A guide to the QAMC obstetric risk prediction web page

D. R. Lovell, M. Niranjan, R. W. Prager,
K. J. Dalton and R. Derom

CUED/F-INFENG/TR301

September 15, 1997

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
United Kingdom

A guide to the QAMC obstetric risk prediction web page

D. R. Lovell, M. Niranjan, R. W. Prager, K. J. Dalton and R. Derom*
{drl,niranjan,rwp}@eng.cam.ac.uk,
kjd5@cam.ac.uk

September 15, 1997

Abstract

The Quality Assurance in Maternity Care (QAMC) project has produced a web page to demonstrate the potential benefits in combining three elements: interactive access to information via the World Wide Web; statistical risk models and data mining techniques; large databases of medical information.

This document describes the context in which this web page was developed, as well as the strengths and weaknesses of this approach to medical risk prediction.

1 The aims of the QAMC project

The QAMC project began in 1995 with the primary aim being to explore artificial neural networks for obstetric risk prediction. This investigation required neural networks to be meaningfully compared to alternative risk models and a detailed description of this research is given in [6, 7]. The initial motivation behind the project was

1. to try to develop systems that could provide clinicians early warning of cases with a high risk of having specific *adverse pregnancy outcomes* (APO — see Section A).
2. to see whether useful models of risk could be constructed from large databases of routinely collected perinatal information;
3. to address theoretical issues, such as how to select good predictors of outcome using the information contained in large medical databases.

*The first three authors may be contacted at the Cambridge University Engineering Department, Trumpington St., Cambridge CB2 1PZ, UK. K. J. Dalton is with the Cambridge University Department of Obstetrics and Gynaecology, Rosie Maternity Hospital, Cambridge CB2 2SW, UK. R. Derom is with the Katholieke Universiteit Leuven, Center for Human Genetics, Herestraat 49, B-3000 Leuven, Belgium.

As the project evolved, it became clear that there was an additional reason for doing this research: to report on the factors which currently restrict the usefulness of medical risk prediction systems. As we shall show, these factors range from difficulties in data collection and problem specification, through to medico-legal issues and interpretability of results.

Before we discuss these matters, we shall describe the QAMC project's risk prediction web page in detail. The page provides access to the observed and estimated incidence of failure to progress in labour (see Section A.1) within a portion of the *SMR2 database* (see Section B).

2 Modelling the risk of failure to progress in labour

The QAMC web page[1] allows a user to enter a description of a particular case[2]. The case can be described in terms of

- maternal age (in five year groups);
- number of children now living;
- maternal height (less or greater than 155cm);
- previous obstetrical history (previous Caesarian sections or neonatal deaths);
- current health state as represented by 30 different ICD-9 codes [12].

The reasons behind this particular choice of descriptive variables are discussed in detail in [6, 7]. Suffice to say that they were found to provide good discrimination between cases in which a failure to progress occurred, and cases in which labour proceeded normally.

After entering a specific patient profile, a user can (after agreeing to a legal disclaimer) receive information about the observed and estimates incidence of failure to progress for a case with those characteristics. This information is presented in three parts.

2.1 Prior probability of failure to progress

Before we know any information about a specific case, we can express our degree of *prior* belief that a labour will fail to progress. This prior can be formed in a variety of ways but it is most convenient for us to give the rate of incidence of failure to progress observed in the SMR2 data from 1980-88 (9.61%).

[1] See `http://svr-www.eng.cam.ac.uk/projects/qamc/FailProgress.05.html`

[2] In this document, we use the terms "case", "patient" and "patient profile" for convenience. We would not normally refer to mothers-to-be in this way. Similarly, the term "user" is a convenient way to refer to a person accessing the web page via a web browser.

2.2 Observed rate of incidence

Given a specific patient profile, we can report the number of cases in the SMR2 data that exactly match that description, and also the number of matching cases in which an adverse outcome occurred. If we were to ignore our prior belief, we could use the ratio of these numbers to estimate the probability of adverse outcome associated with the patient profile. The variance of this *maximum likelihood* estimate depends on the total number of cases matching the specified profile: the more cases, the lower the variance.

