AN ASSESSMENT OF THE QUALITY OF PRACTICE-GENERATED COMPUTERIZED MEDICAL RECORDS AS A DATA SOURCE FOR RESEARCH INTO PREGNANCY LOSS IN DAIRY COWS

Carol Mulder, Brenda Bonnett, Paul Page
Department of Population Medicine
Ontario Veterinary College
University of Guelph
Guelph, Ontario, Canada

SUMMARY

This paper discusses the suitability of medical records as a source of data for research into the problem of pregnancy loss in dairy cows. The process of converting the computerized veterinary practice records into analyzable data is described. Various measures of data quality are discussed including the percent of missing, incorrect and outlier values in the final data set. Finally, descriptive statistics are calculated to demonstrate the similarity between the population of dairy cows described by the medical records and the target population of commercial dairy cows. This paper demonstrates that the positive aspects of practice-generated computerized medical records (volume of data, similarity to the target population) compensate for the disadvantages of poor case definition and missing data. The analysis of these data to investigate pregnancy loss in dairy cattle is currently under way and will be reported elsewhere

RESUMEN

El proposito de este articulo, es discutir si las historias clinicas son una fuente adecuadad de datos para investigar la perdida de prenez en vacas lecheras. Los procesos de conversion de las historias clinicas computarizadas en datos analizables son descritos. Varias medidas en la calidad de los datos son discutidas, incluyendo el porcentaje de valores no hallados, valores incorrectos y fuera de los rangos. Finalmente, la estadistica descriptiva ha sido calculada con el objeto de demostrar la similaridad entre la poblacion de vacas lecheras descrita por las historias clinicas y la poblacion de vacas lecheras comerciales a la cual va dirigido el estudio. Este documento demuestra que los aspectos positivos de las historias clinicas computarizadas (volumen de datos, similaridad con la poblacion a la cual va dirigido el estudio) compensan las desventajas de una pobre definicion de los casos y la ausencia de datos. El analisis de estos datos con el proposito de investigar la perdidad de prenez en vacas lecheras sera reportado posteriormente.

INTRODUCTION

In recent years, dairy practitioners have begun using veterinary computer software programs to enhance the reproductive health monitoring services they provide for their dairy clients (1). While the immediate value of the data generated by these programs is becoming more widely recognized by dairy veterinarians and their clients, the potential value of these same data for research has been somewhat overlooked. The use of routinely-collected data for research has been advocated primarily for economic reasons (2,3). Because the cost of data collection under these circumstances is either shared with or borne entirely by the primary user, the cost of these data to the researcher is minimal. Another major advantage in using routinely-collected field data is the similarity between the sample and target populations. Because of this, associations found in the sample population will be more valid for the target population than would findings from non-field data. However, there are limitations to using routinely-collected data. These include missing data and variable diagnostic rigour which can lead to information bias (2,3). The prime objective of this study is to examine the quality of data from practice-generated computerized medical records for use in research. The problem of pregnancy loss in commercial dairy cows was chosen as the research subject. Data concerning reproductive events are likely to be more complete than data concerning any other body system in this recording program. (This is because the software system was initially designed to monitor

reproductive performance.) The second reason pregnancy loss was chosen as the research subject is that the current diagnostic rate in bovine abortion is relatively low (4,5) and thus warrants further investigation.

