Hidden diabetes in the UK: use of capture–recapture methods to estimate total prevalence of diabetes mellitus in an urban population

Geoffrey V Gill MD Aziz A Ismail PhD Nicholas J Beeching MB Sarah B J Macfarlane MSc
Mark A Bellis PhD[1]

J R Soc Med 2003;**96**:328–332

SUMMARY

An early requirement of the UK's Diabetes National Service Framework is enumeration of the total affected population. Existing estimates tend to be based on incomplete lists. In a study conducted over one year in North Liverpool, we compared crude prevalence rates for type 1 and type 2 diabetes with estimates obtained by capture–recapture (CR) analysis of multiple incomplete patient lists, to assess the extent of unascertained but diagnosed cases. Patient databases were constructed from six sources—a hospital diabetes centre; general practitioner registers; hospital admissions with a diagnosis of diabetes; a hospital diabetic retinal clinic; a research list of patients with diabetes admitted with stroke; and a local children's hospital. Log linear modelling was used to estimate missing cases, hence total prevalence.

The crude prevalence of diabetes was 1.5% (95% confidence interval [CI] 1.41, 1.52), compared with a CR-adjusted rate of 3.1% (CI 3.03, 3.19). Age-banded CR-adjusted prevalence was always higher in males than in females and the difference became more pronounced with increasing age. Among males, CR-adjusted prevalence rose from 0.4% at age 10–19 years to 18.3% at 80+ years; in females the corresponding figures were 0.4% and 9.3%.

The gap between crude and CR-estimated prevalence points to a rate of 'hidden diabetes' that has substantial implications for future diabetes care.

INTRODUCTION

Concerned about the morbidity and mortality of diabetes mellitus, the UK Government has published a Diabetes National Service Framework aimed at improving outcomes over the next decade. A major and early requirement is to enumerate all people with diabetes. This is difficult, first because type 2 diabetes commonly goes undiagnosed and, second, because many patients attend for health care irregularly if at all. Standard methods of diabetes prevalence estimation include population surveys, postal questionnaires, house-to-house surveys, clinic or hospital records, and computerized diabetes registers.[1–4] All these are labour-intensive and tend to undercount. Thus 'ascertainment correction' has been used to adjust crude prevalence rates for patients not counted by the customary methods.[5] The technique, known as capture–recapture (CR), was originally developed by zoologists to count animal populations but is applicable in specific diseases.[6,7] Independent lists or registers of patients are used as 'captures'. Those people who appear on two lists or more are, in effect, 'recaptured'.[6–8] CR has been used for type 1 (insulin dependent) diabetes and other diseases,[9–11] but has seldom been applied to prevalence estimation of both type 1 and type 2 diabetes. We have used CR to determine total diabetes prevalence in North Liverpool.

METHODS

When applied to epidemiology, the CR technique assumes that lists of people with the disease in question will not include every person with the disease. When various lists are examined, high degrees of overlap suggest that most people with the disease are being 'captured' and the reverse may point to a substantial number of 'missing' cases. By use of multiple lists and statistical software, accurate estimates of the 'true' diabetic population (with confidence intervals) can be obtained. Provided the total general population is known, these results can be transposed to prevalence figures as a percentage of the local population.

The target population for this study was the district of South Sefton in North Liverpool, an urbanized area with a

Division of Tropical Medicine, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA; [1]School of Health, Liverpool John Moores University, 79 Tithebarn Street, Liverpool L2 2ER, UK

Correspondence to: Dr G V Gill

E-mail: g.gill@liv.ac.uk

stable population of about 177 000, mainly of European origin (Merseyside Information Services, Central Operation Group, personal communication). Six lists of cases were constructed—a list from 25 computerized general practices; patients attending the local hospital diabetes centre; hospital admissions with a discharge diagnosis of diabetes; diabetic patients attending the local hospital retinal clinic; a research list of patients with diabetes admitted to hospital with stroke; and patients attending the diabetes service at the local children's hospital. Lists were obtained from computer printouts obtained over one year. The diagnoses of diabetes were based on World Health Organization criteria applicable at the time.[12] Apart from the local stroke research database (which was small), all the lists were of a kind likely to be available in other areas of the UK, and our methodology should be transferable elsewhere.

