Refereed paper

Identifying the optimal search strategy for coronary heart disease patients in primary care electronic patient record systems

Krish Thiru MPH PhD
Health Informatics, Pan Thames PICAnet Co-ordinator, Great Ormond Street Hospital, Cardiac Critical Care Services, London, UK

Peter T Donnan PhD
Professor of Epidemiology and Biostatistics, Division of Clinical and Population Sciences and Education, Dundee, UK

Peter Weller PhD
Senior Lecturer, Centre for Health Informatics, City University London, London, UK

Frank Sullivan PhD FRCP FRCGP
Head of Division/Director, Scottish School of Primary Care, Dundee, UK

ABSTRACT

**Objectives** General practitioners are increasingly required to practice in a paperless environment and to collect clinical data electronically on electronic patient record (EPR) systems. A principal step in meeting general practice information needs continues to be the establishment of disease registers and consequently the identification of patient populations within primary care databases is a prerequisite. This study aims to identify and validate the optimal search strategy for coronary heart disease (CHD).

**Methods** A multiple logistic regression model for the identification of CHD patients was developed in one site using electronic data, the receiver operating characteristic (ROC) curve and Bayesian statistics. The model was tested on two trial sites.

**Results** Young male CHD patients are more easily identified by generic searches than older females. The optimal search strategy for CHD was found to be the diagnostic code for CHD, nitrate and digoxin but this was dependent on the disease description, age and sex of the study population and the coding system used within the database. Diagnostic code for CHD identified 80.6% (95% confidence interval (CI) 77–83%), 90.0% (CI 88–92%) and 95.9% (CI 94–97%) of local, national and international definitions respectively, with 100% positive predictive values (PPVs) for all definitions.

**Conclusion** Generic queries may inadvertently perpetuate inequalities in health care. Queries should be bespoke and mindful of the conceptualisation of disease by the clinicians recording these data.

**Keywords:** clinical coding, coronary heart disease, database query, electronic patient record system, primary care

Introduction

Computerisation of UK primary care is ubiquitous, driven by the need for improvements in clinical and business processes. Computerisation of the UK’s National Health Service (NHS) continues to expand with the intention of establishing an EPR for each UK resident. Primary care, which has led this computerisation programme, will act as one of the prime sources of clinical information. As a result, family practitioners are increasingly required to practice in a ‘paperless’ or ‘paper-light’ environment to facilitate
the collection of clinical data electronically. Underpinning governmental ambition is the need to derive management data from routine clinical information. Disease registers are an indispensable component of practice management. However, increasingly these are becoming virtual registries (i.e. embedded within EPRs). A number of studies have investigated methods of identifying heart disease patients within EPRs but these have tended to rely on paper records to confirm disease status.

Read codes (Clinical Terms) are used to code clinical information within primary care EPRs. These data support practice remuneration mechanisms, implement evidence-based medicine (EBM) and enhance quality improvements. In a systematic review of data quality within primary care EPRs, we found much of the literature focused on this basic need to identify patient groups and establish disease registers using coded data. Clinical codes and associated terms are like words in that they imply meaning or sense. These terms can be used to describe the nature or the essence of CHD. As such, the codes' meanings depend on who uses them and the context in which they are used. An episode of chest pain experienced by a patient with ischaemic heart disease can be described in a number of ways (e.g. CHD, myocardial infarction, angina, heart attack etc.). As a result, due to the granularity, structure, prioritisation and use of the available codes, establishing a sensitive (complete) and positively predictive (correct) search for patient populations is complex. The aim of this study is to identify and validate the optimal search strategy for CHD in UK primary care whilst being attentive to description of disease, coding granularity and population characteristics.

Method

We identified no published work that explored different interpretations of the disease state and its impact on search strategies. Search strategies usually focus on expected diagnostic and prescribing codes and pay little attention either to the range of other terms available, or to the age and gender of patients.

The model training and testing sites were purposively sampled through the Northern Regional Research Network for their high level of computerisation and expected high levels of electronic data gathering. The training site, a rural family practice of 13,000 patients, used the Egton Medical Information System (EMIS) (www.emis-online.com/) and used 4-byte Read codes. The clinical, administrative and attached community staffs were all competent computer users.