Note that this section of the web page reports the cases that *exactly* match the description provided. This can come as a surprise to uses who expect the page to report cases that match *any* of the specified characteristics. For example, the results of a query regarding a 20-24 year old mother, who has 2 children, is taller than 155cm and is obese (ICD-9 code `2780`) will report all cases in the SMR2 data with those *and only those* characteristics. The query *will not* cause a report which aggregates the number of cases with the specified characteristics and *any* other conditions.

Different aspects of this section of the web page are best demonstrated using four test cases. (Readers of the printed version of this document will have to find some means to access the QAMC web page to get the full benefit of these examples.)

1. **Age 20-24, 0 children, taller than 155cm, no previous obstetric history or ICD-9 codes.**
   This is the modal patient profile in the SMR2 data. Note the observed incidence of failure to progress is about 5% higher than the incidence across the entire database.

2. **Age 15-19, 4+ children, taller than 155cm, no previous obstetric history or ICD-9 codes.**
   ISD Scotland has asked that we suppress precise details of profiles for which there are fewer than 10 cases. The observed incidence of failure to progress is rounded to the nearest 5%.

3. **Age 15-19, 4+ children, taller than 155cm, no previous obstetric history, obese (ICD-9 code `2780`).**
   No instances of this profile are recorded in the SMR2 data. Recall that we are searching for and exact match of all details, including the presence of ICD-9 code `2780` *and the absence of all other codes.*

4. **Age 25-29, 0 children, shorter than 155cm, no previous obstetric history, poor fetal growth (ICD-9 code `6565`) AND excessive fetal growth (ICD-9 code `6566`).**
   This case suggests the presence of inconsistent information in the SMR2 data — it is very unlikely for both poor *and* excessive fetal growth to occur. A quality audit of 651 SMR2 records (presented in [6]) showed that ICD-9 codes of maternal condition were particularly prone to error. These errors can cause incongruities in prediction systems derived from the data [6].

2.3 Estimated rates of incidence

This section of the web page shows the *predicted* risk of adverse outcome according to three different risk models: logistic regression [3], an ensemble of neural networks [1, 10] and a look-up table smoothed with a Dirichlet[3] prior [8]. Details of these models and how they were trained are given in [6, 7].

These models estimate the probability of adverse outcome associated with (*i.e.*, conditional upon) a specific patient profile. They *do not* provide information about the joint probability of patient characteristics, that is to say, they cannot directly estimate whether a given profile is common or unusual. Some degree of *novelty detection* is possible by looking at how well these prediction models agree. If the *mean absolute deviation* of the three predictions is greater than 5% (as in case 4 above), a warning message is printed with the predictions.

3 Advantages

As we mentioned in Section 1, part of the motivation behind developing this web page was to explore the shortcomings of this approach to risk prediction. Before discussing these drawbacks, there are two major benefits worth mentioning.

3.1 Easy access

The World Wide Web paradigm facilitates access to database information and risk predictions. At the same time, the database and prediction systems remain under the control of the web page authors. Maintenance or alteration to these systems is a straightforward matter, as opposed to situations where software is supplied on portable media (*e.g.*, diskette or CD-ROM).

3.2 Mining the data mountain

The QAMC web page demonstrates a number of practical ways to allow large amounts of data to be "mined" for useful information. Simple database look-up enables clinicians to determine rates of incidence of failure to progress across over 770 000 cases. With a little extra effort, more powerful search engines could provide information on more complex queries. Further analyses, such as risk prediction, can be supplied if necessary.

Note that identifying information can easily be excluded from such database and prediction systems. ISD Scotland and other database providers have ensured that patient anonymity has been preserved at all stages of the QAMC project.

A great many databases of patient information have been, and continue to be collected. The collection of this information is often funded by public money, yet the data is seldom made publicly available. Patient confidentiality is often cited as a primary reason for this

[3]Technically speaking, in this two-class problem, we should speak of using a Beta prior.

secrecy but, as the QAMC web page demonstrates, access to useful information can be provided without compromise to privacy.

4 Problems

Successful application of technology is often an incremental process. It takes a number of iterative refinements to make an idea work, and a thorough exposition of the problems encountered at each step is a necessary part of that evolution.

The QAMC project has encountered several interesting problems in developing a web based system for risk prediction in obstetrics. Surprisingly few of them are related to shortcomings in the risk prediction models themselves. As we shall see, far more difficulties arose in relation to the raw data and the interpretation of results.

