PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data
AUTHORS	Rupert A Payne, Gary A Abel and Colin R Simpson

VERSION 1 - REVIEW

REVIEWER	Simon Capewell Professor of Clinical Epidemiology Division of Public Health, University of Liverpool United Kingdom
REVIEW RETURNED	19/12/2011

GENERAL COMMENTS	An interesting and potentially important paper. No major deficiencies. But it would be even better after attention to a few minor things Page 3, ABSTRACT Basically fine. Line 31. “Cases of myocardial infarction, ischaemic heart disease and cerebrovascular disease were identified” Better to say “Cases of myocardial infarction, ischaemic heart disease and stroke (cerebrovascular disease) were identified...” Line 33. “Patient characteristics and incidence rates were assessed for all 3 clinical outcomes, based on GP, hospital, paired GP/hospital (similar diagnoses recorded simultaneously in both datasets), or combined GP and/or hospital records.” “Incidence rates were highest using combined GP and/or hospital data, “And/or” is a rather inelegant phrase which potentially confuses readers, Please phrase both sentences more clearly, and better define HOW this group was identified.

	Page 4, Line 35 “Limitations of this study include restricting our analysis to four coding groups, uncertainty as to whether GP and hospital events could be considered to be recorded simultaneously, and potential diagnostic coding inaccuracies.” Please also mention the relatively small number of GP practices, and the consequent possibility of non-representativeness. METHODS Page 6, line 14 please amend: “considered REASONABLY representative of the Scottish population”. Line 57 “CVD including haemorrhage and TIA” please mention the word "stroke". Please spell out more clearly how the “GP and/or hospital” differs from the GP alone or hospital alone group. This is important, and currently potentially confusing. Did you limit or otherwise standardise the period to be examined prior to 1/1/2005? If not you will need to discuss potential biases in the Limitations paragraph. current smoking status, presence of hypertension, presence of diabetes, ...” would be better described as RECORDED current smoking status, RECORD of hypertension, RECORD of diabetes, “ because of the many false negatives. Page 7, line 51 Charleson index. Can you say anything about completeness or validity in any of these data sets? Chi square test for proportions is an inherently conservative test , particularly for relatively small patient groups. Thus some of the NS results in Table may be false negatives (a Type II error). Caution is therefore required when interpreting the apparent lack of statistically significant differences. RESULTS Figure 1 appears to have an error. Patient D has three GP episodes, and no hospital admissions, but is not categorised as a GP patient Table 3, and narrative text. These are "case fatality rates", not mortality rates (which are population based). Please amend. DISCUSSION, para 3. Page 12, line 51 onwards Please say "case fatality", not "mortality". Better to use standard phrases to distinguish acute myocardial infarction admissions (high case fatality) , post MI patients (have

	survived beyond discharge, much lower case fatality subsequently) Page 12, Line 35 “lower prescribing rates of statins and antiplatelet agents in the hospital group may echo inadequate communication at the primary-secondary care interface. “ This seems to be contradicted by the higher prescribing rates in paired patients (Table 3). Please rephrase. Line 51. “Furthermore, the less clear cut nature of “heart attack”, due to the introduction of highly sensitive cardiac enzyme assays, has led to overlap between the diagnoses of angina and myocardial infarction[14].” Not quite. Please re-read reference 14, then the relevant Framingham paper (Circulation 2009;119;1203-1210); then rephrase. Ideally using terms such as "acute coronary syndrome, STEMI , Non-STEMI and unstable angina" etc Page 13, Limitations paragraph. Basically fine. Please also mention the relatively small number of GP practices, and the consequent possibility of non-representativeness. Also the relatively small numbers of patients, risking type II errors. (a statistical referee might insist on 95% confidence intervals to make thnings clearer) Did you limit or otherwise standardise the period to be examined prior to 1/1/2005? If not you will need to discuss potential biases here in the Limitations paragraph. Funding. Not very informative. Perhaps sensible to say where the authors salaries actually came from, just to reassure some sceptical readers. As well as Circulation 2009;119;1203-1210; Please also mention three other key Simpson References. And please critique & link your findings to theirs: 1. Brian S. Buckley, Colin R. Simpson,David J. McLernon, Philip C. Hannaford, Andrew W. Murphy,Considerable differences exist between prevalent and incident myocardial infarction cohorts derived from the same population. Journal of Clinical Epidemiology December 2010, Volume 63, Issue 12, Pages 1351-1357. 2. Matt P McGovern etal; Family Practice (2008) 25 (1): 33-39. doi: 10.1093/fampra/cmm073 The effect of the UK incentive-based contract on the management of patients with coronary heart disease in primary care (a potentially relevant issue). 3. N F Murphy, et al Prevalence, incidence, primary care burden and medical treatment of angina in Scotland: age, sex and socioeconomic disparities: a population-based study. Heart 2006;92:1047-1054 doi:10.1136/hrt.2005.069419

