Title: Modelling of errors in databases

Abstract

A lot of time and energy are expended assembling national databases containing information about health care processes and outcomes. Unfortunately, given the complexity of the data gathering procedures involved, errors occur. This inevitably leads to problems when it comes to the analysis of data from such sources. Indeed, sometimes it is very much a matter of faith that summary statistics represent a true reflection of the facts. On the assumption that one knows the rates at which different forms of errors occur, mathematical modelling methods can be used to obtain estimates of the effects of such errors on the estimates that would be derived for summary statistics associated with an erroneous data base.

Key Words: Databases, Mathematical Modelling, Errors

1. **Introduction**

A major problem faced by those who rely on database systems is that data quality is often questionable. It is relatively common that data are missing or miscoded; indeed, unless large amounts of time and resources are expended, such errors are almost inevitable. Various measures can be taken to try to promote data quality, however it is very difficult to ensure that databases are completely accurate. Often all that can be achieved is a belief that most of the data are correct, but with relatively little hard evidence about the exact proportion of records that are erroneous. This is clearly less than satisfactory.

It is worth giving an example of the extent of this problem. Various authors have reported error rates in the miscoding of patient factors affecting diagnosis and hence risk estimators (e.g. [1,2,3,4,5]), with one study on a carefully audited Californian database reporting at least one diagnostically relevant error in 63% of patient records [1]. Evidence presented to the Bristol Royal Infirmary Inquiry, a major inquiry into mortality rates at a leading UK hospital [6], showed considerable differences between mortality rates calculated from two national databases. Depending on the class of operation concerned, the ratio of mortality rates from the two sources varied between 0.46 and 1.76. This is a major issue since it is difficult to determine acceptable mortality standards given such a hazy notion of what outcomes actually are. Such large scale divergence between national data sources also makes the task of statistical inference unrealistic, given the need to correct for case mix [1,7].

In view of the errors that may exist within databases, analyses that make use of such data are often conducted with a degree of optimism that the results being derived are sound (see for example [8]). Whether such optimism is warranted is debatable since it is far from clear what effect data errors have. Also, given an acceptance that there are possibly major errors, what analysis methods are appropriate? Clearly the standard methods of statistics are not, since they are based on the assumption that the underlying data are correct. The publication of risk-adjusted mortality rates for individual hospital units is already widespread, and the introduction of the outcomes for an individual surgeon is expected soon [9]. In a recent paper discussing the benefits and disadvantages to publishing outcome data, Mason and Street [9] give the effect of miscoding errors only small consideration and assume that large enough databases will mitigate their effect. We consider it timely to explore the validity of such assumptions in this paper.

In a study of 2x2 contingency tables, Bross [10] set up the standard framework for misclassification studies by considering binary misclassification of data. Valuable later work has focused on the interpretation of estimates derived from databases in the presence of such errors using Bayesian techniques (e.g. [11,12]) or direct comparison (e.g. [1]). Here, the focus is on a complementary problem suggested to the authors by Jaroslav Stark, an eminent paediatric cardiothoracic surgeon with particular interest in the topic of clinical data bases. He suggested an investigation whereby one would take an existing clinical data base, known to be accurate, and then deliberately seed it with errors of different sorts, using simulation methods, to examine the consequences in terms of consequent errors in the resulting statistics. Sutherland and Botz [2] have reported such a simulation study where they set probabilities for misclassifying the case complexity of a patient and considered the resulting impact from a health economic point of view on the cost weightings given to hospitals.

The purpose of the present paper is to describe a mathematical model that can be used to examine the effect of such miscoding of errors, including data omission, on estimates of mortality but without using a specific database or simulation. Relatively simple probability theory suffices, albeit resulting in quite complex formulae.

Before describing specific analysis related to errors, it is useful to establish a general mathematical result related to the stochastic properties of certain set transformations.

2. Analysis of a general set transformation process

Suppose there are two collections of K sets indexed $S_0, \cdots, S_K$ and $\hat{S}_0, \cdots, \hat{S}_K$. Suppose further that initially that there are N_i elements in set S_i, $1 \le i \le K$ and that the sets $\hat{S}_0, \cdots, \hat{S}_K$ are empty.

Consider transitions that occur whereby elements from the sets $S_0, \cdots, S_K$ transfer to the sets $\hat{S}_0, \cdots, \hat{S}_K$, as illustrated in Figure 1.

