
Robert W. Aldridge1, Hannah E. R. Evans2, Alexei Yavlinsky1, Alireza Moayyeri1, Krishnan Bhaskaran3, Rohini Mathur4, Kelvin P. Jordan5, Peter Croft5, Spiros Denaxas1, Anoop D. Shah1, Ruth M. Blackburn1, Henrik Moller6, Edmond S. W. Ng3, Andrew Hughes7, Sebastian Fox7, Julian Flowers1, Jurgen Schmidt7, Andrew Hayward8, Ruth Gilbert1, Liam Smeeθ1, Harry Hemingway1

1Institute of Health Informatics, University College London, London, England, NW1 2DA, UK
2University of London, St George’s University, London, SW17 0RE, UK
3Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
4Wolfson Institute of Population Health, Queen Mary University of London, London, England, EC1M 6BQ, UK
5School of Medicine, Keele University, Staffordshire, England, ST5 5BG, UK
6Cancer Epidemiology & Population Health, King’s College London, London, England, WC2R 2LS, UK
7Department of Health and Social Care, Office for Health Improvement and Disparities, London, SW1H 0EU, UK
8Institute of Epidemiology and Health Care, University College London, London, England, WC1E 6BT, UK

First published: 21 Jun 2023, 8:262
https://doi.org/10.12688/wellcomeopenres.19470.1

Abstract

Background: Electronic health records (EHRs) have the potential to be used to produce detailed disease burden estimates. In this study we created disease estimates using national EHR for three high burden conditions, compared estimates between linked and unlinked datasets and produced stratified estimates by age, sex, ethnicity, socio-economic deprivation and geographical region.

Methods: EHRs containing primary care (Clinical Practice Research Datalink), secondary care (Hospital Episode Statistics) and mortality records (Office for National Statistics) were used. We used existing disease phenotyping algorithms to identify cases of cancer (breast, lung, colorectal and prostate), type 1 and 2 diabetes, and lower back pain. We calculated age-standardised incidence of first cancer, point prevalence for diabetes, and primary care consultation prevalence for low back pain.

Results: 7.2 million people contributing 45.3 million person-years of active follow-up between 2000-2014 were included. CPRD-HES combined and CPRD-HES-ONS combined lung and bowel cancer incidence estimates by sex were similar to cancer registry estimates.
Linked CPRD-HES estimates for combined Type 1 and Type 2 diabetes were consistently higher than those of CPRD alone, with the difference steadily increasing over time from 0.26% (2.99% for CPRD-HES vs. 2.73 for CPRD) in 2002 to 0.58% (6.17% vs. 5.59) in 2013. Low back pain prevalence was highest in the most deprived quintile and when compared to the least deprived quintile the difference in prevalence increased over time between 2000 and 2013, with the largest difference of 27% (558.70 per 10,000 people vs 438.20) in 2013.

**Conclusions**: We use national EHRs to produce estimates of burden of disease to produce detailed estimates by deprivation, ethnicity and geographical region. National EHRs have the potential to improve disease burden estimates at a local and global level and may serve as more automated, timely and precise inputs for policy making and global burden of disease estimation.

**Keywords**
burden of disease, electronic health records

Robert W. Aldridge1, Hannah E. R. Evans2, Alexei Yavlinsky1, Alireza Moayyeri1, Krishnan Bhaskaran3, Rohini Mathur4, Kelvin P. Jordan5, Peter Croft5, Spiros Denaxas1, Anoop D. Shah1, Ruth M. Blackburn1, Henrik Moller6, Edmond S. W. Ng3, Andrew Hughes7, Sebastian Fox7, Julian Flowers1, Jurgen Schmidt7, Andrew Hayward8, Ruth Gilbert1, Liam Smeeth3, Harry Hemingway1

1Institute of Health Informatics, University College London, London, England, NW1 2DA, UK
2University of London, St George’s University, London, SW17 0RE, UK
3Faculty of Epidemiology & Population Health, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
4Wolfson Institute of Population Health, Queen Mary University of London, London, England, EC1M 6BQ, UK
5School of Medicine, Keele University, Staffordshire, England, ST5 5BG, UK
6Cancer Epidemiology & Population Health, King’s College London, London, England, WC2R 2LS, UK
7Department of Health and Social Care, Office for Health Improvement and Disparities, London, SW1H 0EU, UK
8Institute of Epidemiology and Health Care, University College London, London, England, WC1E 6BT, UK

Abstract

Background: Electronic health records (EHRs) have the potential to be used to produce detailed disease burden estimates. In this study we created disease estimates using national EHR for three high burden conditions, compared estimates between linked and unlinked datasets and produced stratified estimates by age, sex, ethnicity, socio-economic deprivation and geographical region.

Methods: EHRs containing primary care (Clinical Practice Research Datalink), secondary care (Hospital Episode Statistics) and mortality records (Office for National Statistics) were used. We used existing disease phenotyping algorithms to identify cases of cancer (breast, lung, colorectal and prostate), type 1 and 2 diabetes, and lower back pain. We calculated age-standardised incidence of first cancer, point prevalence for diabetes, and primary care consultation prevalence for low back pain.

Results: 7.2 million people contributing 45.3 million person-years of active follow-up between 2000-2014 were included. CPRD-HES combined and CPRD-HES-ONS combined lung and bowel cancer incidence estimates by sex were similar to cancer registry estimates. Linked CPRD-HES estimates for combined Type 1 and Type 2 diabetes were consistently higher than those of CPRD alone, with the
difference steadily increasing over time from 0.26% (2.99% for CPRD-HES vs. 2.73 for CPRD) in 2002 to 0.58% (6.17% vs. 5.59) in 2013. Low back pain prevalence was highest in the most deprived quintile and when compared to the least deprived quintile the difference in prevalence increased over time between 2000 and 2013, with the largest difference of 27% (558.70 per 10,000 people vs 438.20) in 2013. 

