Title: “A novel approach to estimate the local population denominator to calculate disease incidence for hospital-based health events in England”

Running Title: Denominator paper

Authors: James Campling¹, Elizabeth Begier², Andrew Vyse¹, Catherine Hyams³⁴⁵, Dave Heaton⁶, Jo Southern², Adam Finn³, Harish Madhava², Bradford D. Gessner², Gillian Ellsbury¹

Author affiliations:

1. Pfizer Limited, Walton Oaks, Dorking Road, Tadworth Surrey, KT20 7NS, United Kingdom
2. Vaccines Medical Development, Scientific and Clinical Affairs, Pfizer Inc, Collegeville, PA, USA
3. Bristol Vaccine Centre, Schools of Cellular and Molecular Medicine and of Population Health Sciences, University of Bristol, Bristol, BS2 8AE, UK
4. Academic Respiratory Unit, Learning and Research Building, Southmead Hospital, Bristol, BS10 5NB, UK
5. Acute Medical Unit, Southmead Hospital, Bristol, BS10 5NB, UK
6. Harvey Walsh, Open Health Group, Marlow, Buckinghamshire, SL7 2FF, UK

First corresponding author: James Campling

Email: james.campling@pfizer.com

Telephone: +44 (0) 1304 616161

Word Count: 3,055

This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.
Summary

While incidence studies based on hospitalisation counts are commonly used for public health decision-making, no standard methodology to define hospitals’ catchment population exists. We conducted a review of all published community-acquired pneumonia studies in England indexed in PubMed and assessed methods for determining denominators when calculating incidence in hospital-based surveillance studies. Denominators primarily were derived from census-based population estimates of local geographic boundaries and none attempted to determine denominators based on actual hospital access patterns in the community. We describe a new approach to accurately define population denominators based on historical patient healthcare utilisation data. This offers benefits over the more established methodologies which are dependent on assumptions regarding healthcare seeking behaviour. Our new approach may be applicable to a wide range of health conditions and provides a framework to more accurately determine hospital catchment. This should increase accuracy of disease incidence estimates based on hospitalised events, improving information available for public health decision making and service delivery planning.
INTRODUCTION

When considering the introduction of an immunisation programme, it is paramount that the incidence of the diseases of interest are estimated as accurately as possible. Calculating annual incidence rates (expressed as number of cases per 100,000 population) depends on accurate estimation of two parameters: (1) the number of people diagnosed with the disease during a specified time interval, (2) the size of the population from which the cases originated at the start of the time interval of interest. Measuring each parameter has its own challenges, but here we focus on challenges associated with estimating the size of local populations within England, hereafter referred to as the denominator. For national datasets where the catchment area is determined based on clear geographic boundaries, the denominator can be estimated using census data which are maintained through annually adjusted estimates. However, many surveillance studies use health centres such as clinics and hospitals, and in these cases the denominator population usually is not clearly defined.

To estimate healthcare facility catchment populations, a few map-based approaches have previously been proposed (e.g., defined urban conurbation area, crow-fly, road distance and road time access), [1-5] all of which rely on census data to provide population estimates based on where the boundary is drawn on the map from the given approach. However, in England, and for several reasons, geographically defined denominators may provide a poor estimate of the population accessing care at a particular health centre. The National Health Service (NHS) provides healthcare free of charge for all residents in England and allows patients to choose where they receive medical care, which is an important principle of the English healthcare system. Although geography plays an important role in influencing this choice, other factors may be important including public transport, parking, waiting times, traffic considerations both for patients and visiting family members, experience with a particular
hospital, GP recommendation, ambulance preference, hospital capacity, specialist services and hospital reputation.[6] Moreover, while it might be expected that those who live close to a hospital would preferentially choose that location, many people live equidistant to more than one hospital (both in terms of distance and travel time). In summary, no standardised methodology exists to estimate incidence based on person seeking healthcare at a given facility.

In this report, we describe a novel methodology to estimate local population denominators for the Bristol AvonCAP study — a study set up with the specific aim of measuring the burden of hospitalised respiratory disease in England, to provide evidence for informed decision making for public health interventions including vaccines, that have the potential to alleviate some of this burden. The study was designed to measure the incidence of hospitalised community-acquired pneumonia (CAP) and other acute lower respiratory tract diseases (aLRTD) in two large secondary care hospitals located in Bristol. We think this methodology could be replicated for other health outcomes and other regions in England (or elsewhere if a high level of formal primary care practice registration exists), which could substantially improve disease incidence estimates and thus accurate public health decision-making.