The information was entered onto computer using the database software Epi-Info version 6.04.[13] The Statistical Package for the Social Sciences (SPSS)[14] and the Generalized Linear Interactive Modelling Package (GLIM 4)[15] were used for analysis. Cases were matched between lists by surname, first name, date of birth and postcode, by use of a sort and aggregate command in SPSS. Records without surname, first name, date of birth and the first part of the postcode were removed to ensure accurate matching. The postcode inclusion ensured that only patients resident in South Sefton were enumerated. To avoid errors due to list dependency, the final CR analysis was performed on three pooled lists, composed of all the general-practice lists (list *a*); the diabetes centre and the children's hospital list (list *b*); and the hospital admission list, retinal clinic list, and the stroke database (list *c*). Such combinations reduce interdependence errors,[16,17] and use of more than three lists does not significantly improve the accuracy of the population estimate.[18] A stepwise selection procedure of the various combinations of lists was used, interdependence being tested by goodness of fit (log-linear modelling). The number of missing cases was estimated from a $2 \times 2 \times 2$ contingency table (for the three combined lists) by use of GLIM software. Asymptotic confidence intervals were estimated with the same program. Full details of statistical methods have been published elsewhere.[17–21]

The ascertainment rate—a measure of completeness of case identification—was calculated by dividing the number of cases identified in each source by the aggregated number of cases identified from all sources. The prevalence of diabetes was determined by dividing the number of cases (aggregated cases for crude rate and estimated cases for CR-adjusted rate) by the total population in the group and subgroup (1991 Census).

The study was approved by the research and ethics committee of the Aintree Hospitals NHS Trust and Sefton Health Authority. Confidentiality of information was maintained at all times, according to the UK Data Protection Act,[22] and the information was anonymized after the matching procedures.

RESULTS

A total of 2585 known diabetic patients were identified through the six lists. Table 1 shows the numbers on each list, with age and sex ratio details, and the case ascertainment. For the three combined lists used for CR analysis, the distribution was 1469 for list *a*, 1314 for list *b* and 710 for list *c*. These details are shown, together with overlap, by Venn diagram in Figure 1. Log linear modelling by the GLIM package showed the estimated number of missing cases to be 2907 yielding a total diabetic population of 5492 (95% confidence interval [CI] 4870, 6285).

From the known diabetic patients appearing on all the lists (*n*=2585) and the population of 176 682, the crude prevalence rate was 1.5% (CI 1.41, 1.52). However, use of the CR-adjusted number of cases gave a prevalence of 3.1% (CI 3.03, 3.19).

Table 1 **Number of diabetic patients identified by each list and their characteristics**

List	*Total*	*Ascertainment (%)*	*Female/ male ratio*	*Mean age (SD)*
General practitioner	1468	57	50/50	62 (17)
Diabetes centre	1252	48	47/53	58 (16)
Hospital admission	454	18	49/51	64 (16)
Retinal clinic	351	14	46/54	65 (14)
Children's hospital	64	3*	58/42	12 (4)
Stroke database	38	2	42/58	75 (10)
Total aggregated	2584	100	48/52	60 (18)

*This figure is overall ascertainment (all ages); for patients <15 years the ascertainment was 60%

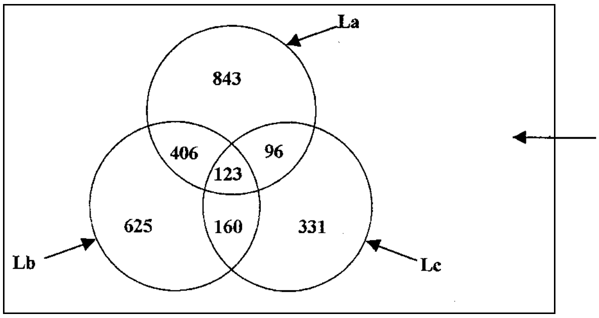

Fig. 1 **Number of diabetic cases identified by three sources**
La=general practice, Lb=diabetes centre and Alder Hey Children's Hospital, Lc=hospital admission, stroke database and retinal clinic, m=missing cases

Table 2 shows crude and CR-adjusted prevalence rates for age bands by decade. Males were consistently ahead of females, and the excess of CR-adjusted rates over crude rates was particularly evident for men over 70. Finally, we estimated age-adjusted rates using 1997 demographic data.[23] This showed similar figures for crude rates—1.4% (CI 1.2, 1.6) in females, and 1.6% (1.4, 1.9) in males. Age-adjusted rates by CR were slightly lower at 2.4% (2.2, 2.7) in females and 3.1% (2.8, 3.4) in males.