For the independent variables a multiprofessional subgroup of the South Thames Primary Care Research and Development Network used the Nominal Group Technique to identify a core dataset for the implementation and evaluation of evidence-based CHD management. Details of the methodology are provided elsewhere. The group selected 55 essential pieces of information (independent variables). These were Read coded and stratified into four groups for modelling (Table 1). For the dependent variables we categorised the patients as having or not having CHD based on three definitions of CHD. These definitions include a clinical governance focused definition established for local audit purposes, a World Health Organization orientated definition that relies on the International Classification of Diseases version 10 (ICD-10) coding parameters and a national definition of CHD based on the Primary Care Information Services definition (Figure 1).

Statistical methods

The resulting data were analysed using Statistical Package for Social Sciences (Version 11). The Spearman rank correlation coefficient (r) was used to assess correlations between Read codes. Logistic regression is a method for determining the relationship between predictor variables and dichotomously coded dependent variables, while the ROC curve is able to graphically present the sensitivity against (1 – sensitivity) a search method. Forward stepwise multiple logistic regression models (MLRM) using 0.05 and 0.1 entry and exit P-values respectively were used to identify significant codes. Only data for patients ≥35 years of age were analysed. Multiple logistic regression models were developed for a variety of population groups: 1) 35 to 75 years of age; 2) over 75 years of age; 3) males; 4) females; and 5) all patients. The area under the curve (AUC) of an ROC curve was used as an indicator of the effectiveness of search strategies. The AUC (predictive probability) provided a summary index of each model’s performance, while β/SE (standardised regression coefficient) values directed attention towards ‘important’ codes. Bayesian statistics (sensitivity, specificity and yield statistics) were used to further gauge the effectiveness of the model.

Using model data from the training site, the optimal three-code search strategy was refined for each CHD definition from codes highlighted through MLRM. Selected codes were assessed for their cumulative sensitivities, specificities and yield statistics. The selected search strategy was then tested on each other CHD definition to identify the most effective search strategy across the set of three CHD definitions. This was conducted to identify the most effective generic search strategy.
### Table 1 Stratified CHD management 4-byte Read codes used for regression modelling at model training site

<table>
<thead>
<tr>
<th>Codes</th>
<th>Stratified code groups</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G4*</td>
<td>Identification codes strata</td>
<td>Ischaemic heart disease (IHD)</td>
</tr>
<tr>
<td>G41</td>
<td></td>
<td>Acute myocardial infarction</td>
</tr>
<tr>
<td>G42</td>
<td></td>
<td>Acute/sub-acute IHD NOS</td>
</tr>
<tr>
<td>G43</td>
<td></td>
<td>Old myocardial infarction</td>
</tr>
<tr>
<td>G44</td>
<td></td>
<td>Angina pectoris</td>
</tr>
<tr>
<td>3213*</td>
<td></td>
<td>Exercise ECG</td>
</tr>
<tr>
<td>322*</td>
<td></td>
<td>ECG shows myocardial ischaemia</td>
</tr>
<tr>
<td>323*</td>
<td></td>
<td>3232-ECG: old myocardial infarction, 3233-ECG: antero-septal infarct, 3235-ECG: sub-endocardial infarct</td>
</tr>
<tr>
<td>5744*</td>
<td></td>
<td>Isotope dynamic heart scan</td>
</tr>
<tr>
<td>G451</td>
<td></td>
<td>Coronary arteriosclerosis</td>
</tr>
<tr>
<td>771*</td>
<td></td>
<td>7714-heart valve replace-prosthesis, 7716-heart valve replacement NOS</td>
</tr>
<tr>
<td>7732*</td>
<td></td>
<td>Coronary artery venous graft</td>
</tr>
<tr>
<td>440*</td>
<td>Risk factors codes strata</td>
<td>4401-blood sent for serum lipids, 4404-serum lipids high</td>
</tr>
<tr>
<td>44P*</td>
<td></td>
<td>44P-serum cholesterol, 44P3-serum cholesterol raised, 44P5-HDL:total cholesterol ratio</td>
</tr>
<tr>
<td>44P5</td>
<td></td>
<td>Serum HDL cholesterol level</td>
</tr>
<tr>
<td>Egton LD1*</td>
<td></td>
<td>LDL cholesterol level (Egton LD1)</td>
</tr>
<tr>
<td>22K*</td>
<td></td>
<td>Body mass index</td>
</tr>
<tr>
<td>2469*</td>
<td></td>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>246A</td>
<td></td>
<td>Diastolic blood pressure</td>
</tr>
<tr>
<td>900</td>
<td></td>
<td>Anti-smoking advice</td>
</tr>
<tr>
<td>6791*</td>
<td></td>
<td>Health education – smoking</td>
</tr>
<tr>
<td>13B3*</td>
<td></td>
<td>Low cholesterol diet</td>
</tr>
<tr>
<td>Egton 350*</td>
<td></td>
<td>Ideal weight</td>
</tr>
<tr>
<td>6799*</td>
<td></td>
<td>Health education – diet</td>
</tr>
<tr>
<td>Egton 418*</td>
<td></td>
<td>Alcohol intake</td>
</tr>
<tr>
<td>137*</td>
<td></td>
<td>Smoking status</td>
</tr>
<tr>
<td>8H44*</td>
<td></td>
<td>Cardiological referral</td>
</tr>
<tr>
<td>FUNDGRE4*</td>
<td></td>
<td>Cardio surgery referral</td>
</tr>
<tr>
<td>(8H5G)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18*</td>
<td></td>
<td>Cardiovascular symptoms</td>
</tr>
<tr>
<td>182</td>
<td></td>
<td>Chest pain symptom</td>
</tr>
<tr>
<td>Egton AS2*</td>
<td></td>
<td>Aspirin discussed</td>
</tr>
</tbody>
</table>
The test sites were of a similar size and constitution to the training site.

