, or the employer). The implications of your results are that, compared with surgeon follow up, GP follow-up involves a transfer of resources from the health care provider to the patient and the employer. Are health care providers under budget pressure likely to support this? I mention these issues not as criticisms of your study or analysis, but as interesting implications you might like to think about and to mention in your discussion on p. 20.

VERSION 2 – AUTHOR RESPONSE

Reviewer: Dimitri Aristotle Raptis, MD, MSc
Surgeon
Department of Surgery
University Hospital Zurich
Switzerland

I have no competing interests to declare.

This paper would benefit from some editing by the production editors as there are still several grammatical errors.

Response: We will do our best to remove all grammatical errors.

The authors should be congratulated for the efforts and the excellent study they conducted.

Response: Thank you

Reviewer: Dr. Bellinda King-Kallimanis
Research Fellow
Department of Medical Gerontology, Trinity College Dublin

In Figure 2, the authors state that 474 patients were excluded and 321 did not meet the inclusion criteria, in the CONSORT check list, this says criteria is stated on page 11, however I could not find clear inclusion/exclusion criteria, can the authors please elaborate on why 321 patients were excluded.

Response: The 474 patients that did not meet the inclusion criteria belonged to four groups:
1) > 75 years old, i.e. patients that should not (according to national guidelines) be included in a colon cancer follow-up program (n=199)
2) Disseminated cancer diagnosed between surgery and the 1 month baseline appointment (Dukes D) (n=122)
3) Health care trust not participating in the trial (n=121)
4) No informed consent (n=32)
Please see our revised CONSORT flow diagram (Figure 1) and CONSORT checklist.
We have added the following sentence to methods:

Inclusion criteria were age less than 75 years with recent surgery for colon cancer with Dukes' stage A, B or C. Patients receiving postsurgical adjuvant chemotherapy (some Dukes' B and all Dukes' C) were also eligible. Exclusion criteria were patients older than 75 years old, patient belonging to health care trust not participating in the trial, not able to provide informed consent and Dukes D.

CONSORT checklist has items under the results section marked OK, please refer to where in the manuscript this information is located.

Response: Please see our revised CONSORT checklist describing where in the manuscript the items are found.

See general comments with respect to issues around clarifying missing data, this is why I am concerned about appropriate conclusions.

Response: Please see our reply below.

In addition only mean costs are presented, there is no information given to know how skewed these costs were and whether the mean is an appropriate way to report these costs, could authors please provide some additional information with respect to this.

Response: Thank you for addressing this central methodological question (skewed cost and reporting of mean cost). When analyzing data and reporting results we have used previously published papers as a "gold standard", please see:

Beaver K, Hollingworth W, McDonald R, et al. Economic evaluation of a randomized clinical trial of hospital versus telephone follow-up after treatment for breast cancer. Br J Surg 2009; 96: 1406–15.

Grunfeld E. Follow-up of breast cancer in primary care vs specialist care: results of an economic evaluation. Br J Cancer 1999; 79: 1227–33.

Kimman ML, Dirksen CD, Voogd AC, et al. Economic evaluation of four follow-up strategies after curative treatment for breast cancer: Results of an RCT. European Journal of Cancer 2011; 47: 1175–85.

These papers report mean cost with standard deviation or 95% confidence interval.To adjust for skewness cost are bootstrapped with 1000 replications to estimate bias corrected confidence intervals. Please see table 6 were the bootstrapped confidence intervals are provided. We have added the following sentence under methods (and added two citations by Drummond and Desgagne):

Economic evaluation data are invariably positively skewed, and it requires an alternative analysis. We used a bootstrapping technique, which makes no assumptions regarding the equality, variance or shape of the distribution, and takes into account skewness. To adjust for skewness cost were bootstrapped with 1000 replications to estimate bias corrected confidence intervals. The bootstrapping rechnique was undertaken using IBM SPSS Statistics v 19.0

The authors do not mention anything about ethical approval, can a sentence please be included.