METHODS

Methodology Overview

The conceptual distinction between previously proposed approaches to determine population denominators and our methodology is that the former are based on assumptions about which hospitals patients are expected to use. Our new methodology attempts to minimise use of assumptions by utilising multiple data sources to assess which hospitals these populations actually have used in the past.

The NHS in England allocates an annual budget to local geographically defined clinical commissioning groups (CCGs) broadly based on population numbers and utilisation in prior years. In April 2021, there
were 106 CCGs across England and their boundaries were drawn to complement local healthcare resources.[7]

Robust systems are used by CCGs to reimburse hospital care, therefore we hypothesised that CCG geographical regions may be helpful in determining hospital catchment areas and local populations. To test our hypothesis, we utilised Hospital Episode Statistics (HES) data which were re-used with the permission of NHS Digital via Harvey Walsh Limited. aLRTD admissions at the study hospitals between April 2017 – March 2020 were linked to aggregated general practitioner (GP) data to understand from which CCG the hospitals’ patients came (Methods Part 1). Then, we estimated the proportion of patients hospitalised at the study hospitals among all patients hospitalised with LRTD for each practice and multiplied that by count of patients registered at that GP practice to calculate the Bristol hospital catchment population (Methods Part 2).

In England, all hospitalisations in NHS hospitals are captured in Hospital Episode Statistics (HES) and all acute care is provided by NHS hospitals. HES contains information on bed days, length of admission, outpatient appointments, attendances at Accident and Emergency Departments at NHS hospitals in England, discharge diagnoses and hospital death.[8] The primary diagnosis and other clinical conditions are specified using the tenth revision of the International Classification of Diseases version 10 (ICD-10).[9] Furthermore, in England a high proportion of the population are registered with General Practice where it is not possible to be registered at two practices concurrently.[10, 11]

**Method step 1 – defining GP practices associated with patients treated at study hospitals**

To understand from where patients treated at the study hospitals originated (i.e., to which CCG the patients’ GP practices belong), HES data were extracted for all adult patients coded for aLRTD between April 2017 – March 2020 and filtered to include only patients treated at the study hospitals: North
Bristol NHS Trust (NBT), and University Hospitals Bristol NHS Foundation Trust & Weston NHS Foundation Trust (UHBW). Finally, data were analysed to determine in which CCG area the patients lived based on their GP registration. There are 6 CCG regions in the South West of England within a 1-hour drive of the study hospitals, as illustrated in Figure 1.

Figure 1. South West England Clinical Commissioning Groups Map

Map showing CCGs described in the results pie chart along with location of relevant hospitals.

**Method step 2 – defining the catchment population of study hospitals**

As patients registered in the CCG might seek care at a different hospital for a variety of reasons, we did not assume every patient registered with a GP in the Bristol, North Somerset and South Gloucestershire
(BNSSG) CCG used the study hospitals. Therefore, we estimated the proportion of patients from each GP practice treated at the study hospitals among all BNSSG CCG patients, stratified by age group. This proportion was used to calculate the study hospitals’ catchment population. All aLRTD hospitalisations (based on ICD-10 codes; Appendix 1) occurring between April 2017 – March 2020 among patients registered in the BNSSG CCG were analysed by GP practice. For each GP practice, the percent of hospitalisations occurring at study hospitals was calculated within each age-group (18-34, 35-49, 50-64, 65-74, 75-84 and ≥85 years). The percentage of hospitalisations occurring at study hospitals was the number of patients at each GP practice who were admitted for aLRTD at study hospitals (study hospital aLRTD patients) divided by the total number of patients at that GP practice who were hospitalised for aLRTD at any English hospital in the time period (overall aLRTD inpatients). This proportion (i.e., percent of aLRTD inpatients using study hospitals) was multiplied by the practice population for each GP practice by age strata to provide an expected Bristol hospital catchment population contribution for each GP practice (once all age groups summed). GP populations were obtained from NHS Digital "Patients Registered at a GP Practice" data for October 2019. Finally, the catchment population contribution for each GP practice in the BNSSG CCG was combined to provide an expected total Bristol hospital catchment population. In summary, if:

