Acta Genet Med Gemellol 28:139–148 (1979)
The Mendel Institute/Alan R. Liss, Inc.

Received 5 June 1978
Final 18 December 1978

Optimization Procedures in Twin Zygosity Diagnosis by Genetic Markers
A Cost-Effectiveness Analysis

Seppo Sarna, Jaakko Kaprio

Department of Public Health Science, University of Helsinki

More and more genetic markers usable for twin zygosity determination have become available. A relatively small number of markers is sufficient to achieve a satisfactory probability of correctly classifying a twin pair. Previously only the genetic properties of markers have been considered when choosing the markers to be determined. A cost-effectiveness analysis, which considers both genetic properties and relative determination costs of markers, is presented and illustrated with data from the Finnish Twin Registry studies.

Key words: Twin zygosity diagnosis, Genetic markers, Optimization procedures, Finnish twin registry

INTRODUCTION

The use of multiallelic genetic markers offers the most accurate way of deciding zygosity. A twin pair discordant for at least one genetic marker is by definition dizygous (DZ), while a twin pair concordant for genetic marker phenotypes is classified monozygous (MZ), but with only a certain probability [1]. The determination of a relatively small number of markers suffices to achieve a probability of misclassification of a twin pair's zygosity [4, 7], which in most studies is small enough to no longer affect any inferences made. The effect of misclassification errors on epidemiologic twin data has been considered by Friedman [3]. Further increasing the number of markers determined has only a minor effect on decreasing the probability of misclassification.

As more genetic markers are becoming available for routine testing procedures, a choice of the markers to be used can be made. One criterion for this choice has been the genotypic efficiency of the markers [8], which presupposes that information on genotypes is available. Besides considering the genetic efficiency of the markers available, their determination costs should also be taken into account, ie, two cheaply determined markers may together be more efficient than one highly efficient marker that is expensive to determine.

In this paper, a cost-effectiveness analysis for optimizing the choice of markers to be used in zygosity determination procedures is presented. The analysis takes into account both the genetic properties of the markers and their relative costs of determination. Also, a method for determining a priori the minimum sample size necessary to achieve the desired *sample* accuracy in zygosity diagnosis is presented.

GENERAL PRINCIPLES

Suppose that markers $M_1, \ldots, M_k$ are available. Let M_i be any one of these markers and P_i the probability that a DZ twin pair is phenotypically concordant in respect to marker M_i. In this paper these quantities will be called *concordance probabilities* and their complementary probabilities, $Q_i = 1 - P_i$, *discriminating powers* of the markers M_i, $i = 1, \ldots, k$. Let us denote the vectors formed by these quantities with $P = (P_1, \ldots P_k)$ and $Q = (Q_1, \ldots, Q_k)$. Numerical values for these probabilities in any set of twin pairs can be computed, for example, with the formulas or with the algorithm presented by Sarna [6], assuming that the population gene frequencies or their estimates from a sample are known.

In order to be able to operate easily with combinations of indexes of markers, we will define the following sets: $I_1 = \{i\}_1^k$, $I_r = \{(i_1, \ldots, i_r)\}$, where $i_1, \ldots, i_r \in I_1$. The last one of these index sets contains all the $\binom{k}{r}$ combinations of the elements of the set I_1. The probabilities associated with any combination of markers $M_{i_1}, \ldots, M_{i_r}$ is denoted as:

$$P_{(r)} = \prod_{i \in I_r} P_i \quad \text{and} \quad Q_{(r)} = 1 - \prod_{i \in I_r} P_i \tag{1}$$

The last expression measures the effectiveness of the combination of markers $M_{i_1}, \ldots, M_{i_r}$.

Let $C = (C_1, \ldots, C_k)$ be a parameter vector that gives the blood group determination costs for markers $M_1, \ldots, M_k$. If $N_{(r)}$ is the number of pairs whose zygosity is determined using r markers out of a total of k markers, then the total cost of determination is

$$N_{(r)} \cdot \sum_{i \in I_r} C_i \tag{2}$$

Let $(i_1^e, \ldots, i_r^e)$ be the index vector of the most effective combination of r markers and $Q_{(re)}$ its discriminating power. For comparisons with other $\binom{k}{r} - 1$ possible combinations of r markers we define a parameter vector $Q_c = (Q_{1c}, \ldots, Q_{kc})$, where Q_{rc} specifies how great a difference between the maximum value $Q_{(re)}$ and the value Q_i is allowed, when comparing alternative combinations with respect to cost. Thus the parameter vector Q_c defines the range in which the optimization between different combinations occurs.