Response: Certainly. The following sentence is added:
"The Regional Committee for Medical Research Ethics, North Norway approved this protocol in 2006 (P REK NORD 79/ 2006). Patients provided written consent before entering the trial."

This is a 1:1 randomized trial of standard care post curative colon cancer resection (care lead by surgeon) compared to an alternative mode for delivery of care (GP lead). The primary outcomes were health-related quality-of-life and cost. Secondary to this the author examined differences in rate of serious clinical events, false positive tests, time to recurrences and frequency of metastases surgery. This paper addresses an important question in a timely way that has the potential to guide care guidelines for this patient group.

Response: thank you

The paper is clearly written, however, I had trouble following the missing data patterns and would like the authors to clarify this. There were 110 patients enrolled in the study and 600 completed follow-up questionnaires, there is no mention of how many forms were missing and what percent were due to drop out and what percent were related to a missed follow-up.

Response: We have tried to describe the missing pattern better in our new manuscript version i.e.
Expected questionnaires (i.e. with 100% response rate):
n=657 (Surgeon n=330 vs. GP n=327)
Received questionnaires:
n=636, i.e. response rate 96%, (Surgeon n=319 vs. GP n=317).
Missing questionnaires (not returned):
n=21 (4%), (Surgeon n=11 vs. GP n= 10).
Excluded questionnaires (not able to ID responder):
n=36, i.e. 6% (not n=28 as reported in first manuscript version, i.e. caused by a datafile error)
(Surgeon n=18 vs. GP n=18)
Questionnaires included in final analyses of cost and QoL:
n=600 (91%), (Surgeon n=301 vs GP n=299)
We have added the following sentence:

We received 636 of the expected 657 questionnaires (response rate 96%), of those 600 (91%) questionnaires (GP 299 vs. surgeon 301) were included in final cost and QoL analyses. 21 (4%) of questionnaires (surgeon 11 vs. GP 10) were not returned and 36 questionnaires (surgeon 18 vs. GP 18) were excluded from analyses due to insufficient identification.

Also, the authors report that data from these forms were imputed by using the last observation carried forward method. Did the authors look at the profiles of the patients who dropped out to try and understand whether these observations were missing completely at random or missing at random ?

Response: The peer-reviewer rises an interesting question, as a skewed distribution of dropouts have the potential to bias the analyses. Patient related trial dropouts were either caused by recurrent cancer disease, severe concomitant disease, dementia, no wish to further participate in a surveillance program. Patient related dropouts are defined in the CONSORT flow diagram, and are equally distributed between groups, i.e. 17 in the surgeon group and 15 in the GP group. We argue that the patient related dropouts (i.e. recurrent cancer or comorbidities) leads to a “missing completely at random” (MCAR) pattern. Please see table 1 showing no significant difference in baseline demographics between the two study arms.

While the sample size is small, limiting more complex analyses, I would recommend the authors looking at Fairclough, Thijs, Huang et al, (2008). Handling missing quality of life data in HIV clinical trials: what is practical. Quality of Life Research, 17; 61-73 and Graham. (2009). Missing data analysis: making it work in the real world. Annual Reviews Psychology, 60; 549–76.

Response: We have read the suggested papers with great interest, and believe, to the best of our

knowledge, that we have followed these recommendations

The authors also state that missing items were treated as missing, what was the final N for the p-values reported in Table 4 from the adjusted general linear model?

Response: n=600, i.e. (GP=299 and Surgeon= 301).

In the results section the authors report that 11 people withdrew, with no wish for follow up, and that 84 participants were followed for 12 months and 58 were followed for 24 months, this comes to 142 completers and 153 at baseline, but it appears data was collected on 110 patients. Perhaps I am missing something, but could the authors please clarify how these numbers came about.

Response: Please see our CONSORT flow diagram, providing more details of trial flow. Note that 110 patients were recruited and randomized at baseline, i.e. 55 in each group. These patients were enrolled in postoperative follow-up cycles as described in table 1. 85 (of the originally 110 patients included at baseline) completed 12 months follow-up as described in table 1. 58 patients (52% of the originally 110 included at baseline) completed 24 months follow-up.