DISCUSSION

Since the data collection some 5 years ago, diabetes pr‘evalence may have increased, but our purpose is to demonstrate that simple case collection in a given area and population may seriously underestimate total diabetes prevalence. Indeed, the main finding of our study is that capture–recapture adjustment of crude diabetes prevalence rates increases the figure considerably, and reveals a large number of 'hidden' cases: the crude prevalence was 1.5%, but the CR-adjusted rate was 3.1%. Sex-stratification showed an increase amongst females from 1.4% to 2.7% (crude to CR), and 1.6% to 4.0% in males. Our statistical methodology also ensured that we minimized co-dependence of lists, which reduces the accuracy of CR[18–21] (for example, we separated the diabetes centre and retinal clinic lists, since many patients would clearly appear on both). Finally, our results are in accord with the only similar study, in which Bruno and colleagues in northern Italy reported a CR-adjusted total diabetes prevalence rate of 2.8% (95% CI 2.4, 3.1).[16] Equally, the age-banded data showed large excesses for CR with advancing age, as has been described in other studies.[24,25] The lower total and age-specific prevalence rates in females were likewise in agreement with other data,[2,24,25] and in particular with a careful epidemiological study from South Wales.[26]

In our study, we chose to estimate total, rather than separate type 1 and type 2, diabetes. There are several reasons for this. CR estimation of type 1 diabetes has already been applied widely—for example in Britain,[9] Australia,[27] Spain,[28] Italy[29] and Holland.[30] Because of the smaller numbers of type 1 and the greater ease of diagnosis and identification, lists of type 1 diabetic persons have a high degree of ascertainment, and CR methods are especially applicable and appropriate. Type 2 diabetes is more problematic since ascertainment is much less complete. Additionally, many of the existing lists do not specify the type of diabetes (though it can sometimes be inferred from treatment, age and duration of disease[17]).

Table 2 **Age-banded diabetes prevalence rates for males and females (crude and capture–recapture)**

	Prevalence % (95% CI)			
	Crude prevalence		***CR prevalence***	
Age (years)	***Female***	***Male***	***Female***	***Male***
0–9	0.1% (0, 0.1)	0.1% (0, 0.1)	NA	NA
10–19	0.4% (0.3, 0.5)	0.1% (0.2, 0.4)	0.1% (0.3, 0.6)	0.4% (0.3, 0.5)
20–29	0.4% (0.4, 0.5)	0.5% (0.4, 0.6)	0.8% (0.6, 1.0)	0.9% (0.6, 1.4)
30–39	0.4% (0.5, 0.8)	0.8% (0.6, 0.9)	1.0% (0.8, 1.2)	1.2% (1.0, 1.4)
40–49	0.8% (0.6, 1.0)	1.2% (1.0, 1.4)	1.2% (1.0, 1.4)	1.8% (1.6, 2.0)
50–59	2.0% (1.7, 2.3)	2.7% (2.4, 3.0)	2.8% (2.5, 3.2)	4.1% (3.7, 4.5)
60–69	3.5% (3.2, 3.9)	4.5% (4.1, 5.0)	5.6% (5.1, 6.0)	6.9% (6.4, 7.4)
70–79	3.7% (3.3, 4.1)	5.6% (5.0, 6.3)	7.6% (7.0, 8.2)	14.1% (13.2, 15.1)
80+	4.2% (3.6, 4.8)	6.0% (5.0, 7.1)	9.3% (8.5, 10.2)	18.3% (16.6, 20.0)
Overall	1.4 (1.3, 1.5)	1.6 (1.5, 1.7)	2.4 (2.2, 2.7)	4.0 (3.9, 4.2)*

NA=Not appropriate for capture-recapture techniques as the number of cases was too small
*Compared with crude rate, $P<0.0001$

These difficulties explain why the Italian study referred to above[16] is the only other attempt at CR estimation of total diabetes prevalence. Obviously, ascertainment rate variability between sources for type 1 and type 2 diabetes may introduce the potential for increased error, but our technique of combining multiple lists[18] will have reduced this hazard.