Conclusions: We use national EHRs to produce estimates of burden of disease to produce detailed estimates by deprivation, ethnicity and geographical region. National EHRs have the potential to improve disease burden estimates at a local and global level and may serve as more automated, timely and precise inputs for policy making and global burden of disease estimation.

Keywords
burden of disease, electronic health records
Introduction

Over 98% of the population in England are registered with a general practice. Almost all general practices use Electronic Health Records (EHRs) during their individual consultations to record clinical diagnoses, symptoms, lab results, tests, referrals to other specialties and prescriptions. In this analysis we also consider data captured during visits to secondary care hospitals, national disease registries and death records as EHRs.

These high coverage data could be used to generate automated, timely and detailed burden of disease estimates for national and local policy makers. The algorithms used to create such estimates could be made openly accessible allowing estimates to be widely, regularly and consistently produced. However, Global Burden of Disease studies in England found that accurate local data on mortality was better than for morbidity - a situation that is problematic for estimating burden for conditions with high levels of morbidity but low mortality such as low back and neck pain, skin and subcutaneous diseases, and depressive disorder. Primary care care records linked to data from secondary care and mortality records may provide a comprehensive single source of consistently collected national data that can be used to give a detailed and consistent picture of disease burden over time for a diverse range of conditions.

National EHRs in England have been used extensively to conduct research studies and these data are available for long time periods (for example, Hospital Episode Statistics from 1997, and Clinical Practice Research Datalink from 1987). Whilst national primary care data coverage is high, there is a great deal of regional variation in the different clinical computer systems used. Additionally, the strengths and weaknesses of using linked national EHRs to produce stratified (e.g. by ethnicity, region or deprivation) disease burden estimates is uncertain.

In this study we aimed to address evidence gaps in estimating disease burden using national EHRs. To achieve this we had three objectives. First, to use a single source (primary care) and linked multi-source (primary care records linked to data from secondary care and mortality records ) national EHRs to produce routine disease burden estimates for conditions with both high mortality (cancer) and high morbidity (diabetes and low back pain). Nearly everyone with these three conditions is registered with a GP and/or uses NHS acute services and as a result, will be included in Hospital Episode Statistics even if some treatment is provided in the private sector. Second, to compare these single source national EHR estimates of disease burden to existing incidence and prevalence data from other source estimates including disease registries and national cross-sectional health surveys. These comparisons enable us to triangulate our estimates with those currently used for public health and healthcare planning purposes. Our comparison estimates were selected based upon their strengths, including: high completion and coverage levels (e.g. National Cancer Registration and Analysis Service, Quality Outcomes Framework); high quality sampling strategies (e.g. Health Survey for England); high quality of the primary care data used as a result of multiple iterations of the data gathering process and training with participating general practices (e.g. Consultations in Primary Care Archive - CiPCA). Finally, to examine the ability of single source national EHR data to produce disease burden estimates over time and stratified by age, sex, ethnicity, deprivation and region.

Methods

Study Design

We conducted a national population-based retrospective cohort study using EHRs to estimate the disease burden of four cancers (lung, colorectal, breast, prostate), diabetes (type I and II) and low back pain in England. Two cohorts were created: first a cohort containing data only from primary care records and second a nested cohort containing linked data from primary care, secondary care and national death certification registrations.

Data resources

For the first cohort we used data from the Clinical Practice Research Datalink (CPRD GOLD), a large primary care database that contains anonymized electronic health records from more than 11 million people. We used records that were classified as acceptable for research having met previously specified data quality standards and shown to be broadly representative nationally. Data within CPRD can relate to a consultation with a healthcare professional within primary care or describe other activities relating to the individual’s health or care including repeat prescriptions, lab test results or diagnoses made outside primary care. These data are recorded in the EHR using the Read code system.

Through the CPRD linkage scheme, and where practices have provided consent, data were linked at person-level to inpatient data from Admitted Patient Care (APC) Hospital Episode Statistics (HES), a dataset containing details of all hospital admissions in England. Datasets were linked with a deterministic data linkage algorithm using date of birth, gender, postcode and NHS number (a unique ten digit numerical identifier assigned to NHS patients at birth or at first interaction with the healthcare system). Linkage was done by the Medicines and Healthcare products Regulatory Agency (MHRA) and NHS England with 88% of patients with research quality records having valid NHS number and eligible for linkage in the CPRD standard linked dataset release.

Data used to estimate cancer disease burden were also linked to mortality data from the Office for National Statistics (ONS) death certification register. Clinical diagnoses and procedures recorded during hospital admissions were recorded using the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) clinical classification system. We used the Index of Multiple Deprivation (IMD), linked by the MHRA, to be able to explore how disease burden estimates varied by relative deprivation in England.
Inclusion/exclusion criteria

Our study population included male and female people registered at English general practices. Using these data, the study period we investigated was from 1st January 2000 to 31st December 2013. People contributed active follow up from the latest of: the date the primary care practice started to provide ‘Up-to-Standard’ data, the current registration date of the person, or the start of the study period. People stopped contributing active follow up from the earliest of: the last date CPRD collected data from the practice, the date the individual transferred out of the practice, date of death (recorded either in CPRD or ONS), or the end of the study. For prevalence estimates, the earliest date relating to a diagnostic read code found in either clinical, test and referral records or ICD code found within the active follow up period was used as the date of diagnosis.

Case definitions for national EHRs. We expanded existing rule-based disease phenotyping algorithms to identify cases of cancer (breast, lung, colorectal and prostate), type 1 and 2 diabetes, and low back pain. Cancer diagnoses were determined using validated Read and ICD-10 terms as outlined by Bhaskaran et al.\textsuperscript{10} Read and ICD-10 terms used to identify type 1 and 2 diabetes cases were based on work described by Eastwood et al.\textsuperscript{11}. Our low back pain estimates were derived using the same method and an expanded code list from Jordan et al.\textsuperscript{12} 2014 with the addition of exacerbation of backache (16C8.00) and lumbalgia (N142.12) codes.

