COPD in England: a comparison of expected, model-based prevalence and observed prevalence from general practice data†

Luis Nacul¹, Michael Soljak², Edgar Samarasundera², Nicholas S Hopkinson³, Eliana Lacerda⁴, Tejal Indulkar², Julian Flowers⁵, Hannah Walford⁵, Azeem Majeed²

¹Foundation for Genomics and Population Health, Cambridge CB1 8RN, UK
²Department of Primary Care and Public Health, Imperial College London, London W6 8RP, UK
³National Heart and Lung Institute, Imperial College London, London W6 0YD, UK
⁴Vale do Sao Francisco University, Petrolina 56304-917, Brazil
⁵Eastern Region Public Health Observatory, Cambridge CB2 0SR, UK
Address correspondence to Dr Michael Soljak, E-mail: m.soljak07@imperial.ac.uk

ABSTRACT

Background Primary care data show that 765 000 people in England have a general practice (GP) diagnosis of chronic obstructive pulmonary disease (COPD). We hypothesized that this underestimates actual prevalence, and compared expected prevalence of COPD for English local authority areas with prevalence of diagnosed COPD.

Methods Cross-sectional comparison of GP observed and model-based prevalence estimates (using spirometry data without clinical diagnosis) from the Health Survey for England. Local underdiagnosis of COPD was estimated as the ratio of observed to expected cases. We investigated geographical patterns using classical and geographically weighted regression analysis.

Results Both observed and expected prevalence of COPD varied widely between areas. There was evidence of a ‘north–south’ divide, with both observed and modelled prevalence higher in the north. The ratio of diagnosed to expected prevalence varied from 0.20 to 0.95, with a mean of 0.52. Underdiagnosis was more pronounced in urban areas, and is particularly severe in London. The inclusion of GP numbers in the analysis yielded a stronger regression relationship, suggesting primary care supply affects diagnosis.

Conclusion Both observed and modelled COPD prevalence varies considerably across England. Cost-effective case-finding strategies should be evaluated, especially in areas where the ratio of observed to expected cases is low.

Keywords chronic obstructive pulmonary disease, epidemiological, prevalence, spatial analysis, statistics

Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by progressive airflow obstruction, which is incompletely reversible.¹ Diagnosed prevalence and morbidity data underestimate the total burden of COPD because the disease is usually not diagnosed until it is clinically apparent, and there is considerable variation in reported prevalence.² Mortality data also underestimate COPD as the cause of death.

COPD case finding can be highly cost-effective. The 2004 NICE COPD guidance states that the incremental cost-effectiveness ratio of opportunistic case finding is a cost per
quality adjusted life year of only £814, which is much lower than many NHS-funded interventions. Moreover, an audit of COPD admissions in England showed that over two of three of winter admissions for COPD were of new patients not admitted in the previous year with the condition. Admission rates could be reduced by better identification and management. Stop smoking prevents the development of COPD, or slows its progress and reduces the risk of hospital admission. Receiving a diagnosis or just an estimate of ‘lung age’ increases quit rates, even in the absence of symptoms, and there is evidence demonstrating the effectiveness of stop-smoking services delivered at the time of COPD diagnosis. After 1 year, quit rates are still higher for people with diagnosed airflow obstruction.

Models using smoking rates to estimate COPD prevalence have been previously proposed but none has direct relevance to the UK, and they fail to take into account other relevant risk factors. We developed a multivariate model to estimate the expected prevalence of COPD in England, based on the data from the Health Survey for England (HSfE), which we then used to produce local prevalence estimates. We present the estimated population prevalences of COPD for local authorities (LAs) in England, and compare them with the prevalence of general practice (GP)-diagnosed COPD, as reported in national GP Contract Quality and Outcomes Framework (QOF) data. We investigated geographical patterns in the association between observed and modelled prevalence using classical and geographically weighted regression (GWR).

