Journal of Epidemiology and Community Health 1991; 45: 107–111

Linkage of biopsy, cancer, and population records aimed at the estimation of family risks in neoplasia: a pilot study

I Barrai, I Nenci, E Guidi, G Dell'Acqua, G Formica, G Barbujani, A Marzola, G Marani, R Barale, M Beretta

Abstract

***Study objective*—The aim was to link individual demographic and medical records into sibships to obtain the sibling distribution of biopsies and cancers, and thereby calculate heritability and recurrence risks in families, thus aiding early diagnosis and prevention of cancers.**

***Design*—The 157 823 individual records of the inhabitants of the town of Ferrara in Italy were automatically linked into 106 821 sibships. A 10% sample (10 842 sibships) was then extracted from the distribution of sibships and tabulated, for linkage to medical records.**

***Patients*—The biopsy records at the Institute of Pathological Anatomy of the University of Ferrara were manually linked to cancer records and then to sibships. It was possible to construct the distribution of 2062 biopsies and of 829 cancers in sibships.**

***Results*—From the distribution of biopsies and tumours in sibships, it was possible to estimate the incidence of tumours in the population (0·052) and in siblings of affected (0·083), and to apply to such distributions current methods for the estimate of heritability ($h^2 = 0{\cdot}246$) and of recurrence risks of tumours in sibships, age independent.**

***Conclusions*—The study shows that the procedure resulting in the estimation of incidences and recurrence risks for tumours could be completely automated, and extended to whole populations and homogeneous subgroups in post industrial cultures.**

University of Ferrara, 44100 Ferrara, Italy:
Institute of Zoology
I Barrai
G Barbujani
R Barale
M Beretta
G Marani
Institute of Pathological Anatomy
I Nenci
A Marzola
Institute of Hygiene
E Guidi
Institute of Medical Clinic
G Dell'Acqua
Computing Centre of the Municipality of Ferrara, Contrada Borgoricco, Ferrara, Italy
G Formica

Correspondence to:
Professor Barrai

Accepted for publication
May 1990

There has been dramatic progress in the genetics of cancer in the past decade, particularly in the elucidation of genetic mechanisms resulting in malignancy.[1] Such progress has mainly been due to the identification of oncogenes in retroviruses, and of their homologues in cellular genomes, through techniques of chemical genetics.[2]

It was observed that malignant transformation may occur when the product of an oncogene is more abundant than necessary in the cell, as occurs when a mutation in the gene promoter favours transcription, and when a product is changed, as when there is a structural mutation in the oncogene.[3] Both types of mutation may occur sequentially.[4]

The molecular approach to the genetics of tumours may eventually lead to the control of malignancy, and this aim may be the most important endeavour of health research in our time.

However, the genetics of tumours may also be studied from the point of view of the family aggregation of cancers, to resolve the predisposition to tumour formation, or its actual onset, into segregational units[5]; or at least to compute the heritability of the disease in such a way that an additional dimension is offered to the physician in dealing with early detection of cancer. The physician will be able to estimate recurrence risks in families where one or more persons are affected, and to propose early detection or preventive schedules according to the size of risk.[6]

Early detection, as secondary prevention, is particularly efficient in the case of tumours which have a clear genetic basis (eg, retinoblastoma); as early as 1978 McKusick[7] had listed about 100 such tumours. Among them, about 60 are transmitted as dominant, 30 as recessive, and a few are sex-linked.

When the transmission of the tumour is not attributable to a single factor, its heritability may be calculated with the methods proposed by Falconer[8 9] for semiquantitative traits. From heritability, recurrence risks may be obtained using several methods, the one proposed by Smith[10] appearing to be the most convincing.

It is our purpose in the present work to report the results of a pilot project for the estimation of recurrence risks of cancer in sibships in the population of the town of Ferrara in Italy. The results indicate that, with the automation of record linkage procedures,[11–13] useful preventive information may be gained from the linkage of medical and demographic files existing in a modern town.