The transfers are assumed to occur at random, $\alpha_{i,\,j}$ denoting the probability that an element in set S_i transfers to set $\hat{S}_j$, $0 \le i \le K$, $0 \le j \le K$. We assume that elements transfer independently of one another. Further, since the $\alpha_{i,\,j}$ denote probabilities, it is implicitly assumed that

$$\sum_{j=0}^{K} \alpha_{i,\,j} = 1 \qquad 0 \le i \le K. \tag{1}$$

- - -
- INSERT FIGURE 1 NEAR HERE
- - -

Let $\hat{N}_j$ denote the random variable corresponding to the number of elements in set $\hat{S}_j$ once transfers have been made, $0 \le j \le K$.

It is of interest to know the mean and variance of $\hat{N}_j$ and also the variance-covariance matrix for pairs $\hat{N}_j, \hat{M}_t$. Let $\hat{W}$ be a $K \times K$ matrix with elements $\{\hat{w_{i,\,j}}\}$ which are random variables where, for $0 \le i \le K$ and $0 \le j \le K$, $\hat{w_{i,\,j}}$ denotes the number of elements from the *i*-th set that are transferred to the *j*-th set. Then

$$P(\hat{w_{i,j}} = r) = \begin{pmatrix} N_i \\ r \end{pmatrix} \alpha_{i,j}{}^{r} (1 - \alpha_{i,j})^{(N_i - r)} \tag{2}$$

Given that the random variables $\{\hat{w_{i,\,j}}\}$ are all binomially distributed, the following are standard results for such variables:

$$E(\hat{w_{i,j}}) = N_i \alpha_{i,j} \tag{3}$$

$$Var(\hat{w_{i,j}}) = N_i \alpha_{i,j} (1 - \alpha_{i,j}) \tag{4}$$

$$Cov(\hat{w_{i,j}}, \hat{w_{s,t}}) = 0, \quad i \ne s \tag{5}$$

$$Cov(\hat{w_{i,j}}, \hat{w_{i,t}}) = -N_i \alpha_{i,j} \alpha_{i,t}, \quad j \neq t . \tag{6}$$

Note that equation (5) follows from the transfer of elements from different sets being independent processes.

The total number of elements $\hat{N}_j$ in set $\hat{S}_j$ is simply the sum of the number of elements transferred to $\hat{S}_j$ from each S_i .

$$\hat{N}_j = \sum_{i=0}^{K} \hat{w_{i,j}} \tag{7}$$

Since the transfer of elements from sets S_i and S_j is independent for $i \neq j$ we can use the following standard results:

$$E(\hat{N}_j) = \sum_{i=0}^{K} E(\hat{w_{i,j}}) \tag{8}$$

$$Var(\hat{N}_j) = \sum_{i=0}^{K} Var(\hat{w_{i,j}}) \tag{9}$$

$$Cov(\hat{N}_j, \hat{M}_t) = \sum_{i=0}^{K} Cov(\hat{w_{i,j}}, \hat{w_{i,t}}), \quad j \neq t \tag{10}$$

The simple form of equation (10) is a result of equation (5) above. Substituting equations (3), (4) and (6) into equations (8), (9) and (10) gives:

$$E(\hat{N}_j) = \sum_{i=0}^{K} N_i \alpha_{i,j} \tag{11}$$

$$Var(\hat{N}_j) = \sum_{i=0}^{K} N_i \alpha_{i,j} (1 - \alpha_{i,j}) \tag{12}$$

$$Cov(\hat{N}_j, \hat{M}_t) = -\sum_{i=0}^{K} N_i \alpha_{i,j} \alpha_{i,t}, \quad j \neq t \tag{13}$$

This result is useful, since these statistics allow one to approximate the distribution of $\hat{N}$ using a multivariate normal distribution. However, in the context of clinical data base errors, we are rather more concerned with obtaining distributions for the ratio of the relative sizes of two sets. To calculate such quantities, one needs to make use of the following lemma:

Lemma 1

If *X* and *Y* are two random variables and if $Z=\frac{Y}{X}$ then

$$E[Z]\approx\frac{E[Y]}{E[X]}+\frac{1}{E^3[X]}\left(E[Y]\ Var[X]-E[X]\ Cov[X,Y]\right)$$

and

$$Var[Z]\approx\frac{1}{\left(E[X]\right)^4}\left(E^2[Y]\ Var[X]-2E[X]\ E[Y]\ Cov[X,Y]+E^2[X]\ Var[Y]\right)$$

Proof

This is a standard result obtained by using a Taylor series expansion of *Z* about the values $E[Y]$ and $E[X]$ (For example, see Rice, p153 [13]).