MATERIALS/METHODS

DATA MANIPULATION

The objective of this study was to evaluate the quality of practice-generated medical records. VETCHECK (Infovet, Inc., Lachute, Quebec) medical records were chosen because the program has gained considerable popularity among dairy practitioners and thus constitutes a potentially large source of uniform data. (It is not the intent of this paper to discuss the relative merits of the VETCHECK system compared to other veterinary software programs.) The VETCHECK medical records of one private dairy practice were down-loaded from the practice microcomputer onto floppy disks. They were then loaded onto the computer network at the Ontario Veterinary College. At this stage, there were 3.4 Megabytes of data. The data structure is as follows. Each herd is a separate sub-directory. Within each herd sub-directory, each cow has a separate file identified by her name or number. The cow files are in ASCII format with data entered in 6 fields: date, event, uterus, left ovary, right ovary and comments. Data are entered chronologically within each cow file ending with the most recent date. This data structure was not readily suited for analysis because it was spread over approximately 4,500 files with over 75,000 records in 72 sub-directories and 2 languages. There were no discrete parameters or intervals in the raw data. In addition, events occurred with different frequencies and in different order within reproductive cycles and within cow files. Consequently, the data were translated into a new format using the Statistical Analysis System (SAS Institute Inc. Cary, North Carolina) programming. The final data set consisted of one file with 9,312 records, each record representing one reproductive cycle and containing 34 variables. (For the purposes of this discussion, a reproductive cycle is defined as the period between two consecutive parturitions, normal or otherwise.) The variables were created from the raw data by measuring intervals between events and identifying specific text entries in the various fields. Numerous logical operators were used to keep text entries pertaining to different events, reproductive cycles and cows separated from each other.

AVAILABLE DATA

Data were available concerning farm and cow identification and cull status. The number of calvings a cow had in the period covered by the computerized medical record can be calculated from the data but the total number of calvings cannot be determined. Age information was missing for the majority of cows because date of birth was recorded only for those cows that were born after the herd was enrolled on the VETCHECK software system. (When the herd is first enrolled, the only information entered is cow identification, the most recent calving and breeding dates and pregnancy status. Virtually all parameters that involve intervals between calvings, palpations and their results, treatments, breedings and pregnancy diagnoses were available. The only reproductive diseases identified were abortion, post-partum genital tract infection (which is defined by a range of pathology and/or treatments) and cystic ovaries.

QUALITY ASSESSMENT

Quality assessment was an ongoing component of the data translation process. Three specific approaches were used. First, cut points were instituted for all the continuous variables based on practical experience, biological sensibility and knowledge of the literature. Values that fell outside these limits were marked for examination.

Secondly, the values of all variables were recalculated manually for a random sample of approximately 1% of the cows. This was done six times at various stages in the development

of the final data set with a different sample each time. This was necessary to identify values that were incorrectly calculated but within the expected 'normal' range.

Finally, descriptive statistics were calculated for each variable. Extreme values were examined for correctness and biological sensibility.

RESULTS

DATA COLLECTION

Conversion of the original set of medical records into a data set suitable for analysis constituted a considerable programming challenge. To illustrate the nature of the difficulties encountered and also demonstrate the constant interplay of the quality assessment protocol in the conversion process, the steps involved in the creation of one variable, 'FIRSTBRED', are detailed below. 'FIRSTBRED' is the number of days between calving and first breeding. The first step was to sort the records of each cow in chronological order. These were then searched for the first occurrence of the event 'CALVED' within that cow's records. The date of this record was established as the calving date and used for all subsequent calculations until the next calving or the end of that cow's records. Then the records were searched for the first occurrence of the event 'BRED'. If this event was not found in the records or if it didn't occur until after another occurrence of the event 'CALVED', the program gave the variable a missing value for that reproductive cycle. If the event 'BRED' was found in the cow's records, 'FIRSTBRED' was set to equal the difference between the date of that record and the calving date. Under this definition, the variable 'FIRSTBRED' had a relatively normal mean (90 days) (6) with a range extending from 0 to over a year. However, approximately 20% of the calculated values fell beyond the expected upper limit of 210 days (7). There were no errors in calculation when the values were checked against the raw data. The only remaining explanations for this unexpected distribution were that this population of cows was truly aberrant or that information was missing from the original VETCHECK medical records. A closer examination of the enrolment of herds onto the program revealed that the latter was true. Only the most recent breeding date was included in the cow's record when she first entered the system. This may or may not have been the first breeding. In order to separate these potentially inaccurate values for 'FIRSTBRED' from valid calculations, the program was adjusted to take advantage of another peculiarity of the initial enrolment process. For cows pregnant at the time of enrolment, pregnancy diagnosis was listed either as a 'miscellaneous' event or was accompanied by the comment 'from previous records'. (This was done deliberately by the practice to differentiate between cows diagnosed pregnant by the practice and those diagnosed by other veterinarians.) The program adjusted the value of 'FIRSTBRED' to missing whenever either of these conditions were true. The resulting distribution of the variable was then much closer to what is expected based on other literature estimates.