If $(i_1, \ldots, i_r)$ and $(i_1', \ldots, i_r')$ are two alternative combinations of r markers and $N_{(r)}$ and $N_{(r')}$ are their minimum sample sizes needed to obtain the level $Q_{(re)} - Q_{rc}$ in discriminating power, then the combination that minimizes the total costs [Equation (2)] will be regarded as the best combination of r markers.

MINIMUM SAMPLE SIZE

A method is given for deriving the minimum sample size satisfying the desired accuracy of zygosity diagnosis, when estimates for the marker concordance probabilities are available.

Suppose that the mean concordance probabilities $P_1, \ldots, P_k$ are estimated on the basis of a sample with N observations. Then, according to the binomial distribution,

$$\hat{E}(\hat{P}_i) = \hat{P}_i \quad \text{and} \quad \hat{Var}(\hat{P}_i) = \hat{P}_i(1 - \hat{P}_i)/N$$

In the case of one marker M_i, an approximate minimum sample size N (pairs) can be derived from the inequality

$$\hat{P}_i + t_\alpha(N-1) \cdot \sqrt{\frac{\hat{P}_i(1-\hat{P}_i)}{N}} \leqslant P_i^c \tag{3}$$

where $t_\alpha(N-1)$ is the value of Student's t distribution, with degrees of freedom $N-1$ and risk level α, and $P_i{}^c$ is the minimum concordance probability set for the marker M_i. The inequality (3) for N is solved:

$$N \geqslant t_\alpha{}^2(N-1) \cdot (\hat{P}_i(1-\hat{P}_i))/(\hat{P}_i - P_i{}^c)^2 \tag{4}$$

Correspondingly, when considering a combination of r markers $M_1, \ldots, M_r$, it can be shown [5] that the upper confidence limit of the cumulative concordance probability of these markers is

$$\exp\left[\sum_{i=1}^{r}\left(\log_e\hat{P}_i - \frac{1-\hat{P}_i}{2N\hat{P}_i}\right) + t_\alpha(N-1)\frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^{r}(1/\hat{P}_i - 1)}\right] \tag{5}$$

By requiring that formula (5) be less than or equal to P_c, the critical predetermined minimum value set for the cumulative concordance probability, and by taking logarithms, we obtain:

$$\sum_{i=1}^{r}\left(\log\hat{P}_i - \frac{1-\hat{P}_i}{2 \cdot N \cdot \hat{P}_i}\right) + t_\alpha(N-1) \cdot \frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^{r}(1/\hat{P}_i - 1)} \leqslant \log_e P_c \tag{6}$$

The minimum sample size N that fulfills this inequality cannot be explicitly solved. We obtain, however, a rough estimate for N by considering $t_\alpha(N-1)$ as fixed and by leaving off the term $(1-\hat{P}_i)/2N\hat{P}_i$ from the left-hand side of inequality (6) and by solving with respect to N. In order to obtain a more accurate estimate for the minimum sample size we must use an iterative method.

The iterative algorithm NMIN given in Appendix I computes the minimum N that fulfills inequality (6) with the given accuracy. After determination of minimum sample size, other factors can be considered. In the following, data on the determination costs of the markers are used as well to obtain the optimal set of markers.

COST-EFFECTIVENESS ALGORITHM

The following algorithm presents a stepped method for determination of the cost-effective combinations of blood markers.

Input parameter vectors:

$P[1:k]$ = concordance probabilities of markers $M_1, \ldots, M_k$

$C[1:k]$ = marker costs per sample

$Q_c[1:k]$ = critical values of comparisons between the combinations of markers

$P_c[1:k]$ = critical values of concordance probabilities

Output parameter vectors:

Indexes of the optimum combinations of markers

$I_1^* = (i_1^*), I_2^* = (i_1^*, i_2^*), \ldots, I_k^* = (i_1^*, \ldots, i_k^*)$

$T_c^*[1:k]$ = Total costs of combinations $I_1^*, \ldots, I_k^*$

$N_c^*[1:k]$ = Number of pairs corresponding to the solutions $I_1^*, \ldots, I_k^*$ and the given accuracy level

Step 1. [Input of parameters]
Read values for vectors P, C, and Q_c.

