Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

DATA DETECTIVES: UNCOVERING SYSTEMATIC ERRORS IN ADMINISTRATIVE DATABASES

Sten Ardal[1] and Sherri Ennis[1]

ABSTRACT

Secondary users of health information often assume that administrative data provides a relatively sound basis when making important planning and policy decisions. If errors are evenly and/or randomly distributed this may have little impact. This assumption is betrayed when information sources contain systematic errors, or when systematic errors are introduced in the creation of master files. The most common systematic errors involve underreporting of activity for a specific population, inaccurate re-coding of spatial information, or differences in data entry protocols. The Central East Health Information Partnership (CEHIP) provides information support for development of public health programs and health system planning through a partnership in Ontario's most populous health planning region. CEHIP has identified a number of systematic errors in administrative databases and has documented many of these in reports distributed to partner organizations. Failures to register births and incorrect assignment of geographic codes in vital statistics files have been studied. Misclassification of cause of death has also been explored, particularly with respect to delays in determining cause of death and the effect this has on official data sets. Differences in data entry protocols for reportable disease data have been researched, raising questions about the consistency of data submitted by different tracking agencies. This paper will describe how some of these errors were identified, and note processes that give rise to such losses in data integrity. The conclusion will address some of the impacts these problems have for health planners, program managers and policy makers.

KEY WORDS: Data Quality; Health Information; Health Planning

1. INTRODUCTION

> *"Winwood Reade is good upon the subject. He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."* (Sherlock Holmes, *The Sign of the Four*)

Population health planning and policy development relies on the "certainty" provided through analysis of large groups of individuals. But the information basis on which we describe the characteristics and measure the actions of populations does not always yield a "true" mean. The observed mean will be influenced by measurement error and will be misplaced in the presence of systematic bias.

In statistics "error" is usually assumed to be randomly distributed, yielding the normal curve used for probability tests. The error is attributed to the precision of the measurement instrument used in investigation (Fisher,1949). While this notion of error can be applied to the interpretation of surveys, it is quite different from the type of error encountered when analyzing administrative data. In describing population health, most data originate in a service encounter. Error generated in the assembly of

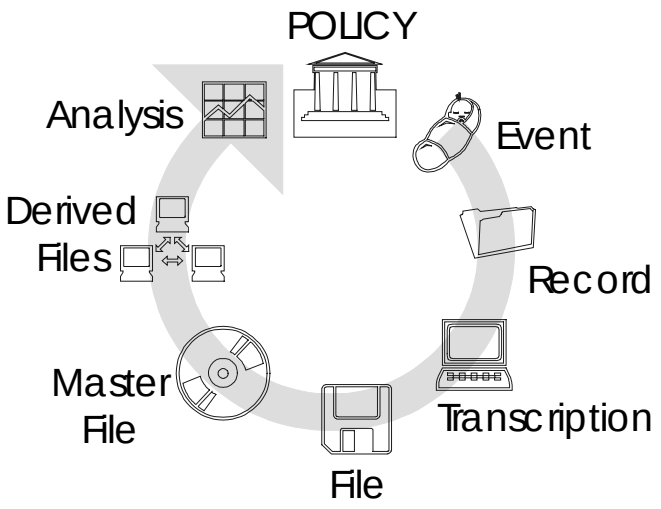
[1] Central East Health Information Partnership
BOX 234 Newmarket Ontario, L3Y 4X1
Phone: 905.764.6346 xt.1211, Fax: 905.895.0848
e-mail: info@cehip.org

administrative data sets reflect the data collection and handling procedures, and are most likely systematic (Wolff & Helminiak, 1996). This paper will demonstrate how the Central East Health Information Partnership (CEHIP) has detected some specific errors in commonly used data sets, and will indicate the procedures that biased key fields. In so doing, descriptions of data process flow will be presented. Data crime scenes will be described, investigations detailed, and culprits identified.

> *"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts." (*Sherlock Holmes, *A Scandal in Bohemia)*

It is compelling to willingly accept the results of analysis, particularly if they are in accordance with expectation. We accept the data as "accurate" when they fit our theoretical framework, but blame the quality of the information if it appears aberrant. The data, however, are always a reflection of the process and procedures that bring it to the analyst's desktop. Before becoming engaged in interpretation, one should be satisfied that there are no important facts being overlooked. Particular attention should be given to those features of collection and handling that are motivated to offence.

