### Cohort profile: Royal College of General Practitioners Research (RCGP) and Surveillance Centre (RSC) sentinel network

<table>
<thead>
<tr>
<th>Journal:</th>
<th>BMJ Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>bmjopen-2016-011092</td>
</tr>
<tr>
<td>Article Type</td>
<td>Cohort profile</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>11-Jan-2016</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Correa, Ana; University of Surrey, Section of Clinical Medicine and Ageing; Royal College of General Practitioners  
                          Hinton, William; University of Surrey, Section of Clinical Medicine and Ageing  
                          McGovern, Andrew; University of Surrey, Section of Clinical Medicine and Ageing  
                          van Vlymen, Jeremy; University of Surrey, Section of Clinical Medicine and Ageing  
                          Yonova, Ivelina; University of Surrey, Section of Clinical Medicine and Ageing; Royal College of General Practitioners  
                          Jones, Simon; University of Surrey, Section of Clinical Medicine and Ageing  
                          de Lusignan, Simon; University of Surrey, Department of Clinical and Experimental Medicine; Royal College of General Practitioners |
| Primary Subject Heading: | Health services research |
| Secondary Subject Heading: | Epidemiology, General practice / Family practice, Research methods, Evidence based practice, Public health |
| Keywords: | EPIDEMIOLOGY, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PRIMARY CARE, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS |
Cohort profile: Royal College of General Practitioners Research (RCGP) and Surveillance Centre (RSC) sentinel network

Ana Correa 1 2
Primary Care Scientist
AC a.correa@surrey.ac.uk

William Hinton 1
Research Fellow
WH w.hinton@surrey.ac.uk

Andrew McGovern 1
Clinical Researcher
AMcG andy@mcgov.co.uk

Jeremy van Vlymen 1
Research Fellow
JvV j.vanvlymen@surrey.ac.uk

Ivelina Yonova 1 2
Practice Liaison Officer
IY i.yonova@surrey.ac.uk

Simon Jones 1
Research Professor
SJ simonjones@surrey.ac.uk

Simon de Lusignan 1 2*
Medical Director RCGP RSC / Professor of Primary Care
SdeL s.lusignan@surrey.ac.uk
+44 0 1483 683089

1 Section of Clinical Medicine and Aging
University of Surrey
GUILDFORD
GU2 7PX, UK

2 Royal College of General Practitioners
30 Euston Square
LONDON
NW1 2FP, UK

*Author for correspondence

Keywords: Cohort studies, sentinel surveillance, primary health care, England

Word count: 3,745
ABSTRACT

Purpose and background
The Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) is one of the longest established primary care sentinel network. In 2015, it established a new data and analysis hub at University of Surrey.

Participants and method
The cohort includes 1,042,063 patients registered in 107 participating GP practices. We compared the RCGP RSC with English national data in the following areas: demographics in the network and within virology sampling practices; geographical distribution; chronic disease prevalence, management, and completeness of data recording; and prescribing and vaccine uptake.

Findings to date
We found a small over representation of people the 25 to 44 age band, under representation of white ethnicity, and of less deprived people; geographical focus is in London, and less in the Southwest and East of England. We found differences in the prevalence of diabetes (National: 6.4%, RCPG RSC: 5.8%), learning disabilities (National: 0.44%, RCPG RSC: 0.40%), obesity (National: 9.2%, RCPG RSC: 8.0%), pulmonary disease (National: 1.8%, RCPG RSC: 1.6%), and cardiovascular diseases (National: 1.1%, RCPG RSC: 1.2%). Data completeness in risk factors for diabetic population is high (Diabetic: 77-99%, Non-diabetic: 26-69%). The RCGP RSC had significantly different prescribing rates and costs for infections (National: 5.58%, RCPG RSC: 7.12%), and for nutrition and blood conditions (National: 6.26%, RCPG RSC: 4.50%). Differences in vaccine uptake were seen in the group of patients aged 2 (National: 38.5%, RCPG RSC: 32.8%). Due to large numbers, most differences were found to be significant (p<0.00015).

Future plans
The RCGP RSC is a representative network, having only small differences across a wide range of variables with the national population, which have now been quantified and can be assessed for clinical relevance for specific studies. In addition to surveillance, this network could also be a rich source for research into routine practice.

Abstract word count: 300

Strengths
• The RCGP RSC network is representative of the national population on a variety of domains, both demographic and clinical, which support the network’s suitability for real-world evidence research.
• Statistically significant differences have been found, due to the large numbers of the sample, but this profile has now quantified them, so that they can be assessed on a study-by-study basis for their clinical relevance.

Limitations
• Some of the analysis compared data extracted using different methodology, which may have skewed the results.
• The network is comprised of volunteer practices, with better management of clinical conditions, which may introduce a bias to research into clinical outcomes.
INTRODUCTION

Data from primary care sentinel networks is widely used in research and surveillance; the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) is one of the longest established.[1] These networks are of particular importance for conditions largely managed in primary care.[2] However, outputs may be biased if the sample is not representative of the monitored population or if there is uncertainty about the denominator.[3] Estimating rates of disease is easiest from countries that have registration-based primary care systems,[4] though useful data can be obtained from health systems that do not require registration.[5]

Demographics and geographical spread are important, and many sentinel networks are involved in the surveillance of infectious conditions, such as influenza, and assessment of vaccine effectiveness. Different age-groups and differing levels of deprivation have different rates of many illnesses, including rates of swab-confirmed influenza.[6] There are also disparities associated with ethnicity, though these are not consistent.[7, 8, 9]

For infectious diseases like influenza, geographic spread of sentinel practices is important, as well as consistent case ascertainment from year to year.[10] It is also important that the characteristics of the participating practices are representative.[11] In particular, when assessing vaccine effectiveness, vaccine uptake needs to be properly recorded within the database, in a way that is representative of the national population rates.[12]

Disease, prescription and vaccination patterns, as well as the underlying population demographics, are interrelated. Any assessment of quality requires an assessment of the rate of case ascertainment followed by an evaluation of the management of the condition.[13] Some sentinel networks have focussed on the management of long term conditions, including diabetes.[14] Chronic conditions, such as diabetes, benefit from improved recording of co-morbidities and risk factors.

The RCGP RSC is a network of general practices, which extracts data from the computerised medical record systems of over 100 practices in England. The network established a weekly returns service in 1964, which has enabled prompt surveillance of infectious diseases and identification of epidemics; with influenza surveillance as a key priority for the network.

The characteristics of the RCGP RSC practice network have previously been compared with population level data to ascertain the representativeness of the sample.[15] However, since the most recent report in 2009, there have been substantial changes within the network, including the commissioning in 2015 of an entirely new data and analytics hub at the University of Surrey.

We carried out this study to describe the representativeness of the RCGP RSC practice network comparing population demographics; geographical distribution; the prevalence and management of chronic conditions, and data completeness taking type 2 diabetes mellitus (T2DM) as an exemplar condition; medication prescribing; and vaccine uptake.

METHODS

We compared four key areas of representativeness with the national population:
• Demographics: age, gender, ethnicity, and deprivation
• Geographical distribution
• Prevalence and management of chronic conditions, and the data completeness of risk factors for T2DM patients
• Medications prescribing rates and costs, and influenza vaccine uptake

We also compared the demographic characteristics of the practices within the RCGP RSC network which provide nasopharyngeal virology swabbing specimens for Public Health England’s (PHE) viral infection surveillance programme, with those that do not.

The data for this study were extracted from practices providing data uploads in March 2015 into the new RCGP RSC network. RCGP RSC member practices are all volunteers who sign a written agreement to participate in surveillance and research. The practice upload includes all registered patients’ coded data.

The data extracted by the RCGP RSC sentinel system are the principal primary care public health surveillance data used by PHE. All data extracted are pseudonymised as close to source as possible, and kept in a secure network database. Patients with flags in their records suggesting they have opted out of data sharing do not have their data retained or analysed. Data used for surveillance are part of direct patient care. However, all planned research use of these data requires review by research ethics committees prior to use.

**Population demographics**
The RCGP RSC population was divided by gender, aggregated into the age bands used in the 2011 Census for England and Wales, and compared with national census data available from the Office for National Statistics. Age was calculated from the date of birth at the point of extraction. Similarly, patients were classified into five ethnic groups (according to census categories), and compared against national data. We used an updated version of an established ethnicity classification mapping to maximise the use of any ethnicity data.[16]

In the process of data extraction, patient postcodes were converted into Lower Super Output Areas (LSOAs),[17] and these were used to assign an Index of Multiple Deprivation (IMD) score to each patient.[18] Postcode data were then removed from the final extract to maintain pseudonymisation. We also compared the demographics (age, gender, ethnicity, and IMD scores) of practices providing viral throat swabs for PHE with practices which do not.

