Estimating the point accuracy of population registers using capture-recapture methods in Scotland

M J Garton, M I Abdalla, D M Reid, I T Russell

Abstract

***Study objective* – To estimate the point accuracy of adult registration on the community health index (CHI) by comparing it with the electoral register (ER) and the community charge register (CCR).**

***Design* – Survey of overlapping samples from three registers to ascertain whether respondents were living at the addresses given on the registers, analysed by capture-recapture methods.**

***Setting* – Aberdeen North and South parliamentary constituencies.**

***Participants* – Random samples of adult registrants aged at least 18 years from the CHI (n = 1000), ER (n = 998), and CCR (n = 956).**

***Main results* – Estimated sensitivities (the proportions of the target population registered at the address where they live) were: CHI – 84·6% (95% confidence limits 82·4%, 86·7%); ER – 90·0% (87·5%, 92·5%), and CCR – 87·7% (85·3%, 90·3%). Positive predictive values (the proportions of registrants who were living at their stated addresses) were: CHI – 84·6% (82·2%, 87·0%); ER – 94·0% (90·9%, 97·1%), and CCR – 93·7% (91·7%, 95·7%).**

***Conclusions* – The CHI assessed in this study was significantly less sensitive and predictive than the corresponding ER and CCR. Capture-recapture methods are effective in assessing the accuracy of population registers.**

(*J Epidemiol Community Health* 1996;50:99–103)

Department of Rheumatology, Aberdeen Royal Infirmary, Foresterhill, Aberdeen AB9 2ZB
M J Garton
D M Reid

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London
M I Abdalla

Department of Health Sciences and Clinical Evaluation, University of York
I T Russell

Correspondence to:
Dr M J Garton.

Accepted for publication September 1995

Population registers support a variety of important activities from disease screening to epidemiological research, and accuracy is therefore important. Unfortunately, like other population records, they may suffer from *inflation* (inappropriate entries), *deflation* (inappropriate absences), and *file errors* (data errors within otherwise appropriate entries).

Inappropriate entries and file errors can be identified through field surveys, and seem to be frequent in general practice[1–3] and family health services authority (FHSA) registers,[4–8] particularly among the elderly[9] and in urban areas.[10] Estimates of deflation are more difficult because the unregistered are by definition hard to find. Capture-recapture methods provide one approach to this problem, requiring comparison of multiple independent samples. Originally developed to estimate the size of animal populations, they have since been adapted for use in human populations.[11,12] They have been used to estimate the point accuracy of paediatric registers[13] and more recently to identify such elusive groups as prostitutes[14] and the homeless.[15]

Capture-recapture methods are used here to estimate the true accuracy of adult registration on the community health index (CHI) in Aberdeen, by comparison with the electoral register (ER) and community charge register (CCR). The CHI is a computerised patient record system, maintained by the primary care division of each Scottish health board; they are equivalent to the FHSA registers in England. The community charge was a local tax which has now been abandoned.

We decided to estimate *sensitivity* (the proportion of a target population who are registered at the address where they live), and *positive predictive value* (the proportion of registrants who live at their stated addresses). These parameters are practical measures of deflation and inflation respectively, and thus characterise the difficulty of contacting individuals registered at addresses other than their actual residence.

Methods

SUBJECTS

Aberdeen is a busy city in north east Scotland which attracts many migrant oil workers and students; 205 000 people live within the city district.[16] We restricted the study to all adults (aged 18 years or more) who on 1 October 1991 (when registration for the community charge was compulsory) were living in properties within the boundaries of Aberdeen North and South parliamentary constituencies, which together comprise about two thirds of the city district; these adults represent our target population. The study was approved by the Grampian Joint Ethical Committee.

OFFICE SURVEY

We estimated the total number of adults with a specific address listed on each parent register. The ER and CCR had to be counted manually, the ER because it was not computerised, and the CCR despite being computerised; in contrast the CHI was counted by its computer. Electors without a specific address or under 18 years old were excluded.