So what are the facts that we must consider? A general schematic of the route traversed by administrative data is provided in Figure 1.



Event: May not be reported.

Record: Poor documentation.

Transcription: Inconsistent entry protocols employed.

File: Structure inadequate to properly describe events.

Master File: Exclusion criteria.

Derived Files: Recoding errors.

Analysis: Tabulations can compound errors.

Policy: Can introduce systematic reporting bias.

Figure 1: *Simplified information flow chart with examples of factors that contribute differentially to errors at identified information processing stages.*

There may be a myriad of complex processes in this route that can be influenced by events within and between the stages shown in Figure 1. Between stages it is most likely that there will be a failure of capture or transmission. But within each stage there are many potential processes at play. An Ontario group, the Clinical Data Quality Task Force, has been trying to understand how some of these processes work within institutions. Figure 2 expands on some of the people, structures, and processes involved in creating a typical institutional record. This model is being used to explore how changes in organizational structures impact on data quality. For example, trends towards de-centralization of patient registration procedures may require different training, monitoring, and interfaces to maintain quality.

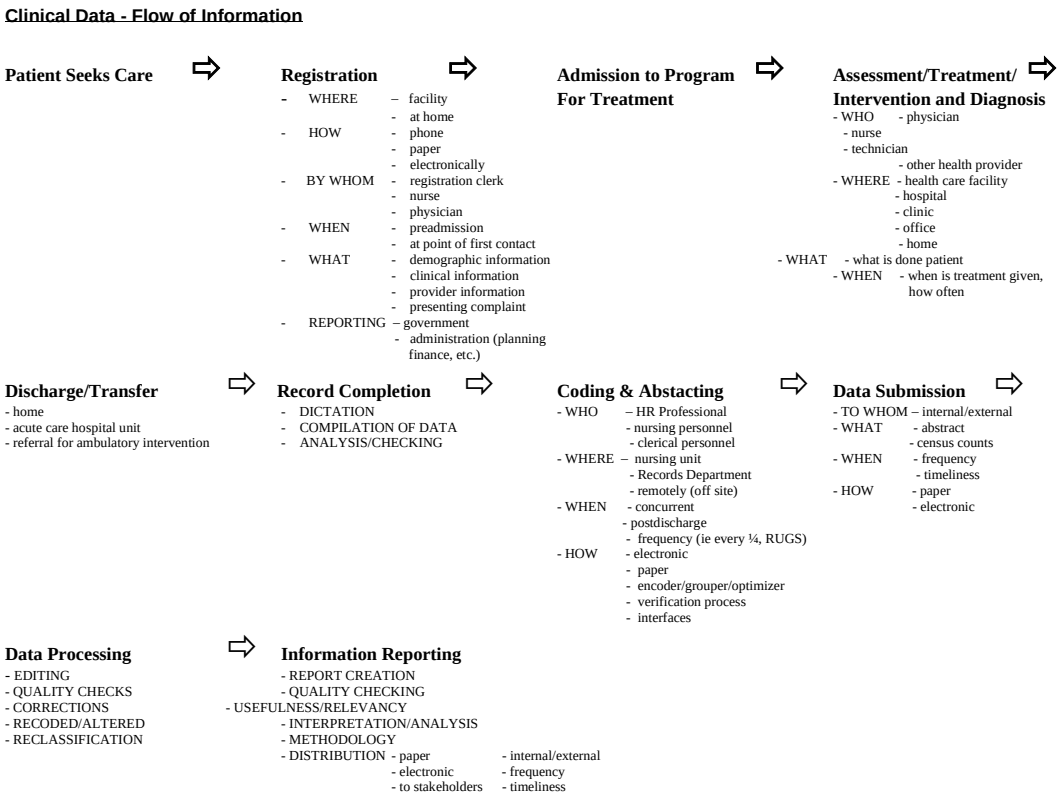


Figure 2: *Clinical data flow diagram for a typical large institution.*

The route traversed by administrative information is clearly complex. At CEHIP we have noted problems with a variety of administrative databases. There are many different kinds of error that can be introduced in a complex system. It is likely that only the most egregious are actually noted, and that identification of systematic errors that can be alleviated though process improvement will assist in appropriate interpretation, but never result in complete accuracy. Three "case studies" are presented to provide examples of investigations conducted at CEHIP.