**Geographic distribution**
We used LSOA data, coordinates, and map files provided by Office for National Statistics geographical survey,[19] to display the geographical distribution of the practices. Using each GP practice postcode, we plotted this against a map of English Clinical Commissioning Group (CCG) regions, using the statistical software R and the package maptools.[20]
Chronic conditions

We compared the prevalence and quality of management of key chronic diseases. We used the data collected for a national pay-for-performance (P4P) scheme, the Quality Outcomes Framework (QOF), to compare the prevalence of chronic conditions that are part of the scheme, for RCGP RSC practices and at the national level. QOF quality points were used as a surrogate of the quality of care and management of these conditions.

Additionally, we selected diabetes as an exemplar condition to demonstrate the level of data quality found in the database for chronic conditions. We assessed the level of completeness in data recording for a number of demographics (age, gender, deprivation) and risk factors (smoking status; systolic blood pressure; Body Mass Index, BMI, or height and weight; ethnicity; cholesterol; and glycaemic control), within the RCGP RSC database.

Medication prescribing and vaccine uptake

Prescription data for England were collated via the Prescription Cost Analysis (PCA) 2014 report, and these were compared with the RCGP RSC data for the same year. We calculated proportions for the most commonly prescribed British National Formulary (BNF) chapters (gastro-intestinal, cardiovascular, respiratory, central nervous, endocrine, nutrition/blood, and infections) and compared these with the national proportions.

The average cost per item of each BNF chapter was multiplied by the quantities prescribed to obtain costs per chapter, for both the RCGP RSC practices and national data. Additionally, we compared influenza vaccine uptake in the RCGP RSC database with national rates published by PHE, focusing on four at-risk age groups (Children aged 2, aged 3, aged 4, and over 65 year olds).

Statistical analyses

We used standard descriptive statistics to compare the RCGP RSC with national data. The chi-square test was used to compare proportions across most variables. Differences in the cumulative frequency distribution of IMD scores were assessed using the Kolmogorov-Smirnov test. The statistical software R was used for all tests of significance, and the p-value was reported. Due to the large number of statistical tests, we applied a Bonferroni correction to the level of significance, resulting in a p-value threshold of 0.00015.

COHORT DESCRIPTION AND FINDINGS TO DATE

The cohort includes data from 1,042,063 patients registered at the time of data upload in one of the 107 participating practices.

Population demographics

Data for age and gender were recorded for the complete RCGP RSC population. The age and gender distribution of the RCGP RSC population was similar to the census distribution (Figure 1 and Supplementary Table 1). There was a significantly higher proportion of both males and females in the 25 to 44 age band when compared to the census and a lower proportion of people in the 0 to 4
age band (p<0.00015). All differences were statistically significant; with the exception of people aged 65 to 74, men aged 45 to 64, and women aged 5 to 14.

Figure 1 – Age-gender profile for the RCGP RSC compared with the census population

The ethnicity of 630,754 (61%) patients from the RCGP RSC cohort was identified from their medical records. The proportions of the five census ethnicity groups were similar to those from census data (Supplementary Figure 1 and Supplementary Table 2). The majority of patients in the RCGP RSC cohort were of white ethnicity (84.4%), similar to the census population (85.4%). Due to the large numbers, the differences between the census and RCGP RSC data were significant (p<0.00015), even when applying a Bonferroni correction.

IMD scores were calculated for all patients included in the analysis, using LSOAs (Supplementary Figure 2). The mean IMD score for the RCGP RSC population was 19.8 (SD 0.00682), which was less deprived than the English population (mean IMD score 21.8; SD 0.00050). This is mostly due to a small over representation of patients in the two least deprived deciles (Supplementary Figure 3). The Kolmogorov-Smirnov test showed that this difference was significant (p<0.00015).

The demographic characteristics of the practices participating in the PHE swabbing programme were compared to non-participating practices. There was a significantly higher proportion of people in the 25 to 44 age band for participating practices, and a lower proportion of people in the 65+ band (Supplementary Figure 4). All differences were statistically significant (p<0.00015), with the exception of men aged 15-24 and women aged 5-14.

Participating practices had significantly less people in the white census group (75%) compared to non-participating practices (83%), and more people in ethnic minority groups (p<0.00015, Supplementary Figure 5). Finally, the population in participating practices was more deprived (mean IMD score 21.5, SD 0.018) than that of the non-participating practices (mean IMD score 19.0, SD 0.010, Supplementary Figure 6), and the Kolmogorov-Smirnov test showed this to be significant (p<0.00015).

Geographic distribution
The RCGP RSC practices are broadly distributed across England, with a higher concentration of practices in London, and a slightly lower number in the Southwest and East of England (Figure 2). The distribution shows a number of practice clusters, probably due to effective recruitment strategies using the support of CCGs.

Figure 2 – Geographical distribution of the RCGP RSC practices in England

Chronic conditions
The prevalence of common chronic diseases was similar to that reported nationally in the QOF scheme for chronic disease management (Figure 3 and Supplementary Table 3). Diabetes (National:

\[ \text{The denominators vary by condition, depending on the QOF definition (all population, population over 18 years, etc.)} \]
6.4%, RCPG RSC: 5.8%), learning disabilities (National: 0.44%, RCPG RSC: 0.40%), obesity (National: 9.2%, RCPG RSC: 8.0%), and pulmonary disease (National: 1.8%, RCPG RSC: 1.6%) are slightly under represented, while cardiovascular diseases (National: 1.1%, RCPG RSC: 1.2%) are slightly over represented.

All differences were significant (p<0.00015), with the exception of asthma, chronic kidney disease (CKD), depression, palliative care, and stroke. The mean proportion of QOF targets achieved by the RCGP RSC network (97.4%; SD 0.0233%) was higher than the national mean (94.7%; SD 0.0006%), due to an overrepresentation of the network in the higher score deciles (Supplementary Fig 7).

*Figure 3 – RCGP RSC recorded prevalence of chronic conditions compared with all English practices, using QOF data*

Diabetes was taken as an exemplar condition to demonstrate the completeness of recorded data within the RCGP RSC database (Table 1). This analysis was performed on patients who have complete data for the past 10 years. The recording of data related to cardiovascular risk factors ranges from 77% to 99% for the diabetic population, and from 26% to 69% for the non-diabetic population.

*Table 1 – RCGP RSC data completeness on key variables for patients with and without diabetes*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Diabetes</th>
<th>Non-Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>39,559</td>
<td>977,445</td>
</tr>
<tr>
<td>Gender</td>
<td>39,559</td>
<td>977,445</td>
</tr>
<tr>
<td>Deprivation score</td>
<td>38,312</td>
<td>935,128</td>
</tr>
<tr>
<td>Smoking status</td>
<td>38,928</td>
<td>669,874</td>
</tr>
<tr>
<td>Blood pressure (Systolic)</td>
<td>38,935</td>
<td>502,188</td>
</tr>
<tr>
<td>BMI (or Height and Weight)</td>
<td>38,225</td>
<td>524,384</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>30,409</td>
<td>585,279</td>
</tr>
<tr>
<td>Cholesterol (HDL, LDL, other)</td>
<td>37,386</td>
<td>293,082</td>
</tr>
<tr>
<td>Glycaemic control (blood glucose/HbA1c)</td>
<td>39,244</td>
<td>253,532</td>
</tr>
</tbody>
</table>

*Medication prescribing and vaccine uptake*

Prescription items and costs by BNF chapters in the RCGP RSC dataset were compared with national data. The RCGP RSC had higher prescribing rates and costs for drugs within the infections chapter (National: 5.58%, RCPG RSC: 7.12%), while the practices in the network prescribed less drugs in the nutrition and blood chapter (National: 6.26%, RCPG RSC: 4.50%). This difference was significant (p<0.00015, Figure 4) and it was also reflected in the prescription costs (Supplementary Figure 6).

*Figure 4 – Prescription items by BNF chapter: National and RCGP RSC*

Influenza vaccine exposure rates within the RCGP RSC network were compared to the national rates published by PHE. Influenza vaccine uptake in the RCGP RSC network is similar to the national rates, although slightly smaller; these are shown in Figure 5. The largest difference was seen in the group...
of patients aged 2 (National: 38.5%, RCPG RSC: 32.8%), while the smallest differences were in the
over 65 year olds (National: 73%, RCPG RSC: 71%).

