Secondary Databases in Equine Research

Data quality and disease measurements

Johanna Penell
Faculty of Veterinary Medicine and Animal Science
Department of Clinical Sciences
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2009

Acta Universitatis agriculturae Sueciae
2009:59

ISSN 1652-6880
ISBN 978-91-576-7406-7

Print: SLU Service/Repro, Uppsala 2009

Secondary Databases in Equine Research – Data quality and disease measurements

Abstract

Knowledge on disease occurrence in the Swedish equine population is lacking. Secondary data (data not produced primarily for research) including medical information offer potential to investigate disease occurrence in populations without primary data collection. This thesis explored the potential use of two nation-wide secondary equine databases for research on diseases in the Swedish horse population.

The data quality in one insurance database and one database from a national equine clinic network was evaluated. For diagnostic information, the agreement in insurance data was 84% whereas the completeness (proportion of problems in the clinical records recorded in the database) and correctness (proportion of recorded disease events in the database truly occurring) of clinic data was 91% and 92%, respectively. Logistic regression showed that agreement/correctness was significantly associated with type of visit (clinic data and veterinary care claims in insurance data), whether diagnostic codes were present in the clinical record and affected system (clinic data). To present disease occurrence in the respective study populations, disease was presented as incidence rates (insurance data) and proportional morbidities (both databases). For insurance data, the most commonly affected system was joints, followed by whole body, skin and digestive system. The most common specific diagnosis was fetlock inflammation. For clinic data 22% of all visits were health visits, and for problems visits, the most commonly affected body system was joints, followed by whole body, respiratory and skeleton system. For both databases, disease occurrence was highly related to demographic factors in the horse population.

The data quality in both databases was found adequate for research purposes, with due consideration of variation in data quality among disease problems. Presentation of disease indices from the two databases provided useful information on disease occurrence in horses throughout Sweden. Importantly however, as many factors affect disease, results from other studies are not directly applicable to Sweden. Thus disease statistics need be obtained from the specific population of interest.

Keywords: secondary data, equine, insurance data, clinic data, data quality, validation, agreement, completeness, correctness, disease measurements, logistic regression

Author's address: Johanna Penell, Division of Ruminant Medicine and Veterinary Epidemiology, Department of Clinical Sciences, , SLU, P.O. Box 7054, SE- 750 07 Uppsala, Sweden. *E-mail:* Johanna.Penell@kv.slu.se

To my beloved family

When you can measure what you are speaking about and express it in numbers, then you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of mediocre and unsatisfactory kind.

Lord William Kelvin

Jag såg något hoppfullt i det här. Oklart vad.

Bruno K Öijer ur Svart som silver

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Penell JC, Egenvall A, Bonnett BN, Pringle J. (2007). Validation of computerized Swedish horse insurance data against veterinary clinical records. *Preventive Veterinary Medicine* 82, 236-251.

II Penell JC, Egenvall A, Bonnett BN, Pringle J. (2005). Specific causes of morbidity among Swedish horses insured for veterinary care between 1997 and 2000. *Veterinary Record* 157, 470-477.

III Penell JC, Bonnett BN, Pringle J, Egenvall A. Validation of computerized diagnostic information in a clinic database from a national equine clinic network (submitted manuscript).

IV Penell JC, Egenvall A, Bonnett BN, Pringle J. Health events in Swedish horses based on data from a national equine clinic network (manuscript).

Abbreviations

CCR	Computerized clinical record
CI	Confidence interval
CID	Computerized insurance data
CR	Clinical record
HYAR	Horse-years at risk
IR	Incidence rate
PM	Proportional morbidity
VCE	Veterinary care event
VCR	Veterinary clinical record

1 Background

1.1 Investigation of disease in populations

The investigation of disease in populations and of factors that determine its occurrence is the foundation of epidemiology. The focus on disease in populations rather than just in the individual will increase knowledge of disease and how to reduce and prevent it. There are different types of epidemiology, for example descriptive and analytical epidemiology (Thrusfield, 2003). Analytical studies usually include comparisons of groups and examples are experimental, cross-sectional, case-control and cohort studies. Cross-sectional, case-control and cohort studies are observational and focus on the natural occurrence of disease as the researcher has little control over the factors (e.g. weight, sex) that are being studied (Thrusfield, 2003). In general, analytical/observational studies are considered the preferred study design if the exposure(s) is more complex, and not easily controllable by the researcher (Dohoo *et al.*, 2003b). However, the other mentioned type of epidemiology: descriptive studies/surveys of population attributes can help answer a variety of key questions, such as what proportion of individuals have the disease and whether the distribution of disease differ between categories of individuals such as age, gender or race.

Defining the disease burden of populations and describing disease patterns are necessary in assessing the health status of a population. Knowledge on disease patterns can help in prioritizing research, improving health care regimens and guidelines, suggesting life style or management changes and, in veterinary medicine, support breeding strategies to ultimately reduce the severity and frequency of health problems in the population. Further, this information is essential for clinicians interested in an evidence-based medicine approach; to assess their cases based on the best evidence available. For example, estimates of prevalence and incidence of

disease and how they are affected by different characteristics (e.g. sex, breed) influence the probability that an individual has a certain condition. Combined with information from diagnostic tests, the posterior probability of disease in the particular individual can be determined. Further, by describing the population in a specific geographic region, comparisons can be made to populations described in published scientific studies and allow determination of the relevance of findings to the clinician's situation. Moreover, thorough published descriptions of disease events (irrespective of the number of cases included) will provide access to information and experience less anecdotal or affected by memory. These descriptions are also important when new diseases emerge or known diseases change characteristics or appearance. Descriptions of disease are especially important for medical problems where valid scientific evidence from well-designed analytical studies is lacking.

Primary data assembled prospectively are considered to offer the most valid information in observational studies, as the investigator can control the type and collection of data. However, gathering this sort of information can be costly in terms of time and finances. Further, concerns regarding the inclusion of only good compliers/recorders or those with personal interest in the particular research question (e.g. farmers/owners having more or less problems with a certain disease in their animal(s) raise issues on selection bias (i.e. if factors are present that affect the selection of study subjects) in the data collection. In the veterinary field, the ideal situation, i.e. a large primary, active database with currently recorded, up-to-date information on all individuals and disease events in the population will rarely be accomplished. Therefore, secondary databases with pre-recorded medical information offer an opportunity to access information that would have been difficult or tedious to retrieve elsewhere, and their potential in providing useful information on disease should be investigated.

1.2 Secondary data in medical research

Secondary databases contain data originally collected for other primary reasons such as patient care than that for which it is being used (e.g. research studies). Nonetheless, they can be a useful tool in medical research, especially if the data are in a form suitable for analysis and include variables that are informative and accurate.

The main feature of secondary medical data in medical research is obviously that they include diagnostic/disease information on individuals.

Depending on the original source, the database might also include information on the (healthy) population (e.g. gender, breed, age and geographic localisation), which greatly potentiates the use of the database by also enabling estimation of disease frequency in the population (i.e. incidence and prevalence).

Secondary medical data can be used to provide thorough description of disease in populations and quickly produce large samples of cases (especially for less common diseases) and controls. They can be used to conduct cross-sectional, case-control and cohort studies rapidly and efficiently and provide useful information on, for example, demographic characteristics and geographical distribution of disease. Concerns on use of secondary data are further discussed in section 1.3.

1.1.1. Human medicine

In human medicine, secondary databases have been used extensively to describe disease occurrence. The reason for this is that traditionally many countries maintain population-based registries. One example of use of population-based registry data in research is a study from Sweden where cases were identified from hospital data and controls from population registries and data then combined with socioeconomic data and home address to investigate the associations between traffic-generated air pollution and the risk of myocardial infarction (Rosenlund *et al.*, 2009). Further, medical databases from hospitals or general practices have been accessed to identify cases of colorectal cancer in the UK (Hamilton *et al.*, 2009), synovial sarcoma in the USA (Sultan *et al.*, 2009) and uveitis in Austria (Maca *et al.*, 2006). Moreover, based on medical information in secondary databases, incidence and prevalence of disease have been estimated, such as the incidence of breast cancer in Italy (Piscitelli *et al.*, 2009) and the prevalence of arthritis in the UK (Watts *et al.*, 2009). Other sources of medical data have also been accessed; for example a national database of forensic medicine for the study of the incidence of sudden cardiac arrest in young adults in Sweden (Wisten *et al.*, 2002) and "The U.S. Navy Aviation Medical Data Retrieval System" for investigation of the association of allergic rhinitis with other disqualifying ear-nose-throat defects in Naval aircrew (Walker *et al.*, 1998). Finally, the cost of a specific disease as been estimated based on secondary data (Perkins *et al.*, 2009).

When evaluating the usefulness of the results from these studies, certain issues need to be considered. Many case-control studies include cases that are not typical or representative of the cases occurring in the population. As well, lack of conformity in assessing the cases between different

practices/clinics/hospitals may affect classification of disease, especially when retrospective data from many medical centres are used. Further, controls should receive the same level of scrutinizing as the cases and there should be formal assessments that the controls are in fact free of disease. Finally, medical records may be incomplete or inaccurate, for example regarding treatment (Sultan *et al.*, 2009) and hence concerns about the data quality may be raised. Often the secondary data have been used without any consideration or evaluation of the types of cases captured (similar or different than the typical cases) or scrutiny of the data quality.

1.1.2. Veterinary Medicine

In veterinary medicine, secondary databases have been used in research for different species. For dairy cattle, many countries have national or regional cattle/dairy production databases maintained to support the dairy cattle industry. These databases have been accessed to retrieve information for research studies, either directly (for example retrieving individual or herd data) or indirectly by identifying herds that later were approached for the research study (Bruun *et al.*, 2002; Nyman *et al.*, 2007). Depending on the degree of enrolment in those databases, the representativeness of the individuals included and herds for the whole population in a particular area will vary; thus the extrapolation of research findings to other regions or conditions may be limited. Further, as the management and genetics will differ between regions (possibly even within same breed between different regions), valid studies with results applicable to the area of interest are important. As in human medicine, animal hospital data have been used in cattle disease research to investigate, for example, foot conditions (Nguhiu-Mwangi *et al.*, 2008) and abomasal displacement (Rohn *et al.*, 2004).

In canine medicine, secondary data have been used in numerous studies. Animal hospital data were accessed to identify cases of different diseases such as major abdominal injuries (Gower *et al.*, 2009) and pancreatic carcinoma (Priester, 1974). In North America, the Veterinary Medical Database (VMDB) compiles patient encounter data from many North American veterinary medical colleges. Data from this database have been used in many studies. For example, the prevalence of diabetes mellitus in dogs (Guptill *et al.*, 2003), secondary glaucoma (Gelatt & MacKay, 2004) and the seasonality of canine leptospirosis (Ward, 2002) has been investigated. Further, cases of highly specific diseases such as canine osteosarcoma (Ru *et al.*, 1998) and cataracts (Gelatt *et al.*, 2003) have been identified in this database. In some of these studies the potential selection bias of the cases and the bias derived from the structure of the database (e.g. diagnostic accuracy and lack of

uniform diagnostic criteria) were suggested as limitations. In general, the cases seen at veterinary (teaching) hospitals (i.e. referral centres) will differ from the typical cases seen in the population, but rather rarely is this issue discussed.