Methods

On March 3, 1983, the population of Ferrara was 157 823 persons, all listed in the memory banks of the mainframe computer of the municipality. For each person, the data available included family name, forename/s, date and place of birth, address in town, father's name, family name and forename/s of mother, civil status, occupation, and school level.

The file was sorted by family name, name of father, family name of the mother, and name of mother. All individuals belonging to the same full sibship and residing in town were then grouped together; sorting on date of birth within sibship allowed us to separate from the same sibship adjacent individuals between whom there was a birth interval longer than 20 years.

The sibships were tabulated in alphabetical order, the family name first, then the name of the father, the family name and name of the mother

coming last; then the siblings in birth order. For practical convenience, sibships of size one—namely individuals having no siblings in town—were tabulated separately.

The Institute of Pathological Anatomy of the University has an index file of about 20 000 cards of biopsies of presumed tumours, from citizens residing in Ferrara and outside town, with records beginning in 1956.

Each card in this index file has the name, address, and age of the patient, and the reference number of the file of the patient records in the archives of the local health establishments (Unita' Sanitaria Locale No 31). We obtained permission from the director to gain access to the patients' files, and initiated a programme of record linkage of the index file with the sibships file, limited to surnames beginning with letters A, V and Z, which represent a 10·2% sample of the population of Ferrara. The selection of surnames beginning with these letters was only opportunistic, since they made up almost exactly 10% of the population, and we had planned to achieve linkage of the medical records with the sibships manually in a reasonable time, say less than three years. Surnames beginning with these letters should not identify selectively individuals of any particular ethnic origin.

The main criterion of linkage was the identity of names and year of birth in the index file and in the sibship file. The linkage was done by visual inspection of the files by four of us (GM, EG, GB, and GDA). The patient file number, found on the diagnostic card, was transferred to the printed sibship file. The sibship file, completed by patient's record number, was used to access the document with all medical data pertaining to the patient. The sibship file was then completed by type of tumour per affected individual. In the patient record file, further data (such as present address) were used to confirm match between an individual in a sibship and a diagnosis.

The operation was initiated in April 1983 and terminated in May 1986. After that time, we were in a position to compute the frequency of tumours in the sibship files, their recurrence in sibships, and their heritability. In so doing, we assume that the mode of inheritance is polygenic, with many genes contributing to the occurrence of tumours, and each gene has a small additive effect. In this approach, we ignore two main problems, which may be addressed at a subsequent time, namely (a) the consideration of a major gene, and (b) genetic heterogeneity. This second point arises by definition, since the tumours are classified as "malignant" and "benign". Since we know that specific genes predispose an individual to a specific cancer or group of cancers,[14] our analysis of the occurrence of all cancers is presented here as an example addressing the issue of feasibility of record linkage in single cancers with a view to proposing early detection procedures on a family basis of occurrence.

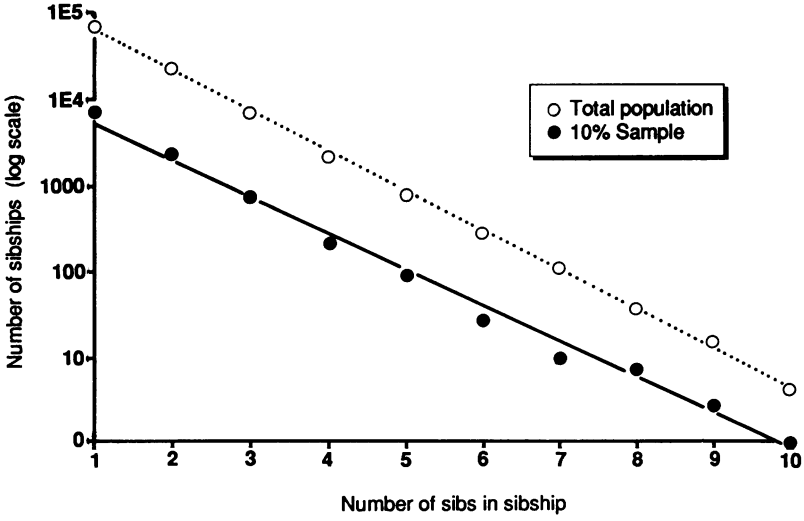

Figure 1 Distribution of sibship size in Ferrara in March 1983. The distribution is independent of age. Note the almost exact fit of the logarithmic regression to the observed frequencies.