Case definitions for comparison EHRs. Cancer incidence estimates were compared to those from the cancer registration statistics produced by the Office for National Statistics using the same method of standardisation (e.g. European Standard Population 2013 in 5-year age-band) to enable direct comparison\textsuperscript{13}. Table 1 summarises the sources of data used, case definitions, coverage and phenotypic depth of these cancer registry data. We used the following groups of ICD10 codes for our comparison estimates from the registry: C34 for lung, C18-C20 for bowel, C50 for breast cancer and C61 for prostate cancers. Our diabetes prevalence estimates were compared to those from estimates from Health Survey for England (HSE) and Quality Outcomes Framework (QOF). We applied HSE methods for estimating the annual diabetes prevalence. In survey year 2010 and for previous years of the survey, diabetes Type 2 was defined as being a self-reported diagnosis at aged 35 or older and not treated with insulin and Type 1 diagnosed before 35 and treated with insulin. For later surveys, 2011 and onwards, time of diagnosis was not used as a proxy measure to classify people by diabetes type. For QOF we used annual point-prevalence (percentage) for type 1 and 2 diabetes. Our low back pain estimates were compared to HSE and the Consultations in Primary Care Archive (CiPCA). The final algorithms used for all conditions are provided in the Extended Data File 2\textsuperscript{14}.

Statistical analysis

We estimated the annual incidence of the four cancers studied, point-prevalence for diabetes and consultation period-prevalence for back pain for the period 2000-2013.

For fair comparison with existing data sources and estimates, where possible we used methodology consistent with these existing data. Annual incidence and prevalence were estimated for all included people.

Population estimates were described for both cohorts and compared with the estimated population of England in 2013 (Table 2). Analyses were stratified by age, sex, deprivation, ethnicity and region. We analysed data sub-regionally by using the nine statistical regions in England as defined by ONS. Quintiles of Index of Multiple Deprivation (IMD), at the person-level, were used as a measure of deprivation. Ethnicity was determined using methods previously described by Mathur et al.\textsuperscript{15,16} with further details provided in Extended Data File 2\textsuperscript{14}. Analyses by strata were restricted to those with 5 incident/prevalent cases or more. An available case analysis approach was used for all analyses. Analyses were performed using Stata version 14 SE and Stata version 15 SE. RWA, HE, AM, KB, RM, LS had access to the database population used for the study.

Cancer estimates: first incidence. Incidence was calculated by dividing the number of incident cases by the total active follow up time (in 100,000 person years) for lung, bowel, breast and prostate cancer. Follow up time ended at the time of the first cancer. Previous work has found evidence to suggest misclassification of prevalent cases, including cancer, as incident cases recorded in the first year after registration\textsuperscript{17}. For example, a healthcare professional registering a new person with their practice may enter the date of the primary care registration rather than the retrospective date of diagnosis, which leads to overestimation of incidence rates because the event occurred before registration and the corresponding years at risk were not included. In order to avoid this potential bias, active follow up was restricted to exclude the first year after registration for all incidence estimates\textsuperscript{18}. People identified with cancer before the active follow up period began were excluded from our analyses. We used direct age-standardization, standardizing to the European Standard Population 2013 in 5-year age-bands.

Point-prevalence (estimates for diabetes). Annual point-prevalence of diabetes among those aged 17 years or older was calculated at mid-year (defined as July 2nd). The analysis was restricted to this age group to be consistent with QOF estimates. For a person to be included in the denominator for a given year they needed to be 17 years or older and be contributing active follow-up at mid-year. For a person to be included in the numerator they needed to be contributing to the active follow up and have at least one diagnosis Read code recorded before or at the mid-year point for that year\textsuperscript{19}. We calculated annual point prevalence by year of age for diabetes type 1 and 2 alone and for both types combined, the latter of which enabled comparisons with HSE and QOF estimates. When calculating prevalence estimates using linked CPRD-HES, we used CPRD dates and if diabetes was not recorded in CPRD but was recorded in HES, then the date of hospital admission recorded in HES was used.
<table>
<thead>
<tr>
<th>Data source</th>
<th>Description and coverage</th>
<th>Phenotypic depth</th>
<th>Case definition used in study</th>
<th>Scale</th>
<th>Strengths*</th>
<th>Weaknesses*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital Episode Statistics (HES)</td>
<td>Routinely collected hospital records, Available since 1987, Emergency, inpatient and outpatient records, Diagnosis using ICD10 coding</td>
<td>Phenotypic</td>
<td>Case definition used in study</td>
<td>National scale</td>
<td>National scale, Includes private practice records when carried out in NHS hospitals, Possibility for linkage to primary care and vital statistics</td>
<td>Availability since 1987, Possibility for linkage to primary care and vital statistics, Administrative information (i.e. waiting time, dates, and methods of admission and discharge)</td>
</tr>
<tr>
<td>Clinical Practice Research Datalink</td>
<td>Primary care electronic health records with linkage to secondary care settings, including disease registries and socio-economic datasets</td>
<td>Phenotypic</td>
<td>Case definition used in study</td>
<td>National - England</td>
<td>Data available from across the UK on approximately 6.9% of the UK population since 1991</td>
<td>Availability since 1991, Stratified random sample of the population, A response rate of around 60% is regularly achieved</td>
</tr>
<tr>
<td>Health Survey for England</td>
<td>Annual NHS-led survey collecting data on health, behaviour, smoking and alcohol consumption and social demographic data</td>
<td>Phenotypic</td>
<td>Case definition used in study</td>
<td>National - England</td>
<td>Data linked for some years to HES</td>
<td>Availability since 1991, Stratified random sample of the population, A response rate of around 60% is regularly achieved</td>
</tr>
</tbody>
</table>