Canine insurance data originating from the same company used in studies I and II in this thesis have been used to describe incidence of mortality in dogs (Bonnett *et al.*, 2005; Egenvall *et al.*, 2005a) as well as incidence or risk of different diseases such as diabetes mellitus (Fall *et al.*, 2007), atopic dermatitis (Nodtvedt *et al.*, 2006b), pyometra (Egenvall *et al.*, 2001), bone tumors (Egenvall *et al.*, 2007) and mammary tumors (Egenvall *et al.*, 2005b). However, the cases captured in this insurance database must be considered. Again, for both hospital data and insurance data the cases included might be different from the cases seen in the general population. Even so, these secondary data can provide information on prevalence and incidence of various diseases; information difficult to retrieve elsewhere in veterinary medicine.

1.1.3. Equine medicine

In equine medicine, studies have also been performed based on data from various secondary sources. Veterinary hospital data have been used to identify cases of specific diseases in horses, such as septic arthritis in thoroughbred foals in the UK (Smith *et al.*, 2004). Further, the VMDB has been accessed to find cases of sarcoids (Mohammed *et al.*, 1992), recurrent airway obstruction (Couetil & Ward, 2003) and cervical vertebral compressive myelopathy (Levine et al., 2008). Moreover, the influence of different factors (e.g. age, gait and sex) upon racing success in standardbreds has been investigated using data from both the breed registry (i.e. the Canadian Standardbred Horse Society) and race records from a trotting association (Physick-Sheard, 1986). The usefulness of the results in these studies varies, as some are limited to a specific breed or usage and others include cases that may not be representative of those occurring in the general population. As well, some studies were performed several decades ago, which might make inferences to current situations more difficult as changes in diagnostic tests and treatments as well as type of cases and case severity over time likely will affect the classification process. Referral patterns and characteristics of the horse population (e.g. breed composition, usage, and management) may also vary over time.

Regardless of species investigated, many studies based on secondary data have failed to address issues of data quality related to the use of pre-recorded data. In particular, the representativeness of the cases included, the accuracy

of the available information and whether data are missing (and if so, what type) should be addressed as such aspects are important to have in order to address the research hypotheses more adequately. Despite potential shortcomings with secondary databases, a well maintained secondary database could be as good as a primary research database. Also, veterinary diagnoses in a secondary database might be the best available secondary medical information the veterinary epidemiology research community can hope for in terms of including a large study material.

1.3 Data quality

Whenever the primary purpose of collecting data differs from the specific goals of the research, the quality of the data must be addressed (Hennessy *et al.*, 2003). A formal validation of the database will provide key details on the structure and the quality of the information in the database, and its strengths and limitations. It is essential that data are accurately and completely recorded for the database to reach its full potential and usefulness, both concerning the primary reason of recording the data, e.g. patient care, clinic management and health service (Pringle *et al.*, 1995), and secondary use for meaningful research studies (Roos *et al.*, 1982)

Several types of errors have been shown to affect the quality of the data in a medical database. For example, data transcription error and entry into the database was detected when data entry into a medical database was assessed (Dambro & Weiss, 1988). Further, lack of transfer of information from the paper file to the summary sheet that was used to record information in the computerized record at a veterinary hospital resulted in error in the computerized record and discrepancy between the two sources (Pollari *et al.*, 1996a). While some of these errors were detected and corrected by the recording technicians the majority were not eliminated. Another study using human patient records reported that the physician was the largest error source (62%), followed by coding (35%) and keypunch (3%) (Lloyd & Rissing, 1985). Failure to record procedures or diagnosis/-es was the most common reason for physician errors. Further, coding was subjective and errors were synergistic with physician error. Also, inaccurate reporting in general practitioners' notes has been reported to affect data quality (Gormley *et al.*, 2008). Interestingly, another study reported that the main reason for data errors was the patient, due to misreporting of information (usually medication) but with transcription errors being relatively uncommon (Wagner & Hogan, 1996).

In human medicine, numerous validations of computerized diagnostic information have been performed. For example, medical databases have been evaluated for studies on specific diseases such as inflammatory bowel disease (Lewis *et al.*, 2002), venous thromboembolism (Lawrenson *et al.*, 2000), fractures (Ray *et al.*, 1992) and autism (Fombonne *et al.*, 2004). The quality of recording has also been found to vary depending on type of problem (Roos *et al.*, 1991; Jordan *et al.*, 2004). This variation could be related to how clear and straight-forward the diagnostic criteria for the specific problem are. In fact, one study reported higher estimates of sensitivity (defined as the proportion of problems in the veterinary clinical records that were recorded in the database) for problems with more objective diagnostic criteria such as diabetes, epilepsy, and glaucoma compared to problems with less objective diagnostic criteria, for example depression and stroke, when examining data quality in general practice data in Scotland (Whitelaw *et al.*, 1996).

In veterinary medicine only a few reported validations of secondary databases have been undertaken. Two evaluations of computerized medical record abstract data, one at a veterinary teaching hospital (Pollari *et al.*, 1996a) and the other of practice-generated computerized medical records (Mulder *et al.*, 1994), both in Canada, showed that the data quality in the two databases varied and were considered to be acceptable in the latter (Mulder *et al.*, 1994) but inadequate for the intended research in the former (Pollari *et al.*, 1996a). Other examples are validations of the diagnostic information in a canine insurance database, both overall (Egenvall *et al.*, 1998) and for a specific disease (Nodtvedt *et al.*, 2006a). The data quality of animal health records has been evaluated by comparing owner-reported to veterinary-reported information. In general, events that were owner-reported were more frequent although varying with the type of event, and some specific disease events were only recorded by the veterinarians (Menendez *et al.*, 2008; Mörk *et al.*, 2009). The findings suggest that neither farmers nor veterinarians can by default be assumed to record complete information on disease in dairy cows.

Evaluations of computerized medical information have usually involved validating a database against either a patient survey or a paper record (Roos *et al.*, 1991; Pilpel *et al.*, 1993). Moreover, validation of electronic patient records based solely on the contents of the clinic database has also been performed, for example by comparing morbidity data to recognised diagnostic standards to confirm a diagnosis and identify further cases (Hassey *et al.*, 2001).

Data quality can be assessed by investigating the data accuracy in terms of completeness and correctness (Pringle *et al.*, 1995; Hogan & Wagner, 1997; Hassey *et al.*, 2001). Completeness can be defined as the proportion of problems in the clinical records (considering the clinical records as gold standard) that were recorded in the database (i.e. epidemiological sensitivity). Correctness can be defined as the proportion of recorded disease events in the database that truly happened (i.e. positive predictive value) according to the clinical record (i.e. the gold standard). Table 1 shows the relationship between the database information and the true health state; thus completeness corresponds to (a/a+b) and correctness to (a/a+c). Reported values of completeness and correctness for diagnostic information have varied greatly between studies. For example, in human medicine general practice data showed 87% completeness and 96% correctness (Hassey *et al.*, 2001) where other studies have presented lower estimates of correctness (83%) (Wagner & Hogan, 1996) and higher estimates of both completeness (90–98%) and correctness (96–99%) (Pilpel *et al.*, 1993). One study estimated the completeness of the computerized hospital records for diabetes mellitus and glaucoma to 97% and 92%, respectively (Pringle *et al.*, 1995).

Table 1. *Relationship between the recorded information in the database and the true health state.*

		Database		
		Disease present	Disease absent	Total
True health state	Diseased	a (truly diseased)	b (falsely healthy)	a+b
	Healthy	c (falsely diseased)	d (truly healthy)	c+d
		a+c	b+d	a+b+c+d

The concordance of information between two sources of data can also be expressed as agreement, especially if there is no clear gold standard. The agreement for diagnostic information presented in different studies in veterinary medicine has varied considerably. For example, the agreement was as low as 33% when comparing farmers' and veterinary records, although varying from 20% to 64% depending on health-related event (Menendez et al., 2008), but considerably higher when veterinary records and canine insurance data were compared (87%) when excluding observations with missing data (Egenvall *et al.*, 1998).

Many studies have been performed on data from secondary sources. This in combination with the few validation studies performed on secondary medical databases in veterinary medicine and the reported problems with errors in the computerized information calls for formal validation of databases that could be used in research studies.

1.4 Disease in equine populations

Traditionally, horses are looked upon as individuals but nevertheless, similar to health care in humans, dogs and production animals, there is a need for population-based studies to increase knowledge on equine health. Population-based studies are needed to orient research and teaching towards the most common disease problems. Studies have been performed to investigate prevalence and risk factors for certain diseases (see section 1.1.2, 1.1.3). However, rather rarely are data available to determine the comparative importance of equine problems (Traub-Dargatz et al., 1991). In general, a combination of (complete) demographic information and medical information is needed to investigate populations. There are few active population-based registries for horses (and some, such as breeding registries, often lack information important for conducting research studies). Thus, other options need to be explored to retrieve information on the composition of and the disease burden in the horse population (Mellor *et al.*, 1999). This is essential so research activities, funding, education, breeding programs and management guidelines can be directed towards decreasing disease and improving the overall well-being of horses.

A few studies have presented valuable information on the size and demographic composition of horse populations (Kaneene *et al.*, 1997b; (NAHMS), 1998; Mellor *et al.*, 1999). In northern Britain, the most numerous breeds were thoroughbred or thoroughbred-cross animals, including 30% of the estimated population of 96,622 individuals (Mellor *et al.*, 1999). In Michigan, the total population was 129,932 horses with the American Quarter Horse, standardbred and arabian being the most numerous breeds (Kaneene *et al.*, 1997b). Both studies retrieved information by questionnaires to equine owners/operators and veterinarians. In the 28-state USA study where data were collected by personal interviews, the most numerous breed, by far, was American Quarter Horse (40% of the estimated population) ((NAHMS), 1998). These simple contrasts (Kaneene *et al.*, 1997b; (NAHMS), 1998; Mellor *et al.*, 1999) highlight how geographic regions can have highly different populations, which can substantially influence the occurrence of disease in each population.

Studies on disease in horse populations are also rather scarce. Lameness, foot problems and skin disease were the most common diseases seen in an owner-reported survey in Australia (Cole *et al.*, 2005). However, in the UK, owners reported that the most frequent health problems attended by veterinarians were traumatic injuries, musculoskeletal, respiratory and skin disorders (Mellor *et al.*, 2001). In contrast, when veterinarians ranked medical problems in adult horses in the USA, the most common medical

problems were (in descending order) colic, viral respiratory tract disease, endometritis, dermatitis and parasitism (Traub-Dargatz *et al.*, 1991). Leg lameness, dermatological problems and colic were the most numerous health problems in a prospective study in Michigan, USA (Kaneene *et al.*, 1997a). Further, the most common problems seen at horse operations in the USA were injury/wound/trauma, followed by lameness, colic and respiratory problems (USDA:APHIS:VS, 2006). The most frequent diagnoses recorded at 12 schools in the USA and Canada about 40 years ago were parasitism and fracture (Priester, 1970). Moreover, the demography and costs of colic in horses in Sweden have been described using insurance data (Egenvall *et al.*, 2008b).

For racing thoroughbreds, racing/jockey club data have been accessed to present, for example, information on the risk and causes of fatality (Boden *et al.*, 2006) and the reasons for retirement (Lam *et al.*, 2007). For this breed, it has also been reported that the greatest number of days lost to training was due to lameness (Rossdale *et al.*, 1985; Bailey *et al.*, 1999).