4.1 Limitations of the data

None of the databases available to the QAMC project were collected with the objective of risk prediction in mind. At this time, perinatal data is gathered for legal reasons, hospital administration reasons, and as a means to provide a quality audit [9]. These objectives have a profound effect on the different kinds of information gathered, and the reliability with which they are recorded (more of which in the next section).

It is important to understand that, even though a large amount of data is available, it might contain very little information relevant to a particular outcome. Consequently, we must accept that the kind of data we collect places a fundamental upper limit on the accuracy of our predictions. This concept was presented in a medical context by Hanley and McNeil [2] and is discussed in relation to the QAMC project in [5, 6, 7].

As an example of this, consider the prediction of birthweight. We can build a prediction model on the basis of maternal information, such as height, weight at different stages of pregnancy, *etc.* However, a far more accurate approach is to take measurements from the baby in utero, perhaps via ultrasound. This is precisely the kind of information which was *not* routinely recorded in the databases we had access to, and our ability to make an accurate forecast would be diminished as a result.

In short, no amount of sophisticated risk modelling can make up for the absence of relevant information in that data that has been collected. This is certainly the limiting factor in predicting the risk of failure to progress (as discussed in [6]). In presenting this web page we hope that clinicians will see the potential benefits of gathering and recording more informative patient details.

4.2 Data reliability

Although quality audits (such as described in [6]) go some way towards reassuring us that database information is fairly consistent with case notes, they cannot establish whether case notes are consistent with reality. Systematic inconsistencies in case notes can have serious consequences for the accuracy of risk prediction models trained on retrospective data.

One way in which case notes could distort the true characteristics of a patient population is by presenting different levels of detail in different patient records. Let us consider how this might happen in the SMR2 data. Case details are completed and entered into the database when the mother is discharged from hospital. A mother who experienced complications at some stage is likely to have more detailed case notes than a mother whose pregnancy and delivery proceeded without problem. Even if both mothers have similar medical histories, it is now probable that the first mother will have her history recorded in more detail that the second. If this is done on a systematic basis, we will observe in the data a spurious association between certain conditions and adverse outcome — even though those conditions are equally prevalent in mothers whose pregnancy had benign outcome — simply because those conditions have usually been recorded when adverse outcomes have occurred.

We suspect this to be the reason behind the following discrepancy in the SMR2 data (reported in [6]). Grand multiparity is a term used to describe a mother who has given birth to four or more children. The ICD-9 code for grand multiparity (`6594`) is recorded in the SMR2 data. A mother's parity is also recorded as a separate item and we would expect agreement between that number and the ICD code.

However, parity 4+ and ICD-9 code `6594` are not identical across the database: there are 15 531 instances of parity 4+, but only 696 instances of code `6594`. Furthermore, code `6594` is associated with a slightly higher incidence of failure to progress (2.9%) than parity 4+ (2.1%). Although these differences are small, the fact that there are differences at all shows that misreporting has the potential to affect the conclusions that we draw from data.

Again, we must keep in mind that the SMR2 database was not compiled for the construction of risk prediction models. However, reliability of data remains an important issue if we wish to make meaningful inferences using large databases of information.

4.3 Definition of ICD-9 and other diagnostic codes

The Achilles' heel of the QAMC risk prediction web page is its use of ICD-9 codes to represent maternal health status. At project review meetings, clinicians have consistently commented that codes such as `6525` (high head at term) or `6429` (unspecified hypertension) are vague and open to a considerable degree of interpretation. This is not just a problem with the SMR2 data; other datasets contain their own variety of imprecise terms. The net effect is to reduce the utility of any systems that use these codes as predictors of outcome.

Difficulties with ICD-9 and other diagnostic codes lead to the development of an alternative risk prediction web page that used only non-ICD code information [6]. Still, the problems caused by vague diagnostic terms raise the question of why this information was recorded in the first place.

One thing is clear: if large medical databases are to be used to develop accurate and meaningful models of risk, the information they record must be defined more rigorously. It is only through attempting to develop such risk prediction models that this issue comes to light.