	Hope these comments are helpful.

REVIEWER	John Robson Reader Centre for Primary Care and Public Health Barts and the London School of Medicine and Dentistry, Queen Mary University of London. London, UK. I have no competing interests to declare.
REVIEW RETURNED	22/12/2011

THE STUDY	1. How many practices in total in the PTI project, and how were they selected - how were the 40 practices then selected? 2. Perhaps make it clearer that this hospital data only relates to in-patient admissions and not outpatient visits. 3. probablistic matching is used but there was no data on the extent to which individuals were correctly or unable to be matched. 4. Could the authors clarify where the prescribing data was obtained from - is there a prescribing data set in both the GP data and a different prescribing dataset for in hospital. If all the data on prescribing is from GP records and a substantive number of inpatients die, the lower prescribing in the hospital groups (which is not a very large diference) may well be due to this. If this is so could the authors clarify in the text? The authors make references to poor communication "better communication" "inadequate communication"- but present no evidence for this in their paper and I dont really feel it deserves the prominence they give it as an explanation for differences. - The differences above are a more likely explanation. 5. The authors say "mortality (in hospital) is generally considerable" - they might say more about the approximate numbers/proportion of inpatient deaths as they will count in hospital data but probably rarely in the GP data - am I correct in thinking this is probably around the 20% 30 day figure they found in the study - and hence may explain a large part of the hospital 'excess' prevalence, which is about 30% higher than GP records. If so this could be expanded in the text. 6. Reference 4 is incorrect date 7. Are any of the references below of relevance? Is variation in both hospital and GP coding adequately considered? Assessment of the under-reporting of diabetes in hospital admission data: a study from the Scottish Diabetes Research Network Epidemiology Group. Anwar H, Fischbacher CM, Leese GP, Lindsay RS, McKnight JA,

	Wild SH; on behalf of the Scottish Diabetes Research Network Epidemiology Group. Diabet Med. 2011 Dec;28(12):1514-1519. doi: 10.1111/j.1464-5491.2011.03432.x. Br J Gen Pract. 2000 Sep;50(458):706-9. An assessment of morbidity registers for coronary heart disease in primary care. ASSIST (ASSessment of Implementation STrategy) trial collaborative group. Moher M, Yudkin P, Turner R, Schofield T, Mant D. Fam Pract. 2003 Dec;20(6):706-10. Optimal strategies for identifying patients with myocardial infarction in general practice. Donnan PT, Dougall HT, Sullivan FM. JRSM Short Rep. 2011 Nov;2(11):83. Epub 2011 Oct 31. Increased length of inpatient stay and poor clinical coding: audit of patients with diabetes. Daultrey H, Gooday C, Dhatariya K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Diabet Med. 2011 Aug 26. doi: 10.1111/j.1464-5491.2011.03419.x. [Epub ahead of print
RESULTS & CONCLUSIONS	For the reasons given above it seems likely that inpatient mortality may have made a substantial contribution to prevalence and presribing in GP survivors - as the study does not provide any evidence on communication it does not follow that differences between groups result from this cause and undue emphasis is placed on this in the paper.
REPORTING & ETHICS	No mention of ethics approval - if it was not necessary does this need mentioning?
GENERAL COMMENTS	This is a useful study which would benefit from mior revision and recent references to similar work in Scotland on diabetes