Mortality in horse populations have also been studied by a number of researchers (Leblond *et al.*, 2000; Wallin *et al.*, 2000; Egenvall *et al.*, 2006; USDA:APHIS:VS, 2007) The mortality rate in insured horses has been estimated to be 415 per 10,000 horse-years at risk (HYAR) (Egenvall *et al.*, 2006). The longevity of Swedish warmblood and coldblood horses has been investigated, showing a longer lifespan for mares compared to geldings for warmblood horses but the opposite for coldblood horses (Wallin *et al.*, 2000). In foals, the most common causes of death during the first 30 days of life were injury/wounds/trauma (19% of the deaths), which are potentially preventable causes of death (USDA:APHIS:VS, 2007). For older horses the most common causes of death/euthanasia were joint problems (Egenvall *et al.*, 2006), musculoskeletal problems (Wallin *et al.*, 2000), old age (USDA:APHIS:VS, 2007) and foaling (Leblond *et al.*, 2000). These comparisons (Leblond *et al.*, 2000; Wallin *et al.*, 2000; Egenvall *et al.*, 2006; USDA:APHIS:VS, 2007) underscore that results are strongly dependent on the studied material (i.e. included horses) and methods such as data collection and classification issues (e.g. who classified the disease event).

The variation in types of horses included in the reported studies (e.g. breed, usage) as well as climate, geographical differences and management factors affect the relevance of the results from other studies to the Swedish situation. Thus disease statistics need be obtained from the specific population of interest.

1.5 The situation in Sweden

Sweden has a tradition of maintaining different types of registries for both humans and animals. The most complete information is available for humans, with various types of registries such as the twin registry and a population census including for example individual socioeconomic data. The registries have been important for the development of population-based epidemiological studies on human health and disease.

For animals, there are several types of databases/registries. For dairy cattle, the Swedish animal disease recording system and the Swedish official milk recording scheme include information on individual animals (e.g. demographic information, production, and reproduction parameters) and herds (e.g. location, size, number of animals, production level). For other species, there are various registries. For example, many dogs are registered with the Swedish Kennel club. This registry includes, for example, demographic and pedigree information on the dog and results from some medical examinations (e.g. radiological examination for hip dysplasia) and data from this registry has been used in research (Malm *et al.*, 2008).

In Sweden over 75% of the horse population is insured (Report of the Commission on Equine Policy, 2000a; 2000b), and the animal insurance company Agria (Agria Insurance, P.O. Box 70306, SE-107 23 Stockholm, Sweden, www.agria.se) insures almost one third of the total horse population. All types of horses can enrol in insurance although the coverage is less complete for active racing horses. Unique features with nation-wide insurance data are that it includes a base population (i.e. the insured animals) with all breeds represented and also information on cases of disease and death/euthanasia. The latter is recorded either when the horse is seen by a veterinarian (by immediate settlement) or later when the owner submit claims for the cost for the veterinary care or for life insurance. Receiving reimbursement for costs related to veterinary care does not affect insurance fees for the owner or result in a general reduction of coverage. This encourages use of the insurance and subsequent recording of disease events in the insurance database. All these factors suggest that this database has a unique potential for describing disease and death in Swedish insured horses and should definitely be explored.

Another unique source of information on both individual horses and disease events is a nation-wide clinic database including information on all visits to a network of equine clinics. At present (2009), the company (ATG Equine Clinics Ltd, Hästsportens Hus, SE-161 89, Stockholm, Sweden, www.hastklinikerna.se) includes 25 clinics. With the exception of the clinic in Skara, which is an equine hospital, they are day clinics with

predominantly outpatients and they examine and treat all types of horses. Many clinics are in close proximity to racing tracks. In addition, the centralized management of the clinics increases the likelihood of conformity in assessing the cases. The use of a diagnostic registry for recording diagnostic information (Swedish Animal Hospital Association, 1993) aims at standardizing the diagnostic information, similar to that in the insurance database (where recording of diagnoses is based on the same registry).

Population-based estimates on disease in horses and presentations of large-scale data on the medical reasons for visiting an equine clinic are lacking in Sweden. There is a need for such information to increase knowledge on the disease pattern in horses in Sweden. This information can be used to direct actions resulting ideally in reduction of ill-health, disease and culling/death in horses. There is potential for use of the two databases to address these needs and to thereby prepare for future research using information in these databases. However, as has been outlined above, issues of data quality particular to these two unique sources of secondary data must be explored.

2 Aims

The overall aim of this work was to investigate whether two specific secondary equine databases could be used to estimate the amount and describe the type of disease in the Swedish horse population.

The following specific aims were set:

To validate the correctness of the diagnostic and demographic information in a Swedish animal insurance company database (I).

To present incidence of disease due to specific causes among horses covered by complete insurance for veterinary care by a Swedish animal insurance company (II).

To validate the correctness and completeness of the diagnostic information in a nation-wide equine clinic network database (III).

To describe the horses and types of problems seen at visits to 20 equine clinics in a nation-wide equine clinic network in one year (IV).

3 Methodological considerations

Overall issues regarding the material and methods applied in papers I-IV are outlined in this section. For detailed descriptions of materials and methods, the reader is referred to each individual paper.

3.1 Study populations

Two different populations were studied regarding data quality and disease frequencies; horses insured by the Swedish animal insurance company Agria (I-II) and horses examined and treated at the ATG Equine Clinics Ltd (III-IV). The characteristics of each population are presented in the two following sections.

3.1.1 Insured horses

Horses could have insurance for veterinary care and life, with different types of coverage called complete, limited or racing. Most horses had complete coverage, but actively racing horses could not have such coverage. In brief, complete insurance reimbursed the owner if the horse was injured or became ill for various reasons whereas limited insurance covered only certain problems of mainly traumatic/acute character. Racing insurance resembled the complete type although excluding most limb problems. In addition, there are other types of insurance for veterinary care such as insurance for breeding horses, but they cover relatively few horses. Table 2 shows the breed group composition of some documented horse populations compared with the population insured for veterinary care by Agria in 2000. For all types of veterinary care insurance, the owner paid 20% of the cost over the policy deductible; the maximum that was reimbursed per calendar year was 25,000 SEK (£2,154/€2,469, exchange rates as of Aug 26th 2009).

The use of the insurance did not affect the future insurance fee for the owner. Life insurance reimbursed the owner if the horse died or if it was euthanized due to severe injury or illness as judged by a veterinarian, and the terms varied among types of life insurance. Details on life insurance are available elsewhere (Egenvall *et al.*, 2006).

Table 2. *Breed group composition (%) of documented horse populations compared to the horse population insured for veterinary care by Agria and analysed in study II, including horses ever insured during year 2000.*

	Michigan 1991[1]	Northern Britain [2]	NAHMS 1998[3]	Sweden 2000 [4]	Agria Total 2000	Agria Complete 2000
Warmblood	60†	16	57‡	30	29	36
Thoroughbred	21§	36§	16	2	7	7
Standardbred	1	-	3	27	20	6
Pony, miniature	11	37	7	24	28	35
Coldblood/draft horse	6	2	4	10	8	6
Icelandic horse	-	-	-	7	6	8
Donkey/mule	1	1	5	-	-	-
Other/ breed unknown		8	8		2	2
Total number of horses	129,932	96,622	4,029,000	225,000	>94,000	>70,000

† light horse (except §) and warmblood
‡ appaloosa, paint, Tennessee walker, quarter horse and other (registered)
§ Arabian, thoroughbred and mixes thereof
[1.] Kaneene et al., 1997
[2.] Mellor et al., 1999
[3.] NAHMS 1998
[4.] Report on the Commission on Equine Policy, 2000

The diagnostic information in the insurance database was based on information in claims for reimbursement (veterinary care or life). For veterinary care, each claim could have one or more receipts included. A receipt could have only one diagnostic code associated with it, which was derived from the diagnosis provided by the attending veterinarian. Veterinary care claims were settled by the insurance clerks and were either processed retroactively (the owner sent in receipt(-s) for veterinary care events) or through immediate settlement with the treating clinic, in which the clinic invoice was made so the owner and the insurance company were billed electronically their respective part at discharge from the clinic.

3.1.2 Clinic horses

The nation-wide equine clinic network ATG Equine Clinics Ltd was originally organized in the 1970's to provide qualified veterinary equine clinical service throughout the country. In 2002 the ATG Equine Clinics were owned by a company (ATG) and included 20 horse clinics throughout Sweden. The geographical distribution of the clinics is shown in figure 1.

The ATG Equine Clinics the veterinarians examine and treat all types of horses with any problem that can be dealt with during office hours, including prophylaxis and health examinations. In 2002 approximately half of the clinic horses were leisure/sport horses and half were racing horses (predominantly trotters).

Figure 1. *The distribution of the ATG Equine Clinics in Sweden in 2002 with the size of the ATG logo corresponding to the proportion of visits at each clinic. Source: ATG Equine Clinics Ltd.*

At each visit to an ATG equine clinic, the identity of the horse was either retrieved from the main ATG database, or for new patients in the leisure/sport horse usage category, created at the visit. Also, name, address and identification number of the owner (i.e. person responsible for payment) as well as information on the visit was recorded. The computerized record for the visit further included a unique visit identification number, date of visit, at which clinic the horse was seen and the diagnostic and procedural information related to the visit. At least one diagnostic code and one procedural code were registered in the clinic database for each visit. There was no upper limit of the number of diagnostic codes that could be recorded at each visit.

3.1.3 The diagnostic classification

For both databases, the same diagnostic registry (Swedish Animal Hospital Association, 1993) was used to assign diagnostic codes to each receipt in the claim (I and II) and to record diagnostic information in the clinic database for the clinic visit (III and IV). It is a hierarchically ordered diagnostic registry for horses, dogs and cats and contains approximately 8000 alphanumeric codes, with the following 14 different major organ systems: integumentary, digestive, genital, respiratory, skeletal, auditory, joints, ocular, urinary, cardiovascular, endocrine, nervous, muscular and unspecified/whole body. In study I, II and IV, the integumentary system was divided into skin and hooves. The system 'unspecified/whole body' pertains to signs of disease that cannot be attributed to a specific organ system, and to diseases that are considered to involve the whole animal, such as infectious diseases and parasitic conditions. Within each organ system, except for the endocrine, there are subsystems. For example, in the respiratory system, the three subsystems are upper airways, lower airways and thoracic cavity/thoracic membranes. Ten major process groups can be assigned within each system: symptomatic, developmental, degenerative, circulatory, inflammatory, immunological, neoplastic, traumatic, toxic and idiopathic, and each process can be further divided into sub-processes. The diagnostic registry includes detailed aetiopathological designations, for example, 'tetanus', general symptomatic descriptions, for example, 'impaired trotting gait', and also 'no diagnosis', the last two designations belong to the symptomatic process group.

3.2 Study methods

3.2.1 Validation of the equine insurance database

The main objective of the validation of the equine insurance database (study I) was to assess the agreement between the diagnostic and demographic (i.e. name, gender, breed and birth date) information in the computerized insurance data (CID) and the clinical records (CR). The validation was performed on a random sample of all claims, including approximately 20,000 veterinary care and 4,000 life insurance claims that were processed during one year (1999). To achieve an adequate distribution across clerks, time of year and location of horses, it was decided to sample two veterinary care claims and one life claim per processing day.

The computerized information for the selected claims was accessed and the corresponding paper files at the insurance company were retrieved. If not already in the paper file, a copy of the horse's medical record was requested from the attending veterinarian/clinic. Claims where the paper file was missing at the insurance company, life claims where no veterinarian was involved and claims with an unavailable clinical record were excluded from further analysis (171 claims removed). The information in the available clinical records was transferred to a database without knowledge of the information in the computerized insurance data (i.e. blinding). All records were evaluated solely by the principal investigator. For demographic information, the computerized insurance data had no missing information whereas the clinical records could have missing information. There was no missing information on diagnosis or system in the clinical record or the computerized insurance data. Essentially objective criteria were used when assessing the agreement between the sources for the demographic information, whereas for the diagnostic information, the assessment was based on rather subjective criteria (table 3). For diagnostic information the correctness was assessed for both the specific diagnostic code and on body system-level (e.g. joint, respiratory). Using the clinical record data as gold standard, the observed sensitivity, specificity and positive predictive value were evaluated for four body systems.