Figure 5 – Vaccine exposure: National and RCGP RSC

STRENGTHS AND LIMITATIONS

Strengths

This study shows the representativeness of the RCGP RSC against nationally published data on a
variety of domains, both demographic and clinical. Beyond the areas concerning surveillance, the
RCGP RSC database is also representative in a number of clinically-relevant areas, which support the
network’s suitability for real-world evidence research.

The differences in the demographic characteristics reflect the geographical spread of the network.
Recruitment efforts continue to be expanded, particularly regarding new projects, such as the
Integrate study for gastrointestinal disease surveillance [25] which is focused on the north west of
England; this shift in geographical focus may change the demographic structure of the network.

The prevalence of chronic conditions, and the completeness of recorded data (specifically for
T2DM), are the main strengths of the RCGP RSC database. Due to QOF targets, which incentivise
accurate recording of chronic conditions diagnoses, and of continuous assessments of risk factors for
patients with these conditions, primary care data are particularly well suited for longitudinal
analysis.

The nature of the database as primarily a surveillance network is reflected in the higher prescription
rates for infections. However, this was not reflected in the influenza vaccine uptake; given that
practices in the network receive constant feedback to correctly record vaccine exposure, the
differences in the rates are likely due to the method of data extraction used in the analysis, which
are discussed under the limitations of the study.

Most of the differences found were statistically significant; this is due to the large numbers involved
in the analysis, which means that very small differences can be easily detected, even when a
Bonferroni correction is applied to adjust the significance level. Though the difference introduces a
bias to the cohort, it has nonetheless been quantified in this paper and can be assessed on a study-
by-study basis for its clinical relevance.

Limitations

An important limitation of this study is that a number of areas (ethnicity, prescriptions, vaccine
uptake) were assessed by comparing publically available national rates with data extracted from the
RCGP RSC database. The criteria used to extract the network’s rates may not have been the same as
that used for the published national data, leading to significant differences.

Additionally, the RCGP RSC network is comprised of volunteer practices, with better recording
practices, as shown by the higher QOF scores compared to a national population. As such, while they
may be demographically representative of the English population, they may have better
management of clinical conditions, which could potentially skew the results of research into clinical outcomes.

CONCLUSIONS AND FUTURE COLLABORATIONS

This network provides a representative sample of the population of England in terms of demographics and clinical outcomes. Future recruitment needs to ensure that, as far as possible, any areas of difference are minimised. The RCGP RSC network, in addition to surveillance, could also be used for research into routine practice and the interaction of infectious disease with long term conditions.

The use of large healthcare datasets continues to expand,[26] with many networks having been setup in the UK providing regular updates to large datasets.[27, 28, 29] These networks are recognised as having a vital role in disease surveillance, monitoring response to policy change, identification of health-environment interactions, and medication side-effect monitoring.[30, 31, 32, 33] More recently it has been suggested that these large data networks could also be used to perform large scale clinical trials and ensure real world medication effectiveness for new medications.[34]

This is particularly important as clinical trial populations are often not representative of the real world population, where any interventions or new treatments are ultimately implemented.[35, 36] If clinical outcomes are to be measured using these datasets, it is important that they provide a representative sample of the underlying population, reporting any deviations. The RCGP RSC network is representative of the underlying English population, and key significant differences have been clearly quantified in this paper, making this database a rich source for health outcomes research.

It is expected that this database will be available to researchers on a case-by-case basis. Ethical approval by the NHS Research Ethics Committee is needed for data requests to be considered. RCGP policy is that the patient level data will not leave the secure network at University of Surrey. Either aggregated data or researchers who wish to work within University of Surrey governance arrangements can have access to these data for approved ethical purposes. We encourage interested researchers to attend the short courses on how to analyse primary care data offered by the university twice a year.
ACKNOWLEDGEMENTS
The authors would like to thank the patients whose data we utilise, and those who consent to having microbiological samples taken as part of the network’s surveillance role. We also would like to thank participating practices for allow access to these data for this cohort. Finally, we would like to thank Richard Pebody from Public Health England for his helpful comments.

COMPETING INTERESTS
There are no competing interests. The RCGP RSC is primarily funded by Public Health England.

CONTRIBUTORSHIP STATEMENT
AC led the analysis of data and drafting of the paper. WH and AMcG supported the analysis of data. AMcG contributed to the drafting of the paper. JvV and SJ supported the statistical analysis of the data. IY supported the development of the cohort and contributed to the drafting of the paper. SdeL led the development of the cohort, conceived of the idea, and contributed to the drafting of the paper.
REFERENCES


Figure 1 – Age-gender profile for the RCGP RSC compared with the census population

124x61mm (300 x 300 DPI)
Geographical distribution of the RCGP RSC practices

Figure 2 – Geographical distribution of the RCGP RSC practices in England
50x59mm (300 x 300 DPI)
Figure 3 – RCGP RSC recorded prevalence of chronic conditions compared with all English practices, using QOF data

79x44mm (300 x 300 DPI)
Figure 4 – Prescription items by BNF chapter: National and RCGP RSC
96x38mm (300 x 300 DPI)
Figure 5 – Vaccine exposure: National and RCGP RSC
63x38mm (300 x 300 DPI)
### Supplementary Table 1 – Age-gender profile for the RCGP RSC compared with the census population (with confidence intervals)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Male RCGP RSC</th>
<th>Male Nationally registered</th>
<th>Male Census Nationally registered</th>
<th>Female RCGP RSC</th>
<th>Female Nationally registered</th>
<th>Female Census Nationally registered</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>2.7%</td>
<td>3.2%</td>
<td>3.0%</td>
<td>2.6%</td>
<td>3.0%</td>
<td>2.9%</td>
</tr>
<tr>
<td></td>
<td>38,407</td>
<td>1,789,744</td>
<td>1,728,207</td>
<td>36,998</td>
<td>1,707,006</td>
<td>1,643,268</td>
</tr>
<tr>
<td>5-14</td>
<td>5.8%</td>
<td>5.8%</td>
<td>5.8%</td>
<td>5.6%</td>
<td>5.6%</td>
<td>5.6%</td>
</tr>
<tr>
<td></td>
<td>80,598</td>
<td>3,273,105</td>
<td>3,325,746</td>
<td>77,901</td>
<td>3,121,283</td>
<td>3,170,879</td>
</tr>
<tr>
<td>15-24</td>
<td>5.9%</td>
<td>6.6%</td>
<td>6.1%</td>
<td>5.9%</td>
<td>6.5%</td>
<td>6.1%</td>
</tr>
<tr>
<td></td>
<td>82,132</td>
<td>3,727,178</td>
<td>3,502,890</td>
<td>82,102</td>
<td>3,619,452</td>
<td>3,495,931</td>
</tr>
<tr>
<td>25-44</td>
<td>15.2%</td>
<td>13.6%</td>
<td>14.1%</td>
<td>15.6%</td>
<td>13.8%</td>
<td>13.8%</td>
</tr>
<tr>
<td></td>
<td>212,806</td>
<td>7,640,606</td>
<td>8,034,987</td>
<td>218,224</td>
<td>7,711,168</td>
<td>7,861,316</td>
</tr>
<tr>
<td>45-64</td>
<td>12.5%</td>
<td>12.6%</td>
<td>12.9%</td>
<td>12.3%</td>
<td>12.9%</td>
<td>12.5%</td>
</tr>
<tr>
<td></td>
<td>175,078</td>
<td>7,046,582</td>
<td>7,370,993</td>
<td>171,911</td>
<td>7,216,715</td>
<td>7,158,553</td>
</tr>
<tr>
<td>65-74</td>
<td>4.2%</td>
<td>4.2%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>58,395</td>
<td>2,331,538</td>
<td>2,596,829</td>
<td>63,330</td>
<td>2,521,295</td>
<td>2,759,534</td>
</tr>
<tr>
<td>75+</td>
<td>3.1%</td>
<td>3.1%</td>
<td>3.3%</td>
<td>4.3%</td>
<td>4.6%</td>
<td>4.5%</td>
</tr>
<tr>
<td></td>
<td>43,020</td>
<td>1,764,623</td>
<td>1,875,571</td>
<td>60,438</td>
<td>2,605,617</td>
<td>2,585,283</td>
</tr>
<tr>
<td>All ages</td>
<td>49.3%</td>
<td>49.2%</td>
<td>49.8%</td>
<td>50.7%</td>
<td>50.8%</td>
<td>50.2%</td>
</tr>
<tr>
<td></td>
<td>690,436</td>
<td>27,573,376</td>
<td>28,435,223</td>
<td>710,904</td>
<td>28,502,536</td>
<td>28,674,764</td>
</tr>
</tbody>
</table>