The resulting generic search strategy from the training site was tested on data from the two additional EMIS practices, one using the 4-byte and the other the more granular 5-byte Read codes.

### Ethical considerations

The Local Research Ethics Group was informed of the aims and method of the study. Since the study did not use patient identifiable data it was registered as a data quality audit study and ethical approval waived.

---

**Table 1 Continued**

<table>
<thead>
<tr>
<th>Codes</th>
<th>Stratified code groups</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>662K*</td>
<td>Drug therapy codes strata</td>
<td>Angina control</td>
</tr>
<tr>
<td>2.6.1*</td>
<td>Drug therapy codes strata</td>
<td>Nitres</td>
</tr>
<tr>
<td>2.12*</td>
<td></td>
<td>Anti-hyperlipids</td>
</tr>
<tr>
<td>2.5.1 to 2.5.5*</td>
<td></td>
<td>Anti-hypertensives</td>
</tr>
<tr>
<td>2.3.2, 2.3.3, 2.6.2*</td>
<td></td>
<td>Anti-arrhythmics</td>
</tr>
<tr>
<td>2.4, 2.4.1, 11.6, *</td>
<td></td>
<td>Beta blockers</td>
</tr>
<tr>
<td>2.2.1, 2.2.2, 2.2.3, 2.2.4*</td>
<td></td>
<td>Anti-diuretics</td>
</tr>
<tr>
<td>2.9, 4.7.1, 4.7.1.1*</td>
<td></td>
<td>Anti-platelet agents (aspirin)</td>
</tr>
<tr>
<td>2.8.2*</td>
<td></td>
<td>Anti-coagulation therapy (drugs): Warfarin sodium tablets</td>
</tr>
<tr>
<td>88A5</td>
<td></td>
<td>Anticoagulant therapy</td>
</tr>
<tr>
<td>2.1*</td>
<td></td>
<td>Digoxin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Warfarin monitoring</td>
</tr>
<tr>
<td>12*</td>
<td>Comorbidity codes strata</td>
<td>Family history recorded</td>
</tr>
<tr>
<td>12C2*</td>
<td></td>
<td>Family history of CHD</td>
</tr>
<tr>
<td>1252*</td>
<td></td>
<td>Family history of diabetes mellitus</td>
</tr>
<tr>
<td>1228</td>
<td></td>
<td>History of diabetes</td>
</tr>
<tr>
<td>C2*</td>
<td></td>
<td>Diabetic status</td>
</tr>
<tr>
<td>14A2*</td>
<td></td>
<td>History of hypertension</td>
</tr>
<tr>
<td>G3*</td>
<td></td>
<td>Hypertension</td>
</tr>
<tr>
<td>G7*</td>
<td></td>
<td>Cerebrovascular disease</td>
</tr>
<tr>
<td>G452</td>
<td></td>
<td>Aneurysm</td>
</tr>
<tr>
<td>G823aaa*</td>
<td></td>
<td>G82-aortic aneurysm, G823-abdominal aortic aneurysm, G82Z-aortic aneurysm NOS</td>
</tr>
<tr>
<td>G86, G86Z*</td>
<td></td>
<td>Peripheral vascular disease NOS</td>
</tr>
</tbody>
</table>

* Codes used in MLRM
Results

Of the independent variables selected for the modelling, 14 were excluded from further investigation due to their low prevalence (<1.96 × standard deviation (SD) of expected numbers) or because they had highly correlated partner codes (Table 1). Table 2 details the age/sex standardised distribution of the three CHD definitions.