REVIEWER	Michael Hobbs Emeritus Professor Unnversity Western Australia Ausralia I declare that I have no competing interests in relation to my review of this paper.
REVIEW RETURNED	10/01/2012

GENERAL COMMENTS	This is an interesting study demonstrating the use of linked data from different sources (hospital and GP records), to provide improved estimates of incidence, clinical characteristics, and prevalence of treatment in selected cardiovascular disorders. The objectives of the study have been achieved but the methods are not always clear and require some clarification. The study design is complicated. Could the authors indicate whether the following is correct and if necessary clarify the text accordingly: 1. The study relates to the population of the Practice Team Information (PTI) project, which has an aggregate population of 240,846 patients. 2. All 'incident' events of AMI, total IHD or CVD occurring in the two year period 2005-2007 in the PTI population were identified from linked hospital and GP records with a variable lead-in interval to identify first events. 3. The cases so identified were allocated randomly to four equal sub-groups (presumably assumed to be based on four equal patient populations). 4. The two year incidence of each of AMI, total IHD and CVD in eachsub- group were then identified using different selection criteria as follow: a. Events noted in Present in GP records only b. Events noted in hospital admission records only. c. Events noted in both GP and hospital records within the same 30 day period (assumed to be the same event), referred to as paired records. d. Events present in either GP AND/OR hospital admission records, irrespective of interval between GP and hospital events if both present. 5. Comparisons were made of the distributions of socio/demographic and clinical variables within each diagnostic group according to each selection method. 6. Thirty-day case fatality ratios were estimated for AMI according to each selection method 7. Incident rates were estimated separately for each diagnostic group and each selection method. The following points also need clarification: i. Why was it not possible to apply each case identification method to the full data set? It would then have been possible to identify through direct comparison the differences between the non-concordant groups – for example cases diagnosed from GP data only (no hospital records) and hospital cases (no GP records). This may also have solved some of the problems with statistical power. ii. I did not find Figure 1 easy to follow. I suggest the four identification systems be listed separately in the text and explained in detail. iii. A preferable method of identifying first events in the study period would hve been to use a fixed lead-in interval (say of ten years) so that all incident cases were defined in the same way. iv. It is not clear how hospital records relating to the PTI populations were identified in the SMR. Was this by address codes for patients'

	normal residence or is there a special flag for PTI cases? Can admissions to hospitals outside the study population areas be identified? v. How was the Charleson Index estimated? Was this from other coded comorbidities in the Index records or from data compiled more widely across the data sets? vi. Are drug therapies noted routinely in the SMR? It is noted from comparison of counts in Tables 1 and 2 that in **hospital records**, drug data are missing in 20% of cases of AMI, nearly 10% of IHD and 15% of stroke. This should be mentioned. vii. How were the denominators (person years at risk) for incidence rates in each of the sub-groups determined? Was these assumed to be one quarter of the total multiplied by two? **Results** viii. Comparisons for CF across the groups are not very meaningful as cases dying in hospital (particularly AMI) generally occur well within the 30 day interval and thus cannot have a subsequent GP event that would result in a 'paired event' ix. As counts for each selection method for each diagnostic group are derived from different populations, even if of equal size, they cannot strictly be compared. However it would be of interest to know how much records from either GP or hospital sources add to the overall total based on GP AND/OR hospital records. For instance, Table 1 suggests that hospital records account for over 80% of total cases of AMI and CVD and over 90% of total IHD, compared with 65%, 70% and 60% in the case of GP records. While hospital cases are thus the major source for all diagnostic groups, GP records would nevertheless appear to increase counts of AMI by about 20%, IHD by 10% and CVD by 15% and could make major contributions to case-finding epidemiological studies of AMI or Stroke based on disease registers. x. The counts for paired hospital-GP events are well below the total events and would appear to have no practical value. Is it possible that these results are affected by elective hospital admissions (say for angiography or revascularisation procedures) as these may not necessarily occur within 30 days of the GP attendance when specialist referral was made? **Discussion** xi. It should be stressed that further validation of coding according to predetermined criteria for the diseases of interest is required based on the extraction of further clinical and diagnostic information from the source records. The use of linked data, including laboratory test results could be of great value in facilitating such studies.