- E = Calculated catchment population
- SHP = Number of patients at the GP Practice with an aLRTD hospitalisation at the study hospitals during 2017–2019
- OL = Overall patients with an aLRTD hospitalisation in the GP Practice during 2017–2019
- POP = Local GP population
- i = Each individual practice
Then:

\[ E = \sum \left( \frac{SHP_i}{OL_i} \right) POP_i \]

Drive time methodology

The BNSSG CCG uses a 20-minute drive time for their healthcare utilisation mapping purposes.\[12\] We have included this alternative methodological approach to allow comparison between our methodology and other methodologies in current use. We obtained data from the BNSSG CCG which divides the CCG region into small geographical areas used by the UK census known as lower layer super output areas (LSOA). LSOAs have a population of between 1,000 – 3,000 people or 400 – 1,200 households.\[13\] Data were filtered according to estimated drive time from each LSOA to the study hospitals according to the Automobile Association (AA) route planner, (AA, Hampshire, UK).\[14\] UK population data by LSOA for all ages (0 – ≥90 years) were downloaded from the UK Office of National Statistics census website. Population estimates were derived for the following drive times from the study hospitals 20 mins, 30 mins, 40 mins and 60 mins by matching the LSOA population data with the drivetime data.

RESULTS

In 2019, there were 82 GP practices in the BNSSG CCG. Figure 2 shows the proportion of patients that attended the study hospitals in 2019 that were registered at GP practices in both the BNSSG CCG as well as six other CCGs that, combined, represented where >99% of patients hospitalised at study hospitals were registered. The majority of hospitalised patients (96%) were registered at BNSSG CCG GP practices, with most of the remaining 4% were based in the surrounding CCGs.
Figure 2: Residential Address Stratified by CCG of Patients Attending Study Hospitals 2017-2019

Substantial variability existed by GP practice in the percent of all persons hospitalised for aLRTD who were hospitalised at a study hospital with much less variability by age (Figure 3) (based on a representative sample of 10 anonymised GP practices within the BNSSG CCG). Lower proportions were reported for GP practices that were located either close to the CCG boundary or close to Weston hospital (a non study hospital situated in the BNSSG CCG). Full tables reporting these data for all GP practices located in the BNSSG CCG for 2017, 2018, 2019 and the combined data can be found in Appendix 2.
Figure 3: A Bar Chart Showing the Proportion of Persons hospitalised for Acute Lower Respiratory Tract Disease who were Hospitalised at a Study Hospital, Stratified by Individual Anonymised General Practice and Patient Age Group.

The degree to which the estimates from our methodology compared to estimates produced by other methods varied, including within specific age groups (Table 1 and Figure 4). The total CCG population (the sum of the population of all GP practices in the CCG) overestimated the catchment population compared to our estimates by 15% to 24%. By contrast, the population living within 20 minutes underestimated the catchment population by 10% to 29%. As drive time increased linearly, the estimated population increased non-linearly such that population based on a 60-minute drive time overestimated the catchment population by 276% to 428%. The degree of underestimation or overestimation from other methods did not vary substantially by age group.
<table>
<thead>
<tr>
<th>Age Group</th>
<th>Estimated Catchment (Study Method)</th>
<th>Total CCG Catchment</th>
<th>Estimated based on ≤20 min (Study Method)</th>
<th>Estimated based on ≤25 min (Study Method)</th>
<th>Estimated based on ≤30 min (Study Method)</th>
<th>Estimated based on &lt;40 min (Study Method)</th>
<th>Estimated based on &lt;60 min (Study Method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five Adult Age Groupings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-34</td>
<td>231,342</td>
<td>268,093</td>
<td>208,924</td>
<td>238,301 (↑3%)</td>
<td>295,130</td>
<td>442,590</td>
<td>870,841</td>
</tr>
<tr>
<td>35-49</td>
<td>184,269</td>
<td>211,568</td>
<td>130,881</td>
<td>162,469</td>
<td>211,452</td>
<td>337,781</td>
<td>714,415</td>
</tr>
<tr>
<td>50-64</td>
<td>152,380</td>
<td>178,970</td>
<td>108,404</td>
<td>143,508 (↓6%)</td>
<td>196,307</td>
<td>395,130</td>
<td>732,702</td>
</tr>
<tr>
<td>65-74</td>
<td>74,245</td>
<td>89,015</td>
<td>52,954 (↓29%)</td>
<td>73,368 (↓1%)</td>
<td>102,148</td>
<td>175,757</td>
<td>391,718</td>
</tr>
<tr>
<td>75-84</td>
<td>45,989</td>
<td>55,720</td>
<td>33,712 (↓27%)</td>
<td>46,919 (↑2%)</td>
<td>55,244 (↑42%)</td>
<td>111,109</td>
<td>239,310</td>
</tr>
<tr>
<td>≥85</td>
<td>19,229</td>
<td>23,938</td>
<td>15,280 (↓21%)</td>
<td>20,400 (↑6%)</td>
<td>28,261 (↑47%)</td>
<td>47,108</td>
<td>99,865</td>
</tr>
<tr>
<td>Total</td>
<td>707,454</td>
<td>827,304</td>
<td>550,155</td>
<td>684,965</td>
<td>898,542</td>
<td>1,446,140</td>
<td>3,048,851</td>
</tr>
</tbody>
</table>