We have added the following sentence:

85 patients (75%) (GP 41 vs. surgeon 44) were followed for 12 months, 58 patients (52%) (GP 29 vs. surgeon 29) were followed for 24 months. During the trial 32 patients were defined as lost (surgeon 17 vs. GP 15), of those 14 patients had cancer recurrence (surgeon 8 vs. GP 6). 20 patients (surgeon 9 vs. GP 11) were transferred to the new national colon cancer surveillance program (CONSORT flow figure 1).

Also in the flow chart in Figure 2 suggests that 32 dropped out, were these the people who's data was LOCF?

Response: No, these patients were not included in further analyses. LOCF was used when missing data from patients still enrolled in the trial, i.e. forms not returned (n=21, 4%) or forms with inappropriate ID (n=36, 6%).

I was also wondering what percentage of participants in the GP arm also had a surgical consultation as there were 218 consults (phone or in person), if there were a group who did not have surgical contact, I wonder if their QoL was different, did the authors look at this? While I realize it might not have been part of the SAP to conduct a subgroup analysis, I think it would be helpful to know when thinking about implementing this intervention

Response: The peer-reviewer asks a highly relevant research question. We are working with the analyses and plan to report this subgroup analyses in a separate paper, as this was not part of the primary trial objective.

Reviewer: Kjetil Søreide MD PhD

Professor of Surgery
Department of Surgical Sciences
University of Bergen, Norway

Department of Surgery
Stavanger University Hospital
Armauer Hansensvei 20

POB 8100, N-4068
Stavanger, Norway

Competing interests: None

The randomised controlled trial by Augestad and colleagues investigates how a GP led vs surgeon led follow up of patients operated on for colon cancer with curative intent would interfere with patient reported quality of life on a global health perspective as well as implications for overall cost in following the programme. The programme is based on the national recommendations by the Norwegian Gastrointestinal Cancer Group (NGICG), as defined at the time. The trial was stopped prematurely before reaching the endpoint at 24 months. The authors conclude that there was no difference in follow-up as measured by the outcomes of QoL and costs.

I congratulate the authors for having designed and performed this trial. As they correctly state, very little is known about GP led follow up and, as such, these results are a welcomed remedy to the present knowledge gap in this respect.

Response: Thank you.

However, I have some concerns and remarks that the authors may wish to address to further improve the presentation and to, specifically, acknowledge some limitations:

Major comments:
I lament the fact that the trial was stopped early, as it hampers drawing firm conclusions based on the stated endpoints in the study trial design. While this cannot be rectified at this time, I think the authors should take a more modest view on the interpretation of the results and their conclusions based on this limitation.

Response: We agree with the peer reviewer that early stopping limits the interpretation of data related to recurrence. We have added the following sentence under the limitations section:

52% of included patients were followed for two years. This limits the interpretation of recurrence, as 80% of colon cancer recurrences occurs within three years.

The decision of early stop was based on the following: New national surveillance guidelines were gradually implemented in Northern Norway. These guidelines recommended CT Thorax (as opposed to chest x ray in 2007 guidelines) and 6 months interval between consultations (as opposed to 3 months interval in 2007 guidelines). This would obviously have large impact on our cost-effectiveness analyses. We realized that trial data might be contaminated (if trial continuation) by new guidelines, due to confusion among GPs and surgeons organizing the follow-up program which guidelines (old vs. new) to adhere to. This left us with two choices:
1) Continue collecting data from a trial intervention (i.e. GP follow-up by 2007 surveillance guidelines) that might be “contaminated” by new 2010 surveillance guidelines, or to
2) Perform an interim analyses to assess the probability of showing a significant result if trial continuation (conditional power), and to halt the trial if low conditional power.