### Supplementary Figure 1 – Ethnicity distribution by census groups for the RCGP RSC compared with the census population

### Supplementary Table 2 – Ethnicity distribution by census groups for the RCGP RSC compared with the census population (tabulation)

<table>
<thead>
<tr>
<th>Ethnicity census group</th>
<th>RCGP RSC</th>
<th>Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>84.4%</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>543,452</td>
<td>45,281,142</td>
</tr>
<tr>
<td>Mixed</td>
<td>1.9%</td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td>12,515</td>
<td>1,192,879</td>
</tr>
<tr>
<td>Asian</td>
<td>8.0%</td>
<td>7.8%</td>
</tr>
<tr>
<td></td>
<td>51,502</td>
<td>4,143,403</td>
</tr>
<tr>
<td>Black</td>
<td>4.2%</td>
<td>3.5%</td>
</tr>
<tr>
<td></td>
<td>26,996</td>
<td>1,846,614</td>
</tr>
<tr>
<td>Other</td>
<td>1.5%</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td>9,808</td>
<td>548,418</td>
</tr>
</tbody>
</table>
Supplementary Figure 2 – Cumulative distribution of IMD scores for the RCGP RSC compared with the census population

Supplementary Figure 3 – Distribution of IMD scores for the RCGP RSC compared with the census population, by deciles

Supplementary Figure 4 – Age-gender profile for the practices participating in the PHE swabbing programme compared to non-participating practices

Supplementary Figure 5 – Ethnicity distribution by census groups for the practices participating in the PHE swabbing programme compared to non-participating practices

Supplementary Figure 6 – Cumulative distribution of IMD scores for the practices participating in the PHE swabbing programme compared to non-participating practices

Supplementary Table 3 – RCGP RSC recorded prevalence of chronic conditions compared with all English practices, using QOF data (with confidence intervals)
<table>
<thead>
<tr>
<th>QOF Condition</th>
<th>National</th>
<th>RCGP RSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>5.99%</td>
<td>6.01%</td>
</tr>
<tr>
<td></td>
<td>CI 5.98% - 5.99%</td>
<td>CI 5.96% - 6.05%</td>
</tr>
<tr>
<td></td>
<td>3,402,437</td>
<td>62,402</td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>1.63%</td>
<td>1.69%</td>
</tr>
<tr>
<td></td>
<td>CI 1.63% - 1.63%</td>
<td>CI 1.66% - 1.71%</td>
</tr>
<tr>
<td></td>
<td>926,551</td>
<td>17,538</td>
</tr>
<tr>
<td>Cancer</td>
<td>2.26%</td>
<td>2.36%</td>
</tr>
<tr>
<td></td>
<td>CI 2.25% - 2.26%</td>
<td>CI 2.33% - 2.39%</td>
</tr>
<tr>
<td></td>
<td>1,281,811</td>
<td>24,535</td>
</tr>
<tr>
<td>CVD</td>
<td>1.07%</td>
<td>1.25%</td>
</tr>
<tr>
<td></td>
<td>CI 1.07% - 1.07%</td>
<td>CI 1.22% - 1.27%</td>
</tr>
<tr>
<td></td>
<td>337,299</td>
<td>7,227</td>
</tr>
<tr>
<td>CHD</td>
<td>3.25%</td>
<td>3.02%</td>
</tr>
<tr>
<td></td>
<td>CI 3.24% - 3.25%</td>
<td>CI 2.98% - 3.05%</td>
</tr>
<tr>
<td></td>
<td>1,843,813</td>
<td>31,338</td>
</tr>
<tr>
<td>CKD</td>
<td>4.13%</td>
<td>4.18%</td>
</tr>
<tr>
<td></td>
<td>CI 4.12% - 4.13%</td>
<td>CI 4.13% - 4.22%</td>
</tr>
<tr>
<td></td>
<td>1,859,963</td>
<td>34,508</td>
</tr>
<tr>
<td>Dementia</td>
<td>0.74%</td>
<td>0.70%</td>
</tr>
<tr>
<td></td>
<td>CI 0.74% - 0.74%</td>
<td>CI 0.68% - 0.72%</td>
</tr>
<tr>
<td></td>
<td>419,073</td>
<td>7,275</td>
</tr>
<tr>
<td>Depression</td>
<td>7.33%</td>
<td>7.31%</td>
</tr>
<tr>
<td></td>
<td>CI 7.32% - 7.34%</td>
<td>CI 7.25% - 7.36%</td>
</tr>
<tr>
<td></td>
<td>3,305,363</td>
<td>60,371</td>
</tr>
<tr>
<td>Diabetes</td>
<td>6.37%</td>
<td>5.75%</td>
</tr>
<tr>
<td></td>
<td>CI 6.36% - 6.38%</td>
<td>CI 5.70% - 5.80%</td>
</tr>
<tr>
<td></td>
<td>2,913,538</td>
<td>48,228</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>0.79%</td>
<td>0.74%</td>
</tr>
<tr>
<td></td>
<td>CI 0.79% - 0.79%</td>
<td>CI 0.73% - 0.76%</td>
</tr>
<tr>
<td></td>
<td>357,096</td>
<td>6,149</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>0.72%</td>
<td>0.69%</td>
</tr>
<tr>
<td></td>
<td>CI 0.72% - 0.73%</td>
<td>CI 0.67% - 0.71%</td>
</tr>
<tr>
<td></td>
<td>410,783</td>
<td>7,168</td>
</tr>
<tr>
<td>Hypertension</td>
<td>13.79%</td>
<td>13.26%</td>
</tr>
<tr>
<td></td>
<td>CI 13.78% - 13.80%</td>
<td>CI 13.19% - 13.32%</td>
</tr>
<tr>
<td></td>
<td>7,833,779</td>
<td>137,704</td>
</tr>
<tr>
<td>Learning Disabilities</td>
<td>0.44%</td>
<td>0.40%</td>
</tr>
<tr>
<td></td>
<td>CI 0.44% - 0.45%</td>
<td>CI 0.38% - 0.41%</td>
</tr>
<tr>
<td></td>
<td>252,446</td>
<td>4,119</td>
</tr>
<tr>
<td>Obesity</td>
<td>9.24%</td>
<td>7.97%</td>
</tr>
<tr>
<td></td>
<td>CI 9.23% - 9.25%</td>
<td>CI 7.91% - 8.02%</td>
</tr>
<tr>
<td></td>
<td>4,286,727</td>
<td>67,763</td>
</tr>
<tr>
<td>Palliative Care</td>
<td>0.31%</td>
<td>0.33%</td>
</tr>
<tr>
<td></td>
<td>CI 0.31% - 0.31%</td>
<td>CI 0.32% - 0.34%</td>
</tr>
<tr>
<td></td>
<td>177,329</td>
<td>3,430</td>
</tr>
<tr>
<td>Pulmonary Disease</td>
<td>1.82%</td>
<td>1.64%</td>
</tr>
<tr>
<td></td>
<td>CI 1.82% - 1.82%</td>
<td>CI 1.62% - 1.66%</td>
</tr>
<tr>
<td></td>
<td>1,034,578</td>
<td>17,034</td>
</tr>
<tr>
<td>Stroke/TIA</td>
<td>1.73%</td>
<td>1.71%</td>
</tr>
<tr>
<td></td>
<td>CI 1.72% - 1.73%</td>
<td>CI 1.68% - 1.73%</td>
</tr>
<tr>
<td></td>
<td>981,836</td>
<td>17,752</td>
</tr>
</tbody>
</table>
Supplementary Figure 7 – RCGP RSC practices QOF scores compared with national scores, by decile

Supplementary Figure 8 – Prescription costs by BNF chapter: National and RCGP RSC
IMD deciles distribution

34x18mm (300 x 300 DPI)
QOF scores 2014-2015 (by decile): National and RCGP RSC

Percentage of practices vs. Decile (1 - lowest QOF score, 10 - highest QOF score)

34x18mm (300 x 300 DPI)
52x18mm (300 x 300 DPI)
# Cohort profile: Royal College of General Practitioners Research (RCGP) and Surveillance Centre (RSC) sentinel network