* Data and phenotypes are based on ICD10 codes building upon previously validated phenotypes.
<table>
<thead>
<tr>
<th>Data source</th>
<th>Description and coverage</th>
<th>Phenotypic depth</th>
<th>Case definition used in study</th>
<th>Scale</th>
<th>Strengths*</th>
<th>Weaknesses*</th>
</tr>
</thead>
</table>
| National Disease Registration Service (NDRS) | • NDRS systematically collects data about cancer and tumour diseases in England from multiple data sources. NDRS includes data from histopathology and haematology services, HES, primary care including CPRD, radiotherapy, hospices, independent hospitals, screening services and death certificates. | • Patient personal data  
• Type of cancer  
• Stage of cancer  
• Treatment received  
• Histopathology and haematology records  
• Medical records (from GPs, national and independent hospitals)  
• Radiotherapy records  
• Hospice records  
• Screening service records  
• Death certificates  
• Data from other cancer registries | ICD10 codes were used for comparison estimates from the registry: C34 for lung, C18-C20 for bowel, C50 for breast cancer and C61 for prostate cancers. | • Approximately 300,000 cases annually | • Availability since 1990  
• High completion and coverage | • England only  
• Time lag between collection and data availability |
| CIPCA | • Database of anonymised medical record data from a subset of general practices in North Staffordshire, UK, maintained by Keele University. | • Diagnoses  
• Laboratory test results  
• Socio-demographic data  
• Prescriptions  
• Clinical referral information | The case definition used Read codes that included codes with the terms ‘backache’ and ‘back pain’ with no region (lower or upper) stated, with the addition of exacerbation of backache and lumbalgia codes. | | • High quality data as a result of multiple iterations of the data gathering process and training recording of practices  
• Available since 2000 | • Coding based on Read system which is not used widely outside of UK and New Zealand and will be phased out in future  
• Low volume of data  
• Prescription (i.e. the morbidity Read Code) is not directly linked to specific problems or events  
• Referral is not always directly linked to a specific problem  
• Under-recording of multiple problems presented at one contact |
| Data source                        | Description and coverage                                                                 | Phenotypic depth | Case definition used in study                                                                 | Scale                                                                 | Strengths*                                                                                       | Weaknesses*                                                                                       |
|-----------------------------------|------------------------------------------------------------------------------------------|------------------|------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| Quality Outcomes Framework (QOF)  | • QOF is a quality management and analysis system for primary care                       | • Very limited as data only available in aggregate | Annual point-prevalence (percentage) for type 1 and 2 diabetes.                                     | • UK wide and whilst voluntary coverage is high                              | • Many conditions managed only in primary care and therefore QOF data can provide source of information for these conditions | • Accuracy is dependent on coding of the disease registers  |
|                                   | • Primary purpose if for collection of data used to pay primary care practices against delivery for a set of milestones. |                  |                                                                                                  | • Primary care practices reported in QOF will also be recording in CPRD | • Coverage is relatively complete as most people are registered with primary care  | • Data only available in aggregated form                                                                 |
|                                   |                                                                                         |                  |                                                                                                  |                                                                          | • Data extracted monthly                                                    | • Comparison across primary care practices not possible as different sizes and population characteristics affect the data. |
|                                   |                                                                                         |                  |                                                                                                  |                                                                          | • Available since 2004                                                        |                                                                                                                                               |
| Office for National Statistics (ONS) death certification register. | • Legal requirement for all deaths to be registered in 5 days                           | • Cause of death | ICD Codes used for Cancer (C34 for lung, C18-C20 for bowel, C50 for breast cancer and C61 for prostate and diabetes (E10, E11, E13, E14, N08, C10). | • National - UK with complete coverage due to legal status of reporting     | • ICD 10 coding done by ONS, uniformly high quality                              | • Clinical code less accurate for older patients with co-morbidities  |
|                                   |                                                                                         | • Date and place |                                                                                                  |                                                                          | • Complete and timely                                                      | • ICD classification varies over years                                                                 |
|                                   |                                                                                         | • Dob, name, occupation |                                                                                                  |                                                                          | • Relatively accurate with uniform coding                                    | • Risk of bias in social class due to occupational advancement                       |
|                                   |                                                                                         | • Details of spouse |                                                                                                  |                                                                          |                                                                             |                                                                                                                                               |
|                                   |                                                                                         | • Usual address   |                                                                                                  |                                                                          |                                                                             |                                                                                                                                               |

