A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000723
Article Type:	Research
Date Submitted by the Author:	07-Dec-2011
Complete List of Authors:	Payne, Rupert; University of Cambridge, GP and Primary Care Research Unit Abel, Gary; University of Cambridge, GP and Primary Care Research Unit Simpson, Colin; The University of Edinburgh, Centre for Population Health Sciences
<b>Primary Subject Heading</b>:	Epidemiology
Secondary Subject Heading:	Cardiovascular medicine, Health informatics, Research methods
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Coronary heart disease < CARDIOLOGY, EPIDEMIOLOGY, PRIMARY CARE, STROKE MEDICINE, STATISTICS & RESEARCH METHODS
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
STROBE_checklist_BMJ-Open_cohort-studies.doc	

A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data

Rupert A Payne[1], Gary A Abel[1], Colin R Simpson[2]

1. GP and Primary Care Research Unit, University of Cambridge, Cambridge, UK
2. eHealth Research Group, Centre for Population Health Sciences, The University of Edinburgh, Edinburgh, UK

Corresponding author:
Dr Rupert A Payne
GP and Primary Care Research Unit
University of Cambridge
Institute of Public Health
Forvie Site
Robinson Way
Cambridge
CB2 0SR

Tel. 01223 746545
Email. rap55@medschl.cam.ac.uk

Key words: Data linkage, primary care, secondary care, routine data, cardiovascular, coding

Word count: 2987

ABSTRACT

Objectives:

Data linkage combines information from several clinical datasets. We examined whether coding inconsistencies for cardiovascular disease between components of linked datasets, result in differences in apparent population characteristics.

Design:

Retrospective cohort study.

Setting:

Routine primary care data from 40 Scottish General Practice (GP) surgeries linked to national hospital records.

Participants:

240,846 patients, age 20 years or more, registered at a GP surgery.

Outcomes:

Cases of myocardial infarction, ischaemic heart disease and cerebrovascular disease were identified from GP and hospital records. 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.

Results:

For all 3 outcomes, we found evidence ($p<0.05$) of different characteristics when using different methods of case identification. Prescribing of cardiovascular medicines for ischaemic heart disease was greatest for cases identified using paired records ($p\leq0.013$). For all conditions, 30-day mortality rates were higher for cases identified using hospital compared with GP or paired data, most noticeably for myocardial infarction (hospital 20%, GP 4%, $p=0.001$). Incidence rates were highest using combined GP and/or hospital data, and lowest using paired data.

Conclusion:

Differences exist in patient characteristics and disease incidence for cardiovascular conditions, depending on the data source. This has implications for studies using routine clinical data.

ARTICLE SUMMARY

Article focus

- Data linkage allows information to be combined from different routine clinical data sources
- Previous work has shown differences between sources of data, but has not examined this at the patient-level

Key messages

- Patients' apparent characteristics, and disease incidence and severity, vary depending on whether primary care, hospital or combined definitions of cardiovascular events are used
- Use of isolated routine primary care or hospital data may result in biased patient selection
- This has implications in the public health arena, clinical trial patient recruitment, and validity and reliability of secondary data in clinical trials

Strengths and limitations

- The strengths of this study are the novel analytical approach, using a large routine dataset linked at individual patient level from multiple GP surgeries
- 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

BACKGROUND

Primary care datasets are commonly used for assessment of cardiovascular outcomes. Such events often are associated with hospitalisation[1]. However, it is possible that the manner in which outcomes are coded and recorded in electronic health records may differ between primary and secondary care. This may result not only in differences in the apparent incidence of a condition, depending on whether primary or secondary care records are used, but also differences in the observed characteristics of patients. Studies have observed that variations in diagnostic criteria can affect estimates of disease prevalence[2], and the complexities of clinical coding systems for electronic health care records can lead to inconsistent data recording[3]. This will lead to uncertainties with respect to disease prevalence and mortality[4], impact on clinical care, have additional health service implications such as affecting funding[5], and potentially influence identification of patients for clinical trials. Previous studies have compared general practice coding and disease prevalence with other unlinked data sources including paper notes[6,7]. However, the effect of combining information from two sources has not been previously examined. This study used linked individual patient electronic health records collected from primary and secondary care to examine the effect of using data from different parts of the health care service on the incidence rates, mortality rates and patient characteristics of myocardial infarction (MI), ischaemic heart disease (IHD) and cerebrovascular disease (CVD).

METHODS

Data sources

We used primary care data from 40 general practice (GP) surgeries taking part in the Practice Team Information (PTI) project. Practices involved in the PTI project provide routine, central recording of clinical activity and morbidity from a sample of GP surgeries considered representative of the Scottish population. Practices are reimbursed to ensure data recording is optimal. Clinical coding used the Read code system. Data are used to calculate national estimates, and used by various organisations (e.g. NHS Boards, Scottish Government) to inform policies and better understand health in Scotland.

The PTI data were linked using probabilistic matching to Scottish national hospital data (the Scottish Morbidity Record, SMR-01). The linkage was carried out by the Information Services Division, NHS National Services Scotland. For the 2004 to 2006 period, SMR-01 data are considered to be 88% accurate[8]. SMR-01 records are generated for all hospital medical discharges and transfers. Coding is based on the International Statistical Classification of Diseases and Related Health Problems (ICD) system (ICD9 prior to 2000, ICD10 thereafter), with up to six in-patient diagnoses per record. Accident and emergency, maternity and psychiatric admissions are not recorded in SMR-01. SMR-01 itself is also routinely linked to national mortality data (General Registrar's Office for Scotland, GROS). SMR-GROS data are also used to generate Scottish national statistics.

Identification and classification of cases

We first identified all records of MI, IHD and CVD from both GP and hospital datasets using the following Read codes (MI: G30%/35%/38%, Gyu34/35/36; IHD: G3%, Gyu3%; CVD: G6%, Gyu6%, F4236; where % indicates a "wildcard" match) and ICD codes (MI: ICD10 I21-22; ICD9 410. IHD including MI: ICD10 I20-25; ICD9 410-414. CVD including haemorrhage and TIA:

ICD10 I60-69, G45-46; ICD9 430-438). Hospital events were identified from any of the six diagnostic positions.

We then found all episodes of a similar GP and hospital event type occurring within a 30-day period, and made the assumption that these pairings represented the same clinical event. Where the GP and hospital dates differed for these paired episodes, the first of the two dates was taken. The choice of 30 days was a pragmatic one, but supported by visual evaluation of the distribution of time gaps between similar hospital and GP event types over a two-year period.

Analysis was carried out over the period 1/1/2005 to 1/1/2007. The total population was randomly allocated to one of four methods of identifying cardiovascular events: those based on GP events only; those based on hospital events only; those based on GP and/or hospital events (not necessarily occurring within 30 days); and those based on paired GP/hospital events (those recorded in both GP and hospital data within 30 days). An episode was included as an incident event only if there was no record of a similar clinical event prior to 1/1/2005 coded in the same dataset(s). For example, an event coded in the GP (but not hospital) dataset in 1990 would not preclude an episode coded in the hospital dataset in 2006 counting as a hospital incident event. This method of identifying incident events is shown graphically in Figure 1.

{INSERT FIGURE 1}

For each incident event, we determined the patient's age, sex, socioeconomic status (Scottish Index of Multiple Deprivation quintile, SIMD)[9], current smoking status, presence of hypertension, presence of diabetes, and Charlson index[10]. Co-morbidities were determined as the presence of any relevant diagnostic Read code prior to the incident episode date; the list of codes used is available from the authors on request. Death from any cause within 30 days of the event was ascertained from linked GROS data. Drug therapy, starting prior to or within 30 days after the event, and continuing for any period of time after the event, was ascertained for patients alive at 30 days. Drug classes included were angiotensin converting enzyme (ACE) inhibitors (including angiotensin receptor blockers), beta-blockers, calcium channel blockers, diuretics (including potassium sparing and combination diuretics), nitrates, statins and antiplatelet agents (aspirin or clopidogrel for MI or IHD; aspirin or dipyridamole for CVD).

Statistical analysis

Incidence rates were calculated excluding patients with events in the relevant dataset(s) prior to 1/1/2005. Incidence rates are expressed per 100,000 patient years. Statistical differences in patient characteristics (including drug treatment) between coding categories were evaluated using Chi-squared tests (for proportions) and Kruskall-Wallace non-parametric ANOVA (for continuous data). The association between coding and 30-day mortality was assessed by logistic regression, including the covariates age, sex, deprivation, smoking status, hypertension, diabetes and Charlson index. Differences in the four incident rates obtained were examined using Poisson regression.

Data management was carried out using Microsoft SQL Server 2000. Statistical analysis was performed using SPSS v17 (SPSS inc, Chicago, Illinois, USA).