DATA QUALITY

Approximately two hundred of the values of continuous variables had biologically impossible but correctly calculated values. These values were set to missing. This effectively eliminated these reproductive cycles from any analyses involving that variable but permitted inclusion in all other analyses as the presence of an impossible value for one variable in a reproductive cycle did not negate the reliability of other variables in that cycle. Four of the continuous variables had biologically possible but highly unlikely values. Many of these values were greater than 3 standard deviations above the mean and thus were considered outliers. The percentage of outliers ranged from 1.5% to 4.6% (for days open and gestational age at pregnancy diagnosis respectively). The relatively higher rate of outliers for the gestational age variable is due to the fact that in cows not diagnosed pregnant by rectal palpation before calving, a positive pregnancy diagnosis is not recorded until the day she calves. Under these

circumstances, the relatively high values for this variable probably do not constitute errors, regardless of their outlier status. (Note that these values are labelled as outliers based on the assumption that the variable has a normal distribution, which is not true for this variable.) Because these values were biologically possible, they could not justifiably be set to missing. Therefore they can be included in subsequent analyses. However, because they are outliers and few in number, their impact on the associations found is expected to be minimal.

The assessment of the reliability of the binary variables was less formal. Programming errors resulting in miscalculation or mis-assignment were corrected such that the final manual recalculation revealed no errors. Where possible, the distribution of variables was compared to previously published estimates. (Table 1) Because of the prevalence of missing data, (28% of the final data set) the number of reproductive cycles entering into subsequent analyses were considerably less than the total of 9,312 cycles. The minimum information present in all 9,312 reproductive cycles is conception date, number of breedings, parturition date and abortion status.

DISCUSSION

The conversion of these medical records into analyzable data was a complex but not impossible task. It required expert computer programming in addition to familiarity with the operation of the VETCHECK system in the field and practical experience in dairy reproductive health management. There is one major problem with the quality assessment protocol used on this set of medical records: there is no gold standard against which to measure the quality. Unlike computerized records in veterinary or human hospitals, the VETCHECK records are the only medical records kept on these cows. The paper data collection sheets are not retained because unlike detailed medical records that have been abstracted into computerized records, these data sheets contain no additional information to that already entered into the VETCHECK computerized record. There are no other recording systems outside the veterinary practice suitable for data-checking purposes. With nothing to compare to, these data simply have to be accepted as the true state of nature just as the hard copy detailed medical records of veterinary and human hospitals have to be accepted as the underlying truth when assessing data quality of abstracted computerized records (2,8).

Data quality as assessed by the methods used in this study is reasonably good. There were 236 data entry errors identified by examining the data for biologically impossible values. The presence of biologically possible but highly unlikely values indicate that there may be at least 179 more data entry errors. Together, these constitute less than 0.8% of the data in these variables. Mullooly (9) has shown that error rates below 1.0% will result in less than 10% attenuation in the measures of association found in the analysis of the data.

Data quality is affected by the prevalence of missing data. In experimental research, the most important consequence of missing data is decreased sample size and resulting difficulties in achieving statistically significant results. In this data set, sample size per se is not a concern. Even in the most restrictive analysis, there are 150 reproductive cycles in the sample. The question is whether these 'complete' cycles are representative of the much larger group of 'incomplete cycles'. It is possible that cows with specific characteristics are likely to be monitored more or less closely and are more or less likely to have abortions. The extent of the bias introduced by the use of complete records only can be estimated by comparison of the associations found in the complete and incomplete records for those variables present in both sets. In true scientific fashion, this exercise will answer one question and create another: if a difference exists between the complete and incomplete records, which one is 'better'? Will associations found on the complete records still be valid reflections of the population of cows in this medical record data set? Missing data has historically been considered 'bad', implying that complete data is 'good'. However, when using routinely collected data instead of