District diabetes registers are often constructed from multiple patient lists[31,32]—in particular, hospital clinic and general practitioner lists. Computerized summation of all patients on such lists is sometimes known as electronic data linkage, and can be used to assess crude prevalence rates in individual districts.[33] This was essentially the technique used by ourselves to estimate crude prevalence in North Liverpool (1.5%). A very high quality group of lists has been used in Tayside, Scotland (the DARTS/MEMO database), and gave an estimated diabetes prevalence of 1.9%.[33] This register has so far not been examined by CR, but our results suggest that the true prevalence will be higher.

Further validation of the CR technique in diabetes epidemiology is needed. It is said to be rapid and inexpensive, but a detailed cost comparison with other epidemiological tools has not been done. A direct comparison with a standard technique such as house-to-house survey would be valuable.[21] Nevertheless, there is already sufficient evidence to support the use of the technique in diabetes epidemiology, provided that multiple lists are used[18,34] and that these are combined in a way that minimizes co-dependency and allows similar chances of capture.[35] The population studied must also be stable in terms of migration and mortality.[36]

The large number of patients with 'hidden diabetes' indicated by this survey has great implications for health resource allocation, particularly in view of the rising age of the general population. Standard estimates of prevalence may lead to gross undercounting, and wider use of capture–recapture techniques should be seriously considered.

Acknowledgments We thank Ms Liz Harnsape of Sefton Health Authority, staff at the Walton Hospital Diabetes Centre, Mr Keith Jones for assistance with the hospital admission data from Aintree Hospitals, Ms Susan Kerr and Dr Colin Smith for providing data from the Alder Hey Children's Hospital Diabetic Clinic, and Dr Anil Sharma and Mr Kevin McDonald for providing data from the Aintree Hospital Stroke Research Unit database. We also thank Professor Ronald LaPorte for his suggestions and comments.

REFERENCES

1 Malins JM, Fitzgerald MG, Gaddie R, Cross KW, Mall M, Allen AM. A diabetes survey. A report of a working party appointed by the Royal College of General Practitioners. *BMJ* 1962;**1**:1497–507

2 Neil HAW, Gatling W, Mather HM, *et al*. The Oxford Community Diabetes Study: evidence for an increase in the prevalence of known diabetes in Great Britain. *Diabet Med* 1987;**4**: 539–43

3 Mather H, Keen H. The Southall diabetes survey: prevalence of known diabetes in Asian and Europeans. *BMJ* 1985;**291**:1081–4

4 Burnett SD, Woolf CM, Yudkin JS. Developing a diabetic register. *BMJ* 1992;**305**:627–30

5 McCarty DJ, Tull ES, Moy CS, Kwoh CK, LaPorte RA. Ascertainment corrected rates: application of capture–recapture methods. *Int J Epidemiol* 1993;**22**:559–65

6 LaPorte RE, McCarty D, Bruno G, Tajima N, Baba S. Counting diabetes in the next millennium: application of capture–recapture technology. *Diabetes Care* 1993;**16**:528–34

7 LaPorte RE. Assessing the human condition: capture–recapture techniques. *BMJ* 1994;**308**:5–6

8 Wittes J, Colton T, Sidel V. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chron Dis* 1974;**27**:25–36

9 Wadsworth E, Shield J, Hunt L, Baum D. Insulin dependent diabetes in children under 5: incidence and ascertainment validation for 1992. *BMJ* 1995;**310**:700–3

10 Squires NF, Beeching NJ, Schlecht BJM, Ruben SM. An estimate of the prevalence of drug misuse in Liverpool and a spatial analysis of known addiction. *J Publ Health* 1995;**17**:103–9

11 Devine M, Syde Q, Tocque K, Bellis M. Capture–recapture estimates of whooping cough in the Merseyside area. *Commun Dis Publ Health* 1998;**1**:121–5

12 WHO. *Diabetes Mellitus: Report of a WHO Study Group*. (*WHO Tech Rep Ser 727*). Geneva: WHO, 1985

13 Dean AG, Dean JA, Coulombier D, *et al*. *EpiInfo Version 6: a Word Processing Database and Statistics Program for Epidemiology on Microcomputers*. Atlanta: Centers for Disease Control and Prevention, 1994