RESULTS

Differences in identification of incidence events

There were a total of 240,846 patients, evenly distributed between the four coding groups. Numbers of incident events are shown in Table 1. Incidence rates for the three conditions are shown in Figure 2. There was strong evidence ($p<0.001$, Poisson regression) that the incidence rates for all three clinical conditions depends on which dataset(s) are used to identify cases. In all cases, the GP and/or hospital data produced the highest incidence rates (376, 1089 and 767 per 100,000 patient years for MI, IHD and CVD respectively), and the paired GP/hospital data gave the lowest incidence rates (188, 489 and 272 per 100,000 patient years respectively). There was no evidence that the incidence rates based on only GP data differ from those of the hospital data for either MI ($p=0.14$) or CVD ($p=0.27$), but there was strong evidence that they were higher for IHD (975 and 673 events per 100,000 patient years for hospital and GP respectively, $p<0.001$). The GP and/or hospital data produced slightly higher incidence rates than hospital data alone for CVD ($p<0.001$) and marginally so for MI ($p=0.048$) and IHD ($p=0.066$).

{INSERT TABLE 1}

{INSERT FIGURE 2}

Patient characteristics

Patient characteristics are shown in Table 1 for all three clinical conditions. There was no evidence that rates of diabetes and hypertension, or the distribution of sex or deprivation, varied between coding groups. Greater numbers of smokers were found in the paired GP/hospital group for patients with MI (45% in the paired group compared with 28 to 34% in the other groups, $p=0.028$) and IHD (35% compared with 24 to 27%, $p=0.021$). The level of co-morbidity for all conditions, as measured by the Charlson index, is lower in the paired GP/hospital group (1.8, 1.3 and 1.9 for MI, IHD and CVD respectively) and higher in the hospital group (2.2, 1.7 and 2.4 respectively, p≤0.014). For IHD and CVD, there is evidence that patients identified using solely GP or solely hospital data were slightly younger.

Prescribing

Differences in prescribing rates were observed between coding groups (Table 2). These were most marked for IHD, where rates of prescribing of ACE inhibitors, beta-blockers, nitrates, statins and antiplatelet agents were higher in the paired group (p≤0.013). However, this finding did not appear to be replicated for MI specifically. For CVD, prescribing rates for statins and antiplatelet agents were lower in the hospital group (p≤0.022).

{INSERT TABLE 2}

Mortality

Considerable 30-day mortality rate differences exist for all three conditions depending on the coding used (p≤0.002, Table 3). Rates for all conditions are highest in patients coded only in hospital, and lower in the GP and paired GP/hospital groups. The most striking differences were observed for MI, with a 30-day mortality rate of 20% for the hospital group but only 4% for the GP group.

{INSERT TABLE 3}

DISCUSSION

In a world where electronic healthcare data are becoming increasingly used for the purposes of clinical trials and epidemiological research, there is a need for researchers to understand whether additional information can be gained by linking two (or indeed more) electronic health record data sources together. However, where there is overlap between the constituent datasets, such as with coding of clinical conditions, the researcher needs to decide which dataset to rely on for identifying cases, or indeed whether combining information from both the two datasets may be of value. Our study demonstrates that the method of coding MI, IHD and CVD appears to result in identification of different types of patient, in particular as characterised by prescribing and mortality rates. Incident rates of disease also vary depending on the coding method used.

Reasons for differences in incidence rates and patient characteristics

Our data do not allow us to determine the exact cause of our findings, but a number of hypotheses may be proposed. Incident disease is reassuringly similar between GP and hospital groups for MI and CVD. The lower incidence of IHD for the GP group reflects the fact that many patients will have had relatively stable coronary disease for a number of years but not necessarily required acute hospital admission. Thus, many GP episodes of IHD do not count as true incident cases as they have had prior contact with the GP, whereas a higher number of hospital episodes are incident cases as these patients have never been previously admitted. The lower incidence rates for the paired GP/hospital group, and higher incidence rates for the combined GP and/or hospital group, are inevitable consequences of the way in which the two datasets are united, although the magnitude of these differences will nonetheless reflect the degree of inconsistency in coding between the two.

The discrepancies in death rates are probably relatively straightforward to explain. Mortality is generally considerable in these patients[1], and if death occurs during a hospital admission, it is possible that the GP may code the patient as deceased but fail to code the cause of death. In contrast, those who survive the hospital admission are picked up by their GP, hence the lower mortality rates in the paired GP/hospital coding group. Furthermore, it is possible that patients coded only by the GP may represent “less serious” illness, where hospitalisation is not deemed necessary by the GP. It is recognised that many patients suffering relatively minor strokes may not be admitted to hospital[11], resulting in lower mortality for CVD in the GP group, although with the growing availability of active treatment options for ischaemic stroke in the form of thrombolysis, this may well change.

The higher prescribing rates for IHD in the paired coding group are probably due to GPs responding appropriately to communications from secondary care, reflected in appropriate treatment. That such differences were not observed for MI, may be due to better communication and awareness for this specific condition compared with other IHD such as angina, meaning that prescribing in the hospital group appears just as good as for the paired GP/hospital group. However, fewer MI events may have left us underpowered to detect differences. The lack of difference in the GP and paired groups for CVD may reflect poorer awareness of stroke management guidelines[12] in comparison with coronary heart disease, and so prescribing rates are consequently no higher in the paired group; the lower prescribing rates of statins and antiplatelet agents in the hospital group may echo inadequate communication at the primary-secondary care interface. The differences in other patient characteristics – specifically smoking and co-morbidity – are less easy to understand, but may represent increased disease severity and mortality in hospitalised smokers and multi-morbid patients. The small differences in age (<3 years) seem unlikely to be clinically relevant, although may be pertinent from the public health perspective. Finally, it may be that miscoding of diagnoses may explain some of the above differences; for instance, heart failure may be used as an alternative but incorrect code for myocardial infarction[13]. 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].

Limitations

This study has highlighted important issues related to patient coding and linked data, but although it has the advantage of using a large routine dataset, linked at the individual patient level, a number of issues and limitations should be considered. We restricted our analysis to four simple coding groups – GP, hospital, paired and combined GP and/or hospital. However, it is clear that there are many further ways of categorising events, including the presence or absence of prior or subsequent coding based on the alternative half of the dataset. For instance, an incident GP event with a historical hospital event, may be coded differently to a GP event with no previous hospital record. However, we found that many of these theoretical categories have only a handful of cases. Furthermore, even when we examined six or seven separate smaller coding categories, similar differences in patient characteristics persisted between groups (data not shown). Our choice of four main groups was therefore a pragmatic one which reflects the choice that would face a researcher dealing with a similar linked dataset. The decision to use a 30-day limit for pairing data could also be questioned; we are unable to prove that these two events are truly the same clinical episode. The choice was again, therefore, partly pragmatic, although supported by examination of the distribution of time gaps between the GP and hospital data. Our study used routine GP data, and it is possible that such profound differences may not be found with research-standard databases such as GPRD[15]. Nonetheless, work linking primary care research databases to hospital (and other) records is ongoing, and the issues raised by our study must be acknowledged. The SMR dataset only records hospital events in Scotland, and thus fails to capture events in elsewhere in the UK or abroad. Similar issues face the English equivalent Hospital Episode Statistics (HES), and a UK-wide hospital events dataset would be valuable. SMR (and HES) also provide multiple diagnostic codes for a single event. We elected to use all six diagnostic positions to ensure maximum capture of relevant hospital events. However, the robustness of low-priority diagnoses might be questioned. Nonetheless, we found similar results when we used only two diagnostic positions (data not shown). We also did not examine miscoding of events – e.g. a code of angina being used rather than the code for MI. Coding of SMR is considered 99% complete and 88% accurate[8], but these data are not available for PTI data. Furthermore, the two datasets use different coding systems, so completely reliable comparison is not possible. However, we used relatively broad definitions, and the Read code system is based on ICD. Nonetheless, we may in particular have missed some administrative Read codes which might have enabled identification of additional cases in the GP group. Finally, our 30-day limit for prescribing was selected from a pragmatic perspective. However, it is possible that patients who were admitted for over 30 days would not have had a new prescription issued by the GP within the 30-day post event period, resulting in an apparent underestimation of prescribing. We believe these numbers will be relatively small, however, and unlikely to alter the overall interpretation of our findings.

Research and policy implications

These results have significant implications for linked data; the drug management, disease severity, and to some degree the patient characteristics, vary depending on how the disease cohort is defined. They also have implications for the use of unlinked routine data – use of isolated primary or secondary care data may result in a biased selection of patients. This may affect patient recruitment as well as the validity and reliability of such information sources as secondary data in clinical trials, including clinical outcomes. It is similarly relevant to the public health environment. Using linked data allows one to have a more robust definition, by using pairs of GP and hospital codes only, but it is clear that the apparent incidence of a disease will be considerably lower. Alternatively, linked data enable a looser but more inclusive disease definition, using both GP and hospital data, but not relying on the coding occurring simultaneously. When using separate data from only one source, one needs to take into account that patient characteristics may not be representative of the wider population. It is difficult to recommend one coding approach over another, however, and the decision will need to be based on the specific question being posed.