Results

The distribution of assumed sibships is given in table I and its logarithmic transformation is graphed in fig 1. The distribution is almost exactly exponential. The goodness of fit of the log transform of the frequency is simply unexpected, and visual inspection in this case may be as convincing as any statistical test.

Table I Distribution of sibship size in Ferrara in March 1983, age independent. A = total population, B = 10·2% of population

Sample size	*Sample frequency* A	B
1	71 429	7267
2	25 076	2465
3	6999	763
4	2078	210
5	780	87
6	285	28
7	115	10
8	39	8
9	16	3
10 and more	4	1
Totals	106 821	10 842

$\chi^2_9 = 15{\cdot}216$
$p < 0{\cdot}10$
Σ obs ln (obs/exp) = 13·725

The whole age pyramid of the town population is reflected in this distribution, with the effects of birth, death, and migration; therefore, it is not immediately possible to interpret it in genetic terms. However, we predict that in an Italian town of similar size the distribution will be very similar to the one observed in Ferrara. At present, we observe that the probability of transition from a sibship of size s to a sibship of size (s + 1), age independent, is 0·35 in Ferrara, and fairly constant from sibship size one to size 10.

The distribution of sibships in the 10% sample we selected is also given in table I, and its log transform is graphed under the general distribution. The parallelism is very good; in any case, the independence χ^2 between sample and population is 15·216 with nine degrees of freedom. This value is just below the 10% significance level; however, the χ^2 from the log of the ratio observed/expected is 13·725, with $p > 0{\cdot}20$.

BIOPSIES

We linked 2062 index cards, each corresponding to one person, with the sibships. It is not specified

Table II Distribution of the number of sibships by size and number of siblings with at least one biopsy

Sibship size	Siblings with biopsy in sibship					
	0	*1*	*2*	*3*	*4*	*Total*
1	6285	982				7627
2	2002	412	51			2465
3	498	198	60	7		763
4	118	73	19			210
5	34	37	14	2		87
6	11	11	4	1	1	28
7	5	4		1		10
8	3	3	2			8
9		1	2			3
10	1					1
Totals	8957	1721	152	11	1	10 842

Table III Distribution of the number of biopsies per person

Number of biopsies	*A priori*	Type of tumour		
		No tumour	*Benign*	*Malignant*
1	1335	880	359	96
2	378	181	110	87
3	155	77	30	48
4	85	43	15	27
5	43	19	10	14
6	22	10	1	11
7	12	5	1	6
8	10	7	0	3
9	4	2	0	2
10	4	1	0	3
Totals	2048*	1225	526	297

* The other 14 cases were as follows: one uncertain diagnosis; eight persons with no tumour and more than 11 biopsies; five persons with malignant tumour and more than 11 biopsies.

