Dairy Cattle Disease Data from Secondary Databases
Use with Caution!

D.F. Kelton, B.N. Bonnett and K.D. Lissemore
Department of Population Medicine
University of Guelph, Guelph, Ontario, Canada, N1G 2W1

1. Introduction

The demand for dairy cattle disease incidence and prevalence data is increasing. Interests include well known infectious and metabolic diseases, as well as less obvious conditions resulting in sub-optimal production or performance. Data are needed for animal health risk assessments to facilitate world trade, benefit cost analyses of health maintenance programs and genetic evaluations for the enhancement of disease resistance. For each of these purposes, high quality, representative data from large populations of dairy cattle are needed. The more complex the relationships that are being investigated, the larger the required database. The challenges of identifying and aggregating data from the required hundreds of thousands of dairy cows in thousands of herds are considerable.

Whatever the purpose, there are some general requirements for data that need to be satisfied. The first of these is data *quality*, which encompasses accuracy, consistency, and completeness. *Accuracy* refers to how closely the data reflect the true state of nature. While some inaccuracies, such as the etiologic misclassification of mastitis caused by minor pathogens, may be tolerable, a mechanism to assess the overall frequency and magnitude of inaccuracies in the data is imperative. *Consistency* in defining and recording disease events is also very important. Important differences in classification of disease events are common in aggregate databases, particularly when information is provided or recorded by many people, often with varied backgrounds, experiences and training. It is important to realize that consistency and accuracy are not synonymous, since individuals can consistently make the incorrect (inaccurate) disease diagnosis. High quality disease data must be *complete*. The challenge is to determine whether the absence of recorded disease event(s) during a lactation indicates that the cow remained healthy, or that disease was present but not recorded. Disease events that cannot be tied to a specific animal or herd are not acceptable for use in genetic evaluations or health status determinations. Data accuracy may be affected by its perceived usefulness by, and the motivation of, the person(s) responsible for collecting and recording it.

The most complete, accurate and consistent data is worthless if it is not accessible in a form that allows data to be aggregated and analysed. The physical structure of the database in which the disease data resides must be considered in assessing overall data quality.

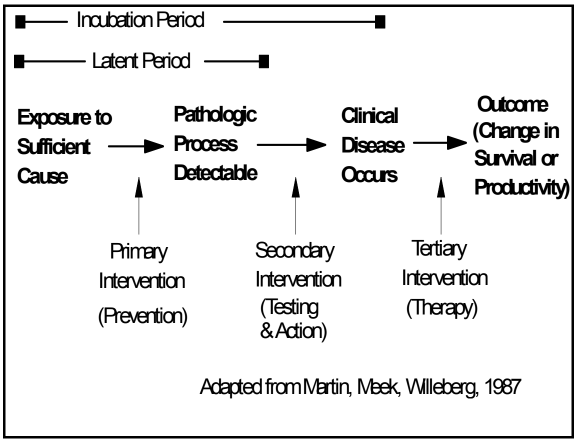
The second requirement for disease data is *representativeness*. In order to extrapolate information beyond the individuals from whom the data originated, the *study group* or *population* must be representative of the larger *reference* or *target population*. This implies that either the entire population needs to be included in the analysis, or that appropriate sampling strategies are employed to select the study animals and herds.

Large quantities of disease data can be obtained in one of two ways. The first is to make disease data collection a primary activity, and establish a mechanism by which to obtain accurate data from either the entire population of interest, or a representative sample of that population. This has been done through extensive health surveys, such as those conducted by the National Animal Health Monitoring System in the United States (Hueston, 1990), or through intensive longitudinal studies such as the Ontario Dairy Monitoring and Analysis Program (Kelton, 1995). Unfortunately, due to the considerable cost of such endeavours very few primary disease databases exist.

