Allocating census data to general practice populations: implications for study of prescribing variation at practice level

Glen Scrivener, David C E F Lloyd

Abstract

Objectives—To assign census data to general practice populations and to test accuracy of different procedures for estimating the proportion of patients aged over 64.

Design—Patients' postcodes from patient register of one family health services authority and the directory linking postcodes to census enumeration districts were used to locate patients in their census area of residence. With different levels of census geography and four different allocation procedures, proportion of patients aged over 64 in each area was used to predict proportion of patients aged over 64 in each general practice. Predicted figures were compared with real figures from each practice register to assess accuracy of allocation methods.

Setting—Data from 1991 census and from 73 practices administered by one family health services authority.

Main outcome measures—Actual and predicted proportions of patients aged over 64 in general practice populations.

Results—Correlations between actual and predicted proportions of patients aged over 64 were significant for all four allocation procedures—values of 0.66, 0.7, 0.84, and 0.84 were achieved ($P < 0.0005$). Predicted ranges of proportions of patients aged over 64, however, were well short of those that actually existed, and significant differences existed between predicted percentages and actual figures for all four methods.

Conclusion—Although predicted values correlated with actual values, the failure of the allocation procedures to correctly predict values, especially at the extremes, casts doubt on the validity of similar techniques for allocating census variables to general practice populations.

Introduction

In order to study the role of patient morbidity in explaining variations in general practitioners' prescribing, proxies have been sought from datasets such as the national census since no clear measure of morbidity has emerged from general practice itself. At practice level, however, assigning census based indicators is not straightforward since practice populations are often dispersed across a wide area, while census data relate to populations located in defined census enumeration districts. The only link between the two populations is that individual patients can be located within their enumeration district of residence by means of their postcode.

Census data are released for groups of households called enumeration districts that were drawn up before the census (each district containing 183 households on average¹). These enumeration districts fit exactly into local authority wards, forming part of a hierarchical structure that builds up into local authority districts,

counties, and beyond. If a general practice population originates from several different enumeration districts and the census indicator of interest varies greatly between these districts, it is difficult to match this indicator to the practice population as a whole.

There is, however, a way of locating individuals in their enumeration district of residence by means of their postcode. Postcodes are primarily a tool of the postal delivery service (each referring to about 15 neighbouring residential addresses) but are useful elsewhere in that they are geographically referenced by the 100 m Ordnance Survey grid reference. Census enumeration districts are similarly referenced, and for the 1981 census the matching of postcodes to enumeration districts was based on the proximity of these grid references. This procedure had serious inaccuracies,² but the 1991 census was adjusted to create a more accurate way to link postcodes and enumeration districts. This was achieved by listing postcodes alongside the enumeration districts in which they fell, creating a postcode to enumeration district directory (see table I).

A typical way of associating characteristics of an area or its population to individuals is to express the characteristic as a rate and to apply that rate to individuals linked to the area. This assumes that the characteristic is evenly distributed in the area and that individuals are equally likely to possess it. The validity of this assumption is difficult to test, but efforts have been made to verify allocations made with this approach. Ward *et al* compared characteristics derived from the census with others obtained from questionnaires,³ while Majeed *et al* compared basic population structure derived from the census at ward level with the actual population structure that exists in practices.⁴

The aims of our study were, firstly, to make use of the new directory to place patients of general practices in their enumeration district of residence and, secondly, to test the accuracy to which the population structure for practices can be predicted from census data using proportional allocation techniques at different geographical levels. If the population structure of a practice cannot be predicted well with such methods then other measures derived from the census cannot be used as a proxy for morbidity with any confidence.

Methods

The study was based on data from the patient register of one family health services authority; this included only the patients living in the area administered by the authority so that not all of the patients attending the authority's general practices were included. Patients' ages and postcodes were taken from the register, grouped under their general practitioner, and then aggregated to form the practice populations. Practices with a large proportion of patients (over 5%) from outside the area administered by the family health

Prescribing Research Unit,
Academic Unit of General
Practice, University of
Leeds, Leeds LS2 9NZ
Glen Scrivener, *social
geographer*
David C E F Lloyd, *applied
research statistician*

Correspondence to:
Mr Scrivener.

BMJ 1995;311:163-5

TABLE I—Simplified representation of part of postcode to enumeration district directory for 1991 census

Enumeration district code*	Postcode	Pseudo-enumeration district code†	No of households
ZZAA01	AB1 1XX	ZZAA01	20
ZZAA02	AB1 1XA	ZZAA02	19
ZZAA02	AB1 1XB	ZZAA03	4
ZZAA03	AB1 1XB	ZZAA03	17
ZZAA03	AB1 1XC	ZZAA03	5

*First four letters give ward code.

†As households with identical postcodes can lie in different enumeration districts this code indicates in which enumeration district most of a postcode's households fall. Postcode AB1 1XB occurs in two enumeration districts, ZZAA02 and ZZAA03, so the pseudo-enumeration code is ZZAA03 where most of its 21 households lie.

services authority were removed from the study, leaving 73 practices (370 198 patients) from a total of 99. A further 1251 patients were excluded because their records lacked their postcode or their age.