Step 2. [Most cost-effective combinations]
Determine with the procedure COSTBENEFIT (Appendix II) the index vectors $I_1^*, I_2^*, \ldots, I_k^*$ corresponding to the most cost-effective combinations of markers.

Step 3. [Minimum value for sample size]
Determine with the procedure NMIN (Appendix I) for the optimum combinations of markers the minimum sample sizes $N_c^*[i], i = 1, \ldots, k$ corresponding to the accuracy level determined by P_c.

Step 4. [Total costs]
Determine the total costs $T_c^*[i], i = 1, \ldots, k$ for each combination of markers corresponding to the sample sizes $N_c^*[i]$.

Step 5. [End of the algorithm]

AN APPLICATION OF THE COST-EFFECTIVENESS ALGORITHM

To illustrate the function of the optimizing algorithm, data from a study of the Finnish Twin Registry is used. Zygosity determination was based on a questionnaire method and verified using blood marker determinations on a sample chosen at random [7]. As the order of the markers with respect to the discriminating power was not the same as the order with respect to costs of determining the marker in a blood sample, the method described was applied to study which combination of markers in the Finnish population would be the most cost-effective.

The order of markers with respect to their discriminating power $\hat{Q}_i$ and the estimated cost per DZ pair tested are given in Table 1. The actual costs for each marker were estimated on the basis of the official price of the comprehensive Finnish paternity testing battery. Since only the relative costs are significant, the actual monetary unit used (Fmk) is not necessary but is included to illustrate the scale of costs. Because costs per individual marker were not available, they were estimated using reagent costs and labor time per sample. The share of reagent costs in the total cost of each marker varied greatly as seen in the last column of Table 1. The labor cost was the same per unit work time in all markers, though in estimating the work time per sample the possibility of simultaneous analysis of samples was taken into account.

Table 2 summarizes the results given using the cost-effectiveness algorithms for the blood markers used in zygosity testing. The indexes of the most cost-effective combina-

https://doi.org/10.1017/S000156600000917X Published online by Cambridge University Press

TABLE 1. Order of Markers With Respect to Their Discriminating Power and the Estimated Costs per DZ Pair

	Marker	(No.)	Discriminating power Q_i	Costs/sample C_i (Fmk)	Share of reagent costs of total cost per sample (%)
1	MNSs	(2)	0.5579	4.71	78.6
2	Rh	(3)	0.5215	13.19	38.7
3	Gm	(7)	0.4535	26.27	7.6
4	A_1A_2BO	(1)	0.4397	1.31	22.9
5	AP	(10)	0.4411	17.08	5.3
6	AlAT	(11)	0.4062	17.08	5.3
7	Hp	(5)	0.3882	18.60	2.1
8	PGM	(9)	0.3244	34.16	5.3
9	Gc	(6)	0.2864	27.77	12.6
10	Fy^a	(4)	0.2326	7.07	14.1
11	AK	(8)	0.0843	35.41	8.6
			Total	202.65	11.2

tions of markers are given in column 2 for an increasing number of combinations r shown in column 1. The indexes of the combination of markers are listed in decreasing order of discriminating power. Column 3 shows the cumulative discriminating power of the corresponding combination in column 2. This naturally increases as the number of markers increases. After eight markers the increase in cumulative discriminating power is very slight.

As previously indicated the cost-effectiveness algorithm includes a parameter vector Q_c that controls the selection procedure of the combination of markers in each phase. The parameter determines the range in which comparisons between different combinations of markers in respect to the costs will be done. The selection order can be weighted in favor of either costs or effectiveness with this parameter vector. If the values of Q_c are chosen small compared to the values $1 - \max\{Q_{(r)}\}$ in each phase, then more weight will be given to effectiveness and in opposite case to costs. Therefore, in the extreme case, when all the elements of Q_c are chosen as zero, we obtain the order in respect of effectiveness, measured by discriminating power. As an example about the meaning of parameter Q_c let us consider the first element of Q_c in our numerical example, that was chosen to be 0.15 (column 4). It means that when we are selecting the most cost-effective single marker we allow into the scope of comparison of costs all those markers that are inside the range $\max\{Q_{(1)}\}-0.15$. In our numerical example the values of Q_c were chosen by giving more weight to effectiveness and simultaneously aiming at sample sizes of the same magnitude as the samples actually tested.