VERSION 1 – AUTHOR RESPONSE

REVIEWER: PROF. SIMON CAPEWELL

a) Abstract
Pg 3, line 31: Better to say "Cases of myocardial infarction, ischaemic heart disease and stroke (cerebrovascular disease) were identified..."

This has been changed as recommended.

b) Pg 3, line 33: "And/or" is a rather inelegant phrase which potentially confuses readers, Please

phrase both sentences more clearly, and better define HOW this group was identified.

The revised term “pooled GP/hospital records” has been used for this group. This new phrase has been changed throughout the paper, including the abstract and figures/tables.

c) Pg 4, line 35: [limitations] Please also mention the relatively small number of GP practices, and the consequent possibility of non-representativeness.

Added to the article summary box as suggested.

d) Methods
Pg 6, line 14: “considered REASONABLY representative of the Scottish population”.

Amended as suggested.

e) Pg 6, line 57: “CVD including haemorrhage and TIA” please mention the word "stroke".

The word stroke has been added.

f) Pg 7, para 3 (“Analysis was carried out…”): Please spell out more clearly how the “GP and/or hospital” differs from the GP alone or hospital alone group. This is important, and currently potentially confusing.

We agree this is potentially difficult to follow. As mentioned above, we have now elected to use the term “pooled GP/hospital records” throughout the manuscript. We have explicitly stated that the pooled data may included events from the GP data only OR hospital data only OR both datasets together (although not necessarily within 30 days as required for the paired group).

g) Pg 7, para 3: Did you limit or otherwise standardise the period to be examined prior to 1/1/2005? If not you will need to discuss potential biases in the Limitations paragraph.

The period prior to 2005 was not limited. The sentence on line 33 that read “…clinical event prior to 1/1/2005…” has been modified to read “…clinical event AT ANY TIME prior to 1/1/2005…” A comment on this has been added in the limitations on page 15.

h) Pg 7, para 4: current smoking status, presence of hypertension, presence of diabetes, ...” would be better described as RECORDED current smoking status, RECORD of hypertension, RECORD of diabetes, “ because of the many false negatives.

Agreed – changed accordingly.

i) Page 7, line 51: Charlson index. Can you say anything about completeness or validity in any of these data sets?

The implementation of Charlson was a pragmatic one, and no formal assessment of performance has been conducted. Khan et al. have published an adaptation of Charlson using Read/OXMIS codes (BMC Family Practice 2010, 11:1). Our code list matches 87% of clinical events (and 91% of codes) identified by the Khan method, based on the 2009 release of GPRD. Although not ideal, we believe this is sufficient to give a reasonable quantification of co-morbidity. A sentence to this effect has been added to the text.

j) Pg 8, line 21: Chi square test for proportions is an inherently conservative test, and there is thus a risk of false negatives (Type II error).

We have acknowledged this issue in the section on limitations on page 15 in the discussion.

k) Pg 22, Figure 1: The figure appears to have an error. Patient D has three GP episodes, and no hospital admissions, but is not categorised as a GP patient.