<table>
<thead>
<tr>
<th>Journal:</th>
<th>BMJ Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>bmjopen-2016-011092.R1</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Cohort profile</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>03-Mar-2016</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Correa, Ana; University of Surrey, Section of Clinical Medicine and Ageing; Royal College of General Practitioners  
Hinton, William; University of Surrey, Section of Clinical Medicine and Aging  
McGovern, Andrew; University of Surrey, Section of Clinical Medicine and Ageing  
vан Vlymen, Jeremy; University of Surrey, Section of Clinical Medicine and Ageing  
Yonova, Ivelina; University of Surrey, Section of Clinical Medicine and Ageing; Royal College of General Practitioners  
Jones, Simon; University of Surrey, Section of Clinical Medicine and Ageing de Lusignan, Simon; University of Surrey, Department of Clinical and Experimental Medicine; Royal College of General Practitioners |
| Primary Subject Heading: | Health services research |
| Secondary Subject Heading: | Epidemiology, General practice / Family practice, Research methods, Evidence based practice, Public health |
| Keywords: | EPIDEMIOLOGY, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PRIMARY CARE, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS |
Cohort profile: Royal College of General Practitioners Research (RCGP) and Surveillance Centre (RSC) sentinel network

Ana Correa 1,2
Primary Care Scientist
AC a.correa@surrey.ac.uk

William Hinton 1
Research Fellow
WH w.hinton@surrey.ac.uk

Andrew McGovern 1
Clinical Researcher
AMcG andy@mcgov.co.uk

Jeremy van Vlymen 1
Research Fellow
JvV j.vanvlymen@surrey.ac.uk

Ivelina Yonova 1,2
Practice Liaison Officer
IY i.yonova@surrey.ac.uk

Simon Jones 1
Research Professor
SJ simonjones@surrey.ac.uk

Simon de Lusignan 1,2
*Author for correspondence
Medical Director RCGP RSC / Professor of Primary Care
SdeL s.lusignan@surrey.ac.uk
+44 0 1483 683089

1 Section of Clinical Medicine and Aging
University of Surrey
GUILDFORD
GU2 7PX, UK

2 Royal College of General Practitioners
30 Euston Square
LONDON
NW1 2FP, UK

*Author for correspondence

Keywords: Cohort studies, sentinel surveillance, primary health care, England

Word count: 3,113
ABSTRACT

Purpose
The Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) is one of the longest established primary care sentinel networks. In 2015, it established a new data and analysis hub at the University of Surrey. This paper evaluates the representativeness of the RCGP RSC network against the English population.

Participants and method
The cohort includes 1,042,063 patients registered in 107 participating GP practices. We compared the RCGP RSC data with English national data in the following areas: demographics; geographical distribution; chronic disease prevalence, management, and completeness of data recording; and prescribing and vaccine uptake. We also assessed practices within the network participating in a national swabbing programme.

Findings to date
We found a small over representation of people in the 25 to 44 age band, under representation of white ethnicity, and of less deprived people. Geographical focus is in London, with less practices in the Southwest and East of England. We found differences in the prevalence of diabetes (National: 6.4%, RCPG RSC: 5.8%), learning disabilities (National: 0.44%, RCPG RSC: 0.40%), obesity (National: 9.2%, RCPG RSC: 8.0%), pulmonary disease (National: 1.8%, RCPG RSC: 1.6%), and cardiovascular diseases (National: 1.1%, RCPG RSC: 1.2%). Data completeness in risk factors for diabetic population is high (77-99%). We found differences in prescribing rates and costs for infections (National: 5.58%, RCPG RSC: 7.12%), and for nutrition and blood conditions (National: 6.26%, RCPG RSC: 4.50%). Differences in vaccine uptake were seen in patients aged 2 (National: 38.5%, RCPG RSC: 32.8%). Due to large numbers, most differences were significant (p<0.00015).

Future plans
The RCGP RSC is a representative network, having only small differences with the national population, which have now been quantified and can be assessed for clinical relevance for specific studies. This network is a rich source for research into routine practice.

Abstract word count: 300

Strengths
- The RCGP RSC network is representative of the national population on a variety of domains, both demographic and clinical, which support the network’s suitability for real-world evidence research.
- Statistically significant differences have been found, due to the large numbers of the sample, but this profile has now quantified them, so that they can be assessed on a study-by-study basis for their clinical relevance.

Limitations
- Some of the analysis compared data extracted using different methodology, which may have skewed the results.
- The network is comprised of volunteer practices, with better management of clinical conditions, which may introduce a bias to research into clinical outcomes.
INTRODUCTION
Data from primary care sentinel networks is widely used in research and surveillance;[1] the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) is one of the longest established.[2] These networks are of particular importance for conditions largely managed in primary care.[3] However, outputs may be biased if the sample is not representative of the monitored population or if there is uncertainty about the denominator.[4] Estimating rates of disease is easiest from countries that have registration-based primary care systems,[5] though useful data can be obtained from health systems that do not require registration.[6]

Demographics and geographical spread are important, and many sentinel networks are involved in the surveillance of infectious conditions, such as influenza, and assessment of vaccine effectiveness. Different age-groups and differing levels of deprivation have different rates of many illnesses, including rates of swab-confirmed influenza.[7] There are also disparities associated with ethnicity, though these are not consistent.[8, 9, 10]

For infectious diseases like influenza, geographic spread of sentinel practices is important, as well as consistent case ascertainment from year to year.[11] It is also important that the characteristics of the participating practices are representative.[12] In particular, when assessing vaccine effectiveness, vaccine uptake needs to be properly recorded within the database, in a way that is representative of the national population rates.[13]

Disease, prescription and vaccination patterns, as well as the underlying population demographics, are interrelated. Any assessment of quality requires an assessment of the rate of case ascertainment followed by an evaluation of the management of the condition.[14] Some sentinel networks have focussed on the management of long term conditions, including diabetes.[15] Chronic conditions, such as diabetes, benefit from improved recording of co-morbidities and risk factors.

The RCGP RSC is a network of general practices, which extracts data from the computerised medical record systems of over 100 practices in England. The network established a weekly returns service in 1964, which has enabled prompt surveillance of infectious diseases and identification of epidemics; with influenza surveillance as a key priority for the network.

The characteristics of the RCGP RSC practice network have previously been compared with population level data to ascertain the representativeness of the sample.[16] However, since the most recent report in 2009, there have been substantial changes within the network, including the commissioning in 2015 of an entirely new data and analytics hub at the University of Surrey.

We carried out this study to describe the representativeness of the RCGP RSC practice network comparing population demographics; geographical distribution; the prevalence and management of chronic conditions, and data completeness taking type 2 diabetes mellitus (T2DM) as an exemplar condition; medication prescribing; and vaccine uptake.

METHODS
We compared four key areas of representativeness with the national population:
• Demographics: age, gender, ethnicity, and deprivation
• Geographical distribution
• Prevalence and management of chronic conditions, and the data completeness of risk factors for T2DM patients
• Medications prescribing rates and costs, and influenza vaccine uptake

We also compared the demographic characteristics of the practices within the RCGP RSC network which provide nasopharyngeal virology swabbing specimens for Public Health England’s (PHE) viral infection surveillance programme, with those that do not. The data for this study were extracted from practices providing data uploads in March 2015 into the new RCGP RSC data upload system. RCGP RSC member practices are all volunteers who sign a written agreement to participate in surveillance and research. The practice upload includes all registered patients’ coded data.

The data extracted by the RCGP RSC sentinel system are the principal primary care public health surveillance data used by PHE. All data extracted are pseudonymised as close to source as possible, and kept in a secure network database. Patients with flags in their records suggesting they have opted out of data sharing do not have their data retained or analysed. Data used for surveillance are part of direct patient care. However, all planned research use of these data requires review by research ethics committees prior to use.

**Population demographics**

The RCGP RSC population was divided by gender, aggregated into the age bands used in the 2011 Census for England and Wales, and compared with national census data available from the Office for National Statistics. Age was calculated from the date of birth at the point of extraction. Similarly, patients were classified into five ethnic groups (according to census categories), and compared against national data. We used an updated version of an established ethnicity classification mapping to maximise the use of any ethnicity data.[17]

In the process of data extraction, patient postcodes were converted into Lower Super Output Areas (LSOAs),[18] and these were used to assign an Index of Multiple Deprivation (IMD) score to each patient.[19] Postcode data were then removed from the final extract to maintain pseudonymisation. We also compared the demographics (age, gender, ethnicity, and IMD scores) of practices providing viral throat swabs for PHE with practices which do not.