The second alternative for obtaining dairy cattle disease data is to access secondary disease data. These are collected as a by-product of dairy industry programs collecting other primary data, such as milk production or farm financial information. While the disease data contained in many of these systems are relatively inexpensive to obtain (since the primary data gatherer incurs the cost), they are not necessarily extensive, and the quality and structure of the data are variable (Willeberg, 1986). It is important that the features and limitations of these secondary data are clearly understood and that caution is exercised in their use.

2. Dairy cattle health and disease

The complex nature of health and disease must be borne in mind whenever existing data are used. The presence or absence of disease, and by extension health, is often considered binary in nature. However, disease is more appropriately described as a continuum, as depicted in Figure 1 (Martin et al., 1987). Theoretically, the aim of preventive medicine programs is to intervene as early as possible in the disease development process, subject to economic constraints. In terms of understanding health and disease from secondary data sources, the format and quality of information recorded on risk factors, diagnostic tests, clinical occurrences, interventions, and outcomes affects the usefulness of the data.


Incubation Period
Latent Period
Exposure to Sufficient Cause
Pathologic Process Detectable
Clinical Disease Occurs
Outcome (Change in Survival or Productivity)
Primary Intervention (Prevention)
Secondary Intervention (Testing & Action)
Tertiary Intervention (Therapy)

Adapted from Martin, Meek, Willeberg, 1987

Figure 1. The spectrum of disease development.

Crucial to the computation of disease prevalence (the presence of the condition at a point in time) and incidence (the rate of occurrence of the condition over time) is establishing a clear and comprehensive case definition. Some diseases are relatively easy to define because they have a short subclinical phase, present one or more pathognomonic indicators (signs, substances, responses or tissue changes that are absolute predictors of the presence of the disease or disease agent) that are readily identified with accurate and available tests, have a clear start and finish and seldom occur more than once in the animal's lactation or lifetime. Left displaced abomasum (LDA) is an example of such a relatively simple disease.

Other diseases are more difficult to define. These complex diseases tend to have prolonged subclinical phases, eventually manifest vague or no clinical signs, have no standardized tests that accurately and consistently identify their presence, may involve more than one etiological agent, may affect multiple organ systems, and recur sporadically throughout the lactation and in subsequent lactations. Bovine mastitis caused by *Staphylococcus aureus* is a prime example of such a complex disease.

Both preceding examples represent disease conditions that are economically important enough to warrant routine identification and recording. A third category of diseases, endemic diseases of negligible and/or artificial economic importance, may be simple or complex in nature but are seldom identified or recorded. An interesting example may be Enzootic Bovine Leucosis, a viral disease considered endemic and of little economic significance in North America, but with major political and trade implications in Europe.

Since diseases of dairy cattle vary from the simple to the complex, the identification and classification of these diseases also varies. A number of classification systems based on etiology, severity, epidemiology, duration and target system(s) have evolved. Ultimately, it is important to understand and refine the level of classification relative to the intended use of the disease data.

Disease coding and standardization of nomenclature is an important area of discussion both in human and veterinary medicine (Case, 1994). Less attention has been directed towards the standardization of disease definitions and recording protocols. The International Dairy Federation (IDF) has established a set of international guidelines for bovine mastitis (Osteras et al., 1996), the American Association of Bovine Practitioners has made recommendations for reproductive performance (Fetrow et al., 1994) and standard definitions for eight clinically and economically significant diseases of dairy cattle are currently under discussion in Canada (Kelton et al., 1997). While some classification guidelines are being developed, there is still a general lack of utilized standard disease definitions and recording guidelines.

3. Disease data recording systems

Globally, there are very few broad-based and comprehensive primary disease recording systems in the dairy industry (Davies, 1985; Ekesbo, 1994). In most cases, disease data are recorded as a secondary component of farm-based management systems, veterinary bureau monitoring and billing systems or regional/national milk recording, regulatory or diagnostic laboratory systems. While each of these can provide useful disease incidence and/or prevalence data, the attributes and deficiencies of each system must be understood to correctly utilize and interpret the resulting information.