Table 3. *Definition of agreement criteria used for comparison of observed agreement on diagnostic information between computerized insurance data (CID) and clinical records (CR) in a sample of horse claims at Agria during 1999 (study I).*

Agreement	Minor disagreement	Major disagreement
If the same diagnostic code was used	When the diagnosis in CID was too precise compared to the information in the CR	When a diagnosis seemingly unrelated to the condition described by the CR was recorded in CID
If the diagnostic term(s) could be used inter-changeably	When the diagnosis in CID had a reasonably similar meaning as the information in the CR	Exact diagnosis in CID and vague diagnostic information in CR
When the diagnosis in the CID was similar but less exact than in the CR	When "no diagnosis" was recorded in CID despite clear information in CR	The registered code had correct body system but incorrect on the exact position (e.g. which joint)

3.2.2 Disease measurements based on insurance data

Study II included all horses with complete veterinary care insurance during 1997–2000. The exact time at risk was calculated for each horse (i.e. the base population). After data cleaning (e.g. by establishing the number of individual horses in the database which excluded approximately 15% of the initial number) a final dataset including over 100,000 horses covered by complete insurance was created. Breeds were classified into breed groups. The owner's postal code was used as a proxy to assign where the horse was kept (south, central and north of Sweden and in area with higher or lower human population density; urban/other).

Horses insured for complete veterinary care were included in the denominator for the calculation of morbidity. In constructing overall crude rates of morbidity (Egenvall *et al.*, 2005c) horses were counted once per year as having at least one veterinary care event (VCE) for which the cost of veterinary care was greater than the deductible and for which the owner was reimbursed; these horses constituted the numerator for the risk of morbidity. In the analysis in study II, diagnoses were taken from the receipts that either alone or together constituted reimbursed claims. Horses were counted as having one or more disease events within each diagnostic category (system, process or specific diagnosis), and they could thus be in the numerator for several system or process calculations if they had several receipts with different systems or processes assigned. A specific diagnosis (or system or process) could be counted only once per horse each year. Descriptive statistics were calculated. Incidence rate calculations were used with the

exact time at risk as the denominator. The annual incidence rates were averaged over the four years, and they were calculated by system, process and five joint conditions and also with further stratification (e.g. by breed group, sex age, geographical location and human population density). Proportional morbidity was calculated by system, by system within breed group, and diagnosis within a system.

3.2.3 Validation of the equine clinic database

The evaluation on the data quality for diagnostic information in the clinic database (study III) was performed on a random sample of visits to 18 of the ATG Equine Clinics in 2002 with the computerized list of visit identification numbers as sampling frame. The validation of the clinic database included investigation of the completeness and correctness of the computerized clinical record (CCR) and was done by comparing the computerized information to the hand-written veterinary clinical records (VCR)(used as gold standard).

Completeness was defined as the proportion of problems in the veterinary clinical records that were recorded in the database (i.e. epidemiological sensitivity). Correctness was defined as the proportion of recorded disease events in the database that truly happened (i.e. positive predictive value). The evaluation of correctness and completeness was performed by the principal investigator and the rather subjective evaluation criteria are shown in table 4.

For four selected body systems (joints, skin and hooves, respiratory and skeletal), the presence/absence of each system relative to each visits was determined for both sources. For four chosen joint diseases (fetlock, carpal, stifle and hock joint inflammation), the correctness of the computerized affected limb information was investigated by assessing the proportion of the computerized affected limb data (front/hind and left/right/both or all four) that truly happened, according to the veterinary clinical record.

Table 4. *Definition of criteria used for assessing correctness of diagnostic information in computerized clinical records (CCR) compared to the corresponding veterinary clinical records (VCR) and number of cases in each category in a sample of visits (in 396 visits including 491 individual diagnoses in the CCR) at 18 ATG Equine Clinics in Sweden during 2002, (study III).*

Correctness		Absence of correctness
Complete	Partial	
The same diagnostic code was used (n=294)	Slightly more precise information in the CCR (n=12)	A diagnosis unrelated to the condition described by the VCR had been recorded in the CCR (n=12)
The diagnostic code in the CCR corresponded to the diagnostic information in the VCR (n=90)	"No diagnosis" in the CCR and non-specific information in the VCR (n=39)	Specific diagnosis in CCR despite non-specific diagnostic information in VCR (n=26)
Diagnosis in the CCR was similar but less exact than the VCR (n=1)	"No diagnosis" in the CCR and specific information in the VCR (n=17)	

3.2.4 Disease measurements based on clinic data

The clinic data used in studies III and IV originate from a large Oracle database (www.oracle.com) that is operated by ATG. The database includes information relating to racing horses in Sweden (e.g. demographic information on all racing horses in the country, racing related information such as trainer, racing starts, finished races, placings, race times, personal best, prize money won), and information on all visits (including the few ambulatory visits) to any of the ATG Equine Clinics. The demographic information on most racing trotters and racing thoroughbreds enter the database as foals (i.e. during the year they were born) by registration of their unique identity (based on demographic information on the horse such as name and exact birth date and also information on parents). In this process, a registration number and a horse identification number specific for each horse is produced. Non-racing horses enter the database at the first visit to an ATG equine clinic and are then given a unique horse identification number.

In study IV, the computerized information on all visits registered at the 20 ATG Equine Clinics during 2002 was accessed. For all horses, available demographic information included name, sex, breed and year of birth. Usage was categorized as racing trotter, racing thoroughbred or leisure/sport horse. Breeds were classified into breed groups; seven for leisure/sport horses, two for racing trotters and one for racing thoroughbreds. Clinics

were classified into south, central or north of Sweden depending on the geographical localization. Age groups were created based on the year of birth.

After data cleaning when visits with incomplete information on the horse or the visit as well as horses with names suggesting a group (e.g. riding school) and horses older than 30 years old were removed (corresponding to 1,837 visits being excluded in total), a total of 21,496 horses with 50,150 visits to the ATG Equine Clinics during 2002 were included in the dataset used in study IV. Descriptive statistics were calculated describing the horses and the visits at the 20 ATG Equine Clinics in 2002.

3.3 Statistical methods

Simple proportions were used to present agreement/correctness between the CID and the CR and the CCR and the VCR, respectively, for demographic (study I) and diagnostic information (study I, III) in the validation studies. They were also used to present the affected limb information in study IV. Confidence intervals (95%) were constructed in study I, III and IV using the two-tailed exact binominal test. In general, non-overlapping CIs were considered a significant difference between compared groups. The χ^2 test was used in papers I and III to evaluate associations with, for example, the proportion of missing information and the proportion of agreement for demographic and diagnostic information, and to investigate the difference in proportion of correctness on diagnostic information for first and follow up visits, for whether a diagnostic code was explicitly written in the VCR and for first and follow up visits for system joint, respectively. In study II, incidence rates were calculated with the horses with at least one VCE in the numerator and the exact time at risk for horses with veterinary care insurance in the denominator. In study II, the proportional morbidity by system, by system within breed group, and diagnosis within system was calculated as shown in Eq. (1-3).

Proportional morbidity $_{(system)}$= (the average of the annual number of cases with a system diagnosis)/(the annual number of cases with at least one recorded disease event)

Eq. (1).

Proportional morbidity $_{(system\ within\ breed\ group)}$=(the number of cases within a system for a breed group)/(the total number of cases in that breed group)

Eq (2).

Proportional morbidity ($_{\text{diagnosis within system}}$)=(the average of the annual number of cases with a specific diagnosis)/(the annual number of cases with a diagnosis in that system)

Eq (3).

The total proportional morbidity by system or within a breed group could exceed 100% because a horse could have several receipts with different systems assigned. The interpretation is the proportion of the horses with at least one veterinary care event that experienced at least one event in that system. In study IV, the proportional morbidity was calculated with all diagnoses recorded at problem visits in the denominator and the specific diagnostic code or system in the nominator.

Logistic regression models were used in study I and III to investigate factors related to agreement/correctness of diagnostic information. In study I, logistic regression analysis was performed separately for veterinary care and life claims to investigate factors related to the agreement (including categories agreement and minor disagreement) on diagnostic information. Explanatory fixed factors were type of visit (clinic/field), computerized or manual clinical records, whether the clinical record was included in the paper file at the insurance company, if the claim was rejected or reimbursed, system diagnosis and if an immediate settlement was done (for veterinary care claims) or if a death certificate (for life claim analysis) was included in the paper file. Clustering of claims within processing clerks and treating veterinarian/clinic, respectively, was investigated in multilevel logistic regression models. In study III, logistic regression analysis was performed to investigate factors related to correctness (including correctness and partial correctness) on diagnostic information. One diagnosis was randomly selected to represent each visit in the logistic model (n=396). Explanatory fixed factors included whether an explicit diagnostic code was present in veterinary clinical records, type of visit (first/follow up), system diagnosis in computerized clinical records (categorized as joints, skin and hooves, respiration, skeletal, whole body and other), gender and usage category. Clustering of diagnoses within clinics was investigated and clinic was added as a random factor to the fixed effects model. The statistical significance of the random effect was determined by comparing the models with and without the random effect included (Dohoo *et al.*, 2003c). Model reduction was done by successively removing the variable with the highest p-value and rerunning the model until only variables with p-values of <0.05 (the final p-value) were left.

In studies I, III and IV, Stata (Stata Special Edition, version 9.0, StataCorp, College Station,TX 77845, USA) was used for analysing the data. MLwiN, version 2.0, Centre for multilevel modeling, Institute of Education, London WC1H 0AL, UK) was used in studies I and III to evaluate the model variance for the random effects logistic model and in study I to determine model fit of the random effects model. In study II, the statistical software program SAS (SAS Institute) was used to analyse the data.

4 Main results

4.1 Insurance data

The validation of the computerized horse insurance data against veterinary clinical records (study I) included 400 veterinary care and 140 life claims. The overall agreement for demographic information (breed, sex, year of birth and name) was >94% and for system and specific diagnosis information 92% (95% CI 89,94) and 84% (95% CI 81,87), respectively. For the systems evaluated (joints, digestive, skeletal, skin and hooves), sensitivity varied between 62% (skin) and 89% (digestive) whereas the specificity was >96% for all systems. The positive predictive values ranged from 86% (skin) to 97% (digestive). For life claims, in the logistic regression analysis no explanatory fixed effects were significantly associated with agreement on diagnostic information. In the logistic regression analysis for veterinary care claims, type of visit was significantly associated with agreement, with clinic visits generating better agreement than field visits. When treating veterinarian/clinic was added as a random effect (p=0.04), this variable accounted for 14% of the model variance. With treating veterinarian/clinic as a random factor in the model, a clinic visit had an OR (for agreement) of 4.4 (95% CI 1.3, 17.3) compared to a field visit. Processing clerk as random factor was non-significant and accounted for <1% of the model variance.