Conclusions

In conclusion, patient characteristics vary depending on whether GP, hospital or combined definitions of cardiovascular events are used. In particular, disease severity as measured by mortality varies considerably. This has important implications for studies using linked routine primary and secondary care data, and for studies where information is only available from one of these sources. These issues should be acknowledged by studies using routine data as a secondary data source, and further work is merited to examine whether similar discrepancies exist for other clinical conditions or within primary care research databases.

TABLE 1 – PATIENT CHARACTERISTICS

		GP	Hospital	Paired GP/ hospital	GP and/or hospital	p value
Myocardial infarction						
N		145	171	105	209	
Males (%)		65%	59%	60%	64%	0.68
Age, mean (SD)		68 (13.8)	67 (13)	68.4 (13.8)	68.8 (14.9)	0.51
Deprivation	1	19%	11%	10%	12%	
quintile	2	15%	25%	26%	17%	
	3	26%	17%	29%	31%	0.55
	4	15%	23%	21%	22%	
	5	24%	24%	14%	17%	
Smokers (%)		33%	34%	45%	28%	0.028
Diabetes (%)		15%	12%	8%	11%	0.29
Hypertension (%)		39%	44%	38%	44%	0.52
Charlson index, mean (SD)		2.5 (1.7)	2.2 (1.6)	1.8 (1.4)	2.0 (1.6)	<0.001
Ischaemic heart disease						
N		362	529	270	585	
Males (%)		56%	55%	61%	56%	0.38
Age, mean (SD)		66.2 (12.7)	65.8 (11.6)	66.9 (13.4)	68.4 (12.8)	0.007
Deprivation	1	17%	13%	11%	13%	
quintile	2	18%	20%	20%	21%	
	3	29%	23%	27%	26%	0.25
	4	17%	22%	24%	20%	
	5	20%	23%	19%	19%	
Smokers (%)		27%	27%	35%	24%	0.011
Diabetes (%)		11%	15%	13%	10%	0.091
Hypertension (%)		42%	47%	44%	45%	0.51
Charlson index, mean (SD)		1.5 (1.6)	1.7 (1.6)	1.3 (1.3)	1.5 (1.5)	0.002
Cerebrovascular disease						
N		302	330	153	424	
Males (%)		48%	47%	46%	47%	0.97
Age, mean (SD)		70.3 (14.1)	70.8 (13.6)	72 (12.9)	73 (13.6)	0.031
Deprivation	1	9%	12%	8%	11.6%	
quintile	2	23%	18%	22%	19.1%	
	3	29%	29%	32%	23.6%	0.72
	4	24%	22%	24%	23.3%	
	5	15%	20%	14%	22.3%	
Smokers (%)		26%	28%	29%	25%	0.68
Diabetes (%)		13%	16%	13%	13%	0.47
Hypertension (%)		46%	49%	53%	46%	0.40
Charlson index, mean (SD)		2 (1.7)	2.4 (1.7)	1.9 (1.6)	2.1 (1.7)	0.014

Patient characteristics for myocardial infarction, ischaemic heart disease and cerebrovascular disease, identified using general practice (GP), hospital, paired GP/hospital, and combined GP and/or hospital data. SD, standard deviation. Deprivation quintile 1 is least deprived. Significant differences are calculated by Chi-squared test or Kruskall-Wallace ANOVA.

TABLE 2 – DRUG THERAPY

	GP	Hospital	Paired GP/ hospital	GP and/or hospital	p value
Myocardial infarction					
N	139	137	99	173	
ACE inhibitor / ARB	68%	77%	77%	71%	0.30
Beta-blocker	68%	61%	59%	61%	0.50
Calcium channel blocker	10%	10%	8%	15%	0.29
Diuretic	32%	32%	28%	29%	0.87
Nitrate	46%	61%	59%	55%	0.065
Statin	79%	81%	77%	76%	0.70
Antiplatelet agent	84%	82%	85%	78%	0.43
Ischaemic heart disease					
N	353	484	262	541	
ACE inhibitor / ARB	48%	48%	58%	45%	0.013
Beta-blocker	57%	54%	62%	49%	0.005
Calcium channel blocker	21%	21%	25%	19%	0.28
Diuretic	35%	30%	34%	33%	0.57
Nitrate	40%	43%	60%	40%	<0.001
Statin	67%	67%	82%	63%	<0.001
Antiplatelet agent	71%	71%	87%	66%	<0.001
Cerebrovascular disease					
N	285	278	145	381	
ACE inhibitor / ARB	38%	33%	31%	36%	0.42
Beta-blocker	25%	19%	22%	19%	0.16
Calcium channel blocker	20%	15%	13%	17%	0.27
Diuretic	32%	33%	32%	33%	0.99
Nitrate	15%	14%	15%	13%	0.94
Statin	56%	41%	53%	50%	0.006
Antiplatelet agent	54%	44%	50%	55%	0.022

30-day prescribing rates for myocardial infarction (MI), ischaemic heart disease (IHD) and cerebrovascular disease (CVD), identified using general practice (GP), hospital, paired GP/hospital, and combined GP and/or hospital data. ACE, angiotensin converting enzyme. ARB, angiotensin receptor blocker. Patients are those alive at 30 days. Significant differences are calculated by Chi-squared test.

TABLE 3 – MORTALITY

	GP	Hospital	Paired GP/ hospital	GP and/or hospital	p value
Myocardial infarction					
N	145	171	105	209	
30-day mortality rate (%)	4%	20%	6%	17%	0.001
Ischaemic heart disease					
N	362	529	270	585	
30-day mortality rate (%)	2%	9%	3%	8%	0.002
Cerebrovascular disease					
N	302	330	153	424	
30-day mortality rate (%)	6%	16%	5%	10%	0.001

30-day mortality rates for myocardial infarction, ischaemic heart disease and cerebrovascular disease, identified using general practice (GP), hospital, paired GP/hospital, and combined GP and/or hospital data. The significance of the differences between coding methods is adjusted for confounding factors using logistic regression (see text for details).

ACKNOWLEDGEMENTS

None

COMPETING INTERESTS

None

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

CONTRIBUTORSHIP

RP conceived the study. RP and GA contributed to the study design, analysis and interpretation, and to the drafting of the article. CRS acquired the data and set up the linked database. All authors contributed to the critical revision of the paper and approval of the final version.

DATA SHARING STATEMENT

There is no additional data available

REFERENCES

1. Scarborough P, Bhatnagar P, Wickramasinghe K et al. *Coronary heart disease statistics. 2010 edition*. British Heart Foundation, London 2010.

2. Erkinjuntti T, Ostbye T, Steenhuis R, Hachinski V. The effect of different diagnostic criteria on the prevalence of dementia. *N Engl J Med* 1997;337:1667–1674.

3. Rollason W, Khunti K, de Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: implications for the quality of care. *Inform Prim Care* 2009;17:113-119.

4. Boyle CA, Dobson AJ. The accuracy of hospital records and death certificates for acute myocardial infarction. *Aust N Z J Med* 199;25:316-323.

5. Cheng P, Gilchrist A, Robinson KM, Paul L. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *HIM J* 2009;38:35-46.

6. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. *Fam Pract* 2004;21:396-412.

7. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010;60:e128-e136.

8. *NHS hospital data quality – towards better data from Scottish hospitals. An assessment of SMR01 and associated data 2004-2006.* ISD Scotland, NHS National Services Scotland, Edinburgh. 2007.

9. *Scottish Index of Multiple Deprivation 2009 Technical Report.* Office of the Chief Statistician, Scottish Government, Edinburgh. September 2010.

10. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chron Dis* 1987;40:373-383.

11. Gibbs RG, Newson R, Lawrenson R, et al. Diagnosis and initial management of stroke and transient ischemic attack across UK health regions from 1992 to 1996: experience of a national primary care database. *Stroke* 2001;32:1085-1090.