To estimate the potential usefulness of recorded disease data, it is critical to understand why and how the data were collected and stored. In some cases, the disease data are used by the dairy farmer and his/her veterinarian in the management of the herd. In such cases the data are deemed to have biological and/or economic importance, and are likely to be of reasonably high quality. Disease data may be collected as a by-product of aggregate billing and drug inventory systems. In these cases, there is more emphasis on billable procedures and drug dispensing than on diagnostic accuracy. Public and private diagnostic laboratories store disease data based on their sample submissions. While they expend considerable energy in arriving at a correct medical diagnosis, their data often represent only a small, potentially biased segment of the population. Regulatory agencies collect disease data pertaining to reportable diseases of national significance. However, many of these diseases occur sporadically, if at all, and these data are of little practical or economic value at the local level.

Whatever the source of the disease data, there are several types of bias (systematic errors) which need to be considered. The first of these is *selection bias*, which may occur in routinely collected data due to selection in admission or inclusion of farms or animals. Selection bias may arise because of differences among farms based on whether or not they are recording on-farm disease information, participating in regular veterinary health care programs and associated local bureau systems, and enrolled in national or regional programs such as milk recording or herd book programs. A second bias commonly encountered in disease recording is *information bias*. An example of information bias is the use of diagnostic test information, such as somatic cell count, instead of clinical case information to establish the occurrence of mastitis. Since there are few, if any, perfect tests, the diagnostic sensitivity (the proportion of diseased individuals correctly identified as diseased) and specificity (the proportion of non-diseased individuals correctly identified as non-diseased) must be considered in the identification of diseased animals (Martin et al., 1987). Depending on the impact of false positive and/or false negative test results, the incidence or prevalence of disease in the animals of interest can be over-estimated or under-estimated.

4. Advantages and limitations of farm-based disease data

Most dairy farms keep records for management purposes. These vary from simple paper systems to complex computer-based programs. The scope and flexibility of computerized herd management programs (HMP) continue to evolve and expand (Parke, 1993). Disease data recorded on farm by dairy personnel are generally believed to have biological and/or economic significance on that farm. The primary recording purpose may be to monitor the effectiveness of preventive programs or to track antibiotic treatments so that appropriate milk and meat withdrawal times can be observed. These data are likely to reflect the producer's and/or the herd veterinarian's perception of disease occurrence on the farm. A recent study of farmer-observed mastitis in dairy cattle suggests that variation in the ability of farmers to classify clinical cases does not adversely affect the validity of their estimates of disease incidence when compared to milk culture (Lam et al., 1993). However, it has been shown that farmer and veterinary estimates of morbidity in the same group of animals may vary substantially (Van Donkersgoed, 1993). For this reason, it may be important to know who is making the primary disease diagnosis. Since farmer-recorded data are used on a daily basis for farm management, there is a reasonable probability that errors in attributing disease events to specific animals will be identified and corrected.

Although the case definition for some diseases may be consistent within the farm, between farm variability often exists, and may be considerable. This makes pooling of the data from multiple farms into a central aggregate database difficult. In some cases, veterinarians are able to impose some uniformity of definition, for some diseases, across their client base. Nonetheless, there is still likely to be some variability in case definition among neighbouring practices.

Disease recording at the farm level is often event driven, with a date and animal identifier tied to the event. The event itself may be the first subclinical or clinical manifestation of a given disease, a subsequent observation of the same disease episode, a recurrence of the disease following a "cure", or one of many actions (treatments) taken in response to the disease. With many diseases, such as mastitis, an animal may have multiple disease events separated by hours, days, weeks, months or years. Depending on the level of information stored with that event, it may be more or less difficult, and sometimes impossible, to determine if the disease event is a new (incident) case or a recurring or continuing (prevalent) case. To further complicate the issue, mastitis may occur in one or more mammary quarters, and can be caused by several different organisms, some of which are very difficult to identify consistently by routine diagnostic methods. While some HMP's facilitate the recording of mastitis on the basis of severity, etiology and duration, others have only one mastitis event category. For these reasons, the amalgamation and use of data from different farm-based HMP's can be extremely challenging.