The study on cause-specific morbidity among Swedish horses insured for complete veterinary care (II) found that the most commonly affected body system was joints with an incidence rate (IR) of 404 events per 10,000 horse-years at risk (HYAR) and a proportional morbidity (PM) of 37%. Problems in the joints was followed by whole body (IR 283, PM 26%), skin (IR 123, PM 11%), digestive (IR 121, PM 11%), skeletal (IR 98, PM 9%), hooves (IR 81, PM 7%), respiratory (IR 54, PM 5%) and muscular (IR 49, PM 4%). The breed group warmblood had the highest risk of morbidity for

the above mentioned systems except skin in which the breed group thoroughbred had higher risk. Geldings had the highest risk of at least one disease incident in system joints, unspecified/whole body, skeletal and respiratory systems, but in the other four investigated systems (skin, hooves, digestive and muscular) the differences compared to mares were marginal. For system joints, horses 5-15 years old had a higher risk of disease when compared to younger and older ages. The process classification showed that inflammatory problems were the most common (IR 519), followed by symptomatic (IR 335) and traumatic (IR 267). The ranking of the specific diagnoses was (in descending order) fetlock joint arthritis, undefined lameness, signs from the locomotor apparatus, traumatic injuries to the skin, arthritis in several joints, and colic. A total lameness estimate was created by combining lameness-related diagnoses. The average incidence rate of lameness was 628 events per 10,000 HYAR. The system-specific incidence rates were affected by geographical location. Within the eight most commonly affected systems, the incidence rates were higher in the urban areas, except for the respiratory system.

4.2 Clinic data

The validation of the computerized diagnostic information in the clinic database (study III) was based on 396 visits/clinical records. The overall completeness was 91% (95% CI 88,93), and the correctness was 92% (95% CI 90,94) and varied between clinics from 69% to 100%. If diagnostic information from the first visit was also included, the overall correctness increased to 97% (95% CI 95,98). There was a significant difference ($p<0.05$) in the degree of correctness for records with and without a specific diagnostic code written in the VCR (99% and 81% correctness, respectively) and for diagnoses recorded at first visits compared to follow up visits (97% and 80%, respectively). The correctness was similar for records from the three types of horses (91% for categories leisure/sport horse and racing trotter, 90% for racing thoroughbred) and also for the three genders (91% for geldings and for mares, 90% for stallions). For the four selected body systems (joints, skin and hooves, respiratory, skeletal) the completeness varied between 71% (respiration) and 91% (joints), the correctness ranged from 87% (skin and hooves) to 96% (respiration) whereas the specificity was >95% for all systems. Logistic regression showed that correctness was associated with type of visit (first visit: OR 3.8, 95% CI 1,10), whether an explicit diagnostic code was present in the veterinary clinical record (diagnostic code present: OR 32, 95% CI 7,146) and body system (OR varied from 0.6 for skin and hooves to 13 for whole body, baseline joints).

The crude correctness for information on affected limb was 95% and varied from 95% (fetlock and carpal inflammation) to 100% (stifle and hock joint inflammation).

The investigation of the health events based on clinic data (study IV) included both description of the horses examined and treated at the clinics and the reasons for the visit (i.e. the diagnostic information). The most numerous breed groups were racing standardbred trotter (41%), warmblood (27%), pony (12%), racing coldblood trotter (6%) and racing thoroughbred (5%). The age of the horses varied from 0-30 years of age (except for racing horses where the maximum age was 15 years old) with the majority of horses being between 3 and 15 years old (79%). Most visits (80%) had one diagnosis recorded at the visit, but the number varied from 0 to 6. Of all visits, 22% were health visits including mainly prophylaxis (36%), without diagnosis (29%), health check-up (18%) and inspection for health certificate (14%). For problem visits, the most commonly affected body systems were (in descending order) joints (PM 36%), whole body (PM 28%), respiratory (PM 7%), skeleton (PM 7%), digestive (PM 6%) and muscular (PM 6%). The ranking of the most numerous specific diagnoses at problem visits (n=48,410) was (in descending order) without diagnosis (20%), fetlock joint inflammation (14%), undefined lameness (5%) and carpal joint inflammation (3%). A total lameness estimate was created by combining lameness-related diagnoses and showed that of the diagnoses recorded at problem visits, 22,553 (47%) were lameness-related. In the analysis of affected limb information for fetlock and carpal joint inflammation, leisure/sport horses had a significantly higher proportion of fetlock problems in right and both fore legs compared to racing trotters, whereas racing trotters had a higher proportion of fetlock problems in hind legs (right, left and both) compared to leisure/sport horses. For carpal joint inflammation, leisure/sport horses had significantly higher proportion of problems in right and left front limbs when recorded as the single affected limb but racing trotters had significantly higher proportion of both front limbs recorded.

5 Discussion

This thesis was written with the focus on the usability of two equine databases of different origin (insurance and clinic data) in research, regardless of aspects concerning the adequacy for their primary use. Therefore, it was important to address issues regarding the secondary use, such as characteristics of the two data sources (e.g. structure, function, data handling/recording) that affect the accuracy of the information and suggest actions for both the insurance and the clinic company to increase the usability in research while maintaining the primary use.

5.1 Data quality

Ideally each secondary database should be continuously monitored for accuracy, considering both its primary and secondary use. However, in-depth validation is costly and time-consuming and therefore a more realistic recommendation is that when conducting studies in secondary data, it is better to include some data quality assessment as part of the research study than not done at all.

In general, the validation studies performed in human and veterinary medicine lack conformity as there is no generally accepted gold standard on how to evaluate data quality. This is, no doubt, related to the databases being very different and any evaluation of data quality needs to consider both the unique features of the database and what aspects of the data quality and/or database are most relevant to investigate and evaluate. In fact, lack of established standards for measuring data quality has been suggested as one of the difficulties in assessing the usefulness of medical data for research (Canner *et al.*, 1983). As for any secondary database, neither insurance data nor clinic data can be expected to have completely accurate recording of

diagnoses, but fortunately neither need to be perfect to still be useful in research studies.

In the field of epidemiology, the terms internal validity and external validity are often used to describe the ability to make unbiased inferences of study results to other populations than the study population. Internal validity relates to the ability to make unbiased inferences to the target population from which the study population was selected, whereas the external validity relates to the ability to make unbiased inferences to populations beyond the target population (Dohoo *et al.*, 2003a). Regarding data quality evaluation, internal validity has been defined as whether the code recorded in the database is correct in terms of reflecting the problem perceived by the clinician (Jordan *et al.*, 2004). In both validation studies in this thesis (I, III), the assessment of internal validity for diagnostic information was based on the available information in the clinical records. Further, external validity has been defined as whether the patient truly had the problem stated by the diagnostic code (Jordan *et al.*, 2004), which was not evaluated in this thesis.

When evaluating data accuracy, estimates of completeness and correctness are complementary – they are both useful and necessary for a comprehensive understanding of accuracy in a system (Hogan & Wagner, 1997). High completeness (similar to sensitivity) and high correctness (similar to positive predictive value) are both necessary for a secondary database to accurately reflect the disease situation in the population. Importantly however, estimations of correctness/agreement will depend upon the definition criteria applied in the assessment (for example exactly the same versus almost the same information). This makes thorough scrutinizing of the study methods essential in order to assess the value of the correctness/agreement estimates. Similar aspects are relevant also when evaluating completeness as there might often be rather subjective criteria applied in determining the number of disease problems present in different sources. Accuracy of diagnostic information in medical databases is often estimated by comparing computerized information against a designated gold standard that could be paper notes (as in study I and III) or diagnostic test and procedures. However, a limitation of this reasoning is that the paper files are likely not totally accurate in specifying the true condition of the patient (i.e. the gold standard), although veterinary diagnosis and CRs may be the best to hope for when using secondary medical data. Moreover, it should be noted that quality of data is not an indicator of the actual quality of treatment and care of the patient, which often is the primary reason for collecting the medical data. Evaluations of data can mainly address how accurate data are but sometimes also estimate to what extent the diagnoses

or procedures are optimal in each case. Further, it has been proposed that data quality in a disease database relies on the return of valuable information back to the clinic/user (Bartlett *et al.*, 1986). Although it has been suggested that transcription and data entry does not have to be perfect in the typical medical setting (Dambro & Weiss, 1988), striving to reduce errors and improve data quality will increase both the primary and secondary use of a medical database.

In the validation of the insurance database (study I), the agreement was concluded to be good (84%) for individual diagnosis and excellent for some investigated systems (>90%). The agreement for diagnosis was similar to that seen when the Agria dog insurance database was validated (Egenvall *et al.*, 1998), which was expected as the routines for recording the information by the clerks and the criteria used for assessment in the two studies were similar. However, when the dog insurance database was validated regarding a canine disease (atopic dermatitis), the correctness varied from 41% to 84% depending on whether a conservative or liberal, respectively, classification was used (Nodtvedt *et al.*, 2006a). An estimation of the completeness of insurance data can be done by contrasting the number of recorded diagnoses in the CID that were correct (including categories agreement and minor disagreement) (n=486) to the total number of problems estimated in the CR (n=740); producing an estimated completeness of 66%. However, it should be noted that this rather low estimate reflects the limitation of the CID to only record one diagnosis for a claim (further discussed in section 5.3). In the univariate analysis for diagnostic information in study I, clinic visits had significantly higher agreement compared to field visits, suggesting more accurate record-keeping at clinics. This could be due to better record-keeping facilities/routines at clinics but maybe reflect different interest in and attitude towards adequate and correct record keeping between veterinarians working in the field versus at clinics. Computerized records may be assumed to have more accurate data (Egenvall *et al.*, 1998), but on the contrary, for diagnostic and system information there was a non-significant tendency of better agreement for manual records compared to computerized records. The agreement on demographic information was excellent and overall higher than seen in the validation of the dog insurance database (Egenvall *et al.*, 1998) although the amount of missing information was larger in study I. The differences seen between the validation studies of the equine and canine insurance data might be due to different working routines for clinicians and clinics depending on species, but could also be related to different study designs (e.g. number of records evaluated or cluster effect of clinics). Computerized CRs are generally accepted to result in

excellent data quality for specific demographic information (Pollari *et al.*, 1996a). However, when comparing computerized to manual records in study I higher agreement on demographic information for computerized records was seen only for name whereas for year of birth it was similar and even less for breed and gender. Further, the non-significant effect of clerk in the logistic regression analysis indicated that the insurance clerks processed the claims similarly, albeit not error-free.

Errors identified in study I were (based on the available information) misclassification bias (i.e. incorrect classification of study subjects due to incorrect information) and transcription error (both classified as major disagreement) and constituted 11% of the claims. An example of misclassification bias is when the type of problem (e.g. joint inflammation) was right but the localization incorrect (e.g. fetlock instead of carpal). Transcription error was seen for example when a diagnosis seemingly unrelated to the condition described by the CR was recorded in the CID. This error estimate in study I was lower than what was seen when physician and coding errors were investigated in patient records (22%) by comparing the medical records to the hospital abstract (Lloyd & Rissing, 1985). However, the study methods differed from study I, for example by investigating failures of the physician such as reporting a procedure or a diagnosis, whereas in study I the CR was considered accurate regarding the investigated (i.e. diagnostic) information. Lack of transfer and miscoding of information have been reported as the most frequent reasons for discrepancies in computerized medical records at a veterinary teaching hospital (Pollari *et al.*, 1996a). Dambro and Weiss (1988) found that errors in data interpretation and entry were attributable to poor handwriting, An estimation of the impact of errors on measures of association has been presented for human medical records (Mullooly, 1990). If the assumptions in human medical records hold true for veterinary medical records, an observed error rate of 0.1-1% of the data would weaken the measure of association with less than 10%. In human medicine, the transferring of data from the CR to the database is becoming more automatic for example by using audiphones that can record and transfer spoken information into the database; methods that likely will or have already been applied to veterinary medicine settings. Some researchers have highlighted the advantages of computerized medical records. For example, Hogan and Wagner (1997) suggested that in theory, computerized medical records should exceed paper-based/hand-written records regarding data quality as the computerized records offer some distinct advantages such as validity checks during data entry, the possibility to improve data by editing rather than

rewriting and standardized transmission and merging of data from different locations into one record. Other advantages of completely computerized medical records could be pre-recorded clinical assessment schemes created for each specific diagnostic code and automatically appearing when choosing a hypothesis/preliminary diagnosis. This could support the clinical work not only by acting as an aid in remembering to perform the relevant assessments but to ensure complete recording of the relevant assessments made. With the increased use of computers in everyday life, future veterinarians will likely be more obliged to use completely computerized systems as they are used to the facilities of the computerized technology.