**Geographic distribution**

We used LSOA data, coordinates, and map files provided by Office for National Statistics geographical survey,[20] to display the geographical distribution of the practices. Using each GP practice postcode, we plotted this against a map of English Clinical Commissioning Group (CCG) regions, using the statistical software R and the package *maptools*.[21]

**Chronic conditions**

We compared the prevalence and quality of management of key chronic diseases. We used the data collected for a national pay-for-performance (P4P) scheme, the Quality Outcomes Framework (QOF), to compare the prevalence of chronic conditions that are part of the scheme,[22] for RCGP RSC
practices and at the national level. QOF quality points were used as a surrogate of the quality of care and management of these conditions.

Additionally, we selected diabetes as an exemplar condition to demonstrate the level of data quality found in the database for chronic conditions. We assessed the level of completeness in data recording for a number of demographics (age, gender, deprivation) and risk factors (smoking status; systolic blood pressure; Body Mass Index, BMI, or height and weight; ethnicity; cholesterol; and glycaemic control), within the RCGP RSC database.

We only analysed coded data to define data completeness; while free text data can be extracted, there are risks of extracting patient identifiable data, so this data is not used within the RCGP RSC network. Coded data in primary care has been widely used in health outcomes research, [23] and has known strengths and limitations.[24]

**Medication prescribing and vaccine uptake**

Prescription data for England were collated via the Prescription Cost Analysis (PCA) 2014 report,[25] and these were compared with the RCGP RSC data for the same year. We calculated proportions for the most commonly prescribed British National Formulary (BNF) chapters (gastro-intestinal, cardiovascular, respiratory, central nervous, endocrine, nutrition/blood, and infections) and compared these with the national proportions.

The average cost per item of each BNF chapter was multiplied by the quantities prescribed to obtain costs per chapter, for both the RCGP RSC practices and national data. Additionally, we compared influenza vaccine uptake in the RCGP RSC database with national rates published by PHE,[26] focusing on four at-risk age groups (Children aged 2, aged 3, aged 4, and over 65 year olds).

**Statistical analyses**

We used standard descriptive statistics to compare the RCGP RSC with national data. The chi-square test was used to compare proportions across most variables. Differences in the cumulative frequency distribution of IMD scores were assessed using the Kolmogorov-Smirnov test. The statistical software R was used for all tests of significance, and the p-value was reported. Due to the large number of statistical tests (69 tests), we applied a Bonferroni-Šidák correction [27] [28] to a level of significance of 0.01, resulting in a new level of significance of 0.00015 (Equation 1). This version of the correction was used since all the tests are independent.

**Equation 1 – Bonferroni-Šidák equation**

\[
\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}}
\]

\(\alpha' = \text{New significance level}\)

\(\alpha = \text{Current significance level}\)

\(m = \text{Number of tests}\)
COHORT DESCRIPTION AND FINDINGS TO DATE

The cohort includes data from 1,042,063 patients registered at the time of data upload in one of the 107 participating practices.

Population demographics

Data for age and gender were recorded for the complete RCGP RSC population. The age and gender distribution of the RCGP RSC population was similar to the census distribution (Figure 1 and Supplementary Table 1). There was a significantly higher proportion of both males and females in the 25 to 44 age band when compared to the census and a lower proportion of people in the 0 to 4 age band. All differences were statistically significant; with the exception of people aged 65 to 74, men aged 45 to 64, and women aged 5 to 14.

Figure 1 – Age-gender profile for the RCGP RSC compared with the census population

The ethnicity of 630,754 (61%) patients from the RCGP RSC cohort was identified from their medical records. The proportions of the five census ethnicity groups were similar to those from census data (Supplementary Figure 1 and Supplementary Table 2). The majority of patients in the RCGP RSC cohort were of white ethnicity (84.4%), similar to the census population (85.4%). Due to the large numbers, the differences between the census and RCGP RSC data were significant, even when applying a Bonferroni-Šidák correction.

IMD scores were calculated for all patients included in the analysis, using LSOAs (Supplementary Figure 2). The mean IMD score for the RCGP RSC population was 19.8 (SD 0.00682), which was less deprived than the English population (mean IMD score 21.8; SD 0.00050). This is mostly due to a small over representation of patients in the two least deprived deciles (Supplementary Figure 3). The Kolmogorov-Smirnov test showed that this difference was significant.

The demographic characteristics of the practices participating in the PHE swabbing programme were compared to non-participating practices. There was a significantly higher proportion of people in the 25 to 44 age band for participating practices, and a lower proportion of people in the 65+ band (Supplementary Figure 4). All differences were statistically significant, with the exception of men aged 15-24 and women aged 5-14.

Participating practices had significantly less people in the white census group (75%) compared to non-participating practices (83%), and more people in ethnic minority groups (Supplementary Figure 5). Finally, the population in participating practices was more deprived (mean IMD score 21.5, SD 0.018) than that of the non-participating practices (mean IMD score 19.0, SD 0.010, Supplementary Figure 6), and the Kolmogorov-Smirnov test showed this to be significant.

Geographic distribution

The RCGP RSC practices are broadly distributed across England, with a higher concentration of practices in London, and a slightly lower number in the Southwest and East of England (Figure 2).
The distribution shows a number of practice clusters, probably due to effective recruitment strategies using the support of CCGs.

Figure 2 – Geographical distribution of the RCGP RSC practices in England

Chronic conditions

The prevalence of common chronic diseases was similar to that reported nationally in the QOF scheme for chronic disease management (Figure 3 and Supplementary Table 3). Diabetes (National: 6.4%, RCPG RSC: 5.8%), learning disabilities (National: 0.44%, RCPG RSC: 0.40%), obesity (National: 9.2%, RCPG RSC: 8.0%), and pulmonary disease (National: 1.8%, RCPG RSC: 1.6%) are slightly under represented, while cardiovascular diseases (National: 1.1%, RCPG RSC: 1.2%) are slightly over represented.

All differences were significant, with the exception of asthma, chronic kidney disease (CKD), depression, palliative care, and stroke. The median percentage of QOF targets achieved by the RCGP RSC network (98.76%; IQR 2.64%) was higher than the national median percentage (97.04%; IQR 5.45%), due to an overrepresentation of the network in the higher score deciles (Supplementary Figure 7).

Figure 3 – RCGP RSC recorded prevalence of chronic conditions compared with all English practices, using QOF data

Diabetes was taken as an exemplar condition to demonstrate the completeness of recorded data within the RCGP RSC database (Table 1). This analysis was performed on patients who have complete data for the past 10 years. The recording of data related to cardiovascular risk factors ranges from 77% to 99% for the diabetic population, and from 26% to 69% for the non-diabetic population. Due to the introduction of QOF targets, the data completeness related to risk factors or clinical management is expected to be higher for patients with a chronic condition. [29]

The comparison with non-diabetic patients is provided to demonstrate the improvement in recording expected for patients with a chronic condition, compared to those without. Since the QOF targets incentivise accurate recording of chronic conditions diagnoses, and of continuous assessments of risk factors for patients with these conditions, primary care data is particularly well suited for longitudinal research into chronic conditions.

Table 1 – RCGP RSC data completeness on key variables for patients with and without diabetes

<table>
<thead>
<tr>
<th>Variables</th>
<th>Diabetes</th>
<th>Non-Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>%</td>
</tr>
<tr>
<td>Age</td>
<td>39,559</td>
<td>100%</td>
</tr>
<tr>
<td>Gender</td>
<td>39,559</td>
<td>100%</td>
</tr>
</tbody>
</table>

1 The denominators vary by condition, depending on the QOF definition (all population, population over 18 years, etc.)
Medication prescribing and vaccine uptake

Prescription items and costs by BNF chapters in the RCGP RSC dataset were compared with national data. The RCGP RSC had higher prescribing rates and costs for drugs within the infections chapter (National: 5.58%, RCPG RSC: 7.12%), while the practices in the network prescribed less drugs in the nutrition and blood chapter (National: 6.26%, RCPG RSC: 4.50%). This difference was significant (Figure 4) and it was also reflected in the prescription costs (Supplementary Figure 8).

Figure 4 – Prescription items by BNF chapter: National and RCGP RSC

Influenza vaccine exposure rates within the RCGP RSC network were compared to the national rates published by PHE. Influenza vaccine uptake in the RCGP RSC network is similar to the national rates, although slightly smaller; these are shown in Figure 5. The largest difference was seen in the group of patients aged 2 (National: 38.5%, RCPG RSC: 32.8%), while the smallest differences were in the over 65 year olds (National: 73%, RCPG RSC: 71%).