4.4 Interpretability of results

Leaving aside the question of whether accurate risk prediction models could be built, it is difficult to say whether there is a place for such systems in modern medical practice. Would clinicians actually find a percentage risk estimate useful in case management? This is an issue on which we hope to obtain some feedback about from visitors to the web page. As we remarked at the beginning of Section 4, the successful application of technology is an incremental process, but if there is little enthusiasm for the ultimate objective of that application, the process will founder.

We noted earlier that the QAMC risk prediction web page requires *all* patient characteristics to be specified before a database query or risk prediction can be made. Bayesian belief networks (see *e.g.*, [11, 4]) offer a promising approach to modelling risk with incomplete information, however, training these systems with large amounts of data remains a challenge. An even greater challenge is that of inferring *causality* from data and this is currently a topic of great interest within the AI community.

We shall return to the issue of interpretability in Section 4.6.

4.5 Generalisation issues

Implicit in our use of SMR2 data is an assumption that the underlying characteristics of the patient population (*i.e.*, training data) do not change over time. This assumption allows us to generalise about the risk of adverse outcome associated with new cases on the basis of cases we have seen in the past. At this stage, we have not assessed whether stationarity assumptions hold in the SMR data.

Stationarity is an important issue in the medical domain, where accepted methods of patient management are constantly evolving to improve patient outcome. One consequence of changing medical practice is that information in medical databases has an implicit "use-by date" with regard to building systems to predict certain outcomes.

Our ability to make predictions about patients outside the catchment area of the SMR2 data is another important issue in generalising to new domains. Standards of medical practice and methods of treatment vary between geographical regions, as do certain physical characteristics of the patient population. Thus, it is not clear whether the prediction systems built using the SMR2 data should be used to make forecasts for cases outside Scotland. In an ideal situation, we would have explored this topic by using data collected elsewhere to validate performance. However, as mentioned in Section B, the types of information recorded in the datasets available to QAMC were too varied to allow valid comparisons to be made.

4.6 Quality indicators *vs* meaningful warnings

The term *quality indicator* refers to a statistic which is thought to be related to the quality of care delivered in a specific medical setting. Caesarian section and perinatal mortality rates are two indicators common in obstetric practice. These statistics allow clinicians and administrators to form rough comparisons between the quality of care delivered in different districts, hospitals, even by different doctors. Naturally, such crude estimates must be combined with other relevant information. For example, a large hospital may have a high perinatal mortality rate simply because it receives more than its fair share of high risk cases.

While these rates and measures might be associated with quality of care, their summary nature often makes them of little use to forecast. (This point is emphasised in Section A in relation to some of the prediction tasks originally proposed in the QAMC project.) It is hardly any use to to make an accurate prediction about risk of perinatal death unless one can identify reasons why individual cases are at risk. Effective intervention in high risk cases demands that the cause(s) of risk be understood.

This matter could be addressed by modelling more specific, well defined outcomes, however, there is a fundamental, and more challenging issue at play. Statistical modelling of risk can have two objectives: one, to achieve an accurate forecast, and two, to achieve interpretable results [7]. At this point in time, it appears as though some trade-off between the two objectives is unavoidable, *i.e.*, the most accurate prediction system makes inscrutable forecasts, and the most interpretable system is not (usually) so accurate[4]. Anyone dealing with models of complex phenomena (such as the relationship between symptoms and disease) ought to bear this in mind.

4.7 Medico-legal issues

Few medical domains are as litigious as obstetrics. Hence, it is with a certain sense of trepidation (and a carefully worded disclaimer) that we present the QAMC web page. But suppose for a moment that this web page was actually meant for forecasting risk in real situations — a plethora of medico-legal questions is raised.

Take the very nature of the information provided: what is the legal status of an estimated probability of adverse outcome? Unless a forecaster asserts something with certainty, it seems highly unlikely that they could ever be prosecuted for providing a duff prediction. What mechanism of checks and balances exists to ensure this does not occur?

Consider the question of liability: who is ultimately responsible for the information provided? Even if that could be ascertained, if the information is made available across the Internet, the question of which jurisdiction prosecution should take place under is a contentious one. As telemedicine becomes more commonplace, more and more such questions are causing consternation in the courts, concern among clinicians, and lucre for the lawyers.