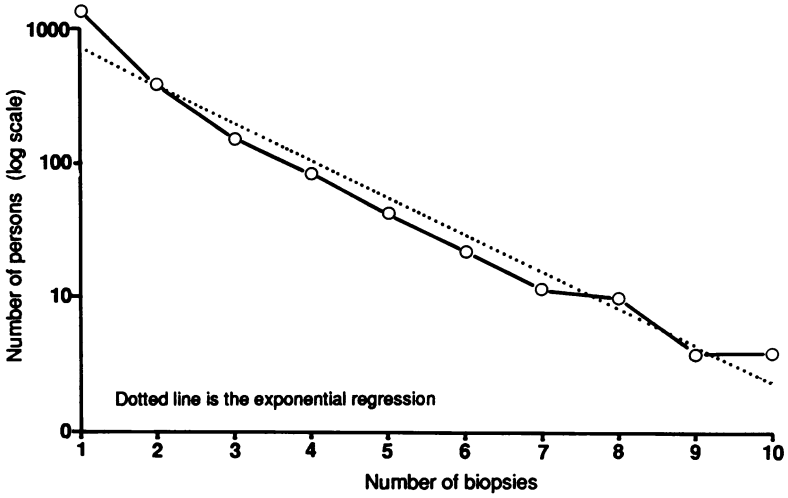

Figure 2 Distribution of the number of biopsies per person without considering the resulting diagnosis. The fit of the logarithmic regression is quite good, indicating exponential decrease of the number of persons with more than one biopsy.

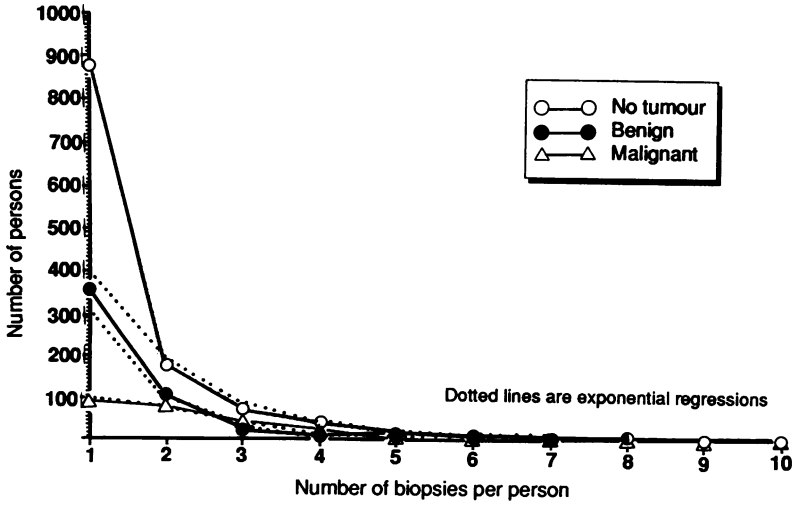

Figure 3 Exponential decay of the number of biopsies per person as a function of the severity of disease. Note the excess of single biopsies in absence of tumour; there is considerable repetition of biopsy when the diagnosis is malignant. The overall diagnostic ratio from biopsies in this population and town is approximately 3:2:1 for No tumour: benign tumour: malignant tumour.

in the index cards whether the biopsy resulted in a diagnosis of no tumour, non-malignant tumour, or malignancy. However, each card lists all the biopsies taken from the same individual at different times, so we were in a position to study the distribution of the number of biopsies per individual, given that he/she has at least one.

In table II, we give the distribution of the number of siblings with at least one biopsy by sibship size. Again, this is the sibship size in Ferrara as ascertained under the restrictions described above. Furthermore, the probability of a person undergoing biopsy is clearly dependent on the health facilities available to a population group; so it would not be meaningful to calculate the heritability of the phenotype "biopsy" in sibships. However, we can examine the distribution of the number of biopsies per individual; it is given in table III. The logarithmic transform of the distribution is given in fig 2; a straight line fits the logarithmic distribution, indicating that, in this area and in this population, with the available medical facilities, the frequency (y) of the number of biopsies (x) per person is of the type:

$$y = ae^{-bx},$$

a common type of decay; in the present case $a = 1316$ and $b = 0{\cdot}633$. The analysis of biopsy distribution by the three subgroups resulting later from the linkage of the pathological records, namely, "no tumour", "benign tumour", and "malignant tumour", is given in the same table III. The association between number of biopsies and subgroup is clearly visible. The same type of exponential decay holds for the three groups, although the decrease in the number of biopsies is of course more rapid in the "no tumour" and "benign" groups than in the malignant tumour group (fig 3). In table IV we give the parameters of the exponential curves fitted to the observed distributions of the number of biopsies per patient.