After discussion among authors we choose the latter. In June 2012 there was a 4 % probability of showing a significant impact (10 units or more) on global health scale. Based on this result we felt it unethical to continue using large resources on a trial that would not prove the primary hypothesis (a moderate improvement of QoL). We acknowledge that early stopping limits interpretation of data on colon cancer recurrence, but argue that results on cost-effectiveness and QoL is solid.

Intro
Based on this Norwegian cohort, I do think it would be in place to discuss previous findings on surveillance and outcome – even though based on smaller cohort studies that would warrant the assumed conditions that the authors have of the effectiveness and compliance to this.

Response: Please see our response below, commenting compliance and effectiveness of follow-up.

Please see results obtained from:
- Körner H, Söreide K, Stokkeland PJ, Söreide JA. Systematic follow-up after curative surgery for colorectal cancer in Norway: a population-based audit of effectiveness, costs, and compliance. J Gastrointest Surg. 2005 Mar;9(3):320-8

This study found no difference in survival between followed and non-followed patients (although not an RCT, it reflects real-life). Please also note the considerable actual difference in compliance to the tools, and the respective costs associated with each asymptoimatic, curable recurrence per test.

- Körner H, Söreide K, Stokkeland PJ, Söreide JA. Diagnostic accuracy of serum-carcinoembryonic antigen in recurrent colorectal cancer: a receiver operating characteristic curve analysis. Ann Surg Oncol. 2007 Feb;14(2):417-23.

Exemplifies the poor diagnostic ability of CEA as a surveillance tool.

- Søreide K, Træland JH, Stokkeland PJ, Glomsaker T, Søreide JA, Kørner H. Adherence to national guidelines for surveillance after curative resection of nonmetastatic colon and rectum cancer: a survey among Norwegian gastrointestinal surgeons. Colorectal Dis. 2012 Mar;14(3):320-4.

Points to the actual differences and change in practice outside guidelines, which actually affects adherence and thus also the assumed “intention to treat” approach in your outcomes analyses. Should at least be discussed in the Discussion.

Response: Thank you for informing us about these surveys, these are absolutely very central. We have added the following paragraphs (citing the references above) under “Comparison with existing literature”:

Surveys have assed cost of follow-up in a Norwegian setting. In a retrospective survey 314 patients were assessed with regards to cost, compliance and success rate of curative surgery. It was concluded that the cost of one successful curative surgery was $ 25 289, and that further implementation of such a program should be debated. 33 Harms and unintended effects of a follow-up program is poorly explored. Especially is the rate of false positive tests in a follow-up program unknown. Current surveillance is often based on serial CEA measurements, this biomarker has several pitfalls and shortcomings. I a recent survey, it is shown that the diagnostic accuracy of serial measurement of CEA is low, and is impacted by the cut off value.34 These aspects are of high importance when designing a follow-up program, as false positive test probably has a negative impact on the patients quality of life. Finally, its there exist considerable variance in follow-up strategies, internationally and at a national level.35 This makes outcome comparison between different follow-up strategies challenging.

Methods
Please give some more info on the non-included cohort as this was a rather large proportion of the entire eligible patients.

Response: Please see our response to peer-reviewer 1 (info of non included cohort) and the CONSORT flow cart (more details provided).

Please remember that surveillance is based on the assumption that this approach will detect asymptomatic, yet curable recurrence and this is the main target approach.

Response: We completely agree. Yet there exist no international evidence on just what a surveillance program should entail (type and combination of test, frequency of tests, frequency of consultations, level of care) to optimize survival (and detection of the asymptomatic metastases). We do think more research is warranted, as CRC cancer surveillance post a great cost burden on society.

The cost analyses should take into consideration the uncertainty with “intention to treat”, f.ex. some tests will have lower compliance than others.

Response: We agree, please see our sensitivity analyses (tornado diagram, figure 4) showing impact of cost variance. Variance in cost of test (and thus impact of test compliance) is rated second. This means that follow-up compliance is a central cost driving element in analyses of overall cost of colon cancer follow-up.

