Content analysis of free-text clinical records: their use in identifying syndromes and analyzing health data

Lam, K. [†1], Parkin, T. [*], Riggs, C. [†2], Morgan, K. [‡]
[†1] Department of Veterinary Regulation and International Liaison, and [†2] Department of Veterinary Clinical Services, Hong Kong Jockey Club, Hong Kong.
[*] Centre for Preventive Medicine, Animal Health Trust, Newmarket, UK.
[‡] Epidemiology Group, Faculty of Veterinary Science, University of Liverpool, UK.

Abstract

The use of a content analysis software package- WordStat and SimStat, Provalis Research, Quebec, Canada, enables a large volume of records to be sorted in a systematic manner with high accuracy and reliability.

The reasons for retirement from racing in Hong Kong for 3727 Thoroughbred racehorses, between the 1992/93 and 2003/04 racing seasons, were categorized into a user-defined dictionary. Identification of the most common causes of retirement from racing is the prerequisite starting point for epidemiological studies into the factors affecting racehorse performance and wastage in Hong Kong.

Introduction

A wide range and volume of information in a database, established in the early 1970's, at the Hong Kong Jockey Club including horse health and racing performance records of more than 6000 horses has provided a great opportunity for a retrospective analysis of data to assess the pattern of retirement from racing of Thoroughbred racehorses at the Hong Kong Jockey Club. This paper describes the methodology employed in the text mining of over 3700 free-text clinical records from this database using a commercial content analysis statistical software package- WordStat and SimStat, Provalis Research, Quebec, Canada.

Materials and Methods

A retrospective descriptive analysis of computerized Thoroughbred horse performance and health records, maintained by the Hong- Kong Jockey Club, was carried out. Content analysis of free-text records was used to identify and classify the reasons for retirement of horses.

Identification and classification of retirement reasons using WordStat content analysis software

The access database containing reasons for retirement was imported into content analysis program WordStat v 4 (Provalis Research, Quebec, Canada) Exploration of the records detailing reasons for retirement was carried out by two veterinarians (experienced in equine clinical medicine and epidemiology) using the phrase finder facility in the content analysis program.

Results

A final selection of 3727 records from the 1992/93 to 2003/04 racing seasons was included in the study.

Content analysis identified a total of 23,181 words in the free text records and 909 of these (3.9%) were unique individual words. A dictionary of 21 retirement categories was established and 96% (3564/ 3727) of records were categorized. The remaining 4% (163/ 3727) were manually assigned to one of the categories. Cross tabulation and Dendrogram function in the WordStat software provided information on the distribution of cross-matched case occurrence and clustering effect among the defined categories.

Removing cross-matched cases from the defined categories in WordStat dictionary by filtering in SimStat produced a final list of case definitions with a single condition for each category. Fifty four percent of records (2021/3727) included a single veterinary reason for retirement and 96% (1949/2021) of these were classified into 16 categories. The remaining seventy three records (3.6%: 73/2021) were unclassified with rare veterinary diagnoses (for example, stomach ulcers, ataxia, etc.). Three hundred and seventeen records (8.5%: 317/3727) appeared in more than one veterinary category. Horses in this group had more than one veterinary problem specified as the reason for retirement (for example, degenerative joint disease and tendon injury). There were 1389 records (37.3%: 1389/3727) with no veterinary reason for retirement. Ninety-three records (2.5%) had nothing recorded as reason for retirement.

Discussion

The most critical and difficult task in this process was the initial definition of the categories. Knowledge of the database, target population and clinical domain were important in screening the database to define these categories. Wordstat and Simstat represent a user-friendly combination of programs for content and statistical analysis respectively. It is easy to import and adequate support for conversions from most common formats (Excel, MS Access, SPSS, etc.) is provided. The automated procedures based on explicitly formulated and unambiguous logical conditions exclude inter-subject variance and present a clear time advantage over manual coding.

The use of WordStat/ SimStat has been employed to derive and validate an optimal search filter for retrieving clinical prediction rules in an attempt to enhance clinical judgment in diagnostic, therapeutic, and prognostic assessment in the US National Library of Medicine's MEDLINE database (Grimsmo, A.E. et al. 2001; Heinze, D.T. et al. 2001; Ingui and Rogers 2001).

The clinical information related to the reasons for retirement of racehorses in the free-text veterinary records has provided valuable information for both Jockey Club clinicians and managers to understand the pattern of different causes of retirement over the past 12 years. Areas of veterinary interest (tendon injury; osteoarthritis; exercised induced pulmonary hemorrhage; and fractures of which greater than 50% affected the proximal sesamoid bones) and non-veterinary reasons for retirement (poor racing ability; old age and compulsory retirement) have been identified in this study for further research investigation.

Acknowledgements

The authors wish to thank The Hong Kong Jockey Club, for full financial funding of this project; Mr Winfried Engelbrecht, the Executive Director of Racing; Dr Keith Watkins, Head of Veterinary Regulation and International Liasion, and Dr Brian

Stewart, Senior Veterinary Officer, Hong Kong Jockey Club, for their support on this project; Ms Iris Yu (Department of Veterinary Clinical Services), Mr Anthony Leung, Mr Danny Kwok and Mr Leo Cheung (Department of Information Technology) for their technical help in the retrieval of veterinary clinical records in this study.

References

Grimsmo, A., E. Hagman, et al. (2001). "Patients, diagnoses and processes in general practice in the Nordic countries. An attempt to make data from computerised medical records available for comparable statistics." Scand J Prim Health Care **19**(2): 76-82.

Heinze, D. T., M. L. Morsch, et al. (2001). "Mining free-text medical records." Journal of the American Medical Informatics Association: 254-258.

Ingui, B. J. and M. A. M. Rogers (2001). "Searching for Clinical Prediction Rules in MEDLINE." Journal of American Medical Informatics Association **8**(4): 391-397.