Simple random samples were drawn on 1 October 1991 from the computer records of

Table 1 Sensitivities of the community health index (CHI), electoral register (ER) and community charge register (CCR)

Register			n*	Subsample	Response			p†	np‡	θ§
CHI	ER	CCR			Present	Absent	Not known			
+	+	+	2164	133	119	2	12	0·99	2131	0·763
−	+	+	243	133	97	14	22	0·86	210	0·075
+	−	+	100	100	45	21	34	0·66	66	0·024
+	+	−	131	131	101	10	20	0·91	119	0·043
+	−	−	161	133	22	59	52	0·27	44	0·016
−	+	−	81	81	35	18	28	0·65	53	0·019
−	−	+	61	61	29	12	20	0·70	43	0·015
−	−	−	x	0	0	0	0	−	128	0·046
Total			x+2941	772	448	136	188		2794	1·000

Estimated sensitivities (95% confidence limits):
CHI = 0·763 + 0·024 + 0·043 + 0·016 = 0·846 (0·824, 0·867)
ER = 0·763 + 0·075 + 0·043 + 0·019 = 0·900 (0·875, 0·925)
CCR = 0·763 + 0·075 + 0·024 + 0·015 = 0·877 (0·853, 0·903)

* The first seven cells of this column subdivide the combined sample of 2941 according to the registers on which they appeared; the unknown number "x" in the eighth cell represents residents absent from all three registers.
† Proportion of individuals within subsample living at their stated addresses on 1 October 1991, adjusted in the manner described in the paper.
‡ Estimated number of individuals within the *effective* combined sample (that is, living at their stated addresses on 1 October 1991).
§ Estimated proportion of target population within each cell.

the CCR and CHI; the target sample size was 1000 for each register. A systematic random (every one hundred and twentieth name starting from a randomly chosen origin) was drawn from the ER compiled on 10 October 1991.

Each sample was cross checked against the other two registers. The three samples were then merged; obvious duplicate entries (that is, same name and address appearing more than once) were identified and reduced to a single entry. Individuals were classified into one of seven possible cells according to which combination of registers included their addresses (CHI only, ER only, CCR only, CHI and ER, ER and CCR, CCR and CHI, or all three).

FIELD SURVEY

A field survey was performed to estimate the proportion of individuals within each cell who were correctly registered. Persons were defined as "correctly registered" if they were living at their registered address. Cells with less than 133 subjects were taken whole, but random subsamples were drawn from larger cells in the interests of economy.

Three survey methods were used to establish whether subjects had actually been living at their stated addresses on the 1 October 1991. Attempts were made to contact all subjects, first by telephone and then by home visit. Those who remained uncontacted were sent a reply-paid letter followed by one reminder. Responses to each method were noted, and individuals coded as definitely present, definitely absent, or unknown.

The proportion of telephone respondents living at their stated addresses on 1 October 1991 (94·7%) was much higher than the corresponding proportions of respondents visited at home (71·2%) and of respondents who replied by post (66·9%). However, there were no significant differences in these last two proportions either in general or within any of the seven distinct cells. We therefore estimated the proportion of those coded as unknown within each cell who were actually living at their stated addresses on the qualifying date by the corresponding proportion of those visited at home or replying by post within the same cell. Given that almost all surveys suffer from non-response some means of imputing the true responses is almost always needed. Our approach was in effect to assume that non-responders are more like late responders than early responders – a more sophisticated approach than that adopted by most surveys.

ESTIMATION OF SENSITIVITY

These proportions were used to estimate the actual numbers of correctly registered people within each of the seven possible cells. The unknown number "x" in the eighth cell of table 1, representing those residents who were absent from all three registers, was estimated by log-linear modelling[17] on the assumption of no second order interaction between the three registers. The details of this technique, first advocated by Fienberg[18] and reviewed by McCarty *et al*,[12] are described in Appendix I.

The sensitivity of each register was then estimated. The population denominator was taken as the sum of all the correctly registered individuals within each cell, plus the estimated value of "x". The proportion of this population present on each individual register provided an estimate of the true sensitivity of each register.

ESTIMATION OF POSITIVE PREDICTIVE VALUE

For each of the original samples in turn, individuals were again classified into cells describing the combination of registers on which they appeared. The estimated proportions of correctly registered subjects were again applied to each cell, leading to an estimate of the number of correct entries within each sample, and hence an estimate of the positive predictive value of each register.