Furthermore, if CEUS is performed (and I assume only some hospitals offered this investigation during the study period), the only benefit from having GPs doing the surveillance is the clinical examination and blood-tests, as the other diagnostics would be performed in-hospital anyway (and as such should incur on the hospital costs, not the GP costs).

Response: The peer-reviewer rises an interesting question regarding the cost driving elements in follow-up. Radiologic examinations and colonoscopy has to be performed at hospitals. However, we argue that other cost driving elements matter by far more in a follow-up program. Please see our reply below and our sensitivity analyses.

It may not be feasible to do this calculation, but then again it should at least be dealt with in the discussion, as it infers with the actual cost-analyses for GP vs. hospital led surveillance.

Response: We have added the following section in the discussion:

Therefore, the cost driving elements in a colon cancer follow-up program have to be critically evaluated. From a societal perspective, this survey has important implications. It may be argued that there are limited benefits from having GPs organising the follow-up program, as the radiological examinations and the colonoscopy have to be be performed in-hospital anyway. However, we believe the most important factors causing a less costly GP follow-up are: Better coordination of care: As shown in table 5, GP organised follow-up leads to fewer hospital travels. We believe this is mainly caused by improved coordination of care, for instance by performing multiple radiological test at the same hospital visit. Interestingly the GP group had fewer extra travels (GP 52 travels versus Surgeon 102 travels) due to poor logistics (table 5). Cost of GP consultation vs. hospital consultation: The societal cost of GP consultations is lower compared to cost of hospital consultations, due to a more costly hospital infrastructure. Complex and chronic conditions: Patients surgically treated often have other chronic illnesses, and there is a trend towards higher involvement of primary care in treating these conditions as described in the chronic care model. 13 Sick leave: Although not statistical significant, patients in the GP group return to work 17 days (mean) earlier compared to patients in the surgeon group.

Results
You mention a lower than expected recurrence rate during fo-up (about 15%). This is no surprise as

you have a median of 1,5 years (=18 mo) of fo-up, and to catch the majority of recurrences you would need at least 3 years (about 95% of all recurrences occur within 3 years of surgery). So this is simply reflecting the short fo-up time rather than improvements in surgery.

Response: We agree, and have deleted this section from the manuscript.

The fo-up time severely hampers drawing conclusions for the 24-months period, as less than 50% had reached this fo-up time.

Response: We agree that this hampers drawing conclusions on survival, as 52% reached 24 months fu. Please remember our primary research question were not survival, as such a trial would not be feasible in Northern Norway (to large sample size needed compared to our population). This patient group imposed a great work burden to our surgical outpatient department and patients were traveling for log distances for short 20 minutes follow-up consultations. Thus the primary research questions were:

1) Does GP follow-up lead to a decrease in the patients quality of life ?
2) Does GP follow-up in any way harm the patient ?
a. By providing poor follow-up program compliance ?
b. By increasing the time to diagnoses of a serious clinical event (and thus cancer recurrence) ?
3) Does GP follow-up increase the societal cost of follow-up ?

A fundamental problem with the existing international surveillance programs, is that there exist no evidence of the best combination of tests, test frequency and level of care that maximizes survival. This is exemplified in the fact that there exist no similar designed follow-up programs at an international level. Different follow up programs will lead to different societal cost spending's. We therefore argue that these patients should be enrolled in the least costly follow-up program until evidence based medicine is unified in what a follow-up program should entail (radiologic modalities ect). In this aspect we argue that our trial (despite clear limitations when it comes to survival assessment) brings new information to the area, as we have observed that a decentralized fu program will not harm patients, it provides similar follow-up compliance to a lower societal cost.

Please, again, remember that surveillance is based on the assumption that this approach will detect asymptomatic, yet curable recurrence and this is the main target outcome – 3 of 6 in the GP group, and 5 of 8 in the hospital group. Recurrence per se is not the main target, and as such not the outcome of interest.

Response: We completely agree with the peer-reviewer that the surveillance programs ability to save patients life (by curative salvage surgery) is the most important feature in a follow-up program .