Other researchers have reported varying quality of data depending on type of disease. For example, it has been suggested that insurance data on diabetes mellitus in dogs has a higher diagnostic validity compared to other diseases due to the relatively clear clinical presentation and readily interpreted clinical biochemistry findings (Fall *et al.*, 2007). In study I, taken together the observed sensitivity, specificity (defined as the proportion of disease negative in the clinical record are so-recorded in the database) and positive predictive value, the data quality for system digestive is rather good. This was considered when the demographics and costs of colic in horses was investigated using the CID (Egenvall *et al.*, 2008b). However, the other evaluated systems had rather low sensitivity, meaning that diseases in those systems will be underestimated if disease prevalence is calculated using the CID. In research studies based on CID, it is recommended to include a horse's entire insurance history which would preserve the correctness while increasing the completeness. This suggestion is supported by the findings that the sensitivity for systems joints, digestive, skeletal and hooves was significantly lower when there was more than one diagnosis listed in the CR compared to one problem in CR. Although this suggestion might increase the size of the analysed database extensively, this should not be a problem with modern computers and software programs. Further, this approach would include more comprehensive information on a disease problem such as cost, recovery and relapse, in particular for chronic/long-term illnesses. In addition, if the complete CR for the selected cases (not only the record for a specific visit) were retrieved it might be that false positives would be even fewer (meaning that the individual actually had the problem), which was observed when cases of canine atopic dermatitis in CID were validated and 98% reclassified as having some allergic skin disease (Nodtvedt *et al.*, 2006a).

In the validation of the clinic database (study III), both the completeness and correctness of the diagnostic information was deemed to have high

values. This means that most cases in the CCR actually had the problem (i.e. internal validity) and overall, only 9% of cases would be missed by the CCR, although varying with type of problem as shown in the system assessment. Reasons for lack of completeness could be that only a procedure (which is necessary for the invoice) is recorded and not the diagnosis. One can speculate that this could be more common to happen for less severe disease problems. In fact, most cases of lack of completeness were due to under-reporting of co-morbidities and the findings suggested that less severe and non-clinically important problems were less often/frequently recorded in the CCR. This was supported by that on several occasions lack of completeness was related to the procedure teeth-floating being recorded without a related diagnosis. A similar level of under-reporting of co-morbidities has been reported in human medical computerized record systems (Brouwer *et al.*, 1995). In conclusion, major reasons for visiting an equine clinic are well captured which supports use of this database in research studies on those types of problems.

Not only co-morbidities can be under-reported in medical databases. For example, it has been described that under-reporting of post-operative complications (POC) of elective surgeries in dogs and cats were substantial when comparing the computerized abstract to the complete paper medical record; the proportion POC was four to seven times higher in the latter (Pollari *et al.*, 1996b). On the other hand, in study III, there were nine claims with more diagnoses noted in the CCR than in the VCR which is likely because only one specific visit was investigated and not the horse's entire clinic history (i.e. these extra diagnoses were likely assigned at other visits). Compared to another validation study performed in Sweden, the completeness was higher than when disease information in a dairy disease database was compared to farmers' disease records in Sweden (71%), and it also varied between systems (Mörk *et al.*, 2009). Further, in study III, the correctness of diagnostic information was higher than seen in the validation of the insurance canine database (Egenvall *et al.*, 1998). Correctness of computerized information could be affected by, for example, personal enthusiasm, personal preference for certain codes and useful tools such as well-designed software programs and user-friendly diagnostic coding systems. Most of these factors also affect the completeness.

The two investigated databases investigated in this thesis have different recording situations. At the ATG-clinics the veterinarians are encouraged to write the diagnostic codes in the VCR and the coder (i.e. the veterinarian) and the person entering the data work more closely together. These are probably the main reasons that the transcription process may yield relatively

few mistakes. This can be compared to the insurance data, where clerks may code disease visits from receipts with scarce information, without any direct contact with the attending veterinarian. However, as discussed elsewhere (section 5.3), for some processed claims the clerk had personal contact with the treating veterinarian/clinic which would likely affect the data accuracy. By combining findings on the correctness and completeness it is suggested that, for the ATG clinic database, problems in body system joints will be quite accurately captured by searches from the CCR. Future research on joint diseases is further supported by the large number of joint problems examined and treated at the clinics, in all types of horses. If research based on the CCR is planned for specific conditions (other than joints) some validation should be considered.

Only a few studies have suggested actions to improve data quality (Wagner & Hogan, 1996), for example by improving the paper record (Mann & Williams, 2003). Computer-based clinical record systems must be well-designed to make the entry of information as easy as possible and the introduction of such a system, and maintenance, must include continuous user education and support. Internal quality of the records has been reported to be affected by the user-friendliness of the recording system (Menendez *et al.*, 2008). Further, it has been suggested that turn-over in personnel may result in changes in procedures which may negatively affect the quality of the recorded data (Hogan & Wagner, 1997). Quality of clinic data would likely improve if the veterinary CR and transcription of diagnostic information by the technician was replaced by a completely computerized medical record system. Based on the results in study III, it is highly recommended that, if a completely computerized system cannot be implemented, the veterinarian explicitly states the diagnostic codes in the VCR to increase data completeness and correctness.

5.2 Disease measurements

The overall morbidity for horses with complete insurance at Agria has been estimated to 1,137 VCE per 10,000 HYAR with geldings having more VCE than mares and stallions (1,398, 1,042 and 780, respectively) (Egenvall *et al.*, 2005d). In study II, IRs of different types of disease were presented, mainly focused on system-level, although the IRs for some individual diagnoses were also calculated. Few other studies have reported on IRs of disease in horse populations. One example is the analysis of morbidity and mortality data in riding school horses based on horses with complete veterinary care from the same insurance database as used in study II but investigated for the

period 1997 to 2002 (Egenvall *et al.*, 2009). The average IR for any VCE in riding school horses was higher compared to all horses with complete insurance (1,584 versus 1,137 VCEs per 10,000 HYAR) but the differences between the usage of insurance among riding schools were large (Egenvall *et al.*, 2009). This suggested that some riding schools seemed able to prevent serious disease and reduce costly VCEs.

The findings in both studies II and IV confirm that lameness is a prominent problem affecting all types of horses in Sweden. In study II the lameness estimate of 628 per 10,000 HYAR corresponded to approximately 55% (628/1,137) of the disease problems and in study IV the PM for lameness at problem visits was 47% (22,553/48,411) of the recorded diagnoses. These findings strongly suggest that future research should be directed towards different types of studies on lameness-related problems. Other studies have reported varying incidence or PM of lameness. For example, a USA study reported a much higher incidence density for lameness (approximately 1,500 events per 10,000 HYAR when combining disease categories lame: leg problems and lame: hoof problems) (Kaneene *et al.*, 1997a) which was probably related to the events being operation-reported and not necessarily veterinary treated cases, and that data were collected prospectively with opportunities for data collectors to stress correct and timely recording of disease events. On the contrary, an Australian study found a lower PM of lameness of approximately 33% when combining foot lameness, other lameness and laminitis for owner-reported events (Cole *et al.*, 2005). This could be explained by differences in the study population such as a large proportion breeding animals and young ("unbroken") as well as retired horses, compared to studies II and IV. Lameness in racing thoroughbreds has also been addressed. The incidence of lameness-related injury and disease in young thoroughbreds in training was strikingly higher, 3,828 events of first occurrence per 10,000 HYAR (when combining shin soreness, fetlock problems, foot problems and carpal problems and converting the estimate in horse-months to HYAR) (Bailey *et al.*, 1999) in a prospective, trainer-reported study. The findings likely reflects different challenges for young racing horses compared to leisure/sport horses (study II), different inclusion criteria (i.e. levels of capturing the problems: trainer versus veterinary-reported) and maybe biased selection/non-representativeness by including a non-random sample of thoroughbred trainers. A three fold higher incidence of musculoskeletal injuries has been found in 2-year-old compared to 3-year-old racing thoroughbreds based on prospective trainer-recorded data (Cogger *et al.*, 2008), pointing out the risk for locomotor problems in this age-group.

For disease problems other than lameness, studies have presented quite different ranking of disease problems. For examples, compared to the findings in study II and IV, eye problems, behavior problems, tying up and weight loss were much more commonly seen in Australia (Cole *et al.*, 2005) whereas in the USA a large proportion of cases were dermatological problems (Kaneene *et al.*, 1997a). Also, when veterinarians in the USA ranked the most common medical problems in adult horses, colic was the most common problem followed by viral respiratory disease, endometriosis, dermatitis and parasitism (Traub-Dargatz *et al.*, 1991). The differences in disease occurrence (incidence/prevalence) and disease patterns between regions have many plausible explanations; for example breed composition of the horse population, usage of the horse, climate/environmental conditions, veterinary expertise and categorization of disease. These variations confirm the need for conducting these types of studies on the populations of interest. Importantly, clear definitions of study populations increase the possibility for comparisons to other populations. Even if results have been categorized differently, comparison between studies will still be possible as long as there is transparency in study methods and application of different research criteria.

In study IV, it was possible to determine the proportion of health visits of all visits to the ATG clinics. This type of veterinary interaction with horses has rarely been addressed. The proportion of horses with health visits (7,246/21,449=34%) was considerably lower than the reported proportion of routine or elective procedures (64%) reported in a study in northern Britain (Mellor *et al.*, 2001). However, that study only included the most recent veterinary visit and did not state whether the horses were examined and treated in field practice where likely a vast proportion of all health interventions are performed or at a clinic. Several procedures were more frequent in the northern Britain study; vaccination constituted the majority of procedures (72%), but also pregnancy checkups (6%) and teeth-floating (14%) were common. In study IV, vaccination also constituted the vast majority of procedures recorded at health visits (41%). On the contrary, a low proportion of pregnancy check-ups and teeth floating were seen, the latter, as discussed elsewhere, due to common under-reporting of this procedure in the clinic database. The clinic database investigated in this thesis offers an opportunity to study the implications of different health recommendations and to capture times trends in this area.

5.3 Methodological considerations

All studies in this thesis (I-IV) are based on pre-recorded data. Obviously, the usefulness of secondary data for research purposes is dependent on access to the information

In both study II and IV, the databases were "cleaned" and some observations removed; rather few in study IV but 15% in study II. This procedure must be done with great caution and traceability/transparency. For dairy herds, it has been suggested that excluding records that are incomplete might lead to selection bias (Bartlett *et al.*, 1986). For example, if low disease rates and good record-keeping are strongly correlated excluding incomplete records might lead to underestimation of disease. However, incomplete records can be removed from analyses if the reasons for exclusions are relatively unrelated to the variables being studied (Bartlett *et al.*, 1986), which was believed to be the case in the studies in this thesis.

For practical reasons, the principle investigator classified the agreement/correctness of the investigated factors in the CID/CCR in both validation studies. This approach might have influenced the classification, especially since the assessment was based on rather subjective criteria. However, one could argue that assessment by one person probably will reduce inter-rater variability. In fact, all claims were assessed and re-assessed in concentrated periods of time to ensure equal interpretation and classification of the records.