ACTUAL POPULATION

For a simple representation of population structure the percentage of each practice population aged over 64 was calculated from the patient register. This figure varied widely among the practices (range 2.7-28.6%).

PREDICTED POPULATION

Four different methods were used to predict the percentage of patients aged over 64 for each practice from census data. All made use of the postcode to enumeration district directory from the 1991 census. Census data were taken for all enumeration districts and wards in the area administered by the family health services authority. Ward populations could be adjusted to account for underenumeration in the census (S Simpson, "Research on the 1991 census" conference, Newcastle upon Tyne, 1993).

Method 1 was based on census wards, and patients' postcodes were not used. Patients were assumed to reside in the ward in which their general practice was located and therefore to have the same characteristics as the population of that ward. The practice postcode was used to locate each practice in its host ward. If practices had two or more surgery locations only the main one was used: as there was no way to determine which surgery patients attended, there was no benefit in treating each surgery individually.

Patients' postcodes were used in methods 2-4, although 11 710 patients had postcodes that were not matched in the directory. The proportion of these patients from individual general practices ranged from 1.8% to 6.6%. The percentage of each practice's population who were aged over 64 was estimated from the census data with the following equation:

$$\sum (\text{No of patients from area } i) \times (\% \text{ of population of area } i \text{ aged } > 64) \\ \text{Population of practice}$$

This calculation was made for each area from which patients originated (see box). Only those areas required to achieve at least 90% of the practice population were

used in each case. The differences between methods 2, 3, and 4 concerned the type of area used.

Method 2—The area used was the census ward, and patients were located in their ward of residence and assigned the characteristics of that ward's population. Postcodes were matched against the first four characters of the pseudo-enumeration district code given in the postcode to enumeration district directory (see table I).

Method 3—The area used was the enumeration district, with patients located in their enumeration district of residence and assigned the characteristics of that district's population. Postcodes were matched against the first six characters of the pseudo-enumeration district code in the postcode to enumeration district directory.

Method 4—The area used was an individual postcode unit. Data from enumeration districts were allocated to postcodes according to the proportion of their households found in each postcode as indicated in the postcode to enumeration district directory. For example, a census count could be assigned to postcode AB1 1XB in the following way. The postcode falls across the boundary between enumeration districts ZZAA02 and ZZAA03, with four households in ZZAA02 and 17 in ZZAA03 (see table I). A census count from district ZZAA02 will include 23 (19+4) households, as indicated in the directory. For district ZZAA03 the number of households is 22 (17+5). The census count for the postcode AB1 1XB can be created by adding proportions of the counts from each enumeration district involved as indicated by the number of households (that is, 4/23 of the count from district ZZAA02 and 17/22 from district ZZAA03). Patients with that postcode were assumed to have the characteristics assigned to that postcode.

Results

The correlations between general practices' actual and predicted percentages of patients aged over 64 for the four methods described were as follows: method 1, 0.66; method 2, 0.70; method 3, 0.84; and method 4, 0.84. All the correlations were significant ($P < 0.0005$). Table II shows the actual and predicted ranges of percentages of patients aged over 64 for practices.

Methods 1 and 2 could never recover the full range of values found in individual practices since the proportion of residents aged over 64 in wards ranged only from 8.37 to 23.82 (values in enumeration district ranged from 0 to 53.81). Clearly, methods 3 and 4 were more successful at reproducing the true variability of the practice populations. The figure shows the scatter plot of actual against predicted percentages of patients aged over 64 for method 4. The method did not predict as wide a range of percentages as that which actually existed. The predicted and observed percentages were compared with the χ^2 test: for all methods the differences were significant ($P < 0.005$).

Discussion

Although the results showed a strong relation between predicted and real values for the percentage of the population aged over 64, the failure to achieve the actual range of values is worrying. The allocation procedures produced values tending towards the average and failed to predict the extreme values. This greater error at the extremes of the ranges is inevitable given that we were using aggregated data, but we think that many users of such methods of allocating data are unaware of this effect. To emphasise the effect that this might have, consider prescribing costs per prescribing unit as used in prescribing analysis and cost (PACT) reports (each patient counts as one unit except for those

TABLE II—Actual and predicted ranges of percentages of patients of general practices aged over 64

	Minimum	Maximum
Actual range	2.7	28.6
Predicted range:		
Method 1	9.0	18.9
Method 2	10.5	18.8
Method 3	8.1	23.8
Method 4	8.2	23.8

Comparison of actual percentages of patients aged over 64 in general practice registers with percentages predicted from census data by method 4 (see text for details)

Estimation of proportion of patients aged over 64 in a general practice from census data

The practice has 1000 patients who reside in three enumeration districts:

500 in district A
300 in district B
200 in district C

Proportion of population aged over 64 in each enumeration district:

2% in district A
4% in district B
8% in district C

The expected proportion of patients aged over 64 in the practice is therefore:

$$\frac{(500 \times 2\%) + (300 \times 4\%) + (200 \times 8\%)}{1000} = 3.8\%$$

Key messages

- Patient based indicators are required to examine variations between general practitioners in prescribing practice
- No clear indicators have emerged from general practice itself, so proxy measures have been sought from data sets such as the census
- We assigned census data to general practice populations and assessed accuracy of estimating proportions of patients aged over 64 in general practice by comparison with census from each practice register
- Correlations between actual and predicted proportions of patients aged over 64 were significant, but predicted ranges of proportions were well short of actual ranges and predicted and actual values were significantly different
- Failure to correctly predict actual values casts doubt on the validity of allocating census data to general practice populations

aged over 64, who count as three). Comparing costs calculated from the observed populations with costs calculated from values predicted by method 4, we find percentage differences of up to 16.9% and cost differences of up to £8.66.