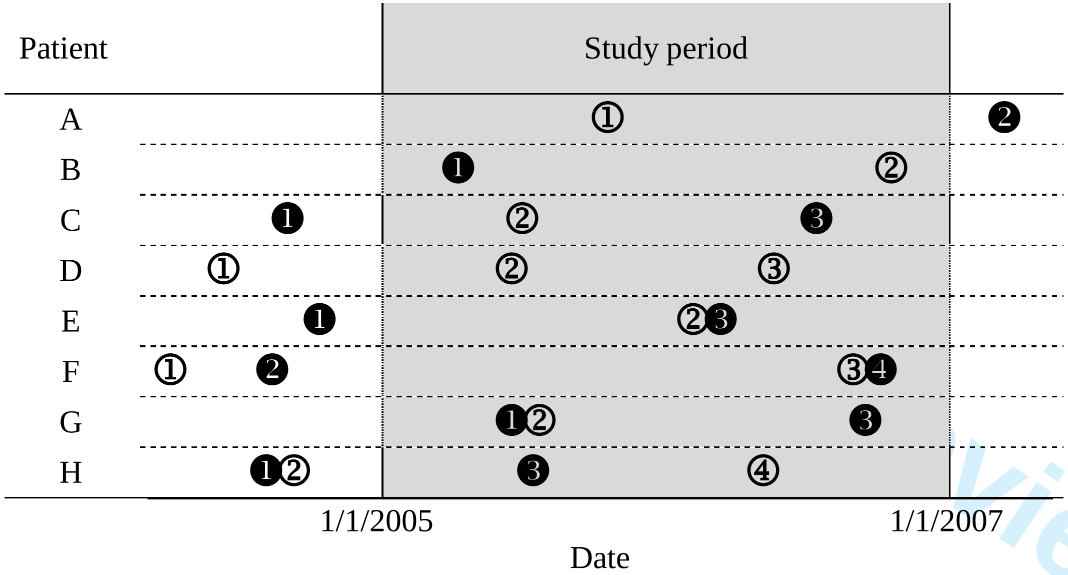
12. Jagadesham VP, Aparajita R, Gough MJ. Can the UK guidelines for stroke be effective? Attitudes to the symptoms of a transient ischaemic attack among the general public and doctors. *Clin Med* 2008;8:366-370.

13. Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. *Am Heart J* 2002;144:290-296.

14. Jishi F, Hudson PR, Williams CP, et al. Troponin I, laboratory issues, and clinical outcomes in a district general hospital: crossover study with "traditional" markers of myocardial infarction in a total of 1990 patients. *J Clin Pathol* 2004;57:1027-32.

15. The General Practice Research Database. http://www.gprd.com/academia/primarycare.asp. Accessed 29/6/11

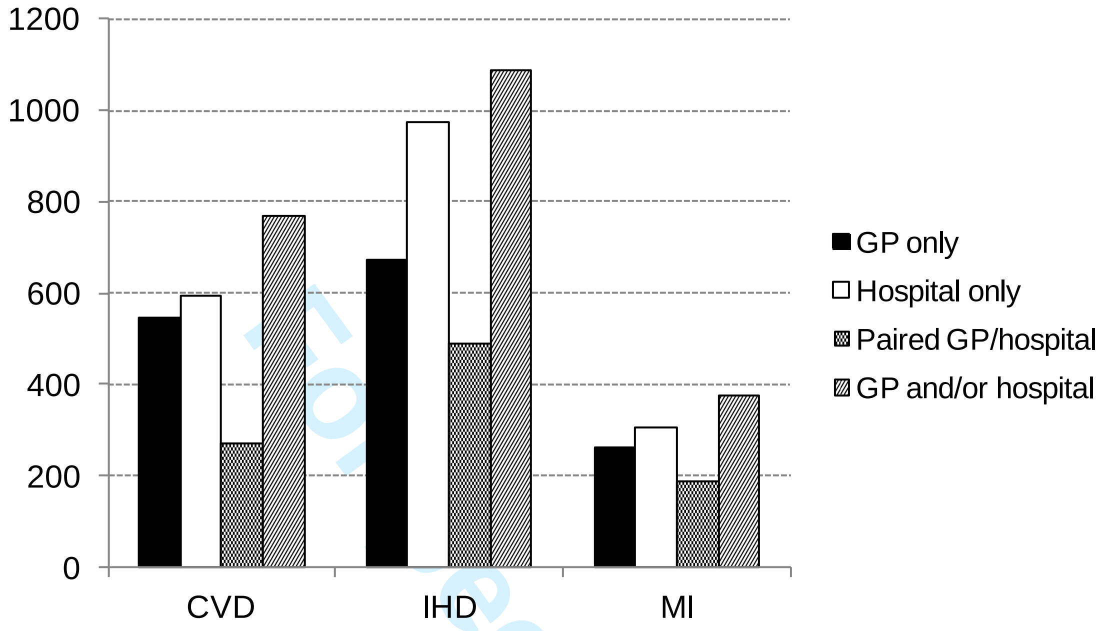
FIGURE 1. IDENTIFICATION OF INCIDENT EVENTS



Patient	GP ○	Hospital ●	Paired ●○	GP ● and/or hospital ○
A	①			①
B	②	❶		❶
C	②			②
D				
E	②		②	②
F			③	
G	②	❶	❶	❶
H				

The figure shows how incident events can be identified from linked GP and hospital datasets, for eight hypothetical patients, illustrating some of the potential coding combinations. Circles correspond to the presence of a GP (○) or hospital (●) clinical code, with numbers illustrating the order. Immediately adjacent circles represent codes occurring within 30 days of one another. It can be seen that, for any given patient, it is possible to classify them as having an incident event in up to four ways: GP data only, hospital data only, paired GP/hospital, and GP and/or hospital; the code which identifies an incident event for each of these methods is shown on the right of the figure. Codes do not count as incident events if a further, similarly classified, event has occurred prior to the start of the study period. In our study, patients were randomly allocated to one of the four coding methods. For instance, if patient E was allocated to “hospital only” coding, they would not be classified as having had an event; in contrast, they would be classified as having had an event if they were allocated to any of the other three coding methods.

FIGURE 2 – INCIDENCE RATES



Incidence rates, expressed per 100,000 patient years, for different clinical conditions over a 2-year time period beginning 1/1/2005, based on general practice (GP), hospital, paired GP/hospital, and combined GP and/or hospital data. CVD, cerebrovascular disease; IHD, ischaemic heart disease; MI, myocardial infarction.

A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000723.R1
Article Type:	Research
Date Submitted by the Author:	02-Feb-2012
Complete List of Authors:	Payne, Rupert; University of Cambridge, GP and Primary Care Research Unit Abel, Gary; University of Cambridge, GP and Primary Care Research Unit Simpson, Colin; The University of Edinburgh, Centre for Population Health Sciences
<b>Primary Subject Heading</b>:	Epidemiology
Secondary Subject Heading:	Cardiovascular medicine, Health informatics, Research methods
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Coronary heart disease < CARDIOLOGY, EPIDEMIOLOGY, PRIMARY CARE, STROKE MEDICINE, STATISTICS & RESEARCH METHODS
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
STROBE_checklist_BMJ-Open_cohort-studies.doc	

A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data

Rupert A Payne[1], Gary A Abel[1], Colin R Simpson[2]

1. GP and Primary Care Research Unit, University of Cambridge, Cambridge, UK
2. eHealth Research Group, Centre for Population Health Sciences, The University of Edinburgh, Edinburgh, UK

Corresponding author:
Dr Rupert A Payne
GP and Primary Care Research Unit
University of Cambridge
Institute of Public Health
Forvie Site
Robinson Way
Cambridge
CB2 0SR

Tel. 01223 746545
Email. rap55@medschl.cam.ac.uk

Key words: Data linkage, primary care, secondary care, routine data, cardiovascular, coding

Word count: 3861

ABSTRACT

Objectives:

Data linkage combines information from several clinical datasets. We examined whether coding inconsistencies for cardiovascular disease between components of linked datasets, result in differences in apparent population characteristics.

Design:

Retrospective cohort study.

Setting:

Routine primary care data from 40 Scottish General Practice (GP) surgeries linked to national hospital records.

Participants:

240,846 patients, age 20 years or more, registered at a GP surgery.

Outcomes:

Cases of myocardial infarction, ischaemic heart disease and stroke (cerebrovascular disease) were identified from GP and hospital records. 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 pooled GP/hospital records (diagnosis recorded in either or both datasets).

Results:

For all 3 outcomes, we found evidence ($p<0.05$) of different characteristics when using different methods of case identification. Prescribing of cardiovascular medicines for ischaemic heart disease was greatest for cases identified using paired records ($p\leq0.013$). For all conditions, 30-day case fatality rates were higher for cases identified using hospital compared with GP or paired data, most noticeably for myocardial infarction (hospital 20%, GP 4%, $p=0.001$). Incidence rates were highest using pooled GP/hospital data, and lowest using paired data.

Conclusion:

Differences exist in patient characteristics and disease incidence for cardiovascular conditions, depending on the data source. This has implications for studies using routine clinical data.

ARTICLE SUMMARY

Article focus

- Data linkage allows information to be combined from different routine clinical data sources
- Previous work has shown differences between sources of data, but has not examined this at the patient-level

Key messages

- Patients' apparent characteristics, and disease incidence and severity, vary depending on whether primary care, hospital or combined definitions of cardiovascular events are used
- Use of isolated routine primary care or hospital data may result in biased patient selection
- This has implications in the public health arena, clinical trial patient recruitment, and validity and reliability of secondary data in clinical trials

Strengths and limitations

- The strengths of this study are the novel analytical approach, using a large routine dataset linked at individual patient level from multiple GP surgeries
- 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, potential diagnostic coding inaccuracies, and the relatively small number of GP surgeries which may not have been representative.