The number of patients identified decreased using local to national to international definitions of CHD at the 5-byte model training site which used the more granular codes (Figure 1). Standardisation of morbidity levels revealed agreement with national statistics for local and national definitions (Table 2). The international definition, however, selected significantly lower numbers of patients than expected. Figure 1 presents the overlap of patients across definitions. MLRM and ROC curve results indicated variation in effectiveness of the stratified code groups to identify the CHD definitions. The technique highlighted significant codes and excluded 75% of the variables from further investigation. (As an example, Figure 2 highlights variation in the ability of strata of codes to identify a CHD population as defined under the national definition (Figure 1).) Diagnostic codes were highly specific but varied in sensitivity. The diagnostic code for ischaemic heart disease (G4) was the most sensitive for all three definitions. It identified 80.6% (95% confidence interval (CI) 77–83%), 90.0% (CI 88–92%) and 95.9% (CI 94–97%) of local, national and international definitions respectively, with 100% PPVs (true positives/true positives + false positives) for all definitions.

Regression modelling of distinct age and sex populations of patients highlighted differences in the strata of codes most valuable for identifying the three CHD populations and also the most effective individual codes. G4 was the most significant predictor for the national population (β/SE = 10.8), but the same code was not significant for the local and international populations. The 35 to 75 year age group yielded the greatest number of identification codes. The code 771 (heart-valve procedure) was significant (P<0.01) for the over-75-year-olds who also had a greater number of negatively correlated codes (e.g. hypertension and cardiological referral). The code 771 and nitrates were most effective for the international definition (AUC = 99.4%) while the local definition selected comorbidity and hypertension codes with an AUC of 96.3%. The national definition relied on medication codes to achieve an AUC of 99.2%.

Figure 1 Venn diagram detailing the three definitions of CHD and the overlap of patients across definitions as identified at the model training site
Search strategies identified male CHD patients more easily than females. For males, 771 and 3213 (exercise ECG) were the most significant criteria for all CHD definitions. Females were also more likely to have negatively correlated independent variables, indicating the absence of code suggestive of having CHD (i.e. 22K =–2.64, smoking status =–3.62). Identification and risk factor codes demonstrated better efficiency for males. For local and international definitions, hypertension performed best (β/SE = 6.2 and 4.6 respectively) while for the national population, nitrates performed best (β/SE = 11.0 females, 7.5 males).

As expected, codes from the identification strata (Table 1) selected 90.8%, 95.8% and 98.5% of local, national and international populations, respectively. G4 was significant for local and national definitions. Nitrates were more effective for the national (β/SE =11.8) and international (β/SE=5.0) definitions while for the local definition digoxin was the most efficient predictor of the existence of CHD (β/SE = 6.1).

Identification and prophylactic strata of codes performed better for the international definition of CHD while management codes performed better for the national definition (Table 1). The spectrum of significant codes increased to include comorbidity codes for the local definition. G4 formed the primary

### Table 2 Summary of definition specific optimal search strategies and summary statistics of effectiveness at training site (G4 = diagnostic codes for CHD)

<table>
<thead>
<tr>
<th>Definition</th>
<th>Codes selected</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local definition</td>
<td>Diagnostic code (G4), anti-hypertensive drugs and hypertension</td>
<td>95.1%</td>
<td>81.0%</td>
<td>35.0%</td>
<td>2.86</td>
</tr>
<tr>
<td>National definition</td>
<td>G4, nitrates and digoxin</td>
<td>99.0%</td>
<td>92.1%</td>
<td>50.4%</td>
<td>1.98</td>
</tr>
<tr>
<td>International definition</td>
<td>G4, anti-arrhythmic and anti-lipids</td>
<td>98.0%</td>
<td>96.4%</td>
<td>71.0%</td>
<td>1.41</td>
</tr>
</tbody>
</table>

**Figure 2** ROC curves showing AUC (predictive probability (%)) of tiers of codes for the identification of the national CHD definition at the training site