Figure 5 – Vaccine exposure: National and RCGP RSC

STRENGTHS AND LIMITATIONS

Strengths

This study shows the representativeness of the RCGP RSC against nationally published data on a variety of domains, both demographic and clinical. Beyond the areas concerning surveillance, the RCGP RSC database is also representative in a number of clinically relevant areas, which support the network’s suitability for real-world evidence research.

This paper provided a clear assessment of the prevalence of chronic conditions, when compared to national data, and the completeness of recorded data for risk factors, comparing patients with T2DM and those without. The purpose of this was to assess the suitability of longitudinal research into chronic conditions. It was shown that the data completeness for risk factors drastically improves for patients with an exemplar chronic condition (T2DM), due to the QOF scheme in English primary care.

This paper also highlighted key differences of this network against a national population, due to the nature of its primary purpose. Prescriptions for infections were higher than national levels (though not vaccine uptake), and there was an overrepresentation of practices in the higher deciles of QOF scores. This is because the RCGP RSC is a sentinel network comprised of volunteer practices and focused around infectious and respiratory disease surveillance. These differences have been evaluated in this paper, and their relevance can be considered for research using this database.
Most of the differences found were statistically significant; this is due to the large numbers involved in the analysis, which means that very small differences can be easily detected, even when a Bonferroni-Šidák correction is applied to adjust the significance level. Though the difference introduces a bias to the cohort, it has nonetheless been quantified in this paper and can be assessed on a study-by-study basis for its clinical relevance.

Limitations
An important limitation of this study is that a number of areas (ethnicity, prescriptions, vaccine uptake) were assessed by comparing publically available national rates with data extracted from the RCGP RSC database. The criteria used to extract the network’s rates may not have been the same as that used for the published national data, leading to significant differences.

It must be noted that recruitment efforts within the network continue to be expanded, particularly regarding new projects, such as the Integrate study for gastrointestinal disease surveillance [30] which is focused on the north west of England; this shift in geographical focus may change the demographic structure of the network. Therefore, the assessment of representativeness presented in this paper may change in future.

CONCLUSIONS AND FUTURE COLLABORATIONS
This network provides a representative sample of the population of England in terms of demographics and clinical outcomes. Future recruitment needs to ensure that, as far as possible, any areas of difference are minimised. The RCGP RSC network, in addition to surveillance, could also be used for research into routine practice and the interaction of infectious disease with long term conditions.

The use of large healthcare datasets continues to expand,[31] with many networks having been setup in the UK providing regular updates to large datasets.[32, 33, 34] These networks are recognised as having a vital role in disease surveillance, monitoring response to policy change, identification of health-environment interactions, and medication side-effect monitoring.[35, 36, 37, 38] More recently it has been suggested that these large data networks could also be used to perform large scale clinical trials and ensure real world medication effectiveness for new medications.[39]

This is particularly important as clinical trial populations are often not representative of the real world population, where any interventions or new treatments are ultimately implemented.[40, 41] If clinical outcomes are to be measured using these datasets, it is important that they provide a representative sample of the underlying population, reporting any deviations. The RCGP RSC network is representative of the underlying English population, and key significant differences have been clearly quantified in this paper, making this database a rich source for health outcomes research.
It is expected that this database will be available to researchers on a case-by-case basis. Ethical approval by the NHS Research Ethics Committee is needed for data requests to be considered. Data requests for aggregated data will be provided by the University of Surrey team. Researchers wishing to directly analyse the patient-level anonymised data will be required to complete information governance training and work on the data from the secure servers at the University of Surrey. We encourage interested researchers to attend the short courses on how to analyse primary care data offered by the university twice a year.
CONTRIBUTORSHIP STATEMENT
AC led the analysis of data and drafting of the paper. WH and AMcG supported the analysis of data. AMcG contributed to the drafting of the paper. JvV and SJ supported the statistical analysis of the data. IY supported the development of the cohort and contributed to the drafting of the paper. SdeL led the development of the cohort, conceived of the idea, and contributed to the drafting of the paper.

COMPETING INTERESTS
The RCGP RSC is primarily funded by Public Health England.

FUNDING
The development of the RCGP RSC dataset is supported by surveillance work funded by Public Health England.

DATA SHARING STATEMENT
The RCGP RSC dataset can be accessed by bona fide researchers on a case-by-case basis. Ethical approval by the NHS Research Ethics Committee is needed for data requests to be considered. Aggregated data tables may be created from the source data to allow specific analyses, for approved research and surveillance projects. Researchers wishing to directly analyse the patient-level anonymised data will be required to complete information governance training and work on the data from the secure servers at the University of Surrey. Patient level data cannot be taken out of the secure servers at the University of Surrey. We encourage interested researchers to attend the short courses on how to analyse primary care data offered by the university twice a year.

ACKNOWLEDGEMENTS
The authors would like to thank the participating practices and patients for providing the data for this cohort.
REFERENCES


For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml


Figure 1 – Age-gender profile for the RCGP RSC compared with the census population
124x61mm (300 x 300 DPI)
Figure 2 – Geographical distribution of the RCGP RSC practices in England
50x59mm (300 x 300 DPI)
Figure 3 – RCGP RSC recorded prevalence of chronic conditions compared with all English practices, using QOF data

79x44mm (300 x 300 DPI)
Figure 4 – Prescription items by BNF chapter: National and RCGP RSC

96x38mm (300 x 300 DPI)
Figure 5 – Vaccine exposure: National and RCGP RSC

Flu vaccine uptake in at risk age groups

<table>
<thead>
<tr>
<th>Risk groups</th>
<th>Percentage of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years</td>
<td>39% National, 33% RCGP RSC</td>
</tr>
<tr>
<td>3 years</td>
<td>41% National, 38% RCGP RSC</td>
</tr>
<tr>
<td>4 years</td>
<td>33% National, 31% RCGP RSC</td>
</tr>
<tr>
<td>Over 65 years</td>
<td>73% National, 71% RCGP RSC</td>
</tr>
</tbody>
</table>

63x38mm (300 x 300 DPI)
SUPPLEMENTARY FILES

Supplementary Table 1 – Age-gender profile for the RCGP RSC compared with the census population (with confidence intervals)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Male RCGP RSC</th>
<th>Male Census Nationally registered</th>
<th>Male RSC</th>
<th>Female RCGP RSC</th>
<th>Female Census Nationally registered</th>
<th>Female RSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>2.7%</td>
<td>3.2%</td>
<td>3.0%</td>
<td>2.6%</td>
<td>3.0%</td>
<td>2.9%</td>
</tr>
<tr>
<td></td>
<td>38,407</td>
<td>1,789,744</td>
<td>1,728,207</td>
<td>36,998</td>
<td>1,707,006</td>
<td>1,643,268</td>
</tr>
<tr>
<td>5-14</td>
<td>5.8%</td>
<td>5.8%</td>
<td>5.8%</td>
<td>5.6%</td>
<td>5.6%</td>
<td>5.6%</td>
</tr>
<tr>
<td></td>
<td>80,598</td>
<td>3,273,105</td>
<td>3,325,746</td>
<td>77,901</td>
<td>3,121,283</td>
<td>3,170,879</td>
</tr>
<tr>
<td>15-24</td>
<td>5.9%</td>
<td>6.6%</td>
<td>6.1%</td>
<td>5.9%</td>
<td>6.5%</td>
<td>6.1%</td>
</tr>
<tr>
<td></td>
<td>82,132</td>
<td>3,727,178</td>
<td>3,502,890</td>
<td>82,102</td>
<td>3,619,452</td>
<td>3,495,931</td>
</tr>
<tr>
<td>25-44</td>
<td>15.2%</td>
<td>13.6%</td>
<td>14.1%</td>
<td>15.6%</td>
<td>13.8%</td>
<td>13.8%</td>
</tr>
<tr>
<td></td>
<td>212,806</td>
<td>7,640,606</td>
<td>8,034,987</td>
<td>218,224</td>
<td>7,711,168</td>
<td>7,861,316</td>
</tr>
<tr>
<td>45-64</td>
<td>12.5%</td>
<td>12.6%</td>
<td>12.9%</td>
<td>12.3%</td>
<td>12.9%</td>
<td>12.5%</td>
</tr>
<tr>
<td></td>
<td>175,078</td>
<td>7,046,582</td>
<td>7,370,993</td>
<td>171,911</td>
<td>7,216,715</td>
<td>7,158,553</td>
</tr>
<tr>
<td>65-74</td>
<td>4.2%</td>
<td>4.2%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>58,395</td>
<td>2,331,538</td>
<td>2,596,829</td>
<td>63,330</td>
<td>2,521,295</td>
<td>2,759,534</td>
</tr>
<tr>
<td>75+</td>
<td>3.1%</td>
<td>3.1%</td>
<td>3.3%</td>
<td>4.3%</td>
<td>4.6%</td>
<td>4.5%</td>
</tr>
<tr>
<td></td>
<td>43,020</td>
<td>1,764,623</td>
<td>1,875,571</td>
<td>60,438</td>
<td>2,605,617</td>
<td>2,585,283</td>
</tr>
<tr>
<td>All ages</td>
<td>49.3%</td>
<td>49.2%</td>
<td>49.8%</td>
<td>50.7%</td>
<td>50.8%</td>
<td>50.2%</td>
</tr>
<tr>
<td></td>
<td>690,436</td>
<td>27,573,376</td>
<td>28,435,223</td>
<td>710,904</td>
<td>28,502,536</td>
<td>28,674,764</td>
</tr>
</tbody>
</table>