BACKGROUND

Primary care datasets are commonly used for assessment of cardiovascular outcomes. Such events often are associated with hospitalisation[1]. However, it is possible that the manner in which outcomes are coded and recorded in electronic health records may differ between primary and secondary care. This may result not only in differences in the apparent incidence of a condition, depending on whether primary or secondary care records are used, but also differences in the observed characteristics of patients. Studies have observed that variations in diagnostic criteria can affect estimates of disease prevalence[2], and the complexities of clinical coding systems for electronic health care records can lead to inconsistent data recording[3]. This will lead to uncertainties with respect to disease prevalence and mortality[4], impact on clinical care, have additional health service implications such as affecting funding[5], and potentially influence identification of patients for clinical trials. Previous studies have compared general practice coding and disease prevalence with other unlinked data sources including paper notes[6,7]. However, the effect of combining information from two sources has not been previously examined. This study used linked individual patient electronic health records collected from primary and secondary care to examine the effect of using data from different parts of the health care service on the incidence rates, case fatality rates and patient characteristics of myocardial infarction (MI), ischaemic heart disease (IHD) and cerebrovascular disease (CVD).

METHODS

Data sources

Sixty general practice (GP) surgeries take part in the Scottish national Practice Team Information (PTI) project, of which 40 self-selected surgeries contributed to the dataset used in this study. Practices involved in the PTI project provide routine, central recording of clinical activity and morbidity from a sample of GP surgeries considered reasonably representative of the Scottish population. Practices are reimbursed to ensure data recording is optimal. Clinical coding used the Read code system. Data are used to calculate national estimates, and used by various organisations (e.g. NHS Boards, Scottish Government) to inform policies and better understand health in Scotland.

Patient details from the PTI dataset were linked to the corresponding admissions recorded in Scottish national hospital data (the Scottish Morbidity Record, SMR-01) using probabilistic matching. Matching was based on Soundex-encoded name, date of birth, sex, postcode and a unique nationwide identifier, the community health index (CHI). Experienced human review was used to set a threshold for linkage. A substantial proportion of patients in this GP cohort have no hospital admissions, and as such it is difficult to know whether the absence of a match is either due to a genuine lack of corresponding hospital record, or a false negative error. Match rates are thus difficult to quantify, although the use of multiple identifiers should improve linkage quality. The linkage was carried out by the Information Services Division, NHS National Services Scotland. The work was approved by the Privacy Advisory Committee of NHS National Services Scotland. For the 2004 to 2006 period, SMR-01 data are considered to be 88% accurate[8]. SMR-01 records are generated for all in-patient hospital medical discharges and transfers. Coding is based on the International Statistical Classification of Diseases and Related Health Problems (ICD) system (ICD9 prior to 2000, ICD10 thereafter), with up to six in-patient diagnoses per record. Accident and emergency, maternity and psychiatric admissions, along with out-patient attendances, are not recorded in SMR-01. SMR-01 itself is also routinely linked to national mortality data (General Registrar's Office for Scotland, GROS). SMR-GROS data are also used to generate Scottish national statistics.

Identification and classification of cases

We first identified all records of MI, IHD and CVD from both GP and hospital datasets using the following Read codes (MI: G30%/35%/38%, Gyu34/35/36; IHD: G3%, Gyu3%; CVD: G6%, Gyu6%, F4236; where % indicates a "wildcard" match) and ICD codes (MI: ICD10 I21-22; ICD9 410. IHD including MI: ICD10 I20-25; ICD9 410-414. CVD (stroke) including haemorrhage and TIA: ICD10 I60-69, G45-46; ICD9 430-438). Hospital events were identified from any of the six diagnostic positions. These were not necessarily first events.

We then found all episodes of a similar GP and hospital event type occurring within a 30-day period, and made the assumption that these pairings represented the same clinical event. Where the GP and hospital dates differed for these paired episodes, the first of the two dates was taken. The choice of 30 days was a pragmatic one, but supported by visual evaluation of the distribution of time gaps between similar hospital and GP event types over a two-year period. Of note, an event recorded by the GP does not necessarily require a face-to-face consultation or a referral to be made; hospital admissions will usually be retrospectively recorded by the GP, using the admission date as opposed to the data-entry date.

Analysis was carried out over the period 1/1/2005 to 1/1/2007. The total population was randomly allocated to one of four methods of identifying cardiovascular events: those based on GP events only; those based on hospital events only; those based on pooled GP/hospital events, with an event in GP data only, hospital data only, or both the GP and hospital data (although not necessarily occurring within 30 days); and those based on paired GP/hospital events (those recorded in both GP and hospital data within 30 days). An episode was included as an incident event only if there was no record of a similar clinical event at any time prior to 1/1/2005 coded in the same dataset(s).

This method of identifying incident events is shown graphically in Figure 1. For example, for an event to be included using only GP data, the first event would have to be recorded by the GP during the 2-year period of interest, with no similar events recorded by the GP prior to 1/1/2005; hospital data is completely ignored in this case. A similar approach is used for identifying events using hospital-only data, with GP records ignored in this situation. For the third method, identifying events using pooled GP/hospital data, the first event needs to be recorded by either the hospital or the GP during the 2-year study period; there must be no similar event recorded in either dataset prior to 1/1/2005. For the final method, the first occurrence of paired (i.e. within 30 days) records in both GP and hospital datasets constituted an incident event, if it occurred during the 2-year period; any unpaired GP or hospital records occurring prior to 1/1/2005 were ignored.

{INSERT FIGURE 1}

For each incident event, we determined the patient's age, sex, socioeconomic status (Scottish Index of Multiple Deprivation quintile, SIMD)[9], recorded current smoking status, record of hypertension, record of diabetes, and Charlson index[10]. Co-morbidities, including Charlson index, were determined from the GP data as the presence of any relevant diagnostic Read code prior to the incident episode date; the list of codes used is available from the authors on request. Although we have not formally evaluated performance of our Charlson Index Read code list, we match 87% of those events identified by the method described by Khan et al[11], and as such believe this represents a reasonable, albeit pragmatic, measure of co-morbidity. Death from any cause within 30 days of the event was ascertained from linked national mortality (GROS) data. Drug therapy recorded in the GP record, starting prior to or within 30 days after the event, and continuing for any period of time after the event, was ascertained for patients alive at 30 days. Drug classes included were angiotensin converting enzyme (ACE) inhibitors (including angiotensin receptor blockers), beta-blockers, calcium channel blockers, diuretics (including potassium sparing and combination diuretics), nitrates, statins and antiplatelet agents (aspirin or clopidogrel for MI or IHD; aspirin or dipyridamole for CVD).

Statistical analysis

Incidence rates were calculated excluding patients with events in the relevant dataset(s) prior to 1/1/2005. Incidence rates are expressed per 100,000 patient years (based on total number of days of follow-up for each patient within each respective group). Statistical differences in patient characteristics (including drug treatment) between coding categories were evaluated using Chi-squared tests (for proportions) and Kruskall-Wallace non-parametric ANOVA (for continuous data). The association between coding and 30-day case fatality was assessed by logistic regression, including the covariates age, sex, deprivation, smoking status, hypertension, diabetes and Charlson index. Differences in the four incident rates obtained were examined using Poisson regression.

Data management was carried out using Microsoft SQL Server 2000. Statistical analysis was performed using SPSS v17 (SPSS inc, Chicago, Illinois, USA).

RESULTS

Differences in identification of incidence events

There were a total of 240,846 patients, evenly distributed between the four coding groups. Numbers of incident events are shown in Table 1. Incidence rates for the three conditions are shown in Figure 2. There was strong evidence ($p<0.001$, Poisson regression) that the incidence rates for all three clinical conditions depends on which dataset(s) are used to identify cases. In all cases, the pooled GP/hospital data produced the highest incidence rates (376, 1089 and 767 per 100,000 patient years for MI, IHD and CVD respectively), and the paired GP/hospital data gave the lowest incidence rates (188, 489 and 272 per 100,000 patient years respectively). There was no evidence that the incidence rates based on only GP data differ from those of the hospital data for either MI ($p=0.14$) or CVD ($p=0.27$), but there was strong evidence that they were higher for IHD (975 and 673 events per 100,000 patient years for hospital and GP respectively, $p<0.001$). The pooled GP/hospital data produced slightly higher incidence rates than hospital data alone for CVD ($p<0.001$) and marginally so for MI ($p=0.048$) and IHD ($p=0.066$).

{INSERT TABLE 1}

{INSERT FIGURE 2}

Patient characteristics

Patient characteristics are shown in Table 1 for all three clinical conditions. There was no evidence that rates of diabetes and hypertension, or the distribution of sex or deprivation, varied between coding groups. Greater numbers of smokers were found in the paired GP/hospital group for patients with MI (45% in the paired group compared with 28 to 34% in the other groups, $p=0.028$) and IHD (35% compared with 24 to 27%, $p=0.021$). The level of co-morbidity for all conditions, as measured by the Charlson index, is lower in the paired GP/hospital group (1.8, 1.3 and 1.9 for MI, IHD and CVD respectively) and higher in the hospital group (2.2, 1.7 and 2.4 respectively, p≤0.014). For IHD and CVD, there is evidence that patients identified using solely GP or solely hospital data were slightly younger.