Two Adult Age Groupings

18–64 567,991 658,631 448,209 544,278 (↑4%) 702,889 1,112,166 2,317,958

≥65 139,463 168,673 101,946 140,687 (↑1%) 195,653 333,974 730,893

Total 707,454 827,304 550,155 684,965 898,542 1,446,140 3,048,851
Figure 4: Comparison of Study Hospital Population Size (≥18yrs) by Methodology

The map in Figure 5 shows the location of the study hospitals and Weston hospital. The BNSSG CCG boundary is shown in black and travel time boundaries are identified by colour to the study hospitals based on shortest travel time to either study hospital.
DISCUSSION

Incidence studies based on counts of hospitalisations from one or a few study hospitals are common, but there is no standard methodology to define a health centre’s catchment population for the purpose of accurately estimating incidence denominators. Traditional geography-based approaches (such as defining a population with a certain drive time to a study health centre) that rely on census data do not account for the nuanced ways in which populations access healthcare and therefore are prone to error. We devised a novel approach for establishing local population estimates in England to support disease incidence studies conducted at single or multiple hospital sites. This approach was made possible because nearly everyone in England is registered with a GP and because of the comprehensive
healthcare data captured by NHS Digital.[15] Moreover, a strength of our approach is that it is uses healthcare utilisation data to calculate specific study hospital usage by GP centre and age group and makes no assumptions about which health centres are used by a population within a particular census area.

Depending on the precise method, the geography-based approaches assessed in our study would have overestimated or underestimated the true catchment population and thus either underestimated or overestimated aLRTD incidence. At the extreme, defining the catchment population as those people living within a 60 minutes’ drive from a study hospital would have overestimated the catchment population by 4-fold to 5-fold and thus underestimated incidence to the same degree. At the other extreme, a drive time of 20 minutes would have underestimated denominators by 20-25% and thus overestimated incidence. Alternatively, the use of the entire CCG population would have overestimated denominators by 15%. The differences between geographically estimated denominators and our method is likely to vary by location and thus, the specific results from our study are illustrative of the principal and cannot be used to make conclusions about the relative accuracy of using an entire CCG population or drive time for other areas. For example, higher density areas with a larger number of hospitals would decrease accuracy of drive time or CCG for defining the catchment area of any particular hospital. This was illustrated in our study by demonstrating that for some practices and age groups, less than 20% of the practice population with an aLRTD hospitalisation presented to a study hospital. Since the only way to document the distortion in catchment population estimate for any particular health centres inherent in traditional estimates would be to first employ the methods described here, we suggest a better approach is simply to use our methods, or some similar approach, to define incidence denominators.
Other issues must be considered when using our approach. For example, the percentage of people with aLRTD hospitalisation who were hospitalised in a study hospital was relatively stable for older age groups and larger practices but varied substantially for younger populations and smaller practices, predominantly because of small absolute case counts for the latter groups. We largely overcame this issue by combining data for multiple years and creating larger age bands for younger populations. This issue will be more problematic for rarer diseases, which may require even larger age bands, greater numbers of study years, or aggregating individual ICD-10 codes into a common outcome.

The AvonCAP study was designed primarily to inform decisions on respiratory vaccine use among older adults, including vaccines to prevent pneumococcal, respiratory syncytial virus, and SARS-CoV-2 infection. Policymakers, including vaccine technical committees, have consistently indicated that disease burden is the number one factor in setting priorities for vaccines.[16, 17] Disease incidence, and usually severe disease incidence using hospitalisation as a proxy, is the cornerstone of disease burden and usually is the key outcome driving cost-effectiveness models. Cost-effectiveness values in turn are often used for policy and pricing decisions. E.g., in England a vaccine must be below a threshold of £30,000 per Quality Adjusted Life Year (QALY) saved to meet the criteria to be recommended for a national immunisation programme.[8] Since disease incidence underlies all these downstream measures, its accurate determination is critical for policy decisions. This requires a focus not just on accurate determination of case counts (that is, numerators) but also the catchment population for the surveillance system (that is, denominators).