14 Norusis MJ, SPSS Inc. *SPSS or Windows Base System User's Guide release 6.0*. Chicago: SPSS, 1993

15 Francis B, Green M, Payne C. *GLIM4. The Statistical System for Generalised Linear Interactive Modelling*. New York: Oxford Science Publications, 1993

16 Bruno AG, Bargero G, Vuolo A, Pisu E, Pagano G. A population-based prevalence survey of known diabetes mellitus in northern Italy based upon multiple independent sources of ascertainment. *Diabetologia* 1992;**35**:851–6

17 Ismail AA, Beeching NJ, Gill GV, Bellis MA. Capture–recapture adjusted prevalence rates of type 2 diabetes are related to social deprivation. *Quart J Med* 1999; **92**:707–10

18 Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture–recapture? *Int J Epidemiol* 2000;**29**:536–41

19 Cormack RM. Log-linear models for capture–recapture. *Biometrics* 1989;**45**:395–413

20 Cormack RM. Interval estimation for mark-recapture studies of closed populations. *Biometrics* 1992;**48**:567–76

21 Ismail AA, Gill GV. The epidemiology of type 2 diabetes and its current measurement. In: Gill GV, MacFarlane IA, eds. *Clinical Epidemiology and Metabolism: Frontiers in Clinical Diabetology*. London: Baillière Tindall, 1999:197–230

22 *Data Protection Act*. London: HMSO, 1984

23 Office for National Statistics (ONS). *Population and Health Monitor*. London: HMSO, 1998

24 Unwin N, Alberti KGMM, Bhopal R, Harland J, Watons W, White M. Comparison of the current WHO and new ADA criteria for the diagnosis of diabetes mellitus in the three ethnic groups in the UK. *Diabetic Med* 1998;**15**:554–7

25 McKeigue PM, Pierpoint P, Ferrie JE, Marmot MG. Relationship of glucose intolerance and hyperinsulinaemia to body fat pattern in South Asians and Europeans. *Diabetologia* 1992;**35**:785–91

26 Morgan CL, Currie CJ, Stott NCH, Smithers M, Butler CC, Peters JR. Estimating the prevalence of diagnosed diabetes in a health district of Wales: the importance of using primary and secondary care sources of ascertainment with adjustment for death and migration. *Diabet Med* 2000;**17**:141–5

27 Verge CF, Silink M, Howard NJ. The incidence of childhood IDDM in New South Wales. *Aust Diabetes Care* 1994;**17**:693–6

28 Rios MS, Moy CS, Serrano RM, Asensio AM. The incidence of type 1 diabetes mellitus in subjects 0–14 years of age in the Communidad of Madrid, Spain. *Diabetologia* 1990;**33**:422–4

29 Mazzella M, Cortellessan M, Bonassi S, *et al*. Incidence of type 1 diabetes in the Liguria region, Italy—results of a prospective study in a 0–14 year age group. *Diabetes Care* 1994;**17**:1193–6

30 Ruwaard D, Hirasing RA, Reeser HM, *et al*. Increasing incidence of type 1 diabetes in the Netherlands. The second nationwide study among children under 20 years of age. *Diabetes Care* 1994;**17**:599–601

31 Gatling W, Hill RD. General characteristics of a community-based diabetic population. *Practical Diabetes* 1988;**5**:104–7

32 Whitford DL, Roberts SH. Registers constructed from primary care databases have advantages. *BMJ* 1998;**316**:472–3

33 Morris AD, Boyle SD, MacAlpine R, *et al*. The diabetes audit and research in Tayside Scotland (DARTS) study: electronic linkage to create a diabetes register. *BMJ* 1998;**315**:524–8

34 Bruno G, LaPorte RE, Merletti F, Biggeri A, McCarty D, Pagano G. National Diabetes Programs—Application of capture–recapture to count diabetes? *Diabetes Care* 1994;**17**:548–54

35 Gill GV, Ismail AA, Beeching NJ. The use of capture–recapture techniques in determining the prevalence of type 2 diabetes. *Quart J Med* 2001;**94**:341–6

36 Wong JSK, Pearson DWM, Murchison LE, Williams MJ, Narayan V. Mortality in diabetes mellitus: experience of a geographically different population. *Diabet Med* 1991;**8**:135–9