This patient has a GP episode prior to 1/1/05, and as such neither of the subsequent GP episodes are

counted as incident events. In response to another referee's comments, we have added clarification of the methods used to identify cases during the lead-in period on page 7 of the methods (paragraph beginning "Analysis was carried out…")

l) Pg 18, Table 3, and narrative text: Use "case fatality rates", rather than the term "mortality rates".

This has been amended in Table 3, and in the text on pages 5, 9, 11, 12, 13 and the abstract

m) Discussion
Page 12, line 51 onwards: Please say "case fatality", not "mortality". Better to use standard phrases to distinguish acute myocardial infarction admissions (high case fatality), post MI patients (have survived beyond discharge, much lower case fatality subsequently)

I think this comment was meant to refer to page 11 of the original manuscript. The term case fatality has been used as mentioned above. We have rephrased the paragraph beginning "The discrepancies in…" as suggested.

n) Page 12, Line 35: "lower prescribing rates of statins and antiplatelet agents in the hospital group may echo inadequate communication at the primary-secondary care interface." This seems to be contradicted by the higher prescribing rates in paired patients (Table 2). Please rephrase.

We agree this could be potentially confusing, and have reworded the statement accordingly.

o) Line 51. "Furthermore, the less clear cut nature of "heart attack", due to the introduction of highly sensitive cardiac enzyme assays, has led to overlap between the diagnoses of angina and myocardial infarction[14]." Not quite. Read [Parikh et al, Circulation 2009;119;1203-1210], then rephrase, ideally using terms such as "acute coronary syndrome, STEMI, etc.

We agree that the previous wording is not strictly true; we have reworded the sentence accordingly, and have also changed the reference to that suggested by the referee.

p) Page 13, Limitations paragraph: Please also mention the relatively small number of GP practices, and the consequent possibility of non-representativeness. Also, the relatively small numbers of patients, risking type II errors. Did you limit or otherwise standardise the period to be examined prior to 1/1/2005? If not you will need to discuss potential biases here in the Limitations paragraph.

These additional limitations have been added to the discussion on page 13 as suggested.

q) Funding. Not very informative

We note the editor's comment that the current standard BMJ Open funding statement is acceptable

r) Please reference, critique and link to the following key Simpson References:
1. Buckley BS et al, J Clin Epi 2010;63:1351-1357
2. McGovern MP et al, Fam Pract 2008;25:33-39
3. Murphy NF et al, Heart 2006;92:1047-1054

We have added a new second paragraph to the discussion, referencing these papers (and two of those mentioned by referee 2), giving a brief overview of previous cardiovascular epidemiology work conducted in Scotland.

REVIEWER: DR JOHN ROBSON
1) How many practices in total in the PTI project, and how were they selected - how were the 40 practices then selected?

There are 60 practices in the PTI project; the 40 used in the linked dataset were self-selected. This clarification has been added to the top of page 6 of the methods.

2) Perhaps make it clearer that this hospital data only relates to in-patient admissions and not

outpatient visits.

Agreed – we have explicit statements to this effect on page 6.

3) probablistic matching is used but there was no data on the extent to which individuals were correctly or unable to be matched.

We have contacted the linkage team in ISD who carried out this work. They are unable to provide a definitive value. Of course, a substantial proportion of patients in this general practice cohort have had no hospital admissions. As such, it is very difficult to know whether the lack of a match to a hospital record is either due to there being no such hospital record, or due to a false negative. The linkage was carried out using human review to determine the threshold score for matching.

The matching rate is therefore difficult to quantify. However, as several identifiers including the unique Community Health Index number were available and the techniques used for linkage including individual assessment of non-linkers, the sensitivity and specificity of the matching was considered by the data linkage team to be high. We have added additional text to paragraph 2 of the methods to clarify this.