In study I the term agreement was used for assessing the concordance of information in the CID and CR. In study III, agreement was replaced by the term correctness when evaluating how well the CCR reflected the information in the VCR. This was due to a preference of the principal investigator of the latter term to adjust to the terminology used in many similar studies in human and veterinary medicine. Also, the use of the term correctness implies that the computerized information was contrasted against a designated gold standard (i.e. the clinical record). The findings in the validation studies (I and III) showed that the agreement/correctness (and in study III also the completeness) was higher for system compared to individual diagnoses. Although aggregating data on a higher level often will improve the accuracy of recorded data, the information could be too general for many research and clinical purposes, and also cause loss of information given that the individual diagnosis was correct.

In study I, after the initial analyses, neither the CID nor the CR was considered to completely reflect the truth (i.e. considered gold standard) whereas in study III the VCR was the designated gold standard. This adjustment was due to indications that the insurance data occasionally was

more accurate and complete because of personal contact made between the processing clerk and the treating veterinarian/clinic, hence modifying the available information. However, for the clinic data, the impression was that the reported way of registering the information in the database made it less likely that any information was added or modified compared to the VCR.

The target population in study II was horses carrying insurance for veterinary care with similar terms and conditions in Sweden, assuming no major differences between the studied insurance company and other horse insurance providers in Sweden. The selection of this study population was based on practical and convenient considerations. The readiness of Agria to support research in horses (and other species) and the size of the insurance database suggested usefulness of the insurance data for research purposes. In general, comparisons between different countries on insurance claim-based measures of disease frequencies will be challenging as the differences in the insured horse population, the terms and conditions and, consequently, the recorded disease events will greatly affect the measurements of disease occurrence, which is also discussed in a review on animal insurance data in research (Egenvall *et al.*). Extrapolations to non-insured horse populations should be done with care, in particular if detailed characteristics of each population are unknown. In addition, there might be concerns that insured animals can be a certain socioeconomic class of horses and therefore may not be representative of the general population. However, with the high penetration of insurance in the Swedish horse population this is less of a concern in this region.

As previously mentioned, a major limitation of the insurance database (study I, II) is that only one main diagnosis was recorded per receipt in each claim. This limitation makes it impossible to estimate the completeness of the diagnostic information and will in most instances underestimate the prevalence of disease in the CID. Consequently, estimates on disease frequency based on CID will always be conservative; in particular when counting only the first event of a certain disease per year as I study II.

Measurements of disease in populations are always highly dependent on study methods and material. In particular, major concerns relate to whether the found cases are typical of all cases, and if not, how useful results are in other scenarios/geographic regions. Another limitation of CID is that only disease events where the cost for veterinary care exceeds the deductible will be recorded in the CID. However, due to the claim period of 100 days within which the cost for disease events will be summarized, it is believed that many minor/low cost events will also be captured in the CID.

In general, measures of disease frequency will be much affected by the level at which disease is being noted, for example owner-reported events versus veterinary reported. The most detailed level would be if the early signs in the body such as physiological changes related to disease could be registered. The next, less detailed level would be to record all events of ill-health experienced by the horse. Further, the owner may detect the signs of disease and then make the decision whether to present the horse to a veterinarian. Factors relating to detecting signs of disease in a particular horse and the readiness to seek veterinary advice will depend on knowledge, experience and interest, personality and financial situation of the owner and thus affect the owner-reported frequency of disease. Therefore it is always important to establish the level of disease-recording when comparing results from different studies and consider many measures of disease occurrence, including those from insurance and clinic data, as conservative (i.e. that more disease events occurs than are captured). Owner-related characteristics have been suggested to affect under-coverage of disease in cattle and swine which may limit the usability of disease recording systems for the purpose of disease monitoring and research (Olson *et al*, 2001). Examples of reasons for under-coverage may be unwillingness of the owner to seek veterinary advice in case of disease, or failure of the veterinarian to keep a clinical record (Olson *et al*, 2001). Further, unwillingness to seek veterinary advice may be related to that an owner judge that a veterinary visit does not add to the prognosis or welfare of the animal. Further, the probability for an event to be captured by CID or CCR depends on the severity of the problem. Many cases of specific, usually severe problems are likely to be presented to a veterinarian, whereas the capture of less malignant disease events may be more affected by insurance-, owner, horse-related factors as well as prognostic and cost considerations about pursuit of veterinary care. Also, currently (2009) the owner still needs to pay the deductible and a percentage of the remaining costs when claiming reimbursement for veterinary care costs. With the rising cost for veterinary care and the current financial concerns; for example, owners may be less motivated to present animals or accept expensive treatments. It has also been shown that the cost per claimed horse has increased 59% from 1997 to 2004 compared to a consumer price index increase of 10% (Egenvall *et al.*, 2008a).

A high percentage of cases with the diagnosis "without diagnosis" were seen in particular in studies III and IV. The reason for the frequent recording of this diagnosis is unknown but may be due to varying interest in veterinarians and staff to record more detailed information when it is available. This was supported by that some clinics in study III had a very

high proportion of this diagnosis recorded. This could be related to individual clinic/veterinarian factors such as reluctance to use the diagnostic registry or record information accurately, or a lack of understanding of (or disagreeing with) the benefits of correctly recorded data rather than that the cases seen at those clinics were more difficult to assess compared to those seen at other clinics. Also, some of the cases in this category may have been classified as without diagnosis due to no suitable diagnosis in the diagnostic registry. In fact, it was seen that when the categorization of causes of death in foals was changed to include two new specific categories, it decreased the number of deaths reported as unknown (USDA:APHIS:VS, 2007). The diagnosis "without diagnosis" might be needed as it reflects the reality of veterinary clinical work (e.g. the difficulty to assign a preliminary, more informative diagnosis) but could also reflect a lack of guidelines on which diagnosis to record for certain procedures or situations. However, the impression from the findings in this project is that the use of this diagnosis should be more restricted and careful.

Lack of a standard diagnostic classification system has been proposed as a limitation of using veterinary medical records for epidemiologic research (Brouwer *et al.*, 1995). However, it has also been suggested that formal coding systems might interfere with rather than improve the usefulness of computerized medical record systems in private practice (Williams & Ward, 1989; Udomprasert & Williamson, 1990). A great advantage with the two studied databases is that both use the diagnostic registry. A common coding system enhances the evaluation of accuracy of electronic records and facilitates comparison of diagnostic information between different databases. Even though the diagnostic registry is valuable it may raise some concerns apart from the mentioned high number of diagnosis included but still not constituting a complete list of problems affecting horses (or other species for which it is also used). For example, some Swedish classifications might cause problems outside the country, such as the diagnosis bronchitis which internationally is more likely the disease "recurrent airway obstruction" (RAO) (personal communication John Pringle). Further, the vast number of diagnoses can make the registry difficult to easily over-view whereas some clinicians might argue that it is not detailed enough (i.e. specific problems missing). This results in insecurity regarding the correctness of information on a detailed level but suggests that information should be analyzed on a more general level. If the diagnostic registry was to be updated regularly with sincere efforts to consult different types of clinicians, the registry might be satisfying for more users, even though it is likely that some veterinarians still would be unwilling or unable to commit. Based on the findings from

the validation studies (I and III) the level of detail concerning correctness of diagnostic information is important. Even if the diagnostic registry fails to include all diagnoses, the recommendation is that (assuming the veterinarian has a diagnosis) the correct diagnosis could always be noted in the veterinary notes (for partly computerized records) or in a text field in the database and a higher-level/more general diagnostic code recorded in the diagnosis window, which is not incorrect but slightly less informative. This would still readily facilitate identifying the cases of the specific disease by scrutinizing all records with the more general code recorded.

The use of secondary data for research purposes imposes several considerations. A major limitation of secondary data is access and cost related to access and analyses. The owner/-s of the databases must be willing to share their information. As epidemiological studies rarely focus on the individual, issues regarding privacy and confidentiality can be avoided, for example by not performing studies on very rare conditions or in very particular (and few) individuals. Lack of control over collection and quality of data are inherent disadvantages of using routinely collected data for epidemiological research (Mulder *et al.*, 1994), although this problem is not necessarily eliminated in prospective data collection scenarios. Using retrospective data eliminates the possibility for researcher to standardize the collection of data and the routines for assessing cases and non-cases (i.e. controls). The information might be nonspecific or not detailed enough to be useful, or be missing. When identifying cases from existing databases, it is important to consider for each source both the type of cases included and those that are not included, in particular if the study periods include different time periods (i.e. years). Cases can also have received different assessment due to differences in clinic/hospital routines, but also due to changes over time with improved diagnostic tools and increased knowledge on treatment and prognosis.

5.4 Insurance data versus clinic data in research

Both investigated databases offer distinct advantages; the insurance database includes a base population and can be used to calculate IRs with the exact time at risk and the clinic database includes information on non-referral clinic cases with a greater detail, i.e. "all" diagnoses and procedural information. Further, both have standardized coding schemes used to record the diagnostic information in the respective database (as discussed in section 5.3). It has been said that if the structural problems are overcome then medical databases are an invaluable source of data on epidemiological studies

(Lawrenson *et al.*, 1999). Data in both sources have been found possible to convert to a form suitable for analysis and with due considerations of improving data quality, there are few structural problems with the two investigated databases. Future research will be highly dependent on the willingness of Agria and ATG Equine Clinics to share their data. To increase the usefulness of the two databases for both primary and secondary use it is recommended that Agria allows multiple recordings of diagnoses per claim and that ATG Equine Clinics develop standards for example for recording group events, selling of medicine and material, prophylaxis events and recording of diagnoses at follow up visit. Further, to reduce the risk of re-entering a leisure/sport horse in the database with a new identification number (if the previous identity cannot be found), the individual identity of leisure/sport horses should be established more thoroughly for example by including micro chip number in the demographic information.

Insurance data can highlight common, severe and expensive conditions in horses. The tradition in Sweden to insure horses (and other animals) has had a positive effect on the development of both medicine and surgery in horses as more cases will be available to fully examine and treat when the owner does not need to cover the full cost. The clinic database can in turn be used to monitor changes in disease patterns, health routines and treatment procedures over time in different horse categories in Sweden, given that a specific type of cases has the same probability of being admitted through time. The procedure information will be accurate and complete as this is the basis for the invoice, which allows further investigations into examinations and treatments of certain diseases, for example observed results for different treatments, recovery time related to case type (severity, treatment etc). Further, the joint management of the clinics and the similarities in education background and clinical training/experience for the veterinarians is presumed to enhance the conformity in assessing the cases and assigning diagnoses. Conformity in assessing diagnosis would increase the usefulness of this clinic database by more accurately and precisely reflect the health problems of the client. This is important both for clinic/company management and to assist the clinician's work by providing relevant information. Moreover, information in the clinic database can support an evidence based medicine approach to the clinical work, for example by conducting quality control of the clinical work and evaluating treatments by investigating information on follow up visits such as the overall frequency of follow up visits, the type of problems that come back, treatment results etc, in total and per clinic to identify each clinic's situation (e.g. type of cases/horses treated, database recording routines, skills/training of the

clinicians, available clinical equipment). In particular, scrutinizing follow up visits which are mostly recorded for joint and lameness problems will increase knowledge on treatment and return to previous performance.