Methods

Model development methods

Model development has been reported previously. The HSfE is a representative population-based annual survey, which in 2001 included the assessment of respiratory function using spirometry, as well as comprehensive data on risk factors. The 2001 data refer to 5269 men (98%) and 6133 women (95%) over 15 years with valid spirometry conducted by an HSfE nurse. Additional data for multivariate analysis were available for 94.3% of the sample. COPD was defined using the British Thoracic Society (BTS) criteria: forced expiratory volume in 1 s (FEV₁) divided by forced vital capacity (FVC) under 0.70, and FEV₁ < 80% of predicted using reference values from the HSfE. However in the HSfE, spirometry was not carried out after bronchodilator challenge. The variables included in the model, based on their association with COPD in logistic regression analysis, were age group, gender, ethnicity, smoking prevalence, area of residence (rural, suburban or urban) and area-based deprivation score.

The baseline odds of COPD in non-smokers under 35 years old were obtained directly from the data set. The strength of association between each explanatory variable and COPD caseness was used to calculate the relative odds, which were applied to the baseline odds to derive the prevalence estimates for subgroups of risk factors. The main results are expressed as expected/predicted prevalence of COPD for population subgroups. We applied the model to obtain the total COPD prevalences for 354 LAs in England. We used 2006 mid-year quinary and ethnic population estimates from the UK Office for National Statistics (ONS). The latter data set provides estimates in wider age bands, so we assumed that the distribution of ethnic populations was uniform across quinary age bands. We stratified LAs into deprivation score quintiles and three categories of urbanity/rurality (urban, suburban and rural) based on the ONS system used in the HSfE. Local smoking prevalence estimates are not available from the HSfE because of small sample sizes, so we used model-based estimates from the neighbourhood statistics website.

We also present data on the GP-diagnosed and registered prevalence of COPD, obtained from the QOF, and COPD-specific directly standardized mortality rates for comparison. QOF COPD prevalence estimates are based on populations registered with GPs. We derived residence-based registered prevalence estimates for LAs using a lookup table—a pooled extract of England GP registers—from the National Strategic Tracing Service (NSTS), which matched GP practice populations to LA areas as of January 2006. We apportioned counts of COPD patients by practice to LAs, in accordance with the proportion of each practice population in that area, assuming that COPD prevalence was geographically uniform across a practice population. We divided this count of COPD patients by the mid-year LA population to give estimated crude prevalence. Where less than 50 patients fell into an LA, the numbers were excluded from the look-up process. Three LAs could not be mapped due to discrepancies between QOF and NSTS data sets.

Model validity

We externally validated the model by comparing COPD expected prevalence results to an alternative model, based on a survey of prevalence studies. The comparative prevalence was similar. We also applied the model to the Belfast, Northern Ireland population, and compared the results with those from a population survey of the same population. Our results were slightly lower, but within the 95%
The ratio of the QOF COPD observed or diagnosed prevalence to the expected prevalence was used to estimate the unmet needs for COPD diagnosis. We compared the ratios across LAs, English regions and within population subtypes (rurality and index of multiple deprivation quintiles), to identify the possible risk factors for underdiagnosis. Simple and multiple linear regression analysis were used to investigate the correlations between diagnosed and expected prevalence; and between these and directly standardized mortality rates for COPD. Geographical clustering in the ratio was identified using the Local Moran's I cluster test, which measures associations with neighbouring areas. The data were analysed using Microsoft Excel 2007, Stata 9 and ArcGIS 9 software.

In classical regression we assume that a modelled relationship holds consistently across the study area. In many situations this is not the case, a phenomenon known as spatial non-stationarity. We used GWR3 software to perform GWR modelling to investigate the spatial variation in the association between diagnosed:expected ratio. GWR is a local version of regression that generates parameters disaggregated by the spatial units of analysis. This allows assessment of spatial heterogeneity in estimated relationships between independent and dependent variables. Mapping GWR local correlation coefficients (strength of association), residuals (the difference between observed and predicted data) and Cook's D values (the effect of removing a single data observation from the data set, a test of model robustness) provides an elegant method for modelling geographical variation in regression relationships. The disaggregation of the full data set into spatial units for analysis is performed using a probability kernel, which moves across the study area.