Column 7 gives the minimum sample size needed to reach the lower confidence limit of the cumulative discriminating power. The lower and upper confidence limits of the cumulative discriminating power for the sample size of column 7 are given in columns 5 and 6 respectively.

The last column gives the total cost of determining the blood markers for the sample size with the combination of markers of column 2. In this example the effect of high relative cost of a marker can be seen for the marker 7 (Gm serum group). Though the discriminating power of this marker is the third-best after markers 2 and 3, it is included in the most cost-effective combination only when five markers are considered.

TABLE 2. Results of the Cost-Effectiveness Analysis for the Blood Markers Used in Zygosity Testing

No. of comb. r	Indexes of the most cost-effective combinations of markers ($i_1, \ldots, i_r$)	Cumulative discriminating power $Q_{(r)}$	Critical values Q_{rc}	95% confidence limits		Sample size (pairs)	Total costs (Fmk)
				Lower	Upper		
1	(2)	0.5579	0.1500	0.4077	0.6765	56	490
2	(2, 1)	0.7523	0.1000	0.6884	0.8150	120	1,444
3	(2, 3, 1)	0.8815	0.0600	0.8243	0.9271	66	2,536
4	(2, 3, 1, 10)	0.9338	0.0400	0.8967	0.9622	64	4,645
5	(2, 3, 7, 1, 10)	0.9638	0.0250	0.9396	0.9804	58	7,256
6	(2, 3, 7, 1, 10, 11)	0.9785	0.0150	0.9640	0.9886	66	10,512
7	(2, 3, 7, 1, 10, 11, 5)	0.9868	0.0100	0.9771	0.9934	62	12,182
8	(2, 3, 7, 1, 10, 11, 5, 9)	0.9911	0.0075	0.9838	0.9958	58	15,358
9	(2, 3, 7, 1, 10, 11, 5, 9, 6)	0.9936	0.0050	0.9888	0.9968	68	21,783
10	(2, 3, 7, 1, 10, 11, 5, 9, 6, 4)	0.9951	0.0040	0.9912	0.9975	66	22,075
11	(2, 3, 7, 1, 10, 11, 5, 9, 6, 4, 8)	0.9954	0.0035	0.9940	0.9977	74	29,992

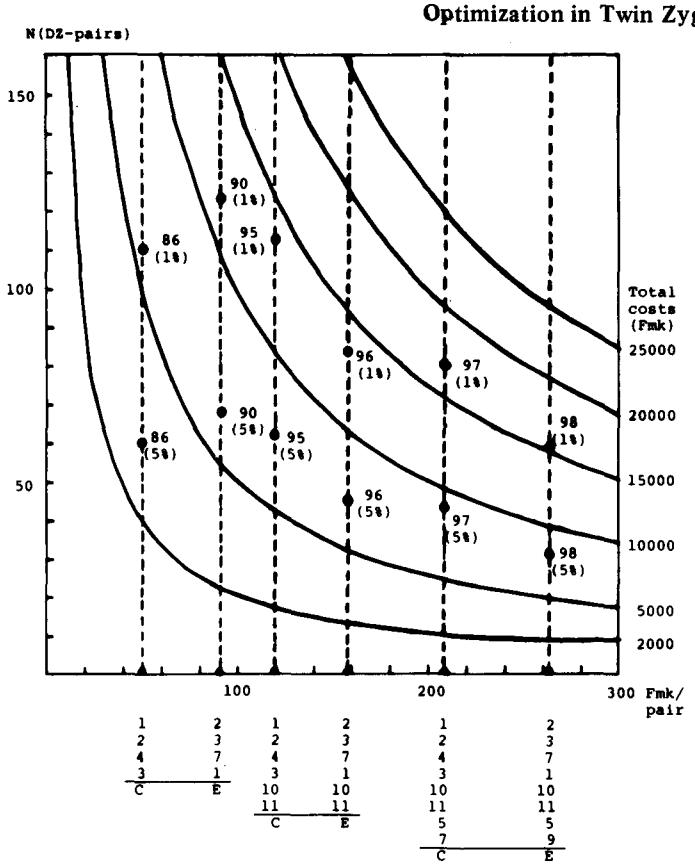

Figure. Family of graphs demonstrating the cost-effectiveness analysis: E) most effective combination of markers; C) cheapest combinations of markers. Markers listed in order of effectiveness or costs. See Table 1 for names of markers.