EXTRAPOLATION TO THE POPULATION LEVEL

The results were extrapolated to the whole population within Aberdeen North and South parliamentary constituencies to estimate the

Table 2 Positive predictive value of the community health index (CHI), electoral register (ER), and community charge register (CCR)

Register	*Register*			*n*	*p*	*np*	*Positive predictive value (95% CL)*
	CHI	*ER*	*CCR*				
CHI	+	+	+	722	0·99	711	
	+	−	+	60	0·66	40	
	+	+	−	57	0·91	51	
	+	−	−	161	0·27	44	
	Total			1000		846	84·6% (82·2%, 87·0%)
ER	+	+	+	735	0·99	724	
	+	+	−	74	0·91	67	
	−	+	+	108	0·86	93	
	−	+	−	81	0·65	53	
	Total			998		937	94·0% (90·9%, 97·1%)
CCR	+	+	+	720	0·99	709	
	−	+	+	135	0·86	117	
	+	−	+	40	0·66	26	
	−	−	+	61	0·70	43	
	Total			956		895	93·7% (91·7%, 95·7%)

likely true adult population present on 1 October 1991 (Appendix II). This estimate was compared to the most recent national census.[16]

CONFIDENCE INTERVALS FOR OUTCOME MEASURES

One thousand replications of the field survey were simulated using Monte Carlo techniques, based on computer generated random numbers.[19] These replicates were generated on the basis of the sample size and proportions of registrants estimated as definitely present, definitely absent or unknown from the actual field survey (Appendix III). For each simulated data set, repeat estimates were made of the parameters of interest, and the distribution of the estimates were used to construct 95% confidence intervals. Estimates of the standard errors of the difference between the sensitivity and specificity of the CHI and that of the other two registers were also derived from these simulated data sets, and then used to test for significant differences between the registers. Because these simulations were based on three specific samples rather than all possible samples, the resulting standard errors will have been slightly underestimated.

Results

The total number of adults registered was 118 845 on the ER, 125 593 on the CHI, and 117 690 on the CCR. Random samples of 1000, 998, and 956 names respectively were drawn from each register; the CHI sample was smaller because of the constraints of existing computer software. No name and address combinations were replicated within any single sample. However 13 appeared in two of the three samples leaving a total of 2941 separate people. Table 1 shows how many of these appeared on each *register* as established by the office survey.

Altogether 772 individuals were sampled for the field survey and useful data were obtained from 584 (75·6%) subjects; the "not known" category comprised 12 explicit refusals (1·6%), 72 non-responders (9·3%), 58 whose addresses were inadequate (7·5%), and 46 current residents (6·0%) who were unaware whether the sampled person had been resident on 1 October 1991. The field survey allowed us to estimate that of the original sample of 2941 individuals, 2666 were living at their specified address on this date. We estimated that another 128 residents (95% confidence limits 64, 192) were absent from all three registers, giving an effective total sample of 2794 (2701, 2887). These estimates were combined to estimate the sensitivities of the three registers (table 1). The sensitivity of the CHI was 5·4% (4·4%, 6·4%) and 3·1% (2·1%, 4·1%) less than that of the ER and the CCR respectively.

Table 2 shows the structure of the original samples and the estimated number of correctly registered subjects present within each. The positive predictive value of the CHI was 9·4% (6·3%, 12·5%) and 9·1% (6·9%, 11·3%) less than that of the ER and CCR respectively.

Finally the total number of adults living within Aberdeen North and South parliamentary constituencies on 1 October 1991 was estimated to be 125 085 (120 121, 130 049).

Discussion

The CHI had a significantly worse sensitivity and positive predictive value than either the CCR or ER. No previous sensitivity estimates are available for the CHI, but predictive values in the range of 75% to 96% have been reported.[20]

It should be emphasised that we considered individuals to be "correctly registered" only if they were actually resident at their stated address. Accordingly some individuals registered within Aberdeen North and South constituencies at legitimate postal addresses would have been excluded. Furthermore, the 188 unknown responses in the field survey were unevenly distributed across the cells. We tried to minimise this potential source of bias by continuing the field survey until the non-respondent category constituted only 9·3% of the field survey sample. The remaining 15·1% were either refusals, genuine "don't knows", or inadequate addresses. Nevertheless, the final results would have been a little different had these subjects been either mostly present or mostly absent.