In order to estimate the optimal kernel bandwidth we used the Akaike information criterion. We compared classical regression with GWR by testing for spatial non-stationarity using a Monte Carlo simulation. The results of the GWR analysis were mapped to visualize any geographic variation in the diagnosed:expected relationship. Two GWR models were run, the first a basic diagnosed:expected regression, and the second a bivariate regression involving diagnosed prevalence as the dependent variable with expected prevalence and primary care supply as the independent variables. The number of full-time GP equivalents per practice, sourced from the information centre, was mapped to geographical areas in the same way as QOF prevalence above, using the NSTS lookup table. This was used as a proxy for primary care supply, to explore any potential association between COPD prevalence and physician availability. Unfortunately national data on practice nurse and other staff supply are only available at the primary care trust level, and so could not be mapped as accurately to geographic areas.

**Results**

**Expected prevalence of COPD**

The overall estimated prevalence of COPD in people over 15 years old in England was 3.58%, i.e. just over 1.4 million. LA prevalences ranged from 1.88 to 6.02%, with a median of 3.14% and inter-quartile range of (IQR) 2.69–3.78. The north west and north east regions had some of the highest prevalence observed. The overall expected prevalence in all age groups was 2.58% (95% CIs = 2.49–2.66). The rate for 15–45 year olds was 1.32%, for 45–64s 4.22% and for those 65–74 and over 75 7.93 and 8.72%, respectively. The mean expected prevalence by region varied from 2.90% in the south east of England to 4.02% in the north east.

**Diagnosed prevalence of COPD**

The mean prevalence of QOF-diagnosed COPD in LAs was 1.37% (95% CIs = 1.33–1.42), and varied between 0.65 and 3.13% (median: 1.29%, IQR: 1.05–1.61). Figure 1 shows the geographical distribution of diagnosed COPD. There is a north–south gradient; the prevalence of diagnosed COPD was generally low in southern England.

**Comparisons between diagnosed and expected prevalences, and mortality**

Figure 1 illustrates the geography of diagnosed and expected prevalence. We found a moderate correlation between diagnosed and expected prevalence of COPD in people over 15 years old ($r = 0.48, P < 0.001$). The expected prevalence also showed a significant association with mortality risk ($r = 0.24, P < 0.001$) and diagnosed prevalence similarly ($r = 0.21, P < 0.001$).
The ratio of diagnosed to expected prevalence in each LA varied from 0.20 to 0.95, with a mean of 0.52 (95% CIs 0.50–0.53). There was pronounced variation ($P < 0.001$) between urban (mean ratio: 0.38, 95% CIs 0.36–0.40) and rural areas (mean ratio: 0.56, 95% CIs 0.54–0.58). The mean ratio of diagnosed to expected cases in London was lowest at 0.31 (95% CIs 0.29–0.33), illustrated in Figure 1. The $\zeta$ scores from the Local Moran’s I analysis (Fig. 2) indicated that there are statistically significant clusters in the diagnosed:expected ratio for London and its hinterland. Positive spatial autocorrelation is observed for London (coloured red) and negative spatial correlation for much of its hinterland (coloured blue), suggesting that the low diagnosed:expected ratios for London is a genuine cluster with sharp contrast to its hinterland. Some other parts of England, particularly in parts of Yorkshire and the north east, also showed significant positive autocorrelation, though unlike London these clusters were associated with relatively high diagnosed:expected ratios.

Comparison of correlation coefficients between classical regression results and those of GWR, as well as the use of analysis of variance and Akaike information criterion methods, confirmed that GWR modelled the data set more accurately than classical regression (Table 1). The Cook’s D and standardized residuals were small, the former indicating robust local regression results and the latter indicating no obvious outliers; neither showed evidence of geographic clustering. The Monte Carlo tests confirmed strong spatial non-stationarity in the diagnosed:expected association ($P < 0.000001$). GWR analyses revealed clear geographic variation in the association at the LA level (Fig. 3) with the weakest associations in the West Midlands, Cornwall and part of Yorkshire, and strongest in the South East and East Anglia.
The inclusion of GP supply as an additional independent variable increased the local correlation coefficient for most LAs in both classical regression and GWR, particularly the latter (Table 1) and especially for the East Midlands.

**Discussion**

**Main findings of this study**

We compared diagnosed COPD prevalences in English LAs with those predicted by a model derived from HSfE data. The main findings were that both diagnosed and expected prevalence varied widely, and that the discrepancy between diagnosed and expected prevalences, as a measure of unmet need, also varied considerably, being significantly greater in urban than rural areas. However, the highly mobile populations of all these urban areas also may be contributing to the low level of diagnosed prevalence. Local Moran’s analysis revealed a statistically significant clustering of similar ratios comprising much of London, surrounded by a hinterland of dissimilar ratio values, indicating COPD underdiagnosis in this metropolitan area. Similar but less extensive clusters are evident in metropolitan areas of Yorkshire and the north east.