5. Advantages and limitations of veterinary hospital disease data

Keeping official veterinary medical records is mandatory in most jurisdictions. While some dairy practices run elaborate bureau systems for their dairy clients (Menzies, 1988; Lissemore, 1989), most have systems that are used primarily for billing and drug inventory purposes. Health data recorded in these systems are attributable to, or representative of, at best, that practice's client base, and may not be tied to a uniquely identified animal, are often difficult to extract, are seldom used or reviewed, and may include only those disease events that resulted in a veterinary consultation and/or a drug dispensation. Mild clinical diseases that may be noted by the farmer, but are deemed not to require veterinary intervention, would not likely be recorded in these systems. While some jurisdictions require that all animal treatments be administered by a licenced veterinarian, many do not. Therefore, basing estimates of disease incidence or prevalence on veterinary hospital data are most likely to underestimate the true state. In fact, very few assessments of quality of veterinary hospital data have been published (Mulder, 1994; Pollari, 1996a; Pollari, 1996b).

As with on-farm systems, bureau and hospital record programs tend to utilize common disease classification schemes, but are less likely to have established specific case definitions. Furthermore, as the number of persons providing, entering and compiling data increases, the level of uniformity may diminish.

Another limitation of some hospital-based recording schemes is the lack of information pertaining to the population at risk. In order to summarize disease information at the herd or population level, one needs not only a standard case definition, a count of incident or prevalent disease events (the numerator), but also a count of animals at risk of the disease. The group at risk may include all animals in the population, or specific sub-groups based on age, parity, breed or previous/current health status. For example, while all lactating dairy cows might reasonably be considered at risk for developing ketosis (acetonemia), only second and greater parity animals are at risk for parturient paresis (milk fever). A more difficult scenario involves the determination of animals at risk for mastitis. If a cow has had a case of mastitis diagnosed in the last 7 days, is she still at risk for mastitis? The answer could be yes, if one is interested in the development of infection in one of the three previously unaffected quarters, or no if one is only concerned with the first case of infection in any given lactation.

6. Advantages and limitations of centralized disease data

The dairy industry has at its disposal many large databases. These databases are often the product of the required aggregation of specific information. The focus may be milk and component production information required for genetic evaluations, animal treatment information required for billing health costs to the government or insurance companies, diagnostic information from laboratory submissions or product quality information for regulatory or payment purposes. Access to large quantities of disease data through these systems is relatively inexpensive, avoids the need to physically aggregate smaller databases and may represent most or all of the animals in the reference population.

Given the relatively narrow breadth of information contained in many of these large systems, it may be necessary to combine particular pieces of information from several sources. In some cases, all of these data have been aggregated into a single database, either physically or through the use of electronic interfaces. Unfortunately, in most countries these data reside in separate systems that are not readily linked. While unique animal and/or premise identifiers may exist in each system, the lack of a common identification system across all systems seriously impairs the use of these data.

Many of the same limitations discussed under veterinary hospital disease data apply here as well. In fact, while some hospital disease data may be scrutinized by clinicians performing outcome assessments, it is unlikely that disease data contained in most national systems will be examined carefully, let alone validated back to the farm. It has been suggested that in human medicine, errors are more likely to be identified in practice databases than in insurance claims based systems, primarily because the data are reviewed occasionally by the physicians (Tierney and McDonald, 1991).

Finally, secondary disease data from large central databases are more prone to the information bias previously discussed. The data stored in larger central databases may not be primary disease data, but surrogate indicators of the presence of disease. For example, milk recording systems will contain individual animal somatic cell counts (SCC), but will not include clinical mastitis events. Defining mastitis based on increased SCC tests requires a leap of faith that exceeds the comfort level of many health professionals.