Table IV Fitting of the exponential model, $y = ae^{-bx}$ to the distribution of the number of biopsies

Group	*a*	*b*	*r*	*df*	*p*
No tumours	794	0·677	0·9798	8	$\ll 1\%$
Benign	841	1·006	0·9817	5	$< 1\%$
Malignant	169	0·463	0·9813	8	$\ll 1\%$

r = correlation between data and model; df = degrees of freedom; p = level of significance

TUMOURS

After the linkage of the biopsies with sibships, it was possible to examine the details of the diagnoses from the individual patient files. Out of 2062 persons undergoing biopsy, 829, or 40%, had a tumour, either malignant or non-malignant.

The distribution of the number of siblings with a tumour per sibship size is given in table V. There were 770 sibships with one tumour. In 28 sibships there were two tumours, and in one, three tumours.

It is then possible to calculate heritability of tumour, independent of degree of malignancy,

Table V Distribution of sibships per number of siblings with tumour and sibship size

Size	No of tumours 0	1	2	3	Total
1	6857	410			7267
2	2256	203	6		2465
3	652	95	16		763
4	176	32	2		210
5	62	23	1	1	87
6	21	5	2		28
7	9	1	0		10
8	7	0	1		8
9	2	1	0		3
Totals	10 042	770	28	1	10 841

with the methods of Falconer, from table V. In the appendix, we describe briefly the methodology proposed by Falconer for data in the form of incidences which refer to "all or none" classification, such as tumour or no tumour.

It is appropriate, since we use the file in which most, if not all, tumours of the Ferrara area are listed, to use a model of truncate selection[5] for the estimation of the segregation frequency, namely of the incidence, of tumours in siblings, to be compared with the incidence among non-related individuals.

In the general population the frequency is:

$$q_1 = 829/16100 = 0{\cdot}0515$$

The estimate of segregation frequency for sibships of size s with r affected siblings under truncate selection is:

$$q_2 = \Sigma r(s-1)/s(s-1), \text{ or}$$

$$q_2 = 62/747 = 0{\cdot}0830$$

Applying to these values the formulas of Falconer, one obtains:

$x_1 = 1{\cdot}635$ for the normal deviate in the general population

$x_2 = 1{\cdot}385$ for the normal deviate in siblings of probands

$z = 0{\cdot}1047$ for the ordinate at the threshold of liability

$a = 2{\cdot}344$ for the average liability of tumours.

The regression coefficient is then calculated as

$$b = (x_1 - x_2)/a = 0{\cdot}123$$

and twice the regression is an estimate of heritability:

$$h^2 = 0{\cdot}246$$

This is the value of heritability to be fed into the algorithm calculating recurrence risks. However, since we observe sibships with more than one case, we can estimate directly some of the recurrence risks and compare them with the risks expected after heritability was calculated.

CALCULATION OF OBSERVED AND EXPECTED RECURRENCE RISKS

From the data in table V, we can calculate risks directly, from the number of sibships having r affected and being of size s, and the number of sibships of the same size having (r + 1) affected. These we call the "observed" risks, since there is no specific model underlying their calculation.

Several observed risks can be calculated from the data. First, the risk of at least one cancer in a sibship, independent of size and age, is 799/10 842 = 0·074; the risk of finding a second affected sibling, given that there is at least one in the sibship, is 28/799 = 0·035; and the empiric risk of a third sibling affected, given that there are at least two in the sibship, is 1/29 = 0·034.