Differing patterns are evident from the Local Moran’s I and GWR analyses, with the Moran’s test highlighting clusters and outliers within regions, whilst the GWR analyses summarize variation between regions. They suggest that regional boundaries may not be optimal for epidemiological purposes and that further analyses at finer geographical scales, e.g. practice level may be warranted. The inclusion of GP supply as an
additional independent variable in the GWR analyses yielded a stronger regression relationship for most parts of the country, supporting the notion of GP numbers as a factor affecting diagnosis levels, with implications for resource allocation.

**What is already known on this topic**
Our findings are consistent with a number of recent small primary care screening studies showing COPD underdiagnosis. In a survey in the Netherlands, the prevalence of self-reported asthma or COPD (9.7%) was almost twice as high as the prevalence based on GP information (5.2%). Vandevoorde et al. screened patients in six Belgian semi-rural practices if they were current smokers between ages 40 and 70, and had a smoking history of at least 15 pack-years. Excluding those with known COPD, spirometry revealed a 29.5% prevalence, with greater underdiagnosis in younger age categories. Tinkelman et al. recruited patients via random mailing from a primary care practice in Aberdeen, Scotland, and Denver, Colorado, who were current and former smokers aged over 40 with no prior diagnosis. Of 818 patients, 18.9% had a post-bronchodilator diagnosis of COPD, which was mild in 57.4%, moderate in 36.8% and severe in 5.8%.

What are the benefits of early diagnosis? There is some evidence that effective management of COPD may reduce the risk of hospital admission. Gotfredsen et al. found a gradient of relative risk of hospital admission with increasing tobacco exposure. Telling smokers their lung age significantly improves the likelihood of them quitting smoking, and practices could be incentivized to deliver smoking cessation advice and prescriptions, or to refer to stop-smoking services. The updated NICE COPD guidelines that are currently out for consultation recommend that pulmonary rehabilitation should be made available to all appropriate patients with COPD, including those who have had a recent hospitalization for an acute exacerbation, and that newer long-acting drugs be prescribed to appropriate patients. In addition, combination treatment with long-acting beta-agonists and inhaled corticosteroids reduces the annual rate of COPD exacerbations and improves health status and spirometric values. A nurse-led self-management programme for COPD patients reduced unscheduled practice attendances.

**What this study adds**
Diagnosed COPD prevalence apparently grossly underestimates true population prevalence, a finding that supports the use of case-finding strategies. Underdiagnosis of COPD varies systematically and is more pronounced in urban areas, where the risks of COPD are also higher. Underdiagnosis appears particularly severe in London, with up to five unknown cases for each case diagnosed in the worst-performing LA. However, relatively wide CIs are involved, and combined with the results of the GWR analyses this suggests the model's results for London are less certain than those for the West Midlands, Yorkshire and north east conurbations.

GP supply appears to be a key factor affecting diagnosis levels. This suggests greater emphasis on GP-based COPD management would be beneficial, particularly in LAs such as in the East Midlands, for which the inclusion of GP supply in our model markedly improved local correlation coefficients. Nevertheless there is considerable unexplained variation in the observed to expected relationship even after the GP supply and the model variables are taken into consideration, especially in Cornwall and the West Midlands.

While opportunistic case finding may be cost-effective, screening of even high-risk groups may not. Practices may lack the resources to extend spirometry to invited patients. The variability in the ratio of diagnosed to expected prevalence may reflect patient and/or provider factors, and highlights the need

---

**Table 1** Comparison of classical regression and GWR results

<table>
<thead>
<tr>
<th>Diagnostics</th>
<th>Classical regression</th>
<th>GWR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O:E</td>
<td>O:E + GP</td>
</tr>
<tr>
<td>Residual sum of squares</td>
<td>0.007210</td>
<td>0.006633</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.004545</td>
<td>0.004366</td>
</tr>
<tr>
<td>Akaike information criterion</td>
<td>-2786.19153</td>
<td>-2813.43807</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.17881</td>
<td>0.24456</td>
</tr>
<tr>
<td>Adjusted correlation coefficient</td>
<td>0.174100</td>
<td>0.23803</td>
</tr>
<tr>
<td>Sum of squares</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>2.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