Still, it seem that these legal problems are not insurmountable. There are a number of medical applications of Bayesian nets available, most notably, a system developed by Knowledge Industries as part of the Microsoft Pregnancy and Child Care package. To

[4] At a recent NATO Advanced Study Institute, Leo Breiman referred to this dilemma as the *Heisenberg uncertainty principle of statistics*. Given some measure of interpretability I, and accuracy A, he suggested that $IA > b$, for some positive b.

the best of our knowledge, this system was trained using expert estimation of conditional probabilities alone, as opposed to a large database of case records.

5 Conclusions

The astute reader will notice that the list of problems encountered by the QAMC project is substantially longer than the advantages offered by the systems it has developed. This is hardly surprising, given the exploratory nature of the project and the honesty that non-commercial research promotes.

Successful development of diagnostic systems trained on large medical databases demands the collection of relevant, reliable, data — free from systematic defects. This is a tall order, as many of those who handle medical data will attest. However, it is important to communicate that the limiting factor in creating accurate prediction systems out of large databases is (in this instance) the data itself, rather than the modelling method.

One *technological* improvement to a system like the QAMC web page would be to implement a more interpretable approach. Bayesian networks seem to have appeal in this regard, but it is not clear (a) whether meaningful conditional independence relations could be found among the variables available, (b) whether training such systems is feasible with over 770 000 cases, and (c) whether any greater diagnostic accuracy could be achieved.

Other issues, like stationarity and generalisation to other patient populations will always be a challenge. As more and more people gain access to the Internet, many of the legal questions we discussed will have to be addressed.

It is our hope that this paper and the QAMC web page will promote thought and discussion about the relevant issues. The reader may be interested to write to us (via the feedback section of the web page[5]), or to read some of the comments[6] that we have received.

[5]see `http://svr-www.eng.cam.ac.uk/projects/qamc/feedback/feedback.html`
[6]see `http://svr-www.eng.cam.ac.uk/projects/qamc/feedback/comments.html`

A The proposed list of APOs

At the beginning of the QAMC project, it was proposed that the following 20 adverse pregnancy outcomes should be modelled:

1. Antibiotics started > 24 hours after term vaginal delivery
2. Inborn term infant with seizures before discharge
3. Maternal hospital stay > 5 days (> 7 days if CS)
4. Maternal re-admission within 14 days of discharge
5. Non-standard reason for induction
6. Delivery unattended by any midwife or doctor
7. Primary CS for fetal distress
8. Primary CS for failure to progress
9. Birthweight < 2.5kg or RDS following induction
10. Postpartum return to delivery suite/theatre
11. Excess blood loss (> 450ml): transfusion or Hb drop > 3.5g/dl
12. Non-standard reason for induction followed by CS
13. Unintended injury to a maternal organ
14. Birthweight < 2.5kg or RDS following a planned repeat CS
15. Neonatal mortality up to 28 days of age
16. Fetal death prior to labour
17. Fetal death during labour
18. Eclampsia
19. Maternal cardio-pulmonary arrest
20. Maternal death.

Each of these outcomes is an important measure of the quality of care provided to mothers and babies. But, while it is certainly desirable to reduce the incidence of any of these APOs, it does not necessarily make sense to do this by predicting the risk of a mother or baby experiencing such an outcome. For example, it would be of little use to forecast a high risk of maternal re-admission within 14 days of discharge *without* identifying the root cause of that outcome. This, and other factors, allowed the list of 20 APOs to be reduced to a smaller number of outcomes that are both meaningful and feasible to predict.

Certain items in this list are rendered infeasible to model accurately by their low incidence (*e.g.*, maternal death), their absence in any of the perinatal databases we had access to (*e.g.*, unintended injury to a maternal organ) or their inability to be precisely defined in terms of database information that we did have access to (*e.g.*, postpartum return to delivery suite/theatre).

Further difficulties were caused by some APOs that actually specified a *combination* of adverse outcomes. Primary Caesarian section for failure to progress is a clear example of this. It is well known that Caesarian section rates vary significantly between different countries, different hospitals and different obstetricians. Any system that tried to predict risk of Caesarian section would have to model the decision process of different obstetricians — a challenging prospect, to say the least. It was decided that these compound APOs should be broken into their component outcomes.