*strengths and weaknesses specifically in relation to disease burden estimation
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>CPRD Active 2000–2013</th>
<th>CPRD-HES Ethnicity Complete (Active 2013)</th>
<th>CPRD-HES Ethnicity Active 2013</th>
<th>CPRD-HES Ethnicity IMD Active 2013</th>
<th>CPRD-HES Ethnicity Active 2013</th>
<th>CPRD-HES Ethnicity IMD &amp; Ethnicity complete (Active 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;16</td>
<td>146,394,220.4%</td>
<td>112,606,660.11%</td>
<td>30,155,257.5%</td>
<td>24,420,655.3%</td>
<td>24,394,731.2%</td>
<td>19,438,966.4%</td>
</tr>
<tr>
<td>16–24</td>
<td>805,153,011.2%</td>
<td>630,643,311.3%</td>
<td>277,566,121.4%</td>
<td>227,942,513.1%</td>
<td>227,245,121.3%</td>
<td>195,910,513.9%</td>
</tr>
<tr>
<td>25–34</td>
<td>1,125,099,316.9%</td>
<td>937,301,211.4%</td>
<td>364,507,314.1%</td>
<td>272,942,513.1%</td>
<td>272,245,121.3%</td>
<td>215,953,121.3%</td>
</tr>
<tr>
<td>35–44</td>
<td>3,120,109,341.5%</td>
<td>2,568,904,152.3%</td>
<td>876,711,214.1%</td>
<td>697,489,214.1%</td>
<td>697,240,814.1%</td>
<td>543,200,814.1%</td>
</tr>
<tr>
<td>45–54</td>
<td>9,844,109,312.8%</td>
<td>8,409,057,312.8%</td>
<td>3,353,933.13%</td>
<td>2,686,421.33%</td>
<td>2,684,109.33%</td>
<td>2,093,109.33%</td>
</tr>
<tr>
<td>55–64</td>
<td>40,109,663,149.1%</td>
<td>39,299,312,149.1%</td>
<td>13,396,149.1%</td>
<td>10,489,312.149.1%</td>
<td>10,487,109.33%</td>
<td>7,783,109.33%</td>
</tr>
<tr>
<td>65–74</td>
<td>4,983,351,697.9%</td>
<td>4,863,060,697.9%</td>
<td>1,639,708.697.9%</td>
<td>1,378,060.697.9%</td>
<td>1,375,809.697.9%</td>
<td>1,005,809.697.9%</td>
</tr>
<tr>
<td>75+</td>
<td>5,811,028,181.1%</td>
<td>5,697,926,181.1%</td>
<td>2,160,798.181.1%</td>
<td>1,782,926.181.1%</td>
<td>1,780,709.181.1%</td>
<td>1,293,798.181.1%</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3,595,680,311.4%</td>
<td>3,455,930,311.4%</td>
<td>1,205,919,311.4%</td>
<td>1,021,460,311.4%</td>
<td>1,021,240,814.1%</td>
<td>731,740,814.1%</td>
</tr>
<tr>
<td>Female</td>
<td>3,595,680,311.4%</td>
<td>3,455,930,311.4%</td>
<td>1,205,919,311.4%</td>
<td>1,021,460,311.4%</td>
<td>1,021,240,814.1%</td>
<td>731,740,814.1%</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North West</td>
<td>937,043,139.3%</td>
<td>879,288,139.3%</td>
<td>297,563,139.3%</td>
<td>241,563,139.3%</td>
<td>241,433,139.3%</td>
<td>187,563,139.3%</td>
</tr>
<tr>
<td>Yorkshire &amp; the Humber</td>
<td>3,725,354,461.6%</td>
<td>3,459,354,461.6%</td>
<td>1,205,919,461.6%</td>
<td>1,021,460,461.6%</td>
<td>1,021,240,814.1%</td>
<td>731,740,814.1%</td>
</tr>
<tr>
<td>East Midlands</td>
<td>3,414,851,479.7%</td>
<td>3,193,851,479.7%</td>
<td>1,093,851,479.7%</td>
<td>917,851,479.7%</td>
<td>917,631,479.7%</td>
<td>691,631,479.7%</td>
</tr>
<tr>
<td>West Midlands</td>
<td>760,045,120.3%</td>
<td>691,080,120.3%</td>
<td>237,658,120.3%</td>
<td>192,658,120.3%</td>
<td>192,438,120.3%</td>
<td>148,438,120.3%</td>
</tr>
<tr>
<td>East of England</td>
<td>870,072,159.2%</td>
<td>804,012,159.2%</td>
<td>270,592,159.2%</td>
<td>222,592,159.2%</td>
<td>222,372,159.2%</td>
<td>168,372,159.2%</td>
</tr>
<tr>
<td>London</td>
<td>1,189,145,162.5%</td>
<td>1,052,145,162.5%</td>
<td>373,635,162.5%</td>
<td>306,635,162.5%</td>
<td>306,415,162.5%</td>
<td>232,415,162.5%</td>
</tr>
<tr>
<td>South East</td>
<td>1,191,845,263.9%</td>
<td>1,052,145,263.9%</td>
<td>373,635,263.9%</td>
<td>306,635,263.9%</td>
<td>306,415,263.9%</td>
<td>232,415,263.9%</td>
</tr>
<tr>
<td>Characteristic</td>
<td>Active 2000–2013</td>
<td>Active 2013</td>
<td>Active 2013</td>
<td>Active 2013</td>
<td>Active 2013</td>
<td>Active 2013</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>South West</td>
<td>733,294 (12.0%)</td>
<td>332,782 (11.0%)</td>
<td>304,289 (12.5%)</td>
<td>240,629 (12.6%)</td>
<td>24,425 (1.1%)</td>
<td>2,439,473 (12.0%)</td>
</tr>
<tr>
<td>IMD</td>
<td>5,588,757 (13.1%)</td>
<td>5,548,979 (12.9%)</td>
<td>2,440,315 (9.9%)</td>
<td>2,394,731 (9.7%)</td>
<td>2,394,731 (9.7%)</td>
<td>1,19,382 (4.7%)</td>
</tr>
<tr>
<td>Q1</td>
<td>1,226,812 (22.2%)</td>
<td>1,226,812 (22.2%)</td>
<td>558,098 (22.6%)</td>
<td>553,257 (22.4%)</td>
<td>553,257 (22.4%)</td>
<td>553,257 (22.4%)</td>
</tr>
<tr>
<td>Q2</td>
<td>1,226,812 (22.2%)</td>
<td>1,226,812 (22.2%)</td>
<td>558,098 (22.6%)</td>
<td>553,257 (22.4%)</td>
<td>553,257 (22.4%)</td>
<td>553,257 (22.4%)</td>
</tr>
<tr>
<td>Q3</td>
<td>1,109,956 (20.0%)</td>
<td>1,109,956 (20.0%)</td>
<td>553,383 (22.2%)</td>
<td>553,383 (22.2%)</td>
<td>553,383 (22.2%)</td>
<td>553,383 (22.2%)</td>
</tr>
<tr>
<td>Q4</td>
<td>989,336 (16.0%)</td>
<td>989,336 (16.0%)</td>
<td>483,249 (19.7%)</td>
<td>483,249 (19.7%)</td>
<td>483,249 (19.7%)</td>
<td>483,249 (19.7%)</td>
</tr>
<tr>
<td>Q5</td>
<td>893,322 (16.0%)</td>
<td>893,322 (16.0%)</td>
<td>488,259 (17.5%)</td>
<td>488,259 (17.5%)</td>
<td>488,259 (17.5%)</td>
<td>488,259 (17.5%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>4,519,253 (10.1%)</td>
<td>4,007,643 (9.1%)</td>
<td>3,551,096 (8.6%)</td>
<td>3,551,096 (8.6%)</td>
<td>3,551,096 (8.6%)</td>
<td>3,551,096 (8.6%)</td>
</tr>
</tbody>
</table>