Prescribing

Differences in prescribing rates were observed between coding groups (Table 2). These were most marked for IHD, where rates of prescribing of ACE inhibitors, beta-blockers, nitrates, statins and antiplatelet agents were higher in the paired group (p≤0.013). However, this finding did not appear to be replicated for MI specifically. For CVD, prescribing rates for statins and antiplatelet agents were lower in the hospital group (p≤0.022).

{INSERT TABLE 2}

Case fatality

Considerable 30-day case fatality rate differences exist for all three conditions depending on the coding used (p≤0.002, Table 3). Rates for all conditions are highest in patients coded only in hospital, and lower in the GP and paired GP/hospital groups. The most striking differences were observed for MI, with a 30-day case fatality rate of 20% for the hospital group but only 4% for the GP group.

{INSERT TABLE 3}

DISCUSSION

In a world where electronic healthcare data are becoming increasingly used for the purposes of clinical trials and epidemiological research, there is a need for researchers to understand whether additional information can be gained by linking two (or indeed more) electronic health record data sources together. However, where there is overlap between the constituent datasets, such as with coding of clinical conditions, the researcher needs to decide which dataset to rely on for identifying cases, or indeed whether combining information from both the two datasets may be of value. Our study demonstrates that the method of coding MI, IHD and CVD appears to result in identification of different types of patient, in particular as characterised by prescribing and case fatality rates. Incident rates of disease also vary depending on the coding method used.

Previous work examining the epidemiology of cardiovascular disease has been conducted in Scotland using routine clinical data. Primary care data has been used to demonstrate that IHD is a common problem associated with male gender, increasing age and socioeconomic deprivation[12]. Yet the recording of IHD data varies in general practice, with different methods used for case detection[13]. Furthermore, external factors such as payment-for-performance have been shown to improve the recording of IHD-related health indicators[14]. Such incentivisation was introduced to UK general practice (but not hospital practice) in 2004, and so it is possible that this may have reduced the discrepancies between hospital and GP data in our study. Interestingly, pooling of GP and SMR records have previously been advocated for detecting MI cases[15], and pooled GP/SMR data from the same dataset we used has demonstrated differences between cohorts of incident and prevalent MI[16]. However, the effect of using only one component of such a dataset has been hitherto unknown.

Reasons for differences in incidence rates and patient characteristics

Our data do not allow us to determine the exact cause of our findings, but a number of hypotheses may be proposed. Incident disease is reassuringly similar between GP and hospital groups for MI and CVD. The lower incidence of IHD for the GP group reflects the fact that many patients will have had relatively stable coronary disease for a number of years but not necessarily required acute hospital admission. Thus, many GP episodes of IHD do not count as true incident cases as they have had prior contact with the GP, whereas a higher number of hospital episodes are incident cases as these patients have never been previously admitted. The lower incidence rates for the paired GP/hospital group, and higher incidence rates for the pooled GP/hospital group, are inevitable consequences of the way in which the two datasets are united, although the magnitude of these differences will nonetheless reflect the degree of inconsistency in coding between the two. Furthermore, it would appear that because the paired GP/hospital data considerably underestimates the true disease incidence, it is probably not a useful method for identifying cases, even though such cases might be more rigorously identified. In addition, the increase in incidence rate using the pooled GP/hospital data, demonstrates the potential advantage of combining two datasets, over use of a single dataset, from the perspective of improving case finding.

The discrepancies in death rates are probably relatively straightforward to explain. Acute myocardial infarction admission has a high case fatality[1], but those surviving beyond discharge have a much lower case fatality subsequently. It seems likely that the GP may fail to record the cause of death in patients who do not survive the hospital admission, thus resulting in the lower case fatality rates observed in the paired GP/hospital coding group. Furthermore, it is possible that patients coded only by the GP may represent “less serious” illness, where hospitalisation is not deemed necessary by the GP. It is recognised that many patients suffering relatively minor strokes may not be admitted to hospital[17], resulting in lower case fatality for CVD in the GP group, although with the growing availability of active treatment options for ischaemic stroke in the form of thrombolysis, this may well change. We used national mortality data to identify deaths from both GP and SMR datasets, so discrepancies in recording of death between GP and hospital are unlikely to explain the differences in case fatality rates observed. Furthermore, the majority of paired events share exactly the same date, suggesting that retrospective date entry by the GP of the hospital event is common, and thus there is no reason why this could not be carried out for fatal events.

The higher prescribing rates for IHD in the paired coding group are probably due to GPs responding appropriately to secondary care instigated intervention, reflected in appropriate treatment. That such differences were not observed for MI, may be due to better communication and awareness for this specific condition compared with other IHD such as angina, meaning that prescribing in the hospital group appears just as good as for the paired GP/hospital group. However, fewer MI events may have left us underpowered to detect differences. The lack of difference in the GP and paired groups for CVD may reflect poorer awareness of stroke management guidelines[18] in comparison with coronary heart disease, and so prescribing rates are consequently no higher in the paired group. The lower prescribing rates of statins and antiplatelet agents in the CVD hospital group may reflect the GP being unaware of these patients' clinical need resulting in under-treatment; this is supported by the higher prescribing rates in the paired group. The differences in other patient characteristics – specifically smoking and co-morbidity – are less easy to understand, but may represent increased disease severity and mortality in hospitalised smokers and multi-morbid patients. The small differences in age (<3 years) seem unlikely to be clinically relevant, although may be pertinent from the public health perspective. Finally, it may be that miscoding of diagnoses may explain some of the above differences; for instance, heart failure may be used as an alternative but incorrect code for myocardial infarction[19]. Furthermore, the introduction of sensitive troponin assays has influenced myocardial infarction detection rates[20]; it is possible that lack of familiarity amongst some clinicians for the resulting terms (e.g. non-ST elevation MI, acute coronary syndrome) may result in inaccurate diagnoses being recorded.

Limitations

This study has highlighted important issues related to patient coding and linked data, but although it has the advantage of using a reasonably large routine dataset, linked at the individual patient level, a number of issues and limitations should be considered. The relatively small number of GP surgeries (40) may not have been fully representative. In addition, the number of events is relatively small, and given the conservative nature of the Chi-squared test, this increases the possibility of Type 2 errors; thus, a larger dataset may have identified more differences between groups. We restricted our analysis to four simple coding groups – GP, hospital, paired and pooled GP/hospital. However, it is clear that there are many further ways of categorising events, including the presence or absence of prior or subsequent coding based on the alternative half of the dataset. For instance, an incident GP event with a historical hospital event, may be coded differently to a GP event with no previous hospital record. However, we found that many of these theoretical categories have only a handful of cases. Furthermore, even when we examined six or seven separate smaller coding categories, similar differences in patient characteristics persisted between groups (data not shown). Our choice of four main groups was therefore a pragmatic one which reflects the choice that would face a researcher dealing with a similar linked dataset. The decision to use a 30-day limit for pairing data could also be questioned; we are unable to prove that these two events are truly the same clinical episode. The choice was again, therefore, partly pragmatic, although supported by examination of the distribution of time gaps between the GP and hospital data. We did not limit the lead-in time period prior to 1/1/2005 in any way. Length of GP records is generally greater and more variable than SMR records, and there is the potential to see a lower number of new incident events amongst persons with longer GP records. Our study used routine GP data, and it is possible that such profound differences may not be found with research-standard databases such as GPRD[21]. Nonetheless, work linking primary care research databases to hospital (and other) records is ongoing, and the issues raised by our study must be acknowledged. The SMR dataset only records hospital events in Scotland, and thus fails to capture events in elsewhere in the UK or abroad. Similar issues face the English equivalent Hospital Episode Statistics (HES), and a UK-wide hospital events dataset would be valuable. SMR (and HES) also provide multiple diagnostic codes for a single event. We elected to use all six diagnostic positions to ensure maximum capture of relevant hospital events. However, the robustness of low-priority diagnoses might be questioned. Nonetheless, we found similar results when we used only two diagnostic positions (data not shown). We also did not examine miscoding of events – e.g. a code of angina being used rather than the code for MI. Coding of SMR is considered 99% complete and 88% accurate[8]; corresponding metrics are not available for PTI data (although the completeness and accuracy of Read coding of morbidity in Scottish general practice has been shown previously to be greater than 91%[22]). Furthermore, the two datasets use different coding systems, so completely reliable comparison is not possible. However, we used relatively broad definitions, and the Read code system is based on ICD. Nonetheless, we may in particular have missed some administrative Read codes which might have enabled identification of additional cases in the GP group. Of course, ideally further validation of the coding should be conducted; linkage to laboratory data might be one way of achieving this. Finally, our 30-day limit for prescribing was selected from a pragmatic perspective. However, it is possible that patients who were admitted for over 30 days would not have had a new prescription issued by the GP within the 30-day post event period, resulting in an apparent underestimation of prescribing. We believe these numbers will be relatively small, however, and unlikely to alter the overall interpretation of our findings.