2. EXAMPLES OF DATA QUALITY STUDIES

2.1 The Case of the Missing Babies

Scene: CEHIP had noted that births recorded in hospital statistics differed from numbers reported by the Ontario Registrar General (ORG). Hospital counts were generally higher, though logically they should have been marginally lower as few births would occur outside Ontario hospitals. In discussion it was found that bothe the Provincial Health Ministry and Statistics Canada had their own concerns about the completeness of Birth Registry Information. It was known that a policy had been implemented some years ago that required both a parental and a doctor's notification of a birth before the master file was updated, but there was no record of the number of notifications that had not resulted in a vital statistics file entry.

Investigation: It was decided that the best way to start an investigation would be to include entries that did not meet the "dual notification" criteria. This required manual entry of thousands of records. In almost all cases the missing notification was the parental form. A new master file was created for analysis that included all birth reports received by the Registrar General. Subsequent analysis revealed that young parents with low birth weigh babies were most likely not to report the birth, and furthermore there was considerable geographic variation. It should be noted that the analysis excluded events that were unlikely to have yielded viable births. Failures to report were found to increase around 1996.

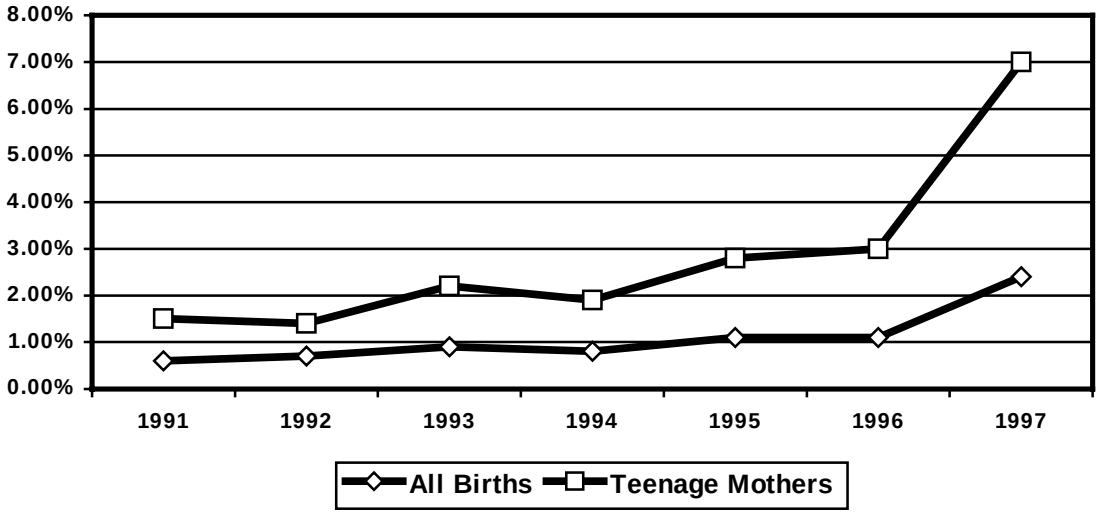


Figure 3: *Percentage of unreported births identified in Ontario 1991-1997 for all births and births to teenage mothers.*

So what happened in 1996? In that year Provincial legislation was enacted to allow cost recovery for parental birth registrations, a service they had previously provided free on behalf of the Provincial Registrar. Every municipality in Ontario was surveyed to find out if they were charging fees, how much they charged, and when the fees were enacted. When this information was included in the analysis it became apparent that the greatest increase in "missing births" occurred in areas that introduced fees. These areas accounted for about 70% of all Ontario births.

Findings: This data quality review shows the introduction of a systematic bias that resulted in under representation of a specific sub-population. As this is a target group for specific health promotion interventions, the undercount has implications for service configurations, key trend analyses and public policy. Uncorrected data would lead to underestimation of low birth weight, teenage pregnancy, and population growth. These results have been forwarded to the Ontario Registrar General, and are expected to impact on registration and data compilation policies.

2.2 The Fluctuating Fertility Mystery

Scene: Fertility rates are an important indicator used to plan a broad range of human services. Changing trends have had a significant impact on local decisions, such as school development, and national policies, such as those governing immigration levels. A routine analysis using an Ontario Birth Registry file supplied through Statistics Canada revealed a sudden loss of hundreds of births in some areas, and gains in others. Fertility rate calculations demonstrated that this was not due to changes in the demographic distribution of potential mothers.