**South West**

- IMD: 5,588,757 (13.1%)
- Q1: 1,226,812 (22.2%)
- Q2: 1,226,812 (22.2%)
- Q3: 1,109,956 (20.0%)
- Q4: 989,336 (16.0%)
- Q5: 893,322 (16.0%)

**Ethnicity**

- White: 4,519,253 (10.1%)
- South Asia: 893,322 (2.3%
- Black: 102,745 (0.3%
- Other: 57,184 (1.3%)
- Mixed: 57,184 (1.3%)

**Table notes:**

- Acceptable patients registered with English GP and standard practices among male and females in our study period (2000 and 2013).
- Patients at mid year are in their first year since current registration; these patients are excluded. Note: these are the patients that contribute to 1-year consultation prevalence and incidence.
- IMD is an index of multiple deprivation for the 2013 census year.
- There isn’t any publicly available estimates for population by ethnicity for mid-2013 so 2011 was used.

**Data sources for national estimates:**

- Age and Gender: [mid-2013](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/series/adhoc/005676/Englishpopulationestimatesanddeathsbysexandsingleyearofage19932013)
- Region: [mid-2013](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/series/adhoc/005676/Englishpopulationestimatesanddeathsbysexandsingleyearofage19932013)
- Ethnicity: [mid-2011](http://www.nomisweb.co.uk/).
Period-prevalence (estimates for low back pain). Period prevalence, defined as a 1-year consultation prevalence, was estimated for lower back pain. The denominator was the total number of active people at mid-year (defined as 2nd July) for each year of the study. Among these people, where one or more consultation for back pain was recorded within an individual’s active follow-up period, then the person would be counted once in the numerator. Consistent with the methodology applied to incidence, active follow-up was restricted to follow-up after one year from registration with the practice. People that were in their first year since primary care practice registration at mid-year were therefore excluded from the numerator and denominator for that year. Primary care records from CPRD were defined as consultation records if they were classified as either a face-to-face or telephone consultation (Extended Data File 2). Primary care consultations were restricted to those that had a low back pain Read code in either their clinical, test or referral records. These additional codes did not exist at local level for this previous study. Secondary care admissions were restricted to those with an ICD-10 code recorded with any diagnosis (code list given in Extended Data File 2).

Results
There were 7.2 million acceptable people contributing 45.3 million person-years of active follow-up between 2000–2013 (Table 2; CPRD Active 2000–2013). Our nested primary care cohort (CPRD-HES Active 2000–2013) linked data to secondary care and death registry data and included 5.6 million people contributing 35.6 million person-years of follow-up between 2000–2013.

At mid-2013 there were approximately 3.0 million active CPRD individuals. This equates to 5.6% (3,015,525 / 53,865,817) of the English population. The median follow-up time for CPRD active people was 9.0 years (IQR 3.5 - 13.4). Of these active individuals, 2.4 million were linked to HES and of these linked people, 99.9% had Index of Multiple Deprivation (IMD) data recorded and 79.6% had ethnicity recorded.

In the following sections we describe our results for each condition in the following order. First, we present a description of our linked and unlinked estimates of disease burden for each disease over time. Second, we compare our new estimates of disease burden to existing incidence and prevalence data. Third, we describe the findings of stratified estimates (e.g. age, sex, region and IMD deprivation groups) from our results to comparator estimates.

Cancer incidence
Estimates of incidence of lung, bowel, prostate and breast cancer using CPRD were consistently lower than CPRD-HES and CPRD-HES-ONS. The smallest difference of 12.4 per 100,000 people (170.4 vs 158.0) in breast cancer estimates in 2006, and the largest difference of 56.53 (182.6 vs 126.1) in prostate cancer in the year 2001 when comparing CPRD-HES-ONS to CPRD alone (Figure 1).

CPRD-HES and CPRD-HES-ONS lung and bowel cancer estimates by sex were similar to cancer registry estimates. For prostate and breast cancer, CPRD-HES and CPRD-HES-ONS provided higher estimates than those produced by the cancer registry, with the greatest difference of 37.6 per 100,000 people (214.7 vs 177.1) found in prostate cancer estimates in 2012 (CPRD-HES-ONS).

Figure 2 provides cancer incidence using national EHR data compared to cancer registry estimates stratified by age. This shows that age-stratified estimates provide similar results to those created using registry data. The 95% confidence intervals of age categorised CPRD incidence estimates overlapped with registry estimates for women aged 35–39, lung cancer in women aged 40–44 and 45–49, breast cancer among women aged 45–50 and 60–64 and prostate cancer among men aged 80–84 (Extended data file 1 Figure S1). All other unlinked CPRD incidence estimates by age were lower than the registry estimates (Extended data file 1 Figure S2). We estimated the incidence of cancer by ethnicity, region and deprivation (IMD) for lung, colorectal, breast and prostate (Extended data file 1 Figures S3 – S6).