experimental data, the completeness of the data may be less an indicator of data quality as it is a reflection of the underlying risk or suspicion of disease. Missing data may also be an artifact of the way data are recorded in private practice. In general, veterinary practitioners record events and treatments only if they occur. Positive data entries are present; negative data entries are missing. Inclusion of prompters in the medical record system may help reduce the amount of missing data. However, in order to encompass all the possible parameters for which data may be required, the medical record system may become cumbersome. The resulting medical records may be more complete but unfortunately, they may also be fewer and less representative of veterinary practice in general since veterinary practitioners will be (justifiably) reluctant to adopt medical record systems they feel are unnecessarily complex.

The absence of specific classes of information clearly affects the quality of the data Age, calving number and lactation number are all missing from this data set. Because of the historical importance of these factors in disease (2), at least age will be manually entered into the data set for as many cows as possible. Inclusion of date of birth in the cow's record at enrolment on the VETCHECK program will overcome this limitation in data quality.

CONCLUSIONS

The amount of data available from these medical records was vast and of acceptable quality as a research data set. More information is needed to set standards for quality in data sets such as these. Acceptable levels of missing, incorrect and outlier values should be recognized so that the quality of information from routinely collected data can be objectively measured.

The task of preparing the medical records for analysis was complex but not insurmountable. In addition, further attempts to use similar medical records will entail less programming difficulty as the programs written for this study can be adapted for use elsewhere.

ACKNOWLEDGEMENTS

We thank Dr. Georges Lemire and L'Hopital Veterinaire Lachute, Quebec for their generous donation of the medical records used in this study. This study was supported in part by the Ontario Milk Marketing Board. Thanks also to Dr. Pilar Donado for Spanish translation.

REFERENCES

1. Lissemore, K., The use of computers in dairy herd health programs: A review. Can. Vet. J. 30:631-636. 1989. 2. Willeberg, P., Epidemiologic Use of Routinely Collected Veterinary Data: Risks and Benefits. Proc. of the 4th ISVEE, Singapore. 1985. 3. Thrushfield, M., Data bases in epidemiology. Equine Vet. J. 18(6):425-431. 1986. 4. Miller, R.B., Diagnosing the Cause of Abortion in Cattle. Bov. Pract. 22:98-101. 1987. 5. Murray, R.D., A field investigation of causes of abortion in dairy cattle. Vet. Rec. 127:543-547. 1990. 6. Lissemore, K., Ontario Dairy Industry Reference Values. Unpublished. 1992. 7. Etherington, W.G., Martin, S.W., Dohoo, I.R., Bosu, W.T.K., Interrelationships Between Postpartum Events, Hormonal Therapy, Reproductive Abnormalities and Reproductive Performance in Dairy Cows: A Path Analysis. Can. J. Comp. Med. 49:261-267. 1985. 8. Roos, L.L., Sharp, S.M., Wajda, A., Assessing Data Quality: A Computerized Approach. Soc. Sci. Med. 28(2):175-182. 1989. 9. Mullooly, J.P., The Effects of Data Entry Error: An Analysis of Partial Verification. Comp. Biomed. Res. 23:259-267. 1990. 10. Thurmond, M.C., Picanso, J.P., Jameson, C.M., Considerations for use of descriptive epidemiology to investigate fetal loss in dairy cows. JAVMA 197(10):1305-1312. 1990. 11. Giri, S.N., Stabenfeldt, G.H., Moseley, T.A., Graham, T.W., Bruss, M.L., BonDurant, R.H., Cullor, J.S., Osburn, B.I., Role of Eicosanoids in Abortion and its Prevention by Treatment with Flunixin Meglumine in Cows During the First Trimester of Pregnancy. J. Vet. Med. A. 38:445-459. 1991. 12. Etherington, W.G., Martin, S.W., Bonnett, B., Johnson, W.H., Miller, R.B., Savage, N.C., Walton, J.S., Montgomery, M.E., Reproductive Performance of Dairy Cows Following Treatment With Cloprostenol 26 and/or 40 Days Postpartum: A Field Trial. Therio. 29)3):565-575. 1988. 13. McLeod, B.J., Williams, M.E., Incidence of ovarian dysfunction in post partum dairy cows and the effectiveness of its clinical diagnosis and treatment. Vet. Rec. 128:121-124. 1991. 14. Sprecher, D.J., Nebel, R.L., Whittier, W.D., Predictive value of palpation per rectum vs milk and serum progesterone levels for the diagnosis of bovine follicular and luteal cysts. Therio. 30(4):701-710. 1988. 15. Halpern, N.E., Erb, H.N., Smith, R.D., Duration of Retained Fetal Membranes and Subsequent Fertility in Dairy Cows. Therio. 23(5):807-813. 1985. 16. Lee, L.A., Ferguson, J.D., Galligan, D.T., Effect of Disease on Days Open Assessed by Survival Analysis. J. Dairy Sci. 72:1020-1026. 1989. 17. Lemire, G.E., Stalheim, P.S., Lemire. M.R., Verdon, L., Tiemann, M., Bruning, T.R., Monitoring reproductive performance of small dairy herds in veterinary practice. Can. Vet. J. 32:551-557. 1991. 18. White, M.E., LaFaunce, N., Mohammed, H.O., Optimal time postbreeding for pregnancy examination in dairy cattle. Can. Vet. J. 30:147-149. 1989.