Our approach has a few limitations. We could not account for people who were not registered with a GP; although, nearly all English residents are registered.[10] Our methodology also did not include the 4% of people that use the study hospitals but are registered with a GP practice outside of the CCG. However, this will be largely addressed in Avon-CAP by excluding from incidence calculations patients
with a study outcome living outside the CCG. Our approach requires a new estimate to be calculated for each disease of interest because some conditions will be disproportionately observed in some hospitals due to therapy area specialism. As discussed above, our approach may not be suitable for rare diseases or surveillance systems with small populations. Lastly, our methodology is appropriate for the particular circumstances of England. The extent to which this approach can be generalised to other countries will need to be evaluated on a case by case basis, but other areas where nearly all persons are formally registered with a primary care provider could consider its use.

We will use the described methodology to define denominators for incidence calculations within the AvonCAP study, which in turn should contribute to providing better data for informing decisions related to adult respiratory vaccine use. A similar approach could be used to refine previous estimates where these are being used to inform respiratory disease vaccine decision making. An historical study reporting disease incidence of hospitalised pneumonia in England was conducted in Hull and the East Riding of Yorkshire.\[5\] This study included 8 hospitals in the region and a geography-based approach was used to define the denominator. Whilst an effort was made to specifically exclude defined postcode areas reflecting a geographic region unlikely to use the study hospitals the accuracy of the denominator used in this study remains uncertain. A more recent study published hospitalised CAP incidence estimates from Nottingham, England and used a denominator based on the entire population of the Greater Nottingham area, but the market share of the two study hospitals used was not formally defined.\[3, 18\] Since the Greater Nottingham area is surrounded by other urban areas with hospitals that also treat CAP, it is unclear how well Greater Nottingham census data matches the hospital catchment population, and this could be formally evaluated by replicating our methodology. More generally, the method we describe may be used for other disease incidence calculations and for relatively common diseases could be extended to focus on specific groups such as those with underlying comorbidities. While the approach we describe takes considerably more human and financial resources than using census data
(through commissioning a specialist vendor that holds an appropriate license to analyse the data), this cost is negligible compared to the inefficiencies introduced when inaccurate disease incidence estimates are used as a core basis for public health decision making.

CONCLUSION

Use of the entire CCG or drivetimes does not account for the nuanced ways that populations access healthcare and may overestimate or underestimate denominators and distort incidence estimates. Our data driven method provides more accurate incidence estimates and thus can improve public health decision-making. Denominators for hospital-based incidence studies should be based on healthcare usage rather than geographical boundaries.

REFERENCES

Author contributions: JC, EB, AV, DH & GE contributed to the initial design of the methodology. All authors contributed to the analysis, interpretation, and discussion of results. We would like to acknowledge the assistance of Qi Yan, PhD (Pfizer, Inc.) who provided indispensable medical writing and literature review support for this manuscript and Harvey Walsh, Open Health Group who performed the denominator calculation. Hospital Episode Statistics (HES) Data were re-used with the permission of NHS Digital via Harvey Walsh, Open Health Group.

Data Availability Statement

The data that support the findings of this study are available from Harvey Walsh, Open Group. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available from the authors with the permission of Harvey Walsh, Open Group.
Conflict of Interest

JC, EB, AV, JS, HM, BG & GE are employees of Pfizer Vaccines and hold stock or stock options. DH is an employee of Harvey Walsh Ltd. CH is Principal Investigator of the Avon CAP study which is an investigator-led University of Bristol study funded by Pfizer and has previously received support from the NIHR in an Academic Clinical Fellowship. AF is a member of the Joint Committee on Vaccination and Immunization (JCVI) and chair of the World Health Organization European Technical Advisory Group of Experts on Immunization (ETAGE) committee. In addition to receiving funding from Pfizer as Chief Investigator of the Avon CAP study, he leads another project investigating transmission of respiratory bacteria in families jointly funded by Pfizer and the Gates Foundation.

Disclosure

This study was conducted as a collaboration between University of Bristol, Pfizer and Open Health Group. Pfizer is the study sponsor.