4a) Could the authors clarify where the prescribing data was obtained from - is there a prescribing data set in both the GP data and a different prescribing dataset for in hospital.
If all the data on prescribing is from GP records and a substantive number of inpatients die, the lower prescribing in the hospital groups (which is not a very large diference) may well be due to this. If this is so could the authors clarify in the text?

The referee is correct in that the prescribing data is entirely from the GP record – this is now explicitly stated on page 8. In terms of in-patient deaths, we only analysed prescribing in patients alive at 30 days, as stated in the methods (para 2, page 8); we also refer to the 30-day limit (and the implications of patients with longer hospital stays) at the end of the paragraph in the discussion on limitations.

4b) The authors make references to poor communication "better communication" "inadequate communication"- but present no evidence for this in their paper and I dont really feel it deserves the prominence they give it as an explanation for differences. - The differences above are a more likely explanation.

We have rephrased the first sentence to remove the word “communication”. However, we have left the single phrase “better communication” as we don’t think this overplays the issue on its own; furthermore, the deaths is less likely to be an issue seeing as we restricted the analysis to survivors. The comment on “inadequate communication” has been replaced by a revised sentence anyway, to address issues raised by referee 1; we have removed the comment about communication in doing so.

5) The authors say "mortality (in hospital) is generally considerable" - they might say more about the approximate numbers/proportion of inpatient deaths as they will count in hospital data but probably rarely in the GP data - am I correct in thinking this is probably around the 20% 30 day figure they found in the study - and hence may explain a large part of the hospital 'excess' prevalence, which is about 30% higher than GP records. If so this could be expanded in the text.

We used national data from the General Registrar’s Office for Scotland to identify deaths; although this will only identify deaths registered (rather than necessarily occurring) in Scotland, it has the advantage of being consistent for both hospital and GP recorded events. Clarification of this has been added at the bottom of page 13.

6) Reference 4 is incorrect date

This has been corrected to 1995

7) Are any of the references below of relevance? Is variation in both hospital and GP coding adequately considered?
1. Anwar et al, Diabet Med 2011;28:1514-1519

2. Moher et al, BJGP 2000;50:706-9
3. Donnan et al, Fam Pract 2003;20:706-10
4. Daultrey et al, JRSM Short Rep 2011;2:83
5. de Lusignan et al, Diabet Med 2012;29:181-189

We have added a new second paragraph to the discussion, including reference to the two papers on coronary disease (Moher and Donnan) suggested above (as well as other papers suggested by referee 1). We have also referenced the accuracy of Scottish GP coding alongside the accuracy for hospital coding in the limitations section.

8) For the reasons given above it seems likely that inpatient mortality may have made a substantial contribution to prevalence and presribing in GP survivors - as the study does not provide any evidence on communication it does not follow that differences between groups result from this cause and undue emphasis is placed on this in the paper.

We acknowledge the referee's concern about placing undue emphasis on the issue of communication, and have attempted to address this – please see our responses to the above comments for details

9) No mention of ethics approval - if it was not necessary does this need mentioning?

This type of work required approval by the Privacy Advisory Committee (PAC) of NHS National Services Scotland – this has been added to paragraph 2 of the methods. PAC is an independent body set up for the purposes of protecting personal data and making data available for research, audit and other important uses, whilst ensuring that any information releases are carefully controlled. See http://www.nhsnss.org/pages/corporate/privacy_advisory_committee.php

10) This is a useful study which would benefit from minor revision and recent references to similar work in Scotland on diabetes

Reference to some of the papers listed above has been made as suggested

REVIEWER: PROF. MICHAEL HOBBS

..) The objectives of the study have been achieved but the methods are not always clear and require some clarification. The study design is complicated. Could the authors indicate whether the following [7 points] is correct and if necessary clarify the text accordingly.

We can confirm that the outline of the study, as summarised by the referee in 7 points, appears correct.

The following points also need clarification:
i) Why was it not possible to apply each case identification method to the full data set? It would then have been possible to identify through direct comparison the differences between the non-concordant groups – for example cases diagnosed from GP data only (no hospital records) and hospital cases (no GP records). This may also have solved some of the problems with statistical power.