Research and policy implications

These results have significant implications for linked data; the drug management, disease severity, and to some degree the patient characteristics, vary depending on how the disease cohort is defined. They also have implications for the use of unlinked routine data – use of isolated primary or secondary care data may result in a biased selection of patients. This may affect patient recruitment as well as the validity and reliability of such information sources as secondary data in clinical trials, including clinical outcomes. It is similarly relevant to the public health environment. Using linked data allows one to have a more robust definition, by using pairs of GP and hospital codes only, but it is clear that the apparent incidence of a disease will be considerably lower. Alternatively, linked data enable a looser but more inclusive disease definition, using both GP and hospital data, but not relying on the coding occurring simultaneously. When using separate data from only one source, one needs to take into account that patient characteristics may not be representative of the wider population. It is difficult to recommend one coding approach over another, however, and the decision will need to be based on the specific question being posed.

Conclusions

In conclusion, patient characteristics vary depending on whether GP, hospital or combined definitions of cardiovascular events are used. In particular, disease severity as measured by mortality varies considerably. This has important implications for studies using linked routine primary and secondary care data, and for studies where information is only available from one of these sources. These issues should be acknowledged by studies using routine data as a secondary data source, and further work is merited to examine whether similar discrepancies exist for other clinical conditions or within primary care research databases.

TABLE 1 – PATIENT CHARACTERISTICS

		GP	Hospital	Paired GP/ hospital	Pooled GP/ hospital	p value
Myocardial infarction						
N		145	171	105	209	
Males (%)		65%	59%	60%	64%	0.68
Age, mean (SD)		68 (13.8)	67 (13)	68.4 (13.8)	68.8 (14.9)	0.51
Deprivation	1	19%	11%	10%	12%	
quintile	2	15%	25%	26%	17%	
	3	26%	17%	29%	31%	0.55
	4	15%	23%	21%	22%	
	5	24%	24%	14%	17%	
Smokers (%)		33%	34%	45%	28%	0.028
Diabetes (%)		15%	12%	8%	11%	0.29
Hypertension (%)		39%	44%	38%	44%	0.52
Charlson index, mean (SD)		2.5 (1.7)	2.2 (1.6)	1.8 (1.4)	2.0 (1.6)	<0.001
Ischaemic heart disease						
N		362	529	270	585	
Males (%)		56%	55%	61%	56%	0.38
Age, mean (SD)		66.2 (12.7)	65.8 (11.6)	66.9 (13.4)	68.4 (12.8)	0.007
Deprivation	1	17%	13%	11%	13%	
quintile	2	18%	20%	20%	21%	
	3	29%	23%	27%	26%	0.25
	4	17%	22%	24%	20%	
	5	20%	23%	19%	19%	
Smokers (%)		27%	27%	35%	24%	0.011
Diabetes (%)		11%	15%	13%	10%	0.091
Hypertension (%)		42%	47%	44%	45%	0.51
Charlson index, mean (SD)		1.5 (1.6)	1.7 (1.6)	1.3 (1.3)	1.5 (1.5)	0.002
Cerebrovascular disease						
N		302	330	153	424	
Males (%)		48%	47%	46%	47%	0.97
Age, mean (SD)		70.3 (14.1)	70.8 (13.6)	72 (12.9)	73 (13.6)	0.031
Deprivation	1	9%	12%	8%	11.6%	
quintile	2	23%	18%	22%	19.1%	
	3	29%	29%	32%	23.6%	0.72
	4	24%	22%	24%	23.3%	
	5	15%	20%	14%	22.3%	
Smokers (%)		26%	28%	29%	25%	0.68
Diabetes (%)		13%	16%	13%	13%	0.47
Hypertension (%)		46%	49%	53%	46%	0.40
Charlson index, mean (SD)		2 (1.7)	2.4 (1.7)	1.9 (1.6)	2.1 (1.7)	0.014

Patient characteristics for myocardial infarction, ischaemic heart disease and cerebrovascular disease, identified using general practice (GP), hospital, paired GP/hospital, and pooled GP/hospital data. SD, standard deviation. Deprivation quintile 1 is least deprived. Significant differences are calculated by Chi-squared test or Kruskall-Wallace ANOVA.

TABLE 2 – DRUG THERAPY

	GP	Hospital	Paired GP/ hospital	Pooled GP/ hospital	p value
Myocardial infarction					
N	139	137	99	173	
ACE inhibitor / ARB	68%	77%	77%	71%	0.30
Beta-blocker	68%	61%	59%	61%	0.50
Calcium channel blocker	10%	10%	8%	15%	0.29
Diuretic	32%	32%	28%	29%	0.87
Nitrate	46%	61%	59%	55%	0.065
Statin	79%	81%	77%	76%	0.70
Antiplatelet agent	84%	82%	85%	78%	0.43
Ischaemic heart disease					
N	353	484	262	541	
ACE inhibitor / ARB	48%	48%	58%	45%	0.013
Beta-blocker	57%	54%	62%	49%	0.005
Calcium channel blocker	21%	21%	25%	19%	0.28
Diuretic	35%	30%	34%	33%	0.57
Nitrate	40%	43%	60%	40%	<0.001
Statin	67%	67%	82%	63%	<0.001
Antiplatelet agent	71%	71%	87%	66%	<0.001
Cerebrovascular disease					
N	285	278	145	381	
ACE inhibitor / ARB	38%	33%	31%	36%	0.42
Beta-blocker	25%	19%	22%	19%	0.16
Calcium channel blocker	20%	15%	13%	17%	0.27
Diuretic	32%	33%	32%	33%	0.99
Nitrate	15%	14%	15%	13%	0.94
Statin	56%	41%	53%	50%	0.006
Antiplatelet agent	54%	44%	50%	55%	0.022

30-day prescribing rates for myocardial infarction (MI), ischaemic heart disease (IHD) and cerebrovascular disease (CVD), identified using general practice (GP), hospital, paired GP/hospital, and pooled GP/hospital data. ACE, angiotensin converting enzyme. ARB, angiotensin receptor blocker. Patients are those alive at 30 days, and this is reflected in lower numbers of patients than in Tables 1 and 3. Significant differences are calculated by Chi-squared test.

TABLE 3 – CASE FATALITY

	GP	Hospital	Paired GP/ hospital	Pooled GP/ hospital	p value
Myocardial infarction					
N	145	171	105	209	
30-day case fatality rate (%)	4%	20%	6%	17%	0.001
Ischaemic heart disease					
N	362	529	270	585	
30-day case fatality rate (%)	2%	9%	3%	8%	0.002
Cerebrovascular disease					
N	302	330	153	424	
30-day case fatality rate (%)	6%	16%	5%	10%	0.001

30-day case fatality rates for myocardial infarction, ischaemic heart disease and cerebrovascular disease, identified using general practice (GP), hospital, paired GP/hospital, and pooled GP/hospital data. The significance of the differences between coding methods is adjusted for confounding factors using logistic regression (see text for details).

ACKNOWLEDGEMENTS

None

COMPETING INTERESTS

None

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

CONTRIBUTORSHIP

RP conceived the study. RP and GA contributed to the study design, analysis and interpretation, and to the drafting of the article. CRS acquired the data and set up the linked database. All authors contributed to the critical revision of the paper and approval of the final version.

DATA SHARING STATEMENT

There is no additional data available

REFERENCES

1. Scarborough P, Bhatnagar P, Wickramasinghe K et al. *Coronary heart disease statistics. 2010 edition*. British Heart Foundation, London 2010.

2. Erkinjuntti T, Ostbye T, Steenhuis R, Hachinski V. The effect of different diagnostic criteria on the prevalence of dementia. *N Engl J Med* 1997;337:1667–1674.

3. Rollason W, Khunti K, de Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: implications for the quality of care. *Inform Prim Care* 2009;17:113-119.

4. Boyle CA, Dobson AJ. The accuracy of hospital records and death certificates for acute myocardial infarction. *Aust N Z J Med* 199;25:316-323.

5. Cheng P, Gilchrist A, Robinson KM, Paul L. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *HIM J* 2009;38:35-46.

6. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. *Fam Pract* 2004;21:396-412.

7. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010;60:e128-e136.

8. *NHS hospital data quality – towards better data from Scottish hospitals. An assessment of SMR01 and associated data 2004-2006.* ISD Scotland, NHS National Services Scotland, Edinburgh. 2007.

9. *Scottish Index of Multiple Deprivation 2009 Technical Report.* Office of the Chief Statistician, Scottish Government, Edinburgh. September 2010.

10. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chron Dis* 1987;40:373-383.

11. Khan NF, Perera R, Harper S, Rose PW. Adaptation and validation of the Charlson Index for Read/OXMIS coded databases. BMC Fam Pract 2010;11:1.