7. Summary: Criteria on which to evaluate the potential usefulness of disease data

The advantages and limitations of disease data collected and stored at the three different organizational levels are summarized in Table 1. Based on the previous discussion, it should be obvious that under ideal circumstances, the disease data being used for health surveillance or genetic evaluation purposes should be collected specifically for that purpose, in a rigorous, prospective manner. Unfortunately, the cost and practicality of doing so are often prohibitive. Therefore, while the analysis of secondary disease data may not be ideal, it is often the only practical option. Under these circumstances, it is important to exercise caution and to understand the strengths and limitations of the data being used.

Table 1. Summary of advantages and limitations of disease data accumulating at three different organizational levels.

Level of data accumulation	Farm	Vet Practice / Local	Regional / National
Data accuracy	Variable	Variable	Variable
Consistency of case definition at the source	Moderate to High	Moderate	Variable
Consistency of case definition when aggregated	Low	Variable	Variable
Reference population	Farm(s)	Client Base	Variable
Event tied to unique farm	Yes	Yes	Occasionally
Event tied to unique animal	Yes	Yes/No	Yes/No
Basis of disease determination	Clinical	Clinical +/- Laboratory	Clinical +/- Laboratory
Degree of validation	Moderate	Low	Low
Information about population at risk	Yes	Some	Rarely
Ease of physical data aggregation	Difficult	Moderate	Variable

Following are five questions that should be answered to better understand the implications of using the available data.

1. ***Why were the data originally collected?*** Disease data collected as a by-product of another initiative may not have been scrutinized or validated to an extent which satisfies a different, secondary, use. Poor disease case definition, under-reporting of mild or unimportant conditions, misclassification of complex diseases and incomplete animal/herd identification are common problems.

2. ***What disease data were collected?*** Recording evidence of disease (actual disease events such as cases of clinical mastitis) rather than inference of disease (surrogate indicators of disease events such as elevated SCC's) is preferable. Some systems are more likely to contain disease treatment data, rather than disease diagnosis data. If single treatments are considered therapeutic and applied only to severe clinical cases, then the treatment rate may underestimate the true disease rate. Preventive treatments applied to all animals at risk of a disease could overestimate the actual disease rate.

3. ***Who was responsible for the identification and recording of the data?*** Diseased animals can be identified by farmers, veterinarians, regulatory officials or others. The severity of the condition, and the criteria upon which the diagnosis is made, may vary significantly among these individuals. If multiple persons are involved in the identification and recording process, there may be poor agreement about disease definition. Additionally, diagnostic biases of each individual involved in the recording process should be considered. Such biases can be based on breed, parity or age predilections for particular conditions, or other attributes of the animal(s).

4. ***When were the disease data collected?*** Disease events may be recorded daily, monthly or annually. Greater separation between occurrence and

recording increases the likelihood of significant recall bias. It is generally accepted that events that have significant positive or negative implications are more likely to be remembered than minor or insignificant events.

5. ***Where are the data collected and stored?*** Data collected, stored and used locally are more likely to be scrutinized and have errors identified and corrected. If the data are later aggregated to a central database, the methods used to merge the data should be clearly understood. If data are entered into more than one system (paper or electronic), there should be concern about possible transcription errors. Validation of data, wherever possible, should be performed on the final database.

8. Conclusion

Innovations in data collection, management and manipulation have created opportunities to answer many animal disease related questions. The temptation to collect and analyse data from large populations is great. However, caution must be exercised in the use of all data, but especially secondary disease data collected as a by-product of a different, primary, initiative. It is crucial that the user of the data understand it's attributes, complexities, biases and limitations, so that the information resulting from the analysis and interpretation is consistent with the original purpose and does not exceed the quality and representativeness of the original data.