Deriving approximate formulae for the proportional split between two sets involves a little more algebraic manipulation, however it is not too difficult to establish the following:

Lemma 2

If X and Y are two random variables and if $W=\frac{Y}{(X+Y)}$ then

$$E[W]\approx\frac{E[Y]}{(E[X]+E[Y])}+\frac{1}{(E[X]+E[Y])^3}\left(Cov[X,Y]\,(E[Y]-E[X])+Var[X]E[Y]-Var[Y]E[X]\right)$$

and

$$Var[W]\approx\frac{1}{(E[X]+E[Y])^4}\left(E^2[Y]\;Var[X]-2E[X]\;E[Y]Cov[X,Y]+E^2[X]\;Var[Y]\right)$$

Proof

Define the random variable $V=(X+Y)$ then it can easily be verified that

$$Var[V]=Var[X]+Var[Y]+2\,Cov[X,Y] \tag{14}$$

and

$$Cov[V,Y]=Var[Y]+Cov[X,Y]\ . \tag{15}$$

Substituting these expressions into the expressions from Lemma 1 gives the required result.

Although rather complex as formulae, the expressions in Lemma 2 are computationally simple.

3. Applying set transformation findings to the analysis of errors in surgical data bases

The expressions derived in Section 2 are generic in nature. In order to apply them in the specific case of clinical database errors, there is a need to introduce additional notation and assumptions related to errors likely to occur in practice.

The motivation underlying this is that there is some method for classifying clinical events into K disjoint categories and that initially the sets $S_1,\cdots,S_K$ correspond to the actual occurrences of each category of event. For example, one set might correspond to a collection of 15 occurrences of an operation to repair an atrial septal defect which resulted in peri-operative death (irrespective of whether records of these procedures subsequently got lost or miscoded). The set S_0 serves a special function discussed below.

The processes involved in recording clinical information and transferring this to a database may lead to errors whereby clinical facts are misinterpreted or miscoded, or indeed information may simply be lost. These errors may be thought of as transforming the sets $S_1,\cdots,S_K$ into sets $\hat{S}_0,\cdots,\hat{S}_K$ the latter corresponding to information that is recorded in the database. The set $\hat{S}_0$ corresponds to information lost from the system, the set S_0 being the empty set.

The consequences of data errors may thus be considered in terms of transformations of the sets and analysed using the results from Section 2.

In order to be able to apply the generic results concerning set transformation to a specific analysis of clinical database errors, it is necessary to introduce notation.

3.1 Notation relating to errors in a clinical database

Suppose that the data of interest concern the occurrence and mortality outcome of a number of different types of surgical procedure during a particular audit period.

Suppose we have Q different operation types indexed $1,\cdots,Q$. Suppose that during the audit period being considered, there have been L_q operations of type q that have resulted in survival, $1\le q\le Q$. Also, suppose that there have been D_q deaths following operations of type q, $1\le q\le Q$.

Let δ denote the probability that a death is miscoded as a survivor.

Let ε denote the probability that a survivor is miscoded as a death.

Let ζ denote the probability that the during the data collection process, an operation that results in a death is omitted from the database.

Let η denote the probability that the during the data collection process, an operation that results in survival is omitted from the database.

We assume that these miscoding and data omission probabilities are independent.

3.2 Estimation of errors associated with the overall mortality rate

Overall, *L*, the number of operations performed during the audit period which result in survival is given by

$$L = \sum_{q=1}^{Q} L_q \tag{16}$$

Also, *D*, the total number of deaths is given by:

$$D = \sum_{q=1}^{Q} D_q \tag{17}$$

Thus the true mortality rate π during the audit period is given by

$$\pi = \frac{D}{D+L} = \frac{\sum_{q=1}^{Q} D_q}{\sum_{q=1}^{Q} \left(D_q + L_q\right)} \tag{18}$$

However, due to miscoding and data omission errors, both the numerator and denominator of this expression might be erroneously estimated.