Supplementary Figure 1 – Ethnicity distribution by census groups for the RCGP RSC compared with the census population

Supplementary Table 2 – Ethnicity distribution by census groups for the RCGP RSC compared with the census population (tabulation)

<table>
<thead>
<tr>
<th>Ethnicity census group</th>
<th>RCGP RSC</th>
<th>Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>84.4%</td>
<td>85.4%</td>
</tr>
<tr>
<td></td>
<td>543,452</td>
<td>45,281,142</td>
</tr>
<tr>
<td>Mixed</td>
<td>19.9%</td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td>12,515</td>
<td>1,192,879</td>
</tr>
<tr>
<td>Asian</td>
<td>8.0%</td>
<td>7.8%</td>
</tr>
<tr>
<td></td>
<td>51,502</td>
<td>4,143,403</td>
</tr>
<tr>
<td>Black</td>
<td>4.2%</td>
<td>3.5%</td>
</tr>
<tr>
<td></td>
<td>26,996</td>
<td>1,846,614</td>
</tr>
<tr>
<td>Other</td>
<td>1.5%</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td>9,808</td>
<td>548,418</td>
</tr>
</tbody>
</table>
Supplementary Figure 2 – Cumulative distribution of IMD scores for the RCGP RSC compared with the census population

Supplementary Figure 3 – Distribution of IMD scores for the RCGP RSC compared with the census population, by deciles

Supplementary Figure 4 – Age-gender profile for the practices participating in the PHE swabbing programme compared to non-participating practices

Supplementary Figure 5 – Ethnicity distribution by census groups for the practices participating in the PHE swabbing programme compared to non-participating practices

Supplementary Figure 6 – Cumulative distribution of IMD scores for the practices participating in the PHE swabbing programme compared to non-participating practices
**Supplementary Table 3 – RCGP RSC recorded prevalence of chronic conditions compared with all English practices, using QOF data (with confidence intervals)**

<table>
<thead>
<tr>
<th>QOF Condition</th>
<th>National</th>
<th>RCGP RSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prevalence</td>
<td>CI</td>
</tr>
<tr>
<td>Ashtma</td>
<td>5.99%</td>
<td>CI 5.98% - 5.99%</td>
</tr>
<tr>
<td></td>
<td>3,402,437</td>
<td></td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>1.63%</td>
<td>CI 1.63% - 1.63%</td>
</tr>
<tr>
<td></td>
<td>926,551</td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>2.26%</td>
<td>CI 2.25% - 2.26%</td>
</tr>
<tr>
<td></td>
<td>1,281,811</td>
<td></td>
</tr>
<tr>
<td>CVD</td>
<td>1.07%</td>
<td>CI 1.07% - 1.07%</td>
</tr>
<tr>
<td></td>
<td>337,299</td>
<td></td>
</tr>
<tr>
<td>CHD</td>
<td>3.25%</td>
<td>CI 3.24% - 3.25%</td>
</tr>
<tr>
<td></td>
<td>1,843,813</td>
<td></td>
</tr>
<tr>
<td>CKD</td>
<td>4.13%</td>
<td>CI 4.12% - 4.13%</td>
</tr>
<tr>
<td></td>
<td>1,859,963</td>
<td></td>
</tr>
<tr>
<td>Dementia</td>
<td>0.74%</td>
<td>CI 0.74% - 0.74%</td>
</tr>
<tr>
<td></td>
<td>419,073</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>7.33%</td>
<td>CI 7.25% - 7.36%</td>
</tr>
<tr>
<td></td>
<td>3,305,363</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>6.37%</td>
<td>CI 6.36% - 6.38%</td>
</tr>
<tr>
<td></td>
<td>2,913,538</td>
<td></td>
</tr>
<tr>
<td>Epilepsy</td>
<td>0.79%</td>
<td>CI 0.79% - 0.79%</td>
</tr>
<tr>
<td></td>
<td>357,096</td>
<td></td>
</tr>
<tr>
<td>Heart Failure</td>
<td>0.72%</td>
<td>CI 0.72% - 0.73%</td>
</tr>
<tr>
<td></td>
<td>410,783</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>13.79%</td>
<td>CI 13.78% - 13.80%</td>
</tr>
<tr>
<td></td>
<td>7,833,779</td>
<td></td>
</tr>
<tr>
<td>Learning Disabilities</td>
<td>0.44%</td>
<td>CI 0.44% - 0.45%</td>
</tr>
<tr>
<td></td>
<td>252,446</td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>9.24%</td>
<td>CI 9.23% - 9.25%</td>
</tr>
<tr>
<td></td>
<td>4,286,727</td>
<td></td>
</tr>
<tr>
<td>Palliative Care</td>
<td>0.31%</td>
<td>CI 0.31% - 0.31%</td>
</tr>
<tr>
<td></td>
<td>177,329</td>
<td></td>
</tr>
<tr>
<td>Pulmonary Disease</td>
<td>1.82%</td>
<td>CI 1.82% - 1.82%</td>
</tr>
<tr>
<td></td>
<td>1,034,578</td>
<td></td>
</tr>
<tr>
<td>Stroke/TIA</td>
<td>1.73%</td>
<td>CI 1.72% - 1.73%</td>
</tr>
<tr>
<td></td>
<td>981,836</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Figure 7 – RCGP RSC practices QOF scores compared with national scores, by decile

Supplementary Figure 8 – Prescription costs by BNF chapter: National and RCGP RSC
Cumulative distribution of IMD scores: Census and RCGP RSC

Percentage of population

IMD score (0 - least deprived, 100 - most deprived)

34x18mm (300 x 300 DPI)
Ethnicity distribution by census groups: Practices participating in the PHE swabbing programme and non-participating practices

- Participating practices
- Non-participating practices

<table>
<thead>
<tr>
<th>Ethnicity Census Groups</th>
<th>Participating Practices</th>
<th>Non-participating Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>75%</td>
<td>2%</td>
</tr>
<tr>
<td>Mixed</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>Asian</td>
<td>11%</td>
<td>9%</td>
</tr>
<tr>
<td>Black</td>
<td>9%</td>
<td>4%</td>
</tr>
<tr>
<td>Other</td>
<td>2%</td>
<td>2%</td>
</tr>
</tbody>
</table>

33x19mm (300 x 300 DPI)
QOF scores 2014-2015 (by decile): National and RCGP RSC

Decile (1 - lowest QOF score, 10 - highest QOF score)

Percentage of practices

34x18mm (300 x 300 DPI)
Prescription costs by BNF chapters: National and RCGP RSC

- Gastro-intestinal system: 7% National, 9% RCGP
- Cardiovascular system: 15% National, 15% RCGP
- Respiratory system: 17% National, 19% RCGP
- Central nervous system: 29% National, 29% RCGP
- Infusions: 4% National, 5% RCGP
- Endocrine system: 19% National, 19% RCGP
- Nutrition & blood: 9% National, 7% RCGP

52x18mm (300 x 300 DPI)
Royal College of General Practitioners
Research and Surveillance Centre (RCGP RSC)
sentinel network: a cohort profile
Ana Correa, William Hinton, Andrew McGovern, Jeremy van Vlymen, Ivelina Yonova, Simon Jones and Simon de Lusignan

BMJ Open 2016 6:
doi: 10.1136/bmjopen-2016-011092

Updated information and services can be found at:
http://bmjopen.bmj.com/content/6/4/e011092

These include:

References
This article cites 30 articles, 8 of which you can access for free at:
http://bmjopen.bmj.com/content/6/4/e011092#BIBL

Open Access
This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections
- Epidemiology (2182)
- Evidence based practice (740)
- General practice / Family practice (678)
- Health services research (1529)
- Public health (2295)
- Research methods (623)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/