The Figure presents for different combinations of markers the sample size needed to reach a set discriminating power (with either 1% or 5% risks), at different levels of total costs. Both most cost-effective and cheapest combinations of markers are presented.

As an example of the use of the Figure, let us consider a situation where 15,000 Fmk are available for zygosity testing. Two different sample sizes are obtained using two different combinations at different levels of discriminating power. Using eight markers (1, 2, 4, 3, 10, 11, 5, 7) with a discriminating power of 0.97 at 5% risk level, the zygosity of 72 pairs can be determined. Using six markers (2, 3, 7, 1, 10, 11) with a discriminating power of 0.96 at 1% risk level, the zygosity of 94 pairs can be determined. Thus it can be seen that, for the same total cost, a greater number of pairs can be studied by accepting a slight decrease in discriminating power.

DISCUSSION

In recent years, an increasing number of multiallelic marker determination methods have become commonly available. It is generally unnecessary to have the laboratory determine

https://doi.org/10.1017/S000156600000917X Published online by Cambridge University Press

all available markers, for, as shown, a smaller set of markers will suffice for achieving adequate zygosity diagnosis.

Selvin [8] compared the efficiency of markers when all genotypes in a system were known. This, however, restricts the use of the markers to those where genotypic identification is possible. In this analysis only phenotypic identification and knowledge of population gene frequencies is required. In addition, determination costs are taken into account, as genetic marker determinations are expensive and may often form a substantial proportion of the total study costs. Optimization of these costs without significant loss in genetic efficiency is possible as presented in this paper. The parameter Q_c used defines the range of effectiveness values permitted, and optimization will occur within this range. The values chosen for the parameters Q_c depend on the purposes of the study, but in every case several runs with different values of Q_c will generally be needed before the final decision can be made.

The possibility of simultaneous analysis of several samples or of different markers will vary from laboratory to laboratory. This, however, can be taken into account in the application of the described algorithm. In some studies (eg [2]), some markers were determined only if the twin pair was concordant for all those previously determined. This will naturally have to be done for all MZ pairs, and for varying numbers of DZ pairs. The total costs of such a procedure were found for our material to be slightly greater than the costs estimated by the cost-effectiveness analysis.

Acknowledgments. This study has been supported by a grant from the Council for Tobacco Research, USA, Inc.

REFERENCES

1. Allen G (1976): Scope and methodology of twin studies. Acta Genet Med Gemellol (Roma) 25: 79–85.
2. Cederlöf R, Friberg L, Jonsson E, Kaij L (1961): Studies on similarity diagnosis in twins with the aid of mailed questionnaires. Acta Genet (Basel) 11:338–362.
3. Friedman GD (1977): A potential pitfall in studying trait-discordant twins. Am J Epidemiol 105: 291–295.
4. Juel-Nielsen N (1958): On the diagnosis of zygosity in twins and the value of blood groups. Acta Genet (Basel) 8:256–273.
5. Sarna S (1977): "Zygosity Diagnosis in Epidemiological Twin Studies." Doctoral dissertation, University of Helsinki.
6. Sarna S (In press). Probabilities of concordance of twins with respect to genetic markers. A general formulation. Acta Genet Med Gemellol.
7. Sarna S, Kaprio J, Sistonen P, Koskenvuo M (1978): Zygosity of twin diagnosis by mailed questionnaire. Hum Hered 28:241–254.
8. Selvin S (1977): Efficiency of genetic systems for diagnosis of twin zygosity. Acta Genet Med Gemellol 26:81–82.

APPENDIX I: ALGORITHM NMIN

Determines the minimum sample size N_{min} for the specified concordance probability P_c.

Step 1. [Values for the parameters]
Give t_α and ΔN (the accuracy wanted for N).