Are salvage surgery thus considered only for asymptomatic, surveillance detected recurrence, or for all recurrences? The same goes for costs, please specify.

Response: Attempts of curative surgery were performed for both symptomatic and asymptomatic recurrences. Of the 7 patients operated on, 4 surgeries were defined as successful (curative surgery) with clear histological resection margins. Cost of all (n=7) attempted curative surgeries were included in the economical analyses. This means that the health care cost of one successful curative surgery (n=4) in this trial (NB with its limitations of follow-up time) is 80 217 £ (i.e. health care cost 24 months follow up one patient 2917 £, 110 patients enrolled, 4 curative surgeries).
Due to obvious uncertainties in this estimate (limited follow-up time and few recurrences) we have not reported these results in our paper.

Conclusions
Should be modified based on the comments above.

Response: We have modified our conclusion i.e.:

However, there exist limitations. 13% (n=14) patients had colon cancer recurrence, this low recurrence rate is most likely caused by limited long term follow-up as most recurrences occur within 3 years. Furthermore, the best combination of consultations, radiological test, blood samples and colonoscopy that optimizes cancer survival is still unknown. We therefore argue that cost driving elements of colon cancer surveillance should be critically evaluated, when designing and implementing follow-up programs, as cancer surveillance represents a huge financial burden for society. Finally, little is known about the potential harms of follow-up, especially when it comes to the impact of false positive tests. Further research is needed to settle these controversies, and new methods of decision-analytic modeling in combination with emerging data from on-going randomised trials must be applied.46

Tables
Table 7 contains info based on a case series approach for each group. This is best left out, or, as a compromise could be included as supplementary info, but has very little to do with the trial as designed and the endpoints investigated.

Response: We agree, and have omitted this table from the paper.

Figures
None of the figures are numbered, at least in my version for review. This needs to be amended. The figure number 1 could go as supplementary info.

Response: We agree, figure 1 is omitted.

Figure 5 is not comprehensible as it now stands in my view, I cannot see any relation between the x-axis and the figures given for each column.

Response: Please remember that in a Tornado chart for cost sensitivity analyses there is no x axis. Most critical variable in terms of impact on cost is listed at the top of the graph, and the rest ranked according to their impact thereafter.

Other minor details:
Several text passages are marked in red text, I assume either as a sign of author revisions before submission or tracked changes. Primary submitted manuscript files should be devoid of authors' marked text.

Response: We agree.

Reviewer: Dr Andrew Hinde
Division of Social Statistics and Demography and
Southampton Statistical Sciences Research Institute
University of Southampton
Southampton
SO17 1BJ
United Kingdom

On p. 10 there are some typographical errors.

In the section headed 'Statistics' , l. 6, 'moths' should be 'months'; l. 8, 'were' should be 'where'; l. 10 'imputated' should be 'imputed'.

Response: Thanks for informing us. The manuscript is corrected.

The only comment I have about the description of the statistical methods relates to the treatment of missing data. What proportion of the forms were missing and were therefore imputed using the last observation carried forward?

Response: n=21 (4%), (Surgeon n=11 vs. GP n= 10).

The cost savings of the GP follow-up compared with the surgeon follow-up are entirely due to the lower costs of travel and sick-leave (i.e. work loss) among the GP group. These more than outweigh the higher costs of serious clinical events and follow-up tests. At least this is my understanding of the (rather complex) Table 5. This raises two questions in my mind. How much would these costs vary if you were to study a less dispersed and remote population than that of northern Norway?