AUC increases as the lines move away from the diagonal (A). The gradient of a straight line from a coordinate of interest to 0,0 gives the likelihood ratio (LHR) positive rate line \(\frac{Ay}{Ax} = \text{true positive rate divided by the false positive rate}=\frac{sen}{1-spec}=\frac{P[+|D]}{1-(1-P[+|D])}\) while a line to 1,1 gives the negative likelihood ratio \(\frac{Dy}{ Dx} = \frac{FN}{TN}=1–\frac{sen}{spec}=1-p[+|D]|/(1–P[+|D]|)=P[–|D|]/P[–|D]|) line. The positive LHR was given prominence in this study where ambiguity existed because it signified the superiority of a search to confirm the presence of a disease.
search criteria for all three CHD definitions. The three optimal codes for each definition and their effectiveness are detailed in Table 2. When each CHD definition search strategy was tested on counter definitions, the codes selected through the national definition (G4, nitrates and digoxin) performed best as a generic search strategy across all three CHD definitions. This query performed with sensitivity, specificity and yield statistics of 86.3%, 91.8% and 1.89 and 98.6%, 91.5% and 1.97 for the local and international definitions of CHD respectively.

Table 3 details the test sites’ age–sex standardised morbidities. The 4-byte site (which used the older Read codes, which were less granular and less descriptive due the number of available codes) had morbidity rates within the expected range, unlike the 5-byte site. There is a progressive increase in the proportion of patients selected under local, national and international criteria using the G4, nitrates and digoxin search strategy (at both sites), while specificity remained in excess of 97% (Table 4). The yield equates to approximately two true patients for every three patients identified through the search.

### Discussion

The optimal search strategy was highly dependent on the definition of CHD, the age and sex of the population and the granularity of the codes used. In this study, the CHD definitions have been used as electronic reference standards (eRS). Read codes (4-byte), 57Kb in size, were developed in the early 1980s to meet the requirements of primary care. As the information gathering requirements of the NHS increased, the 4-byte codes were expanded to form a more descriptive set of 5-byte (125Kb) codes to meet the needs of secondary care and mapping to ICD-9. Although further developments in the granularity of codes have occurred with the evolution of

---

**Table 3** Age–sex standardised CHD levels for three study practices using GPRD rates for Northern and Yorkshire

<table>
<thead>
<tr>
<th>CHD population/definition</th>
<th>Training site</th>
<th>Testing sites</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local</td>
<td>National</td>
<td>International</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4-byte site</td>
<td>5-byte site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>110.2*, se 5.5</td>
<td>101.3, se 6.0</td>
<td>89.1, se 7.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>113.1, se 6.3</td>
<td>105.4, se 5.0</td>
<td>93.8, se 5.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Standardised Mortality Rate (SMR) of 100 implies that the rates are the same for the population of interest and the standard population. An SMR > 100 implies that the rate is greater for the population of interest compared to the standard population. The standardisation population were derived from national general practice statistics.

**Table 4** Performance of the final G4*, nitrates and digoxin-based query on the testing site populations

<table>
<thead>
<tr>
<th>G4*, nitrates, digoxin query</th>
<th>Local</th>
<th></th>
<th></th>
<th>National</th>
<th></th>
<th></th>
<th>International</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>PPV</td>
<td>Sensitivity</td>
<td>PPV</td>
<td>Sensitivity</td>
<td>PPV</td>
</tr>
<tr>
<td>Testing 4-byte site</td>
<td>90.5%</td>
<td>69.0%</td>
<td>97.3%</td>
<td>74.4%</td>
<td>97.7%</td>
<td>66.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing 5-byte site</td>
<td>65.4%</td>
<td>61.0%</td>
<td>74.6%</td>
<td>48.2%</td>
<td>91.2%</td>
<td>39.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* To simplify the text G4 is also used as the diagnostic code for CHD at the 5-byte site
Clinical Terms and SNOMED CT, Read codes continue to predominate in primary care.25

The coded descriptions of CHD in this study are knowledge representation structures (Figure 1). There is implicit and hidden knowledge and purpose within these structures due to the background and aims of the organisations and professionals who developed them and the definitions’ intended use. These less tangible components should be considered when conferring meaning to disease definitions.26 Data collection is context specific. The difficulty in querying this data for disease status is that the information retrieval systems and queries are context independent but content dependent.