The relative performance of the three registers studied may partly reflect incentives to join or leave them. Eligible adults are actively encouraged to join the ER, and regional councils visit or mail all properties in their area. Registration for the community charge was compulsory during its currency, and there was a financial incentive to declare departures from this register promptly. In contrast, registration and deregistration from the CHI require contact with health services and healthy individuals often have little inclination either to register or to notify their departure.

Inaccuracies in population registers can affect health care adversely. Poor sensitivity may reduce the effectiveness of services such as breast and cervical screening,[21,22] while low predictive values may cause resource wastage, for example by exaggerating non-attendance at screening clinics. Furthermore, net expansion

of registers can spuriously increase general practitioner capitation fees.[23]

In this study the precise health consequences cannot be accurately assessed because the distribution of errors across the population is unknown. However, we can say that excess capitation payments seem unlikely as the total number of registrants on the CHI (125 593) exceeded the estimated size of the resident population by under 1%. Comparison with the last census estimate (113 366)[16] might suggest the reverse, but this figure is misleading for several reasons. First, a large number of students (4634) were deliberately excluded from this estimate. Second, a large offshore workforce operates from Aberdeen and some of these individuals may have been difficult to detect and classify for census purposes. Finally, about one million British citizens evaded the last census; the deficit in Aberdeen city district was estimated as 5200, the majority residing within the study area.[24] Thus, our own estimate of the real adult population (125 085) in the area studied is not implausible, and is close to the total number of adults appearing on the CHI. Reassuring though this finding is, it cannot be extrapolated to other parts of the United Kingdom without replicating our capture-recapture methods.

The sensitivity and positive predictive value of the ER were particularly good. There was little evidence of the significant under-registration of the electorate which, some claim, contributed to the loss of Labour-held parliamentary seats (such as Aberdeen South) and perhaps led to Labour's defeat in the 1992 general election. The ER might be useful to health care workers and a recent study demonstrated its value in improving the predictive value of FHSA registers.[25] However, not all residents eligible for health care are enfranchised, and the accuracy of the ER declines between its annual revisions.

How can we improve patient registration? Primary care staff should be encouraged to check patient details when feasible, and patients themselves should be encouraged to register and report changes of circumstance promptly. Furthermore, the separate (although interdependent) patient registers held by general practitioners, their contractors, and the NHS central registry (as well as hospital records) could be linked more effectively to allow rapid updating of core information (such as name and address) each time an individual makes contact with an element of the health service.

Linkage of patient lists with local government registers might also be helpful. Although compulsory national registration was abandoned in Britain in 1952, it remains common in other parts of Europe, particularly Scandinavia. Sweden has done both; her citizens are assigned a single unique identification number at birth which is used for taxation, health care, voting, and social benefits. Although this merger of secular records is reflected in the high quality medical and epidemiological research in Sweden, it is not universally popular.

Similarly, although the British fashion of maintaining multiple independent registers is both expensive and inefficient, it seems unlikely to change in the short term. Record linkage would inevitably generate concerns about civil liberties, even with carefully restricted access, while national registration would be difficult to revive.

In conclusion, the application of capture-recapture methods has suggested that adult registration on the Aberdeen CHI suffers from both substantial inflation and substantial deflation. The potential of these methods to evaluate population and other registers has not yet been fully exploited.

The authors thank Arthur Fearnley, David A Henry, Doris McDonald and Hamish Morris of Grampian Regional Council, and Valerie Leslie and Joyce McAdie of Foresterhill Computing Centre, without whose help this study could not have been undertaken. We also thank Debbie Leslie, Mark Russell, and Marlene Westland for their help in collecting data, and Dr Lewis Reay and two anonymous referees for their comments on the manuscript. MJG was supported by the Wolfson Foundation, and MIA and ITR by the Chief Scientist Office of the Home and Health Department of the Scottish Office. DMR is grateful to the Arthritis and Rheumatism Council for their continued support. However, the views expressed in this paper are those of the authors alone.