Response: This is an important question as cost of travel will decrease in more urban areas of the world. We do think this question is partly addressed in our sensitivity analyses (Tornado chart Figure 4) ranking cost elements according to their impact on societal cost. As shown, cost of sick leave is ranked first, followed by cost of follow-up test, and on third place cost of travel. To increase generalizability we have implemented cost elements from previously reported UK CE analyses (Hill JC, Whitehurst DGT, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice: a randomised controlled trial. Lancet 2011; 378: 1560–71). In this trial cost elements related to follow-up tests, outpatient consultation and GP consultations were reported 30-40% higher than similar Norwegian cost. Subsequently we have included this in the sensitivity analyses by increasing the upper limit by 40% (please see table 2 and figure 4).
This means that cost of sickleave, follow-up tests, hospital travel and overall health care resource utilization influences by far most overall cost. Interestingly, cost of GP travel have a minor impact on overall cost. Even if cost of hospital travel were to be substantially decreased (i.e. more urban areas and thus have less impact on overall cost), the cost of sick leave, follow-up test and overall health care resource use still would have major impact on overall societal follow-up cost.

(2) In the UK for example, the costs of travel would be borne by the patient, the costs of work loss mainly by the employer, and the costs of health care contacts and follow-up tests by the National Health Service. In other words, while assessing the global cost difference is interesting, it may not resonate with any of the individual actors involved in the drama (the patient, the health care provider, or the employer). The implications of your results are that, compared with surgeon follow up, GP follow-up involves a transfer of resources from the health care provider to the patient and the employer. Are health care providers under budget pressure likely to support this?

Response: We certainly agree, this is a drama, where health care providers are debating on how to spend sparse resources to provide the best cancer care. Decision-makers under budget pressure should explore novel ways of performing cancer follow-up, as this represent a huge financial burden to society. Based on our data (although with clear limitations) we argue that a GP organized follow-up must be considered in more urban areas as well: Firstly, these patients often have complex and chronic conditions, where organising cancer surveillance is only a part of the needs for this patient group. In our trial the GP group had less hospital travels, partly caused by better logistics (i.e. performs different tests at the same hospital visit, for instance different radiological examinations). Secondly, if cost of travel is paid by the patient, this is a argument for GP organised follow-up as this means less costly GP office visits and fewer hospital travels.
Thirdly, cost of a GP consultation is less compared to the cost of a hospital consultation (due to a less

costly infrastructure). This will decrease the NHS costs.
Finally, from an employers perspective, it seems like the patients in the GP group are returning to work faster (mean 17 days earlier, but not significant). This will decrease expenses of sick leave for the employer.
We acknowledge that further research is needed and recently a new UK research initiative to explore novel ways of organizing colorectal follow-up was reported:

Beaver K, Campbell M, Williamson S, et al. An exploratory randomized controlled trial comparing telephone and hospital follow-up after treatment for colorectal cancer. Colorect Dis 2012; 14: 1201–9.

Please note that several successful UK based interventions focusing on GP organized follow-up of breast cancer have been reported:

Grunfeld E. Comparison of breast cancer patient satisfaction with follow-up in primary care versus specialist care: results from a randomized controlled trial. British Journal of General Practice 2002; : 1–6.

Grunfeld E. Follow-up of breast cancer in primary care vs specialist care: results of an economic evaluation. Br J Cancer 1999; 79: 1227–33.

Grunfeld E. Routine follow up of breast cancer in primary care: randomised trial. BMJ 1996; : 1–5.

I mention these issues not as criticisms of your study or analysis, but as interesting implications you might like to think about and to mention in your discussion on p. 20.

Response: Thank you for asking these relevant questions which are addressed in the discussion

VERSION 3 - REVIEW

REVIEWER	Prof. Kjetil Søreide MD, PhD Department of Gastrointestinal Surgery Stavanger University Hospital Stavanger, Norway
REVIEW RETURNED	12-Feb-2013

THE STUDY	May benefit from English revision and linguistic work-up.
GENERAL COMMENTS	The suggested work-up in the table of the national guidelines from 2007 suggest contrast enhanced liver sonography (CEUS) - however, this modality has not been used nor introduced in the guidelines before 2010. It needs to be revised to 'liver ultrasonography', which will be the modality performed for all/most patients in the study period. The authors have otherwise recognised several of the limitations well in the discussion. Last paragraph in Discussion (before conclusion) is a mere repetition of several results, and could be deleted to make the discussion more succinct.