Our results improved as the geographical areas used became smaller, with allocation from enumeration districts and postcodes producing the better results. It is not obvious, however, that the improved results were worth the extra work involved in achieving them.

The significant differences between predicted and observed proportions of elderly patients could result from a tendency on the part of patients to choose their general practitioner on criteria that differ with the age of the patients. For example, from discussions with the staff of the family health services authority, we know that one practice has a Polish speaking doctor who is consulted by elderly Polish patients from all over the city, who migrated to the area during the second world war. Long established practices may have a larger number of elderly patients than ones that have been established more recently. Conversely, practices offering a family planning service may attract a disproportionate number of patients who are of childbearing age.

LIMITATIONS OF DATA

The data we used came from the 1991 census, which has well documented limitations,¹ and one family health services authority, the data of which may have unique limitations but are likely to have problems common to many other family health services authorities.

The limitations of the census data include such matters as imputation, data suppression, and timeliness. Although the population figures at ward level were adjusted for underenumeration, those for enumeration districts were not. No account was taken of changes in population between 1991 and 1993, when the extract from the family health service authority's register was made. While estimates of ward populations for 1993 can be obtained commercially, estimates for enumeration districts cannot. Inaccuracies in family health services authorities' registers have been discussed elsewhere (A Lovett *et al*, "GIS disasters" conference, Salford, 1994), but there were clearly incomplete records in the register used in this study, and postcodes might have been inaccurately reported or recorded.

Even if many of the above inaccuracies were rectified a major limitation of the census data is that of small numbers, especially for small areas. Nationally, the size of enumeration districts extends from 16 to 665 households,¹ with the range in population likely to vary accordingly. In the family health services authority in

question some enumeration districts had only about 100 residents. With such small numbers the rates produced with them may be misleading or, given the limitations outlined above, suspect. For example, for one enumeration district in our study the percentage of patients aged over 64 was based on two elderly people in a population of 53. One elderly person fewer would have halved the percentage.

CONCLUSION

The introduction of the postcode to enumeration district directory has made the matching of patients' postcodes to their census area of residence more straightforward and more accurate. This has made it easier to link general practice populations to census areas at a local scale, down to postcode level. Even at such a fine scale, however, the accuracy with which characteristics of practice populations can be predicted from census data is not good. Considering the sophistication of the techniques employed, the results are disappointing. The failure to predict a simple and important characteristic (percentage of population aged over 64) from the census data throws doubt on the accuracy of any other census based variables allocated with similar methods, especially if the characteristic is comparatively rare (such as permanent sickness).

Furthermore, this study has highlighted practical problems with data from the patient register of a family health services authority. Incomplete records were discovered, though they were few in this particular register (1251); the numbers may be greater in the registers of other authorities. There were 11710 patients who had postcodes that could not be matched in the directory: this might in part have been due to some of the postcodes being new, but it might also have arisen from incorrect reporting or recording of postcodes at registration. Furthermore, 26 practices administered by the family health services authority had to be excluded because many of their patients came from outside the area managed by the authority and no postcodes were available for them. A study covering the whole of the area administered by the authority would have required the exchange of patient postcodes between all the neighbouring family health service authorities. Even if these deficiencies were rectified, the results of the allocation procedures would not necessarily improve and would still be suspect.

We therefore conclude that, if time and effort are to be expended in finding patient based indicators to help explain variations in prescribing, it would be sensible to concentrate on collecting data about the actual populations of the practices under consideration rather than predicting characteristics of interest though linkage to other datasets such as the census. The process of linkage, while easier than before, is not straightforward; data are constantly lost throughout the process; and the results eventually obtained are potentially misleading.

We thank Bradford Family Health Service Authority for the use of its patient register and Steve Simpson of Bradford Metropolitan Council for his help with census ward populations.

Source of funding: The Prescribing Research Unit is funded by the Department of Health.

Conflict of interest: None.

- 1 Dale A, Marsh C, eds. *The 1991 census user's guide*. London: HMSO, 1993.
- 2 Gatrell AC. On the spatial representation and accuracy of addressed-based data in the UK. *International Journal of Geographical Information Systems* 1989; 3:335-48.
- 3 Ward P, Morton-Jones AJ, Pringle MAL, Chilvers CED. Generating social class data in primary care. *Public Health* 1994;108:279-87.
- 4 Majeed FA, Cook GD, Anderson HR, Hilton S, Bunn S, Stones C. Using patient and general practice characteristics to explain variations in cervical smear uptake rates. *BMJ* 1994;308:1272-6.

(Accepted 22 May 1995)