Appendix I: Estimation of the number of residents absent from all three registers ("x")

In a three-way table an association between two of the variables may differ in degree or in direction within different categories of the third variable.[17] If there is no such second order interaction, then the association between the first two variables is the same for any category of the third variable. In this example we assumed that any observed association between presence on any two of the registers would not vary if subjects were stratified according to their presence or absence on the third register. In terms of proportions (θ), this relationship can be expressed as follows:

$$\frac{\theta_{+++}\,\theta_{--+}}{\theta_{-++}\,\theta_{+-+}} = \frac{\theta_{++-}\,\theta_{---}}{\theta_{-+-}\,\theta_{+--}}$$

where θ_{+++} = proportion of population present on all three registers; θ_{++-}, θ_{+-+}, θ_{-++} = proportion of population present on two of the three registers; etcetera.

Similarly let x_{+++}, x_{++-}, x_{+-+}, etcetera, represent the number of individuals in the combined sample who are present in any one of these cells. The best point estimate for the unknown x (or x_{---}) is obtained by substituting the θ's by their observed values x_{+++}/n, x_{++-}/n etcetera. Ex hypothesi there is no second order interaction between the three samples.

Hence

$$\frac{x_{+++}\,x_{--+}}{x_{-++}\,x_{+-+}} = \frac{x_{++-}\,x_{---}}{x_{-+-}\,x_{+--}}$$

And

$$x_{---} = \frac{x_{+++}\,x_{--+}\,x_{-+-}\,x_{+--}}{x_{-++}\,x_{+-+}\,x_{++-}}$$

Appendix II: Estimation of the adult population of Aberdeen North and South constituencies ("N")

The number of correctly registered individuals present on each register (R_{CHI}, R_{ER}, R_{CCR}) was estimated by using a predictive value of each register and the total numbers of individuals on each register (table 2).

$$R_{CHI} = R_{+++} + R_{++-} + R_{+-+} + R_{+--}$$
$$R_{CCR} = R_{+++} + R_{++-} + R_{-++} + R_{-+-}$$
$$R_{ER} = R_{+++} + R_{+-+} + R_{-++} + R_{--+}$$

where R_{+++} = number of correctly registered individuals present on all three registers; R_{++-}, R_{+-+}, R_{++-} = number of correctly registered individuals present on two of the three registers etc.

Let $R = R_{CHI} + R_{ER} + R_{CCR}$

$$= 3R_{+++} + 2R_{++-} + 2R_{+-+} + 2R_{-++} + R_{+--} + R_{-+-} + R_{--+}$$

Hence

$$\frac{R}{N} = 3\frac{R_{+++}}{N} + 2\frac{R_{++-}}{N} + 2\frac{R_{+-+}}{N} + 2\frac{R_{-++}}{N} + \frac{R_{+--}}{N} + \frac{R_{-+-}}{N} + \frac{R_{--+}}{N}$$

And

$$N = \frac{R}{3\theta_{+++} + 2\theta_{++-} + 2\theta_{+-+} + 2\theta_{-++} + \theta_{+--} + \theta_{-+-} + \theta_{--+}}$$

Appendix III: Simulation to replicate the field survey

Let S_{+++} denote the sample size used in the field survey to estimate the proportion of individuals appearing on all three registers who were correctly registered. Similarly, let S_{++-}, S_{+-+} and S_{-++} denote the sample size used to estimate this proportion for individuals appearing on any two of the registers, etc.

Let p_{+++}, a_{+++} and u_{+++} denote the number of individuals who were definitely present, definitely absent and unknown out of S_{+++}. These numbers are random variables with the following multinomial distribution:

$$p_{+++}\ a_{+++}\ u_{+++} \sim \text{Mult}\ (S_{+++}\ P_{+++}\ A_{+++}\ U_{+++})$$

where P_{+++}, A_{+++} and U_{+++} are the proportions of individuals who are definitely present, definitely absent and unknown out of the individuals registered on all three registers and $S_{+++} = (p_{+++} + a_{+++} + u_{+++})$. Using estimates of P_{+++}, A_{+++} and U_{+++} provided from the actual field survey, one thousand independent realisations of (p_{+++}, a_{+++}, u_{+++}) were simulated as follows:

(i) Simulate, using Monte Carlo techniques,[19] p_{+++} from a binomial distribution with a sample size S_{+++} and probability of success P_{+++}.