In this case, to use the set transformation analysis from Section 2, we need only consider sets S_0, S_1, S_2, $\hat{S}_0$, $\hat{S}_1$ and $\hat{S}_2$ where S_0 is empty, S_1 corresponds to operations of all types that result in survival and S_2 corresponds to operations of all types that result in peri-operative death. The sets $\hat{S}_1$ and $\hat{S}_2$ correspond to information recorded in a database concerning the number of survivals and deaths and the set $\hat{S}_0$ corresponds to information lost to the system.

In modelling the errors introduced in the data compilation process, the probabilities of transfers between these sets, $\{\alpha_{ij}\}$, are given by

$$\begin{aligned}
&\alpha_{00}=1\\
&\alpha_{01}=0\\
&\alpha_{02}=0\\
&\alpha_{10}=\eta\\
&\alpha_{11}=(1-\varepsilon)(1-\eta)\\
&\alpha_{12}=\varepsilon(1-\eta)\\
&\alpha_{20}=\zeta\\
&\alpha_{21}=\delta(1-\varepsilon)\\
&\alpha_{22}=(1-\delta)(1-\zeta)
\end{aligned} \tag{19}$$

If $\hat{L}$ and $\hat{D}$ are random variables corresponding to the number of operations resulting in survival and peri-operative death respectively, then by Theorem 1,

$$E[\hat{L}]=(1-\varepsilon)(1-\eta)L+\delta(1-\zeta)D \tag{20}$$

$$E[\hat{D}]=\varepsilon(1-\eta)\,L+(1-\delta)(1-\zeta)D \tag{21}$$

$$Var[\hat{L}]=(1-\varepsilon)(1-\eta)(1-(1-\varepsilon)(1-\eta))L+\delta(1-\zeta)(1-\delta(1-\zeta))D \tag{22}$$

$$Var[\hat{D}]=\varepsilon(1-\eta)(1-\varepsilon(1-\eta))L+(1-\delta)(1-\zeta)(1-(1-\delta)(1-\zeta))D \tag{23}$$

$$Cov[\hat{L},\,\hat{D}]=-(\varepsilon(1-\varepsilon)(1-\eta)^2L+\delta(1-\delta)(1-\zeta)^2D) \tag{24}$$

4. Computation of statistical consequences of errors in data bases

The estimates (20), . . ., (24) can be substituted into the expressions given in Lemma 2 to give an approximation for the mean and variance of the overall mortality that would be estimated from the database. In addition the formulae above can be adapted to incorporate errors resulting from the miscoding of operations. These formulae are complex, yet computationally straightforward to evaluate. A simple Excel spreadsheet has been devised to carry out these calculations and this can be downloaded, free of charge, from our website at http://www.ucl.ac.uk/operational-research/downloads.

This spreadsheet model has been used to provide estimates for the following hypothetical worked example, deliberately chosen to be simple for illustration purposes. Suppose we have a small database concerning four types of operation: A,B,C and D and that a primary question of interest is to estimate the mortality rate for operation "A". Table 1 summarises data that a user of the spreadsheet model would supply when investigating the potential effects of errors in the recording of data concerning these operations on the estimated mortality rate for operation "A". Table 2 shows the data summaries as calculated by the spreadsheet model.

The formats of these table are similar to that used in the spreadsheet, although the spreadsheet would allow data for many more operations to be included that might be miscoded as operation "A".

Although the data used in Table 1 is hypothetical, it has been chosen as comparable to what might be found in a data base concerning outcomes for complex paediatric cardiac surgery. The assumed data error rates are not outlandish, indeed for those used to dealing with real clinical data, the choice of such error rates might seem somewhat conservative. As this example shows, the effects of such errors on the estimates of the mortality rate for operation "A" have been substantial. It is worth noting that the 'protective' effect of a relatively large number of operations for "A" assumed by [8] is not apparent.

- - -
- INSERT TABLES 1 & 2 NEAR HERE

-   -

5. Conclusion

This study shows that if we have estimates for the rates at which different types of error occur in clinical data gathering and coding processes, then these can be used to predict the scale of errors likely to occur in the summary statistics that are derived from a database. The requirement to estimate such summaries is of course one of the principal reasons that national databases are established in the first place.

