# Classification of free-text veterinary clinical records using content analysis

## *Reasons for racehorse retirement at The Hong Kong Jockey Club*

Kenneth Lam, [1] Tim Parkin [2], Chris Riggs [1], Kenton Morgan [2]

[1]Department of Veterinary Clinical Services, Hong Kong Jockey Club [2] Epidemiology Group, Faculty of Veterinary Science, University of Liverpool

## Summary

- Content analysis software was used to identify and classify the reasons for cessation of racing using text based clinical data recorded from 3727 racehorses at the Hong Kong Jockey Club between 1992 and 2004.
- More than 95% (3540 out of 3727 records) of the free-text and non-structured veterinary clinical history records were automatically assigned to one of 21 dictionaries.
- This technique enables a large volume of records to be sorted in a systematic manner with high accuracy and reliability.

## Aims and objectives

**Aim**: To identify the reasons for permanent cessation of racing for Thoroughbred horses at the Hong Kong Jockey Club and the risk factors associated with them.

**Specific objective**: To identify and classify the reasons for permanent cessation of racing using clinical text records.

## Background

- Clinical diagnoses are commonly recorded using descriptive text. Coding systems (e.g. SNOWMED, UMLS) have been developed but current evidence suggest that clinicians prefer to use free text.
- The retrieval and classification of this information presents a problem for epidemiological analysis.
- Here we describe the use of content analysis, a technique more commonly used in social sciences, to summarize and classify clinical records held on the Hong Kong Jockey Club Database.
- This is part of a collaborative research project between The Hong Kong Jockey Club (HKJC) and The University of Liverpool, UK.

## Materials and Methods

### Data collection:

- The Computerized Racing Information System (RIS) at The Hong Kong Jockey Club contains 5910 horse records and 3710 data fields (Microsoft Access).
- Data from the racing seasons 1992 – 2004 was used (3727 horse records).
- Horses which had permanently stopped racing were identified in the database and fields with free text records describing the reasons for this were used in this study.

### Content analysis of free-text clinical history records:

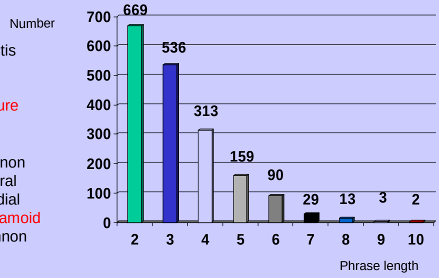
- A content analysis and statistical software package- (WordStat and SimStat, Provalis Research, Quebec, Canada.) was used to automatically select unique words and phrases of specified length from the data.
- The data were categorized using these data.
- Cross tabulation and similarity dendrograms were examined to look for associations and correlations between categories.

## Results

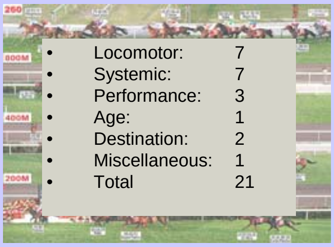
- 23181 descriptive words words were identified, 909 were unique.
- Common phrases, 2-10 words in length were inherent in the text. Their distribution is shown below:

Two word groups
- chronic strain
- chronic tendonitis
- colic surgery
- collapse during
- sesamoid fracture

Three word groups
- fracture LF cannon
- fracture LF lateral
- fracture LF medial
- fracture LF sesamoid
- fracture LH cannon



- Reasons for cessation of racing could be divided into 21 categories, in six major subject areas.below. 95% records were classified automatically.



| | |
|---|---|
| • Locomotor: | 7 |
| • Systemic: | 7 |
| • Performance: | 3 |
| • Age: | 1 |
| • Destination: | 2 |
| • Miscellaneous: | 1 |
| • Total | 21 |

- The categories within the two major areas are shown below:

**Systemic**

1. **HRT** - heart irregularity
2. **CLC** - colic
3. **CLP** - collapsed horse or inability to rise
4. **FBL** - 1st incidence of EIPH recorded
5. **SBL** - 2nd incidence of EIPH recorded - compulsory retirement
6. **LYN** - laryngeal problem
7. **D** - sudden death or severe trauma requiring immediate euthanasia

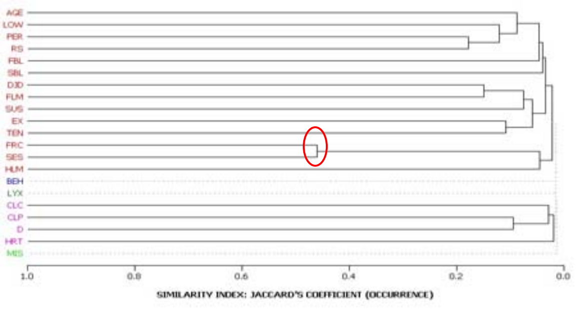
lesions of exercise induced pulmonary haemorrhage

**Locomotor**

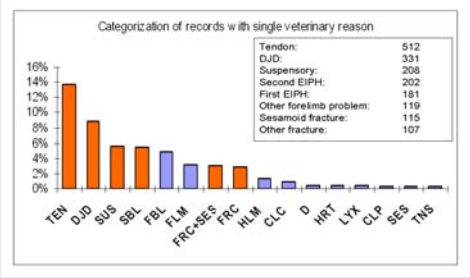
1. **DJD** - osteoarthritis
2. **TEN** - tendon or tendon sheath injury
3. **SUS** - suspensory apparatus injury (including sesamoidean ligaments)
4. **SES** - sesamoid problems including fractures and sesamoiditis
5. **FRC** - fracture of any bony structure
6. **FLM** - forelimb problems, lameness or other unclassified conditions
7. **HLM** - hind limb problems, lameness or other unclassified conditions

medial condylar fracture of the third metacarpal bone

- The dendrograms of similarity shown below identified categories which occurred together. The red circle highlights that more than 40% of the records, which appeared in the FRC category, were also included in the SES category. This informed the creation of a separate category for these records.



- Content analysis software allowed the rapid identification and categorization of data from text based clinical records and their distribution is shown below



## Conclusions

- Content analysis
  - Revealed an inherent standardisation in recording of clinical data (909 words).
  - Informed an improvement in coding; a separate field was added to recording sheets and the database to indicate whether retirement was voluntary or compulsory.
  - Allowed identification of 21categories of reason for cessation of racing
  - Enables research priorities to be identified: tendon injuries, osteoarthritis, sesamoid fractures.

Acknowledgements

The authors wish to thank Mr Winfried Engelbrecht, Executive Director of Racing, to support the on-going 4-year esearch project- Epidemiological study of factors affecting the performance of racehorses at the Hong Kong Jockey Club; Dr Keith Watkins, Head of Veterinary Regulation and International Liaison, and all the clinicians involved in the diagnosis and record sheets; Ms Iris Yu and Mr Leo Cheung for their technical help in data retrieval and analysis.

References

- Heinze, D.T., Morsch, M.L., Holbrook, J., 2001, Mining free-text medical records. J. Am. Med. Inf. Assoc., 254-258.
- Kreis, C., Gorman, P., 1997, Word frequency analysis of dictated clinical data: A user- centered approach to the design of a structured data entry interface. J. Am. Med. Inf. Assoc., 724-728.