Diabetes
Point-prevalence estimates using CPRD and CPRD-HES for combined type 1 and type 2 diabetes increased steadily over time. For CPRD data, this increase was from 2.1% (95%CIs: 2.0 - 2.2%) in 2000 to 5.6% (95%CIs: 5.5 - 5.7%) in 2013 (Figure 3a). Linked CPRD-HES estimates for combined type 1 and type 2 diabetes were consistently higher than those of CPRD alone, with the difference steadily increasing over time from 0.26% (2.99% for CPRD-HES vs. 2.73% for CPRD) in 2002 to 0.58% (6.17% vs. 5.59%) in 2013. Point prevalence estimates of combined type 1 and type 2 diabetes using CPRD-HES increased from 2.3% (95%CI 2.2 - 2.5%) in 2000 to 6.2% (95%CI 6.0 - 6.3%) in 2013. Point prevalence estimates for diabetes type 2 using CPRD-HES increased steadily from 2.0% (95%CI 1.8 - 2.1%) in 2000 to 5.7% (95% CI 5.5 - 5.8%) in 2013.

In 2006 and 2007, QOF estimates were considerably lower than CPRD and CPRD-HES estimates (the largest difference in 2007 of 0.62% and 0.92%, respectively), but from 2008 to the end of the study period QOF estimates were similar to CPRD-HES estimates. We compared our estimates to those from the Health Survey for England (HSE) study to CPRD and CPRD-HES. HSE showed an increase in type 2 prevalence but did not show a similar increase in type 1 (Extended Data File 2). HSE weighted estimates for type 1 and type 2 diabetes were very similar over time to our CPRD-HES estimates. For type 1 diabetes, CPRD-HES estimates were lower than HSE by 0.1% (0.4% vs 0.5%) in 2010 and 2011, and higher by 0.1% (0.5% vs 0.4%) in 2012. For type 2 diabetes, CPRD-HES estimates were lower than HSE in 2010 and 2011 (by 0.1% - 5.1% vs 5.2% and 0.2% - 5.3% vs 5.5% respectively) and the same as HSE in 2012 (5.4%) and 2013 (5.7%).
Figure 1. Directly standardized by age estimates of cancer for male and females (2000-2013) comparing estimates created using the following: 1) primary care records; 2) linked primary care and hospitalisation records; 3) linked primary care, hospitalisations and death records; and 4) National bespoke cancer registry records.

Between 2000 and 2013, prevalence of diabetes type 2 increased over time for each age group (Figure 3d). Prevalence was similar between the 65–74 and 75+ age groups up to 2007 (0.5% greater among 75+ in 2007 - 12.5% vs 12.0%). After 2007 there was a reduction in the relative increase of prevalence over time and by 2013 the difference between these age groups was greatest and 3.5% greater in the 75+ age group (16.9% vs 13.4%). Prevalence of type 2 was highest in the most deprived quintile for IMD and increased more in this group compared to the least deprived, widening the difference between these quintiles over time. Prevalence in the least deprived group increased from 1.5% (95%CI, 1.2 - 1.8%)
Prevalence of Diabetes Type 2 increased in all ethnic groups over time. South Asians had the highest prevalence across all years. The groups with the highest increase during the study period were White, South Asian and Black with a 3.7%, 3.5% and 3.1% increase respectively between 2000 and 2013. Geographically, prevalence was generally higher in the North of England and the Midlands compared to the South and East of England. Prevalence estimates of type 2 diabetes using CPRD and CPRD-HES showed similar patterns by region in 2013 (Figure 3f).

Low back pain
We observed a small increase in low back pain consultation prevalence when comparing CPRD-HES to CPRD estimates alone (Figure 5a). In 2013 the prevalence was 476 (95% CI 464 - 488) for CPRD-HES and 453 (95% CI 442 - 464)
Per 10,000 persons using CPRD alone. Low back pain prevalence increased over time but appeared to plateau in 2008 and was followed by a small decrease from 495 (95% CI 483 - 506) per 10,000 persons in 2009 to 476 (95% CI 464 - 488) in 2013 (CPRD-HES).

Prevalence of low back pain increased with age using national CPRD and CPRD regional data for the West Midlands (Figure 4). Our prevalence estimate of low back pain for the CPRD region of West Midlands were generally consistent with 95% confidence intervals overlapping estimates from...
Low back pain estimates increased up to age group 45-64 for males and females for our national EHR (CPRD-HES) estimates and CiPCA. Prevalence estimated using linked CPRD-HES data was 38% greater among women (prevalence ratio 1.38 (95%CI 1.37 - 1.40) with a prevalence difference of 153.6 per 10,000 people (534.3 vs 380.8) compared to men in 2013 - a trend that was consistent over time. Between 2000 and 2013 there was an overall increase in prevalence among those with white and ‘other’ ethnicity, while there was a decline among those with south asian, black and mixed ethnicity, with black, south asian and white ethnicities becoming similar in 2013 (Figure 5c). The largest difference between any two ethnic groups was 252.5 per 10,000 people, or 67% (PR 1.67 (95%CI 1.54 - 1.82)), between the south asian and mixed ethnicities in 2008 (627.4 vs 374.9). This difference was consistent across time. Prevalence in the most deprived quintile increased over time between 2000 and 2013 (Figure 5f), with the largest difference of 27% (PR 1.27 (95%CI 1.25 - 1.30; 558.7 vs 438.2) in 2013 compared to the least deprived quintile. In 2013, prevalence was higher in the North of England compared to the South regions of England, but confidence intervals were overlapping except for comparisons between the North West and Yorkshire and Humber compared to London. For these categories, the respective CPRD-HES prevalence differences with London were 89.6 per 10,000 people and 156.8, or 20% (PR 1.20 (95%CI 1.18 - 1.23); 527.9 vs 438.3; North West) and 36% (PR 1.36 (95%CI 1.31 - 1.41); 595.2 vs 438.3; Yorkshire and Humber) greater than London (Figure 5e).