Although this establishes the applicability of mathematical modelling to this problem, this is very much a first step. While we have proposed plausible ways in which data may be corrupted and examined the consequences of these, we should stress that we have not considered other potential sources of error. For example our analysis does not include the possibility of deliberate falsification of data, of deliberate systematic data loss nor systematic non-compliance. This is a clear limitation in this analysis.

Another limitation is that, while we have addressed the question "given there are errors, how does this effect the accuracy of mortality rate estimation?", what we have not done is to address the question: "what is the best estimate of mortality rate given there are errors?". The latter seems a more difficult issue. A referee has kindly suggested that statistical methods for modelling measurement error may have relevance here [14, 15] and this may be a promising line for further inquiry, although not an area where the authors would claim to have expertise. It is possible that techniques such as maximum likelihood estimation might be applicable to this (for instance see [11, 12]), although a full discussion is beyond the scope of the present paper. Another referee suggested that data mining techniques might be applicable. Again, we have insufficient experience to add to this suggestion.

An important topic concerns how one might derive estimates for different error rates. In practice, data cleaning exercises shed some light here. In an earlier career, the first author had experience of analysing number plate surveys used to track the journeys that vehicles made through an urban area. Here, observers would be sited at junctions throughout the city and noted the time and registration number of vehicles as they passed. Although observation and recording errors were common, it was found that error rates could be estimated and due allowance made.

An additional limitation of the work is that it is rather theoretical in nature thus it is important to indicate potential areas of practical application. This is research very much intended to underpin other more practical research investigations. An immediately important research question is to assess the extent to which data errors affect mortality rate estimation in practice. This has been investigated by the authors using this analytical framework and the results are disturbing. There is also scope to use such methods to investigate the knock-on effects of data errors on risk modelling and mortality audit which is an area of considerable interest to the clinical community as a whole and to the authors in particular [16, 17, 18]. The results also have potential application to the analysis of a problem faced by those who manage databases, which is how to decide how much time and effort to devote to improving data quality. In principle, an indefinite amount of resources could be devoted to this, but what are the benefits? How accurate does the database have to be? Here the model discussed has considerable relevance since it gives a rationale for considering the pay-off between reducing error rates and improving the accuracy of summary statistics.

Acknowledgements

The Clinical Operational Research Unit receives funding from the UK government Department of Health.

5. References

[1] J. Green and N. Wintfield. How accurate are hospital discharge data for evaluating effectiveness of care? Medical Care. 31(8): 719-731, 1993

[2] J.M. Sutherland and C.K. Botz. The effect of misclassification errors on case mix management. Health Policy. 79:195-202, 2006.

[3] H. Park and Y. Shin. Measuring case-mix complexity of tertiary care hospitals using DRGs. Health Care Management Science. 7:51-61, 2004

[4] M. Wilchesky, R.M. Tamblyn and A. Huang. Validation of diagnostic codes within medical services claims. Journal of Clinical Epidemiology. 57:131-141, 2004

[5] H.A. Khwaja, H. Syed and D.W. Cranston, Coding errors: a comparative analysis of hospital and prospectively collected departmental data, BJU International, 89:178-180, 2002

[6] D.J. Spiegelhalter, S. Evans, P. Aylin, J. Murray. Overview of statistical evidence presented to the Bristol Royal Infirmary Inquiry concerning the nature and outcomes of paediatric cardiac surgical services at Bristol relative to other specialist centres from 1984 to 1995. In: The Bristol Royal Infirmary Inquiry. Learning from Bristol. The Report of the Public Inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995. London, The Stationery Office, Annex B, 2001.

[7] S. Gallivan. Assessing mortality rates from dubious data - When to stop doing statistics and start doing mathematics. Health Care Management Science. 8: 237-241, 2005.

[8] I. Scott, D. Youlden and M. Coory. Are diagnostic specific outcome indicators based on administrative data useful in assessing quality of hospital care? Quality and Safety in Health Care. 13:32-29, 2004

[9] A. Mason and A. Street. Publishing outcome data: is it an effective approach? Journal of Evaluation of Clinical Practice. 12(1): 37-48, 2006

[10] I. Bross. Misclassification in 2x2 tables. Biometrics. 10(4): 478-486, 1954.

[11] M. Ladouceur, E. Rahme, C.A. Pineau and L. Joseph. Robustness of prevalence estimates derived from misclassified data from administrative databases. Biometrics. 63: 272-279, 2007.