Investigation: It was relatively easy to determine that a data crime had been committed. The Markham area increased its births by about 600, or 40%. Meanwhile Vaughan's births were halved, with a drop of

about 850. Furthermore it was noted that birth counts in the areas studied were not consistent with those noted in hospital records. Since virtually all births occur in hospitals concordance was expected. The magnitude of the problem became clear when fertility rates were calculated. This produced comparable indicators that account for the size of the potential child- bearing population. The rates showed divergent trends for the areas studied when geography was determined by the Census Sub-Division (CSD) codes assigned to the files by Statistics Canada. This assignment was based on municipality information contained in the files submitted by the Ontario Registrar General.

The files received also contained postal code information. When CSDs were recoded based on postal codes the trends became more stable and consistent with the hospital record analysis. Meanwhile the Census Subdivision recoding and the Municipality information in the file were in agreement.

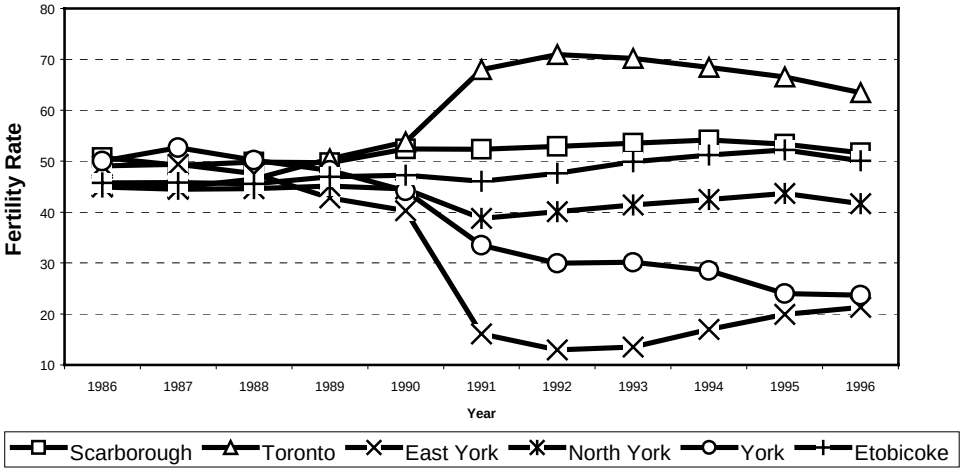


Figure 4: *Fertility rates in select Ontario municipalities 1986-1995.*

Findings: This error would lead to incorrect assumptions about population growth and the need for family and children's services. Having identified the error, and notified those involved in creation of the file, the discrepancy was noted for the benefit of those relying on this information for planning and policy purposes. The reasons for the problem arising in the first place were not fully explored, though it is speculated that residents were far less accurate in reporting their municipality than their postal code. In the Toronto area a contiguous Metropolitan area was subdivided into a number of "municipalities" that have since been amalgamated. Clearly, there were inadequate verification processes at all stages in the data process chain. Municipal information could have been referenced to either street addresses or postal codes during the registration process. The internal consistency of geographic information in the files could also have been determined prior to CSD assignments.

2.3 The Calendar Case

Scene: Prior to 1998 Metropolitan Toronto was divided into six municipalities, each with its own Public Health Department. Amalgamation in 1999 raised questions about the comparability of the former municipality's Reportable Disease Information Systems (RDIS). These systems are supposed to provide comparable and consistent information that is compiled in a Provincial database. Important trends and monitoring of outbreaks relies on accurate identification and tracking of the 60 diseases that health units are mandated to report by the Ontario provincial government. It was suspected that there may be some differences in the way the former municipalities coded information, and so an investigation was conducted to identify any quality issues arising from combining the six RDIS databases in Toronto.

Investigation: Interviews were held with key informants to understand the process used to record information. It was found that the information was often entered by clerks who had little formal training in appropriate entry protocols, and that there were often many different people involved in data entry.

Each field, whether mandatory or optional, was examined to determine if information reported was incomplete, inconsistent, or appeared inaccurate. The investigation focussed on four of the most commonly reported diseases. Errors ranged from virtually 0 for some mandatory fields to almost 100% for some of the optional data entries.

Date fields provide a good example of inconsistency as it is relatively easy to identify dates that would be logically impossible. This was noted in our analysis of Influenza, where 6.4% of cases were reported with onsets occurring after receipt of a medical diagnosis, and 7.6% were shown as reported to the health unit before a diagnosis had been made. The system is designed to record a suspected episode, the subsequent date of diagnosis, and then the subsequent date of a diagnostic report (see table 1). In epidemiological analysis of disease outbreak the accuracy of timing information can be critical. The actual error rate is likely higher than the observed rate of illogical error entry.