(ii) Simulate, using Monte Carlo techniques, a_{+++} from a binomial distribution with a sample size $(S_{+++} - p_{+++})$ and probability of success $(A_{+++}/(1 - A_{+++}))$.

(iii) Calculate $u_{+++} = (S_{+++} - p_{+++} - a_{+++})$.

(iv) Repeat (i) to (iii) 1000 times.

Similarly one thousand realisations of (p_{++-}, a_{++-}, u_{++-}), (p_{+-+}, a_{+-+}, u_{+-+}), etc, were simulated from each of the other six cells.

1 Fraser RC. The reliability and validity of the age-sex register as a population denominator in general practice. *J R Coll Gen Pract* 1978;**28**:283–6.
2 Sheldon MG, Rector AL, Barnes PA. The accuracy of age-sex registers in general practice. *J R Coll Gen Pract* 1984; **34**:269–71.
3 Silman AJ. Age-sex registers as a screening tool for general practice: size of the wrong address problem. *BMJ* 1984; **289**:415–6.
4 Scaife B. Survey of cervical cytology in general practice. *BMJ* 1972;**3**:200–2.
5 Sansom CD, MacInerney J, Oliver V, Wakefield J, Yule R. Recall of women in a cervical cytology screening programme. *Br J Prev Soc Med* 1975;**29**:131–4.
6 Fraser RC, Clayton DG. The accuracy of age-sex registers, practice medical records and family practitioner committee registers. *J R Coll Gen Pract* 1981;**31**:410–9.
7 Eardley A, Elkind AK, Spencer B, Hobbs P, Pendleton LL, Haran D. Attendance for cervical screening – whose problem? *Soc Sci Med* 1985;**20**:955–62.
8 Nathoo V. Investigation of non-responders at a cervical cancer screening clinic in Manchester. *BMJ* 1988;**296**: 1041–2.
9 Bowling A, Leaver J, Hoeckle T. *Survey of the needs of people aged 85+ living at home in City and Hackney.* London: Department of Community Medicine, City and Hackney Health Authority, 1988.
10 Elkind AK, Haran D, Eardley A, Spencer B. Computer-managed cervical cytology screening: a pilot study of non-attenders. *Public Health* 1987;**101**:253–66.
11 McCarty D, Tull E, Moy C, Kwoh C, Laporte R. Ascertainment corrected rates: applications of capture-recapture methods. *Int J Epidemiol* 1993;**22**:559–65.
12 LaPorte RE. Assessing the human condition: capture-recapture techniques. *BMJ* 1994;**308**:5–6.
13 Heward J, Clayton DG. The point accuracy of paediatric population registers. *J R Coll Gen Pract* 1980;**30**:412–6.
14 McKeganey, Barnard M, Leyland A, Coote I, Follet E. Female streetworking prostitutes and HIV infection in Glasgow. *BMJ* 1992;**305**:801–4.
15 Fisher N, Turner SW, Pugh R, Taylor C. Estimating numbers of homeless and mentally ill people in northeast Westminster by using capture-recapture analysis. *BMJ* 1994;**308**:27–30.
16 General Register Office. *1991 Census small area statistics.* Edinburgh: HMSO, 1993.
17 Everitt BS. *The analysis of contingency tables.* London: Chapman and Hall, 1977. p.77.
18 Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrics* 1972;**59**:591–603.
19 Kalow MH, Whitlock PA. *Monte Carlo methods. 1: basics.* New York: Wiley, 1986;50–53.
20 Roworth MA, Jones IG. The community health index – how accurate is it? *Community Med* 1988;**10**:327–30.
21 Chamberlain J. Failures of the cervical cytology screening programme. *BMJ* 1984;**289**:853–4.
22 Bowling A, Jacobson B. Screening: the inadequacy of population registers. *BMJ* 1989;**298**:545–6.
23 Rees MS. The inflation of national health service registers of patients and its effect on the remuneration of general practitioners. *Journal of the Royal Statistical Society* Series A 1969;**132**:526–42.
24 General Register Office. *1991 Census post-enumeration survey.* Edinburgh: HMSO, 1993.
25 Bickler G, Sutton S. Inaccuracy of FHSA registers: help from electoral registers. *BMJ* 1993;**306**:1167.