Our generic search identified the younger age group and males more easily than older females. This corroborates findings that inequalities in care and data recording for CHD patients exist, with women and the elderly less likely to receive procedural and drug interventions.27–30

These inequalities have implications for the use of drug data as disease markers and the possibility that such search strategies may perpetuate inequalities in care by failing to identify patients not on prophylactic agents. Variables such as ethnicity and socio-economic status were not available and so not investigated.31,32 Such data will be of particular significance in ethnically diverse areas but is missing from EPRs.33 Data analysts should be mindful of such spectrum bias when developing queries.

The absence of nitrates from the local and international definition search criteria is a caution against assuming good performance from expected treatment codes. An overreliance on nitrate codes should be discouraged since it has been suggested as a secondary line of care.33,34

The most likely reason for the inflated SMR levels at the 5-byte site was the inter-portability and translation of the 4-byte definition to the 5-byte setting (Table 2).35,36 It is probable that the availability of more granular codes at the 5-byte site allowed clinicians to code in greater detail. Such coding could explain the inflated SMR if the code definitions were ineffective in excluding non-CHD patients.

The use of lead clinicians with experience in informatics to develop CHD definitions enhanced the content and construct validity of our methodology (Figure 1), whilst the selection of highly computerised practices and the use of CHD as a search item ensured the availability of data for research. The study provides a method for identifying the optimal search strategy without a reliance on paper records. The methodology requires further testing on a range of primary care conditions, with the inclusion of continuous biometric measures (e.g. blood pressure), in a larger number of practices.

Plans to manage data centrally presents the UK health service with opportunities to provide tailored advice to practices in developing their disease registers.2 Once effective searches are established, patient status can be further triangulated and validated (i.e. recall nurse clinics, contact the patient or secondary care site, use of clinical notes).37 This will remove the false positives, but will not identify false negative patients. Difficulties in identifying false negatives plague all methods of record keeping, not just EPRs.38

It is proposed that future information systems within the health service will rely on SNOMED CT, a vastly more granular nomenclature than Read codes.39 As organisations are required to collect core datasets and as terminologies become increasingly granular and more like natural language, it may be counterproductive to impose strict coding strategies. Clinicians should be empowered to freely use the codes and nomenclatures available to them. Furthermore, as retrospective datasets become older the likelihood of false positive and negative identification of patient groups is increased. Hence, methodologies to extract information (diagnostic groups) from more intricate and interrelated data are needed. The techniques explored in this study offer mechanisms to deduce information from the emerging sea of coded information in primary care and the health service at large.

Conclusion

The optimal search strategy for CHD was found to be the diagnostic code for CHD, nitrates and digoxin but this may inadvertently perpetuate inequalities in health services due to variation in search performance between age, gender and disease definitions. Although the methodologies used in this study have been successful in other fields of research they have not been sufficiently explored on primary care data sets. We conclude that there is scope to develop this method in primary care data mining and that future search strategies should be bespoke, and be mindful of disease definitions and the transportability of these definitions between versions of codes.

ACKNOWLEDGEMENTS

Thanks to the members of the STaRNet Cholesterol Special Interest Group: Mrs S Brew, Dr B Christie, Dr A Cooper, Dr S de Lusignan, Dr D Goodwin, Dr J Gray, Dr C Kroll, Mrs K Masters, Prof. J Majeed, Dr J Oxenbury, Mrs J Sawyer, Mrs S Wells and Dr D Wheeler. Thanks to members of the Fisher Medical Centre, Bentham Group Practice and the Crosshills Surgery for access to practice data and support.
Krish Thiru was employed by Fisher Medical Centre Research Unit under Northern and Yorkshire NHS Executive R and D funds during the course of this work.

Krish Thiru designed and conducted the study and wrote the article, Peter Donnan advised on statistical methodology, Peter Well commented on the draft paper and advised on statistics. Frank Sullivan supervised the research and commented on drafts.

REFERENCES

31 Scobie S, Basnett I and McCartney P. Can general practice data be used for needs assessment and health

CONFLICTS OF INTEREST
None.

ADDRESS FOR CORRESPONDENCE
Dr K Thiru
Pan Thames PICANet Coordinator
Centre for Health Informatics
School of Informatics
City University
Northampton Square
London EC1V 0HB
UK
Tel: +44 (0)207 040 8370
Email: Thiruk1@city.ac.uk

Accepted October 2009