Interestingly, we tried this approach originally. However, it becomes very difficult to analyse when one considers that a patient may have events identified by more than one selection method. Consider, for instance, a patient who has a GP record of MI 6 months into the study period, and a hospital record of MI 12 months into the study period (and no GP or hospital event prior to the start of the study period). This patient could thus have an incident event identified using three of the four methods we outline – which one do you use? Simply taking the first one risks biasing incidence of disease towards one particular coding system. Thus, the simplest method is the one we have undertaken, where there are equal numbers of patients in four independent subgroups. Of course, a comparison could indeed be carried out between the two groups described in the referee's example (with only GP data or only hospital data), but a majority of patients have coding in more than one group, and so this single comparison is not necessarily helpful on its own.

ii) I did not find Figure 1 easy to follow. I suggest the four identification systems be listed separately in the text and explained in detail.

We added Figure 1 after discussion with others, who felt that offering visual examples helped; a number of iterations of the figure were attempted, before we reached the version that has been included in the manuscript. However, we concur with the referee that there is a lack of detail pertaining to each case selection method in the text, and have added an additional paragraph (6 in the methods) to explain each method in more detail, and to try to complement Figure 1.

iii) A preferable method of identifying first events in the study period would hve been to use a fixed lead-in interval (say of ten years) so that all incident cases were defined in the same way.

We have acknowledged this weakness in the limitations section of the discussion.

iv) It is not clear how hospital records relating to the PTI populations were identified in the SMR. Was this by address codes for patients' normal residence or is there a special flag for PTI cases? Can admissions to hospitals outside the study population areas be identified?

Using a list of all patients in the PTI/GP record, all corresponding hospital records are identified using a probabilistic match. This is based on a number of parameters, including name, date of birth, sex, postcode and a unique nationwide identifier (the community health index, CHI). We have reworded the second paragraph of the methods to clarify this, and added details of the identifier parameters used. Admissions outside Scotland are not identified, as mentioned in the limitations section.

v) How was the Charleson Index estimated? Was this from other coded co-morbidities in the Index records or from data compiled more widely across the data sets?

The co-morbidity data comes from the GP data; this is now clarified in the methods.

vi) Are drug therapies noted routinely in the SMR? It is noted from comparison of counts in Tables 1 and 2 that in hospital records, drug data are missing in 20% of cases of AMI, nearly 10% of IHD and 15% of stroke. This should be mentioned.

We have added clarification that drug therapy is recorded in the GP record, rather than SMR. Only patients alive at 30 days were analysed for prescribing; this explains why there are fewer patients in Table 2 (drug therapy – live patients only) compared with Table 1 (patient characteristics – all patients). This is mentioned in the methods and limitations section, as well as Table 2, although we have now expanded the phrase in the footnote to Table 2 to read "Patients are those alive at 30 days, and this is reflected in lower values of N."

vii) How were the denominators (person years at risk) for incidence rates in each of the sub-groups determined? Was these assumed to be one quarter of the total multiplied by two?

The number of person years at risk is based on the total number of days of follow-up for each patient within each respective group. This clarification has been added to the section on "statistical analysis".

Results
viii) Comparisons for CF across the groups are not very meaningful as cases dying in hospital (particularly AMI) generally occur well within the 30 day interval and thus cannot have a subsequent GP event that would result in a 'paired event'

The point of the comparison is to demonstrate that, because the hospital events are not all being captured by the GP, cases identified using only GP data will not necessarily be representative of the wider population who have experienced, for example, a myocardial infarction. We would thus argue that the comparison is entirely reasonable.