Logical Sequence	*Suspected Episode Date*	*Diagnosis Date*	*Diagnostic Report*
Percent "Suspect"	7.5%	14.4%	8.8%

Table 1: *Percent of date fields for influenza cases reported in Toronto Reportable Disease Information System that are either preceded or followed by a logically impossible dependent date entry.*

Findings: These errors would result in inaccurate date trending and delay the use of this data to rapidly identify epidemics. Entry procedures used widely differing definitions for date fields. For example, "episode date" was interpreted variously as "onset of symptoms (correct), "date of clinic visit", "date of specimen collection", "date of data entry", or "date of report completion". Similarly the "diagnostic date", was found to refer to the date a report was received by the health unit, the date of diagnosis (correct), or the date reported by a client.

These data sets suffered from poor compliance with entry protocols, most likely resulting from inadequate training and an absence of quality control procedures. This type of error has been well documented (e.g., Smulian, et al, 2001), and training interventions have demonstrated positive outcomes (Lorenzoni, et al, 1999).

3. CONCLUSIONS

Healthcare is a significant component of government, business and personal expenditure. Resource allocations continue to expand, and in most jurisdictions this sector is the largest component of government expenditures. Informed decision making and policy development is increasingly critical to address needed reforms. It is generally agreed that healthcare systems in countries like Canada must be more effectively and efficiently organized to cope with increasing needs while containing costs.

There are only a few information systems in place that employ "data professionals" in the collection, transcription, management and analysis of health related information. Most systems were never designed to support the kinds of analysis planners and policy makers are interested in. There has also been little attention given to process improvements, and quality reviews tend to focus only on ad hoc data entry audits. Increasing pressure for legal and financial accountability is helping to shed focus on the importance

of accurate and consistent information handling procedures in health care organizations (Lichtenstein, et al, 1999).

A good analyst must acquire the observational skill of a detective. How did this data get here? What could have happened along the way? What motives could have contributed to systematic errors? Before one can have confidence in the outcome of analysis it is imperative that there be some understanding of the likelihood that our information captured all of the events intended, that the records are properly recorded, that data entry was governed by consistent protocols, that files are well constructed and complete, and that derivations are sound and properly documented. There is little doubt that there will be errors, and various degrees of systematic bias. The data must be analysed, and interpreted in a way that incorporates an understanding of error characteristics. And it is through this recognition that the data detective can point out the culprits, and promote the changes that will enhance data quality.

*"All knowledge comes useful to the detective." (*Sherlock Holmes, *The Valley of Fear)*

REFERENCES

Baltimore County Public Library (2001) "Quotes From Sherlock Holmes", http://www.bcpl.net/~lmoskowi/HolmesQuotes.

Bienefeld, M., Woodward G.L. and S. Ardal (2000), "Underreporting of Live Births in Ontario", unpublished report, Newmarket, Canada: Central East Health Information Partnership.

Dawson, K. (2000), "Data Quality in RDIS: Issues Related to Combining Data Sets", unpublished report, Newmarket, Canada: Central East Health Information Partnership.

Fisher, R.A. (1949), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Kenney, N. (1999), "Identifying Problems with Data Collection at a Local Level: Survey of NHS Maternity Units in England", British Medical Journal, 319, pp. 619-622.

Lichtenstein, D., Materson, B., and D. Spicer (1999), "Reducing the Risk of Malpractice Claims", Hospital Practice, http://www.hosppract.com/issues/1999/07/licht.htm?3b0bce690.

Lorenzoni, L. Da Cas, R. and U.L. Aparo (1999), "The Quality of Abstracting Medical Infomraiton From the Medical Record: The Importance of Training Programmes", International Journal of Quality in Helath Care, 11, pp. 209-213.

Smulian, J. C., et al, (2001) "New Jersey's Electronic Birth Certificate Program: Variations in Data Sources", American Journal of Public Health, 91, pp. 814-816.

Wolff, N. and T.W. Helminiak (1996), "Nonsampling Measurement Error in Administrative Data: Implications for Economic Evaluations", Health Economics, 5, pp. 501-512.

Woodward G.L. and S. Ardal (2000), "Data Quality report: Effect of Residence Code Errors on Fertility Rates", unpublished report, Newmarket, Canada: Central East Health Information Partnership.