To formalise our approach, let us consider the observed probability of occurrence of cancers in sibships of size 1, or $p(r=1|s=1)$. We have

$$p(r=1|s=1) = 410/7267 = 0{\cdot}056$$

to be compared with

$$p(r=1|s>1) = 0{\cdot}074$$

For size 2

$$p(r=1|s=2) = 203/2465 = 0{\cdot}0824$$

$$p(r=2|s=2) = 6/2465 = 0{\cdot}0024$$

but

$$p(r=2|r=1,s=2) = 6/209 = 0{\cdot}029$$

is the observed risk of recurrence of a second cancer in a sibship of size 2 with one sibling already affected.

For size 3

$$p(r=1|s=3) = 95/763 = 0{\cdot}125$$

$$p(r=2|s=3) = 16/763 = 0{\cdot}021, \text{ and}$$

$$p(r=2|s=3,r=1) = 16/111 = 0{\cdot}144$$

is the observed recurrence.

In table VI, the observed risks are compared with the risks calculated with a heritability of 0·246 and a population frequency of 0·0515, using the model and the algorithm of Smith.[10] We prepared a version of the algorithm in FORTRAN. This is available upon request.

It is apparent that the agreement between observed risks and risks derived from heritability is quite variable; there is no consistent trend, and the correlation coefficient between observed and expected is only 0·63, not significant for nine observations ($t = 1{\cdot}798$, $p < 0{\cdot}20$). Should any practical considerations arise from these results, one should rely on the expected risks, since they are not affected by random variation of small number of families, and can be calculated for any sibship size with any number of affected individuals.

Table VI Comparison between observed and expected recurrence risks for a given composition of the sibship

Siblings without proband: Normal	Tumour	Risk: Expected	Observed
0	0	0·049	0·056
0	1	0·080	0·029
0	2	0·110	not available
1	0	0·047	0·082
1	1	0·077	0·144
1	2	0·124	not available
2	0	0·046	0·125
2	1	0·074	0·059
3	1	0·071	0·042
4	1	0·069	0·280 (2/7)
2	2	0·105	0·500 (1/2)

Discussion

The use of record linkage for epidemiological and preventive studies now has a history of more than two decades. The technique was used mainly by geneticists to study historical population structure. We report here the linkage of biopsy and cancer records with demographic records, through which we estimated recurrence risks of tumours in sibships. Knowledge of such risks could be of preventive value in the case of tumours.

In this pilot study, it was possible to automate completely the creation of a sibship file from the individual records of the town municipality. The linkage of biopsies and cancers with sibships was performed by human operators, since the biopsies and cancer records were not on computer readable media. However, once they are on magnetic media, the linkage is not expected to be more difficult than the linkage of siblings from population records, which is already achieved.

The results of record linkage from this study may be summarised as follows: (1) description of the distribution of sibship size in a modern town; (2) description of the distribution of the number of biopsies per individual as a function of subsequent diagnosis; and (3) estimation of incidences and recurrence risks of tumour in sibships. The study shows that it is possible to automate the procedures completely. Given the low cost and high speed of present day computers, a project of record linkage of larger size seems attractive—say, the whole population of one town with specific medical records, under the assumption that (a) the medical records are on computer readable media and (b) the town meets the costs of producing the sibship files necessary for the study. The potential for a programme of preventive medicine seems considerable, particularly when the analysis of familial occurrence within homogeneous subgroups, defined by pathological findings, is made possible by direct automation of record linkage procedures.

This work was supported by the Funds 60% and 40% of the Italian Ministry of Public Instruction. The constructive comments of one Referee were particularly appreciated. The authors acknowledge the cooperation of the Direction of the USL 31 for permitting access to the medical records, and of the Town Municipality of Ferrara, in the person of the incumbent Mayor, Dr R. Soffritti, for permitting access to the town demographic records.

Appendix

THE METHOD OF FALCONER FOR THE CALCULATION OF HERITABILITY FOR "ALL OR NONE" TRAITS

An "all or none" trait, such as the presence or absence of a tumour, may depend on several underlying genetic and environmental factors, called liability factors, which determine its appearance when they are above a given level. The level, the point on the scale of liability above which all individuals are affected and below which all are normal, is called by Falconer the "threshold".