Table 1 Distribution of Selected Variables with Comparisons to Previously Published Values Where Available

Table 1a Binary Variables

VARIABLE	%	SAMPLE SIZE[1]	PUBLISHED ESTIMATES	REFERENCES
% ABORTIONS[2]	7.5	9,312	2 - 20	10, 11
% CULLS	30.8	4,518	24 - 36	6, 12
% PROSTAGLANDIN TREATMENT	19.6	9,312	NA	NA
% GENITAL TRACT INFECTION[3]	13.5	3,264	NA	NA
% NEG PD ERROR[4]	6.0	578	1 - 3.5	13
% POS PD ERROR[5]	0.3	2,232	OVERALL	
% CYSTIC OVARIES	10.7	6,069	4 - 20	6, 12-16

Table 1b Continuous Variables

VARIABLE	MEAN (RANGE)	SAMPLE SIZE	PUBLISHED ESTIMATES	REFERENCES
DAYS TO FIRST PALPATION	43.0 (1-317)	3,369	61	13
DAYS TO FIRST BREEDING	81.5 (0-444)	4,620	81-92	6, 7, 12, 16, 17
BREEDINGS/CYCLE	1.2 (0-11)	9,312	1.7-3.2	12, 14
INTERBREEDING INTERVAL	41.2 (O-369)	2,622	22-42	6, 16
DAYS OPEN	117 (0-582)	4,620	87-126	6, 7, 12, 14, 16, 17
GESTATIONAL AGE AT PREGNANCY DX	60.6 (0-288)	6,007	30-68	18
PP BODY SCORE	2.9 (2-3.7)	169	NA	NA
BODY SCORE	3.14 (1.8-5)	2,018	NA	NA
TOTAL CALVINGS	2.2 (1-9)	9,312	NA	NA

[1] Number of reproductive cycles from which data was available. Total sample size was 9,312 reproductive cycles in 4,158 cows

[2] Includes all entries of 'abortion' in medical records and all gestations <260 days

[3] Defined by one or more of the following entries in the medical record <90 days post-partum: RP, pus, pyometra, metritis, various intrauterine infusions

[4] Negative pregnancy diagnosis error. Cows diagnosed not pregnant that calve are subsequently diagnosed pregnant with no intervening breeding and calve <290 after the last breeding

[5] Positive pregnancy diagnosis error. Cows diagnosed pregnant that calve >270 and <290 days after a subsequent breeding.