We suspect, however, that this comment may reflect a lack of understanding of what GP events can capture. Most GPs record admissions after receiving a summary from the hospital detailing the patient's admission – it does not necessarily require a referral or face-to-face consultation. The admission (fatal or otherwise) is therefore recorded in the GP data retrospectively. Indeed, from our

data, the majority of GP and hospital paired events have exactly matching dates (irrespective of admission duration or fatality) suggesting that retrospective date entry is the norm. We apologise for our failure to explain this properly, and have added clarifications to this effect at the end of paragraph 4 of the methods, and the end of paragraph 3 of the discussion.

ix) As counts for each selection method for each diagnostic group are derived from different populations, even if of equal size, they cannot strictly be compared. However it would be of interest to know how much records from either GP or hospital sources add to the overall total based on GP AND/OR hospital records. For instance, Table 1 suggests that hospital records account for over 80% of total cases of AMI and CVD and over 90% of total IHD, compared with 65%, 70% and 60% in the case of GP records. While hospital cases are thus the major source for all diagnostic groups, GP records would nevertheless appear to increase counts of AMI by about 20%, IHD by 10% and CVD by 15% and could make major contributions to case-finding epidemiological studies of AMI or Stroke based on disease registers.

We disagree with the referee's first sentence – these groups were randomly selected from the total population, and as such comparison between them is entirely legitimate. However, we agree with the rest of the comment, that records from an additional data source can potentially contribute to case finding; we have added a comment to this effect at the end of the second paragraph of the discussion.

x) The counts for paired hospital-GP events are well below the total events and would appear to have no practical value. Is it possible that these results are affected by elective hospital admissions (say for angiography or revascularisation procedures) as these may not necessarily occur within 30 days of the GP attendance when specialist referral was made?

We suspect this comment may reflect a lack of clarity in the methods we employed, for which we apologise. Patients do not necessarily have to visit the GP to have an event recorded; furthermore, event dates are generally retrospectively entered. This should occur regardless of the type of admission; we acknowledge that it is possible, for instance, that coding of an elective angiography admission as something inappropriate – e.g. acute MI – may result in a coding mismatch if the GP (quite appropriately) doesn't record the admission as an MI. Miscoding of diagnoses is referred to under the "limitations" section of the discussion. We agree that, in reality, using pairing to identify cases may not be useful, as it clearly misses cases; however, this was not evident until we had carried out the analysis. We have added a sentence referring to the low event rate identified by the paired data at the end of the second paragraph of the discussion.

Discussion
xi) It should be stressed that further validation of coding according to predetermined criteria for the diseases of interest is required based on the extraction of further clinical and diagnostic information from the source records. The use of linked data, including laboratory test results could be of great value in facilitating such studies.

This is very true. We have added additional sentences in the limitations section on the accuracy of GP coding and the need for validation.

VERSION 2 – REVIEW

REVIEWER	Signed John Robson Centre for Primary Care and Public Health Barts and The London QMUL Yvonne Carter Building 58 Turner Street London E1 2AB Phone: 020 7882 2553 Fax: 020 7882 2552

	I have no competing interest
REVIEW RETURNED	06/02/2012

GENERAL COMMENTS	In the summary the authors say they 'found evidence (P<0.05) of different characteristics.....' is the 0.05 necessary here or could you simply leave it out as it doesnt add anything really and the comment on prescribing does draw attention to an aspect of the statistical analysis about which I felt rather uneasy - ie multiple significance testing - particularly an issue in the prescribing with so many comparisons. I did wonder whether some further statistical correction for multiple testing might set the bar higher for signifcance levels? I felt the paper was on surest ground when describing the different kinds of patient/mortality outcomes - highlighting the rather small differences in general on prescribing which I felt were most striking because of their similarities rather than because of their difference.... might be worth a little further consideration?

REVIEWER	Michael Hobbs Emeritus Professor and Senior Honorary Research Fellow University of Western Australia Australia I have no competing interests relating to this research publication.
REVIEW RETURNED	21/02/2012

The reviewer completed the checklist but made no further comments.