Discussion
Our study uses national EHRs for routine surveillance of disease burden. We used data from more than seven million people to estimate disease burden for three diverse conditions that have either high levels of morbidity, mortality or both. Our results address important evidence gaps in estimating disease burden estimation using a single source of national electronic health records and meet our three study objectives. First, we have used national EHRs to produce estimates for conditions with a high mortality such as cancer and high
Figure 5. Prevalence estimates (per 10,000 person) for Low Back Pain (LBP) by (a) source, (b) gender, (c) ethnicity, (d) age group, (e) region and (f) Index of Multiple Deprivation (IMD). Primary care records used for all figures except (a) and (e) where unlinked primary care and linked primary care and hospitalisation records were used.
morbidity such as diabetes and low back pain. We undertook this work building upon existing validated disease phenotypes and applied these to national EHRs. In future it will be possible to relatively quickly reapply our code to update estimates of these three sets of conditions, but also expand on the number of phenotypes to provide more comprehensive estimates for a wider range of conditions. With the move to SNOMED coding lists in primary care - a structured clinical vocabulary for use in an electronic health record - Read code morbidity lists will need to be converted to SNOMED.

Second, for cancer and diabetes, we have compared our single source national EHR estimates of disease burden to current national incidence and prevalence data that have high completion and coverage (e.g. National Cancer Registration and Analysis Service, Quality Outcomes Framework) and shown that these estimates are broadly comparable. Our prostate and breast cancer were higher than cancer registry estimates, which may reflect the true burden of disease as a result of under-reporting of these common conditions in the cancer register, or they may be false over-estimates. Falsey inflated results may occur due to misclassification bias within primary care records. For example, positive prostate specific antigen results may initially be incorrectly classified as prostate cancer in primary care records, but actually be due to benign prostate enlargement. This misclassification is less likely to occur within the cancer registry data. Read codes used to record diagnoses in primary care can be recorded using non-specific codes; e.g. codes that specify back pain rather than low or upper back pain or diabetes rather than type 1 and type 2 diabetes. In order to estimate specific and more meaningful morbidity prevalence estimates, we encourage GPs to record completely and accurately. Our CPRD-HES estimates for type 1 and type 2 diabetes were very similar over time to those from Health Survey for England which is produced using self-reported data from a stratified random sample of the population. Our prevalence estimates for lower back pain were generally consistent with estimates from CiPCA, which includes high quality primary care data as a result of multiple iterations of the data gathering process and training with participating general practices. Our results are also consistent with more recently published data.

Finally, we have demonstrated that national EHR data from across healthcare settings can produce disease burden estimates over time and stratified by age, sex, ethnicity, deprivation and region. Our results provide national estimates of lower back pain, and in addition to this, provide important data on subgroups. Our data indicate that South Asian and Black ethnicities had highest prevalence estimates for low back pain and that the burden was highest in the lower socio-economic groups.

A weakness of the data used are that they are derived from one primary care computer system, which is distributed unequally across the UK with low representation in the north-east and Midlands. Future work should develop a framework (a governance framework and a harmonised data model) to enable these studies to be run on general practice databases from all major systems. We did not cluster by general practice in the construction of the confidence intervals, which would have resulted in wider confidence intervals than those currently provided. There were high levels of missing data for ethnicity, and our data were overrepresented for white, and underrepresented for south Asians and black populations. Our results showed variation by ethnicity and as a result our national and age group estimates, which did not control for ethnicity, may be biased towards levels found in white ethnic groups. Missing data appeared not to have caused substantial bias in the comparison of EHR-based estimates versus those from national registries or survey in the outcomes being investigated in this work. However, missing data could be an issue with other EHRs, even in high population coverage settings, when certain variables are not missing completely at random, which should be considered when exploring the potential automated derivation of disease burden estimates from EHRs. Secondary, tertiary, or further illness or health conditions in patients with high levels of co-morbidities may not be comprehensively recorded in busy general practice consultations, especially for elderly patients or patients with difficulties with communicating. EHR-based prevalence or incidence estimates of these illnesses or conditions, and for certain subgroups of patients, may be underestimated in the EHR as a result.

Our proposed methods have several strengths over existing disease burden data sources. Compared to cross-sectional surveys, our methods do not rely on self-reported outcomes, are clinician coded, and once validated, could be produced at lower annual costs on a routine basis. The data are national in scale and are detailed enough to produce small local area estimates without the need for substantial statistical modelling.

A growing number of studies are using primary care data linked to HES data. It is encouraging that the data are broadly representative nationally and that estimates from this study provide temporal trends comparable with other work to-date. Our data support a growing body of evidence from other studies for the validity of national EHRs for estimating disease burden from a diverse range of conditions including cardiovascular disease, bronchiectasis, obesity, and learning disabilities and a chronological map of 308 physical and mental health conditions. This growing body of evidence suggests that this approach is generalisable to a wide range of conditions and we envisage that health data providers such as CPRD and NHS England will increasingly be able to provide data within shorter periods of time for pre-approved studies. Expanding this approach could enable continuous collection of registry data for some conditions, instead of using intermittent validation surveys, or augmenting national EHR where needed with individual or clinician reported information. Linked EHRs could be used to produce a picture of the individual journey through primary care and secondary care and these results would allow opportunities for primary, secondary and tertiary prevention to be identified and acted upon. Population level estimates of multiple morbidity, for which we currently lack data in England, could also be produced using these data. In countries with health systems with
high population coverage and complete linked electronic health records, automated, timely and precise disease burden estimates could potentially be available based on use of existing data.

---

Data availability

Underlying data

This study used data provided by the CPRD for health research purposes. Access to CPRD data, including UK Primary Care Data, and linked data such as Hospital Episode Statistics, is subject to protocol approval via CPRD’s Research Data Governance (RDG) Process.

Extended data


This project contains the following extended data:

- Extended Data File 1: Supplementary Figures

References


PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text