O:E of observed to expected prevalence of COPD.
O:E + GP: inclusion of GP supply as an additional independent variable.
for additional case-finding support for practices in low diagnosis areas; many younger patients may attend infrequently and may need to be invited for screening. A strategy we are exploring in London involves applying the model at practice level and providing additional lung function testing alongside active patient recall. There is a strong case for evaluating the feasibility and yield of several such case-finding strategies and using the data to further validate the predictions of the prevalence model, especially at the practice level.

**Limitations of this study**

The estimates need to be interpreted with caution and are subject to a degree of imprecision, particularly when applied to small populations. We used robust census-based demographic data for each LA, but QOF data refers to the PCTs of patients’ GPs, which does not always coincide with their LA area of residence, although they overlap considerably. We attempted to overcome this by apportioning practice populations to LAs, but this may have been erroneous in some areas. Limitations of the study may result from inaccuracies in the model parameters or assumptions, and from the quality of the input data. For example, we used current smoking status as a proxy for lifetime tobacco exposure. However, this information is not readily available and the incorporation of life-time exposure to tobacco in addition to being more difficult to obtain in routine practice, is subject to error and would be unlikely to improve the model significantly.

Post-bronchodilator spirometry is not part of the BTS case definition and was not carried out in the HSfE respondents, so the model results will slightly overestimate...
prevalence. The screening studies above show that in older age groups the proportion with bronchodilator-responsive airflow obstruction is small. We have confirmed this is a case-finding pilot study in London. However, a recent analysis of the BOLD study data found that that the spirometric prevalence of COPD was reduced by 25% after bronchodilator challenge. We carried out a sensitivity analysis by adjusting the expected prevalences accordingly, which changed the minimum value of the observed:expected ratio from 0.015 to 0.020, the maximum from 0.872 to 1.163 and the median to 0.363 to 0.484. This adjustment did not alter the findings of our study, although they slightly reduced the degree of diagnosis underestimation.

Misclassification of diagnosis at the primary care level may also have had additional minor influence on the estimates of underdiagnosis. Some people not meeting the BTS definition may be misdiagnosed as having COPD by their GPs. This would increase the diagnosed but not the expected prevalence, leading to an underestimation of the proportion with undiagnosed COPD. On the other hand, the numerator (diagnosed prevalence) will not include patients with known COPD criteria who do not appear in the QOF statistics, e.g. non-NHS patients or those for whom diagnostic Read codes have not been entered. Although there are financial rewards for GPs to record of COPD, if a GP believes that fulfilling QOF required criteria diagnosis is not in the interests of an individual patient they may avoid entering Read codes, which excludes the patient in the QOF denominator. For a QOF diagnosis such as COPD, where clinical judgement is explicitly sanctioned by NICE, deliberate systematic under-recording is inevitable.

Much of practices’ screening activity is carried out by practice nurses. However, we were unable to examine any effects of the supply of practice nurses or other staff because national data are not collected in sufficiently disaggregated form. Some PCTs commission services delivered by specialist COPD nurses, but these staff mainly provide care for patients with severe COPD, so they are unlikely to affect initial diagnosis levels.

Conclusions

GPs should be challenged to explain how good clinical judgement might account for the substantial differences in observed and expected prevalences we have shown. The distinct geographical variation suggests that these are not simply the results of good clinical judgement. Ultimately the validity and utility of the model depends on its identification of practices where there is underdiagnosis. The COPD prevalence model now available on the APHO website now contains estimates for GP practices’ registered populations, a number of PCTs are already using it as a tool for COPD case finding, and it may also be useful for health planning, and to direct resources to areas of greater need. In other countries national health surveys can be used to develop prevalence models in a similar way.

Acknowledgements

Population data for the model was provided by Eastern Region Public Health Observatory.

Funding

Development was funded by the Department of Health as part of the Informing Healthier Choices Programme, which aims to improve public health information and intelligence.

References