References

Agger, J.F., Bartlett, P.C., Houe, H., Willeberg, P. and Lawson, G.L. 1997. Indicators of incomplete disease surveillance of clinical mastitis on a large national dairy database. *Proceedings of the Society for Veterinary Epidemiology and Preventive Medicine,* pp. 180-186.

Case, J.T. 1994. Disease coding and standardized nomenclature in veterinary medicine. ***In:*** *Proceedings of the 37th Annual Meeting of the American Association of Veterinary Laboratory Diagnosticians,* pp. 1-9.

Davies, M. 1985. Computers and veterinary practice in the United Kingdom. *Vet. Rec. 117*, 161-168.

Ekesbo, I., Oltenacu, P.A., Vilson, B. and Nilsson, J. 1994. A disease monitoring system for dairy herds. *Vet. Rec. 134*, 270-273.

Fetrow, J., Stewart, S., Kinsel, M. and Eicker, S. 1994. Reproduction records and production medicine. *Proceedings of the National Reproduction Symposium,* Pittsburgh, pp. 75-89.

Hueston, W.D. 1990. The National Animal Health Monitoring System: addressing animal health information needs in the U.S.A. *Prev. Vet. Med. 8*, 97-102.

Kelton, D.F. 1995. Monitoring, and investigating the relationships among health, management, productivity and profitability on Ontario dairy farms. *Ph.D. thesis.* University of Guelph.

Kelton, D.F., Lissemore, K.D., Martin, R. and Dekkers, J. 1997. Recommendations for national standards for recording and presenting selected clinical diseases of dairy cattle.

Lam, T.J.G.M., Schukken, Y.H., Grommers, F.J., Smit, J.A.H. and Brand, A. 1993. Within-herd and between-herd variation in diagnosis of clinical mastitis in cattle. *J. Am. Vet. Med. Assoc. 202,* 938-942.

Lissemore, K.D. 1989. The use of computers in dairy herd health programs: a review. *Can. Vet. J. 30,* 631-636.

Martin, S.W., Meek, A.H. and Willeberg, P. 1987. *Veterinary epidemiology: principles and methods.* Iowa State University Press, Ames, Iowa.

Menzies, P.I., Meek, A.H., Stahlbaum, B.W. and Etherington, W.G. 1988. An assessment of the utility of microcomputers and dairy herd management software for dairy farms and veterinary practices. *Can. Vet. J. 29,* 287-293.

Mulder, C.A.T., Bonnett, B.N., Martin, S.W., Lissemore, K.D. and Page, P.D. 1994. The usefulness of computerized medical records for research into pregnancy loss in dairy cows. *Prev. Vet. Med. 21,* 43-63.

Osteras, O., Leslie, K.E., Schukken, Y.H., Emanuelson, U., Forshell, K.P. and Booth, J. 1996. Recommendations for presentation of mastitis related data. International Dairy Federation.

Parke, P. 1993. *Dairy management software review.* Published by BC Dairy Herd Improvement Services.

Pollari, F.L. and Bonnett, B.N. 1996a. Evaluation of postoperative complications following elective surgeries of dogs and cats at private practices using computer records. *Can. Vet. J, 37,* 672-678.

Pollari, F.L., Bonnett, B.N., Allen, D.G., Bamsey, S.C. and Martin, S.W. 1996b. Quality of computerized medical record abstract data at a veterinary teaching hospital. *Prev. Vet. Med. 27,* 141-154.

Tierney, W.M. and McDonald, C.J. 1991. Practice databases and their use in clinical research. *Statistics in Med. 10,* 541-557.

Van Donkersgoed, J., Ribble, C.S., Boyer, L.G. and Townsend, H.G.G. 1993. Epidemiological study of enzootic pneumonia in dairy calves in Saskatchewan. *Can. J. Vet. Res. 57,* 247-254.

Willeberg, P. 1986. Epidemiologic use of routinely collected veterinary data: risks and benefits. In: *Proceedings of the 4th International Symposium on Veterinary Epidemiology and Economics,* Singapore, pp 40-45.