Falconer[8 9] assumes that the variation of such liability factors is normally distributed and has the same variance in the general population and in the consanguineous relatives of the individuals showing the trait, say the affected. It is possible to measure the incidence of the trait in the general population, q_1, and in a group of relatives of affected, q_2.

From the incidences, it is possible to obtain the numbers of the underlying liability factors in units of standard deviations by calculating (or reading from a statistical table) the normal deviates corresponding to q_1 and q_2, say x_1 and x_2.

It is possible to show that x_1 corresponds to the threshold, and x_2 to the level of liability factors in relatives of the affected.

In the same way, it is possible to calculate (or read from a statistical table) the average number of liability factors in units of standard deviations for all individuals above the threshold, a. This is estimated by:

$$a = z/q_1$$

where z is the height of the ordinate of the normal curve at the threshold. Both a and the x's are measured as deviations from the mean, ie, they are standard deviates.

The gain in the mean number of liability factors in relatives compared to the general population is then $x_1 - x_2$ which, divided by a, gives the specific increase of liability in relatives of affected, say the regression of relatives on the affected. Then, the regression coefficient is

$$b = (x_1 - x_2)/a$$

It is possible to show that the regression coefficient is related to heritability, h^2, through the coefficient of relationship between relatives, r, through $b = rh^2$, and then $h^2 = b/r$. For first degree relatives, say parent-offspring and sibling-sibling, $r = 1/2$ and $h^2 = 2b$; for uncles or aunts and nephews or nieces, $r = 1/4$ and $h^2 = 4b$; for first cousins $r = 1/8$ and $h^2 = 8b$.

The incidence of the trait in the general population and its heritability are the key parameters for the estimations of recurrence risks. Smith[10] suggests a convincing algorithm. Both the method of Falconer and of Smith are considered in detail in standard textbooks.[14]

1 Reddy EP, Reynolds RK, Santos E, Barbacid M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* 1982; **300**: 149–52.
2 Temin HM. Do we understand the genetic mechanisms of oncogenesis? *J Cell Physiol* 1984; **suppl 3**: 1–11.
3 Tabin CJ, Bradley SM, Bargman CI, *et al.* Mechanisms of activation of a human oncogene. *Nature* 1982; **300**: 143–9.
4 Klein G, Klein E. Evolution of tumours and the impact of molecular oncology. *Nature* 1985; **315**: 190–5.
5 Morton NE. Genetic tests under incomplete ascertainment. *Am J Hum Genet* 1959; **11**: 1-16.
6 Toti A, Piffanelli A, Pavanelli T, *et al.* Possible indication of breast cancer risk through discriminant functions. *Cancer* 1980; **46**: 1280–5.
7 McKuisk V. *Mendelian inheritance in man.* Baltimore: Johns Hopkins University Press, 1978.
8 Falconer DS. The inheritance of liability to certain diseases, estimated from the incidences in relatives. *Ann Hum Genet* 1965; **29**: 51–76.
9 Falconer DS. The inheritance of liability to diseases, with variable age of onset, with particular reference to diabetes mellitus. *Ann Hum Genet* 1967; **31**: 1–20.
10 Smith C. Recurrence risks for multifactorial inheritance. *Am J Hum Genet* 1971; **19**: 578–88.
11 Acheson ED. *Record linkage in medicine.* Edinburgh: Livingstone, 1968.
12 Baldwin JA, Acheson ED, Graham WJ. *Textbook of medical record linkage.* Oxford: Oxford University Press, 1987.
13 Newcombe HB. *Handbook of medical record linkage: methods for health and statistical studies, administration, and business.* Oxford: Oxford University Press, 1988.
14 Marx J. Eye cancer gene linked to new malignancies. *Science* 1988; **241**: 293–4.
15 Barrai I. *Introduzione alla genetica dei caratteri quantitativi.* Padova: Piccin Editore, 1980.