[12] T. Swartz, Y. Haitovsky, A. Vexler and T. Yang. Bayesian identifiability and misclassification in multinomial data. Canadian Journal of Statistics. 32(3): 285-302, 2004.

[13] J.A. Rice, Mathematical statistics and data analysis, (Duxbury Press, Belmont California, 1995).

[14] M.G. Kendall and A. Stuart. The advanced theory of statistics, volume 2, 4th ed. (Griffin, London, 1979).

[15] C.L. Cheng and J.W. Van Ness. Statistical regression with measurement error, (OUP, New York, 1999).

[16] Lovegrove.J., Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S.
Monitoring the results of cardiac surgeons by variable life-adjusted display (VLAD).
The Lancet. 350: 1128-1130, 1997.

[17] Gallivan S., Davis K.B., Stark J., Early identification of divergent performance in congenital cardiac surgery. European Journal of Cardio-thoracic Surgery . 20: 1214-1219, 2001.

[18] Gallivan S. How likely is it that a run of poor outcomes is unlikely? European Journal of Operational Research. 150: 46-52, 2003.

Legends

Figure 1. Illustration of a general set transformation process.

Table 1. Hypothetical data used to illustrate input required to spreadsheet model for forecasting the effects on estimates of known error rates in data recording.

Table 2. Summary information calculated by spreadsheet model concerning distribution of mortality rates that would be estimated dependent on the occurrence of data recording errors

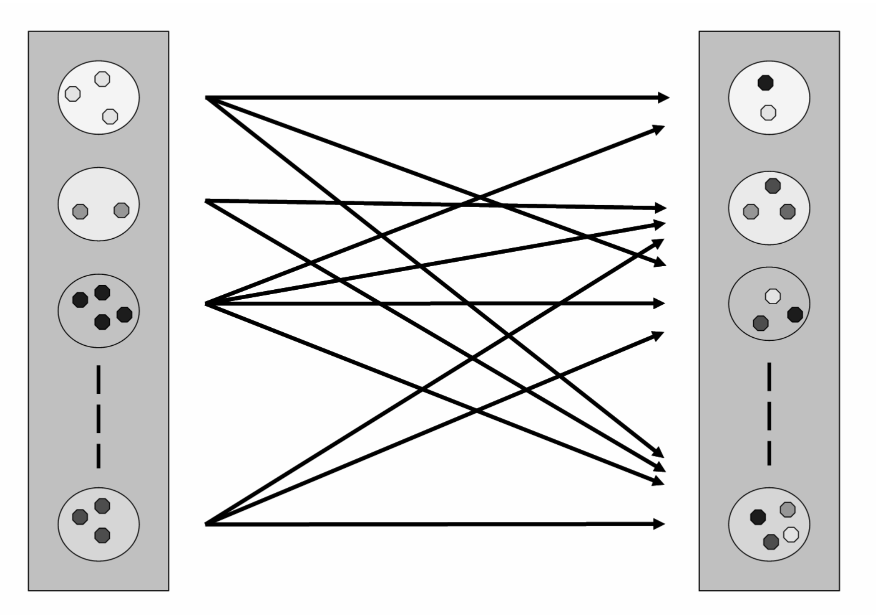

Figure 1 - Illustration of a general set transformation process.

	User supplied data related to operation A				
Op.	True No. survivors	True No. deaths	% deaths coded as survivors	% survivors coded as deaths	% cases lost from database
A	300	9	2.00%	2.00%	3.00%
	User supplied data related to other operations.				
Op.	True No. survivors	True No. deaths	% deaths coded as survivors	% survivors coded as deaths	% cases coded as operation A
B	850	7	3.00%	4.00%	1.00%
C	1250	40	1.00%	1.00%	1.00%
D	500	90	2.00%	4.00%	0.50%

Table 1. Hypothetical data used to illustrate input required to spreadsheet model for forecasting the effects on estimates of known error rates in data recording.

Calculated information concerning operation: “A”		
True mortality rate	**Estimated mortality rate**	**Standard deviation of estimated mortality rate**
2.91%	4.89%	0.84%

Table 2. Summary information calculated by spreadsheet model concerning distribution of mortality rates that would be estimated dependent on the occurrence of data recording errors