So, after a careful review, the list of 20 APOs was reduced to 5 APOs that seemed feasible to tackle:

1. Eclampsia/pre-eclampsia condition
2. Failure to progress in labour
3. Fetal distress
4. Low birthweight
5. Respiratory distress syndrome (RDS).

Of these outcomes, failure to progress in labour had the highest levels of *expected attainable discrimination* [5, 6, 7] (on the basis of the SMR2 database) so it was decided to explore the modelling of that outcome first.

A.1 Definition of failure to progress in labour

Put simply, a failure to progress means that the natural course of labour has stalled before delivery, placing the wellbeing of both mother and child at risk. In terms of the SMR2 data, failure to progress was indicated by the presence of any of the following ICD-9 codes in a case record:

- `6622` prolonged second stage;
- `6620` prolonged first stage;

- `6603` persistent occipitoposterior;
- `6621` prolonged labor not otherwise specified;
- `6591` fail induction labor not otherwise specified.

B Choice of database

The QAMC project is fortunate to have access to a variety of large perinatal databases from all over Europe. Databases from Scotland, Baden-Würtenberg, Wales and Finland provide over 1.2 million case records. However, while it is universally agreed that collecting perinatal data is a good thing, there is considerably less agreement about what information to collect, and no agreement whatsoever about how that data should be stored. Consequently, these databases could not be usefully amalgamated and we had to decide on a particular dataset to work with.

The experiments described in [5, 6, 7] and the QAMC web page make use of the largest dataset available to us: the Scottish Morbidity Record (known as the SMR2) which contains 771 571 singleton births that occurred between 1980–91. Cases in which elective Caesarian section or breech presentation occurred were removed from the dataset[7]. For the training and testing of various risk prediction systems, these records were partitioned into a *learning set* of births occurring between 1980–88 (540 905), and a *testing set* of births occurring between 1989-91 (176 812). While we only have access to *retrospective* data, this partition was used to simulate a *prospective* trial of the prediction systems.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[2] J. Hanley and B. McNeil. Maximum attainable discrimination and the utilization of radiologic examinations. *Journal of Chronic Disease*, 35:601–611, 1982.

[3] D. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, New York; Chichester, 1989.

[4] S. L. Lauritzen, B. Thiesson, and D. J. Spiegelhalter. Diagnostic systems created by model selection methods: a case study. In P. Cheeseman and R. Oldford, editors, *Selecting models from data: AI and statistics IV*, Lecture Notes in Statistics 89, pages 143–152. Springer, New York, 1994.

[7] In almost all cases, these factors will be known before labour commences and indicate that delivery will be by Caesarian section so that failure to progress in labour cannot occur.

[5] D. R. Lovell, C. R. Dance, M. Niranjan, R. W. Prager, and K. J. Dalton. Limits on the discrimination possible with discrete valued data, with application to medical risk prediction. Technical Report CUED/F-INFENG/TR243, Cambridge University Engineering Department, January 1996. Available from `http://svr-www.eng.cam.ac.uk/projects/qamc/Publications/Publications.html`.

[6] D. R. Lovell, B. Rosario, M. Niranjan, R. W. Prager, K. J. Dalton, R. Derom, and J. Chalmers. Design, construction and evaluation of systems to predict risk in obstetrics. *International Journal of Medical Informatics*, 1997. To appear. Available from `http://svr-www.eng.cam.ac.uk/projects/qamc/Publications/Publications.html`.

[7] D. R. Lovell, M. J. J. Scott, M. Niranjan, R. W. Prager, K. J. Dalton, and R. Derom. On the use of expected attainable discrimination for feature selection in large scale medical risk prediction problems. Technical Report CUED/F-INFENG/TR299, Cambridge University Engineering Department, August 1997. Available from `http://svr-www.eng.cam.ac.uk/projects/qamc/Publications/Publications.html`.

[8] D. J. C. MacKay and L. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1995.

[9] M. Maresh, editor. *Audit in Obstetrics and Gynaecology*. Blackwell Scientific Publications, 1994.

[10] C. E. Rasmussen. *Evaluation of Gaussian processes and other methods for non-linear regression*. PhD thesis, University of Toronto, 1996.

[11] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

[12] World Health Organization. *International Classification of Diseases, Ninth Revision*. Geneva: World Health Organization, 1975.