12. Murphy NF, Simpson CR, MacIntyre K, 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.

13. Moher M, Yudkin P, Turner R, et al. An assessment of morbidity registers for coronary heart disease in primary care. ASSIST (ASSessment of Implementation STrategy) trial collaborative group. *Br J Gen Pract* 2000;50:706-709.

14. McGovern MP, Boroujerdi MA, Taylor MW, et al. The effect of the UK incentive-based contract on the management of patients with coronary heart disease in primary care. *Fam Pract* 2008;25:33-39.

15. Donnan PT, Dougall HT, Sullivan FM. Optimal strategies for identifying patients with myocardial infarction in general practice. *Fam Pract* 2003;20:706-710.

16. Buckley BS, Simpson CR, McLernon DJ, et al. Considerable differences exist between prevalent and incident myocardial infarction cohorts derived from the same population. *J Clin Epidemiol* 2010;63:1351-1357.

17. Gibbs RG, Newson R, Lawrenson R, et al. Diagnosis and initial management of stroke and transient ischemic attack across UK health regions from 1992 to 1996: experience of a national primary care database. *Stroke* 2001;32:1085-1090.

18. Jagadesham VP, Aparajita R, Gough MJ. Can the UK guidelines for stroke be effective? Attitudes to the symptoms of a transient ischaemic attack among the general public and doctors. *Clin Med* 2008;8:366-370.

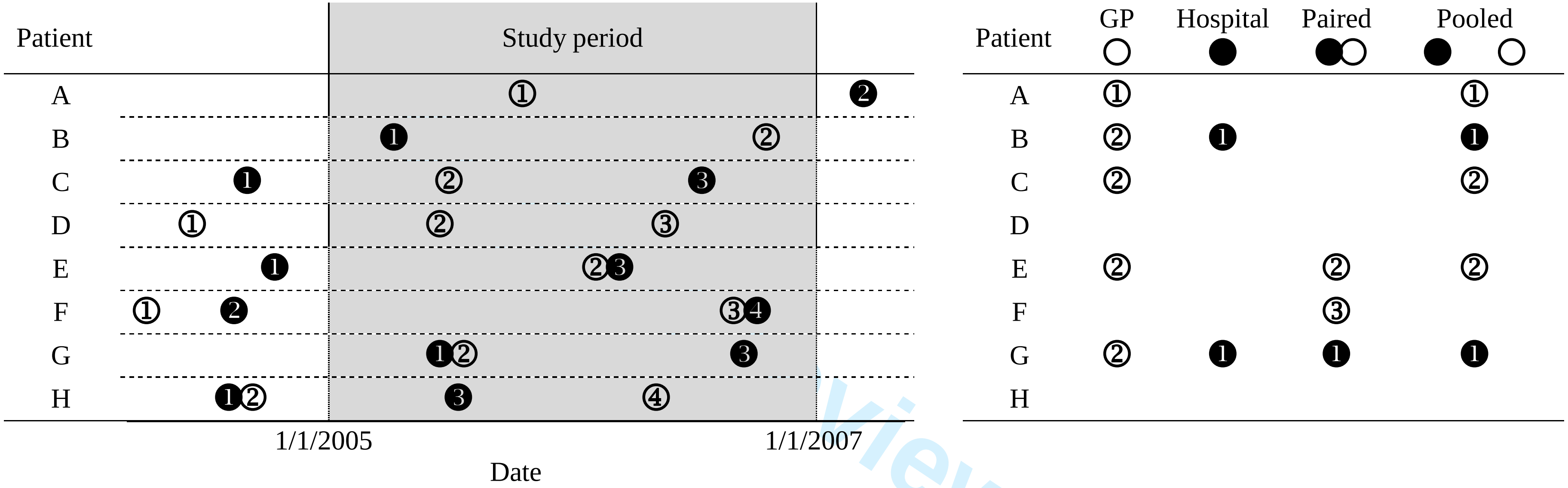
19. Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. *Am Heart J* 2002;144:290-296.

20. Parikh NI, Gona P, Larson MG, et al. Long-term trends in myocardial infarction incidence and case fatality in the National Heart, Lung, and Blood Institute's Framingham Heart study. *Circulation* 2009;119:1203-1210.

21. The General Practice Research Database. http://www.gprd.com/academia/primarycare.asp. Accessed 29/6/11

22. Murphy NF, Simpson CR, McAlister FA, et al. National survey of the prevalence, incidence, primary care burden, and treatment of heart failure in Scotland. *Heart* 2004;90:1129–1136.

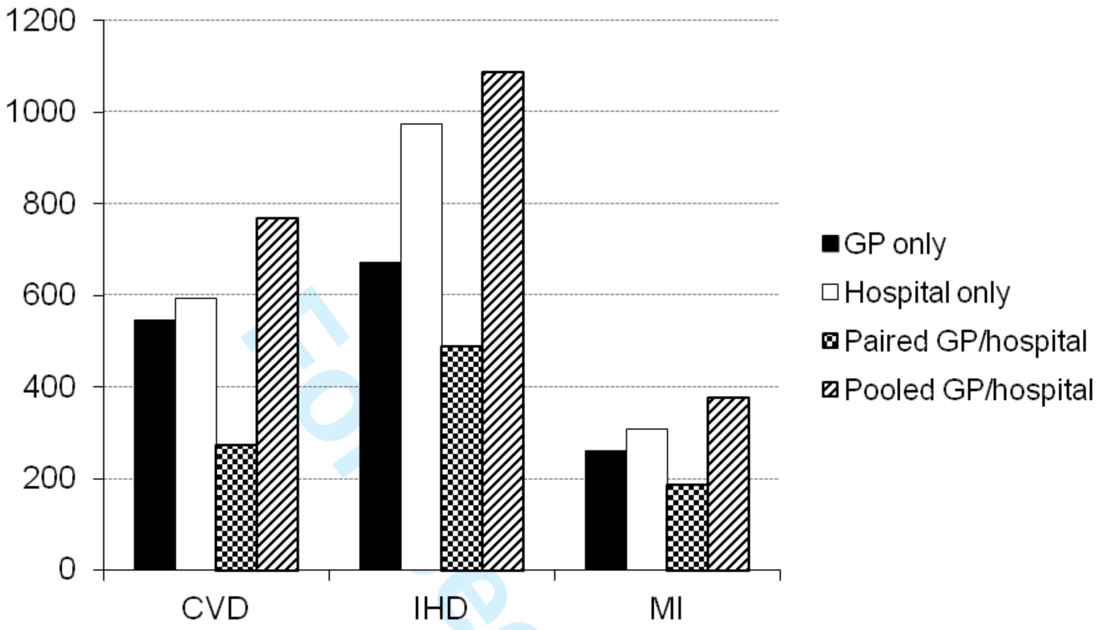
FIGURE 1. IDENTIFICATION OF INCIDENT EVENTS



Patient	GP ○	Hospital ●	Paired ●○	Pooled ● ○
A	①			①
B	②	❶		❶
C	②			②
D				
E	②		②	②
F			③	
G	②	❶	❶	❶
H				

The figure shows how incident events can be identified from linked GP and hospital datasets, for eight hypothetical patients, illustrating some of the potential coding combinations. Circles correspond to the presence of a GP (○) or hospital (●) clinical code, with numbers illustrating the order. Immediately adjacent circles represent codes occurring within 30 days of one another. It can be seen that, for any given patient, it is possible to classify them as having an incident event in up to four ways: GP data only, hospital data only, paired GP/hospital, and pooled GP/hospital; the code which identifies an incident event for each of these methods is shown on the right of the figure. Codes do not count as incident events if a further, similarly classified, event has occurred prior to the start of the study period. In our study, patients were randomly allocated to one of the four coding methods. For instance, if patient E was allocated to “hospital only” coding, they would not be classified as having had an event; in contrast, they would be classified as having had an event if they were allocated to any of the other three coding methods.

FIGURE 2 – INCIDENCE RATES



Incidence rates, expressed per 100,000 patient years, for different clinical conditions over a 2-year time period beginning 1/1/2005, based on general practice (GP), hospital, paired GP/hospital, and pooled GP/hospital data. CVD, cerebrovascular disease; IHD, ischaemic heart disease; MI, myocardial infarction.

A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data

Rupert A Payne, Gary A Abel, Colin R Simpson

Response to referees

We would like to thank all the referees for their helpful, detailed and generally supportive comments. We have attempted to address all of them point-by-point below, and have highlighted changes in the manuscript in yellow.

Reviewer: Prof. Simon Capewell

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.

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.

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.

Methods

Pg 6, line 14: "considered REASONABLY representative of the Scottish population".

Amended as suggested.

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

The word stroke has been added.

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).

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.

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.

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.

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.

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 <u>prior to</u> 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…”)

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

<u>Discussion</u>

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.

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.

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.

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.

Funding. Not very informative

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

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.

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 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.

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.

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

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*

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.