Step 2. [Initial approximation]

$$\text{Set } N = \left[t_\alpha (N-1) \sqrt{\sum_{i=1}^{r} (1/\hat{P}_i - 1)} / (\log P_c - \sum_{i=1}^{r} \log \hat{P}_i) \right]^2$$

Step 3. [Computation of the corrected N]
Compute

$$N_C = \left\{ t_\alpha (N-1) \sqrt{\sum_{i=1}^{r} (1/\hat{P} - 1)} \Big/ \left[\log P_C - \sum_{i=1}^{r} \left(\log \hat{P}_i - \frac{1-\hat{P}_i}{2N\hat{P}_i} \right) \right] \right\}^2$$

Step 4. [Termination criterion of the iteration process]
If $|N - N_C| < \Delta N$ *then* go to step 6 *else* go to step 5

Step 5. [Corrected estimate]
Replace N by N_C and go to step 3.

Step 6. [Final estimate]
Set $N = N_C$ as an output value and stop the procedure.

For most cases it is sufficient to demand that $\Delta N = 1$, because in practice N is an integer. It is reasonable, however, to consider N as a real number. In most cases the convergence of this algorithm is rapid and only two or three iterations will be needed to obtain the accuracy corresponding to the paramete value $\Delta N = 1$.

APPENDIX II: PROCEDURE COSTBENEFIT

Notations

k	Number of available markers $M_1, \ldots, M_k$.
r	Number of markers under consideration.
cprob [1:k]	Probabilities of concordance (P).
costs [1:k]	Costs per sample of blood group determinations (C).
critp	The critical value for comparisons between alternative combinations of r markers (Q_C).
indexes [1:r]	The vector of indexes of the resulting combination of markers.
nmin (t_α, ΔN, P, P_C)	Iterative procedure that determines the minimum sample size.
sort (Q, i)	Procedure that sorts the index vector (1, . . ., k) into the order (i1, . . ., ik) so that Q[i1] ⩾ Q[i2] ⩾ · · · Q[ik].
comb (Q, i, k, r, t, iout)	Procedure that selects the combinations of indexes for a given number r out of possible markers for testing against the best combination formed in the main program. The combinations are selected in decreasing order with respect to the discriminating power vector Q. Parameter t indicates the number of trials; so that with t = 1 the second most cost-effective combination is chosen for comparison; with t = 2, the third most cost-effective combination is chosen; etc. The vector "iout" gives the indexes of the combination to be tested.

```
procedure costbenefit(k,r,cprob,costs,critp,indexes);
real array cprob,costs;
integer array indexes;
real critp,tcosts;
integer k,r;
begin
   real array dpower[1:k];
   integer array ivector,icomb[1:k];
   real cvalue,ccosts,qvalue,qcosts,clevel,nc,nq;
   real tcostsc,tcostsq;label loop;
   integer trial,i;
   comment talfa and deltan are global variables;
   comment sort,comb and nmin are procedures;
   for i:=1 step 1 until k do dpower[i]:=1-cprob[i];
   sort(dpower,ivector);
   cvalue:=1;ccosts:=0;
   for i:=1 step 1 until r do
```

https://doi.org/10.1017/S000156600000917X Published online by Cambridge University Press

```
begin
    comment discriminating power and costs of the most effective combination of r markers;
    cvalue:=cvalue * cprob[ivector[i]];
    ccosts:=ccosts+costs[ivector[i]]
end;
cvalue:=1-cvalue; clevel:=cvalue-critp;
for i:=1 step 1 until r do indexes[i]:=ivector[i];
trial:=1;
loop:comb(dpower,ivector,k,r,trial,icomb);
    qvalue:=1;qcosts:=0;
    for i:=1 step 1 until r do begin
    comment discriminating power and costs of the new combination;
        qvalue:=qvalue * cprob[icomb[i]];
        qcosts:=qcosts+costs[icomb[i]] end; qvalue:=1-qvalue;
    if qvalue>clevel then begin
    comment comparisons of the total costs;
    nc:=nmin(talfa,deltan,1-cvalue,1-clevel);
    nq:=nmin(talfa,deltan,1-qvalue,1-clevel);
    tcostsc:=nc * ccosts;
    tcostsq:=nq * qcosts;
    if tcostsq<tcostsc then begin
        comment change of comparison and output values;
        cvalue:=qvalue;
        ccosts:=qcosts;
        for i:=1 step 1 until r do
            indexes[i]:=icomb[i];
        end;
    trial:=trial+1;go to loop
    end;
end of costbenefit;
```

Correspondence. Dr. Seppo Sarna, Department of Public Health Science, Haartmaninkatu 3, SF-00290 Helsinki 29, Finland.

https://doi.org/10.1017/S000156600000917X Published online by Cambridge University Press