Findings from both study II and IV indicate further studies of joint disease/lameness problems based on both data sources as lameness problems are often investigated at equine clinics and costly enough to exceed the deductible. The clinic database would also be a useful tool in in-depth studies on lameness problems in active racing horses (mainly trotters due to the large proportion of this type of race horses in Sweden) on how different treatments for specific problems affect the recovery and return to racing. Fetlock problems and carpal problems are well suited for analysis using combined Agria and ATG data due to the large total number of cases and the importance of this type of diseases (i.e. lameness) in horse's welfare and performance. As well, for racing horses, the differences in the turf between race tracks might be investigated based on clinic data from clinics situated at or in close proximity to race tracks. Furthermore, the difference in the two databases may be complementary in that the overall incidence can be estimated based on insurance data whereas details on treatment, cost, and rehabilitation (e.g. based on information at follow up visits) could come from the clinic data. Due to issues regarding confidentiality and ownership of information, joint research based on the two databases is likely best performed by a neutral third party. It has been highlighted that if aggregating morbidity data from two data sources is to be done, two particular issues must be addressed before data can be relied upon; use of multiple search strategies and standardized disease inclusion and exclusion criteria (Pringle *et al.*, 1995). For longitudinal studies on CID or CCR, researchers will need to consider that factors relating to disease frequency will be affected by changes in diagnostic criteria over time due to improved diagnostic techniques and also by changes in prevalence/incidence of certain problems. In addition, changes in insurance conditions affecting CID and the cases seen at the ATG Equine Clinics Ltd needs consideration, as previously mentioned.

The findings in the two validation studies (I and III) support the use of these two sources in research studies on horses in Sweden with due consideration to the disease being investigated and the characteristic of the database used. For Agria equine insurance database, the findings suggest that this database is particularly suited to calculate incidence of disease, both overall and for specific problems. For the ATG clinic database, the findings suggest that the database is well suited to investigate health procedures and

the distribution of disease problems (both overall and for specific problems) seen at equine day-clinics in Sweden.

6 Conclusions

- For the insurance database, the correctness of the diagnostic and demographic information was validated and considered adequate for research studies based on the CID. It was concluded that the database can be potentially useful in equine research studies. In particular, the database is useful for calculations of disease incidence but also specific diseases/problems can be investigated, with the caveat that the data quality varies among different disease problems and systems affected.
- Incidence rates of disease and proportional morbidities were presented based on insurance data for horses with complete insurance for veterinary care. Calculations of disease occurrence based on the CID are affected by that the disease problem of interest must be covered by insurance and the cost of the veterinary care at such a disease event must exceed the deductible.
- For the clinic database, the correctness and completeness of the diagnostic information was validated and considered adequate for research studies based on the CCR. It was concluded that the database can be potentially useful in equine research studies. The types of problems treated at these clinics and the varying degree of completeness and correctness for diagnostic information should be considered when using the database for research purposes. The database is well suited to investigate disease events that are predominantly examined and treated at equine day-clinics.
- The horses and types of problems seen at visits to the equine clinics during one year were described and disease occurrence presented as proportional morbidities. Also, the proportion of health visits was determined.

The overall conclusion was that the two investigated equine databases can be used to estimate the amount and describe the type of disease in the Swedish horse population, with due consideration of the characteristics of the two databases.

7 Future research

Both investigated databases can be used to monitor trends and changes in the disease pattern longitudinally for horses in Sweden. Obviously, to perform meaningful time-trend analyses of the disease load in the insured population, the insurance policies need to remain fairly similar. For ATG equine clinic data, changes in the general veterinary access (more or less competing veterinarians/clinics) as well as changes in the ATG Equine Clinics profile may affect the type of cases seen at the clinics.

Insurance data can be used to investigate incidence of diseases that are likely captured by the CID. For diseases where the factors causing disease and the type of cases seen are believed not to differ between insured and non-insured horses, results might be extrapolated to the general leisure/sport horse population in Sweden.

The completeness of the insurance database may be evaluated by combining information on all horses with complete veterinary insurance that visited an ATG clinic during a specific time period. This would increase the knowledge on the insurance database and presents the opportunity to adjust disease frequency measures.

The clinic data could be used to investigate health events further, such as compliance to current health recommendations. Further, changes to health recommendations might be initiated based on study of the horses that regularly make health visits to the ATG Equine Clinics.

For the ATG clinic database, procedure information can be used to investigate if more costly examination procedures and techniques and treatments for certain diseases affect the likelihood of recovery and/or reduce recovery time. For racing horses, the clinic database can be combined with information on the Swedish race horse base population and racing information to conduct studies on for example the association

between performance and the risk for disease/injury, and investigating the time to racing after certain disease/injury events.

Both databases are particularly useful for studying lameness problems as many events will be captured in both databases. For clinic data, the high number of horses with joint disease examined at the clinics, in combination with the detailed information on procedures, cost, and visit type (follow up/first visit) makes the clinic data well suited for in-depth studies on specific joint diseases.

8 Populärvetenskaplig sammanfattning

8.1 Bakgrund

Kunskapen om sjukdom i den svenska hästpopulationen är bristfällig. Sådan kunskap är nödvändig för att till exempel kunna prioritera forskning, förbättra hälsorekommendationer, föreslå livsstilsförändringar och, inom veterinärområdet föreslå förändringar i djurhållning och skötsel samt stödja avelsstrategier i syfte att reducera mängden och allvarlighetsgraden av sjukdom i populationen. Sekundära data (definierat som data insamlat för annat syfte än forskning) som inkluderar diagnosinformation erbjuder en möjlighet att studera sjukdom i populationer utan primär datainsamling, som är tidskrävande och kostsam.

I Sverige finns två unika databaser som inkluderar landstäckande sekundära data på sjukdomsfall hos häst; en försäkringsdatabas som ägs av försäkringsbolaget Agria och en klinikdatabas från ATG Hästklinikerna AB som inkluderar 25 hästkliniker (2009). Syftet med den här avhandlingen var att undersöka om dessa två databaser kan användas för att uppskatta omfattningen och typen av sjukdomsproblem i den svenska hästpopulationen. Mer specifika syften var att undersöka datakvaliteten i de båda databaserna och därefter använda dessa data för att presentera sjukdomsförekomst dels hos hästar med maximal veterinärvårdsförsäkring hos Agria och dels hästar som undersöktes och behandlades vid ATG Hästklinikerna AB under ett år.

8.2 Sammanfattning av studier och resultat

De två valideringsstudierna genomfördes på ett slumpmässigt urval av händelser av veterinärvård (försäkringsdata) respektive klinikbesök (klinikdata). Data för de utvalda händelserna/besöken i respektive databas

jämfördes sedan med information in journalkopior. För försäkringsdata utvärderades hur väl data i de båda källorna (databas och journal) överensstämde för diagnosinformation och demografisk information (namn, ras, kön och födelsedatum). För klinikdata utvärderades hur korrekt (att de sjukdomsproblem som var registrerade i databasen verkligen hände enligt journalen) den datoriserade diagnosinformationen var jämfört med informationen i journalerna samt hur fullständig (inkluderande alla sjukdomsproblem omnämnda i journalen) informationen i databasen var jämfört med diagnosinformationen i databasen.

Överensstämmelsen för den demografiska informationen i försäkringsdatabasen var >94%. För diagnosinformationen var överensstämmelsen 84% för specifik diagnos och 92% för var i kroppen problemet fanns (t ex leder, andningsorganen, matsmältningsorganen). Överensstämmelsen var högre om hästen blev behandlad på en klinik jämfört med i fält. För diagnosinformation i klinikdata var data fullständig till 91% och korrekt till 92% men varierade mellan kliniker från 69% till 100%. Vidare var data mer korrekt för förstabesök jämfört med återbesök; om en diagnoskod var skriven i journalen jämfört med ingen kod samt påverkades av i vilken del av kroppen problemet fanns.

Både försäkringsdata och klinikdata användes för att presentera sjukdomsförekomst. För försäkringsdata beräknades antalet nya fall av sjukdom (insidensrat) samt hur stor del av alla problem som var av en viss typ (proportionell sjuklighet, %). Insidensrat (IR) beräknades som antal fall av en viss sjukdom i täljaren och risktiden (uttryckt som 10,000 häst-år av risk) för sjukdom i nämnaren. Under dessa förhållanden kan IR ungefärligt omvandlas till procent genom att dividera IR med 100. Ledsjukdom var det vanligaste problemet (IR 404, 37%) följt av sjukdom i hela kroppen (IR 283, 26%), huden (IR 123, 11%) och matsmältningsorganen (IR 121, 11%). Exempel på hur dessa resultat kan tolkas är att ungefär 4% av hästarna får ledsjukdom per år, och av alla problem som hästar som blir sjuka drabbas av under ett år var 37% ledsjukdom. Den vanligaste specifika diagnosen var ”kotledsinflammation”. Fördelningen av sjukdom påverkades av flera faktorer. Halvblodshästar hade högst risk för sjukdom i de flesta delar av kroppen utom i huden där fullblodshästar hade högre risk. Valacker hade högre risk för sjukdom i lederna, hela kroppen, skelettet och andningsorganen jämfört med ston, men i övriga delar av kroppen (muskulatur, huden, hovar och matsmältningsorganen) var skillnaden mellan könen liten. Insidensraten för hälta undersöktes genom att kombinera hältrelaterade diagnoser i leder, skelett, hovar, muskulatur och hela kroppen, vilket gav IR 628 per 10,000 häst-risk år.

Av alla hästar som behandlades vid ATG Hästklinikerna under studieåret var ungefär hälften rid- och sällskapshästar och hälften trav- och galoppsporthästar (framför allt travsporthästar). För klinikdata var andelen hälsobesök (inkluderade t ex vaccination, hälsokontroll och besiktning) 22% av alla besök. För problembesöken bestämdes andelen problem i olika delar av kroppen. Vanligast var problem i leder (36%), följt av hela kroppen (28%), andningsorganen (7%), skelettet (7%) och matsmältningsorganen (6%). Den vanligaste diagnosen var ”utan diagnos”. Andelen hältrelaterade problem av alla problem som undersöktes och behandlades vid problembesök uppskattades genom att kombinera hältrelaterade diagnoser i olika delar av kroppen (på samma sätt som för försäkringsdata) till 47% av alla diagnoser. Vidare undersöktes för rid- och sällskapshästar och travsporthästar vilket ben som oftast drabbas av kotleds- repektive karpalleds- (framknä) inflammation. För kotledsinflammation hade rid- och sällskapshästar högre andel problem i höger framben och båda frambenen jämfört med travsporthästar, medan travsporthästar hade högre andel kotledsproblem i bakbenen (höger, vänster, båda bakbenen). För karpalledsinflammation hade rid- och sällskapshästar högre andel problem i höger och vänster fram om endast ett ben var drabbat medan travsporthästar hade högre proportion av båda framben drabbade.

8.3 Sammanfattning

Överensstämmelsen för diagnosinformationen i försäkringsdatabasen ansågs som tillräcklig för forskning baserad på dessa data. Försäkringsdata är särskilt lämpat för beräkning av insidens av sjukdom, även om datakvaliteten varierade mellan sjukdomsproblem och vilken del av kroppen som var drabbad. Insidensrat och fördelningen av sjukdom i olika delar av kroppen presenterades för försäkringsdata.

Diagnosinformationen i klinikdata ansågs tillräckligt korrekt och fullständig för forskning baserad på dessa data. Dock bör typen av problem som undersöks och behandlas vid dessa kliniker beaktas samt den varierande datakvaliteten mellan sjukdomsproblem. Databasen lämpar sig väl för att undersöka sjukdomar som framför allt behandlas vid den här typen av kliniker (öppna dagtid). Hästarna och besöken till hästklinikerna beskrevs och sjukdomsförekomst presenterades. Även andelen hälsobesök bestämdes. Sammanfattningsvis ansågs att båda databaserna kan användas för att uppskatta mängden och typen av sjukdom i den svenska hästpopulationen.

References

(NAHMS), National Animal Health Monitoring System. (1998). Part I: Baseline reference of 1998 equine health and management, N280.898. . In: United States Deartment of Agriculture Web site, http://nahms.aphis.usda.gov/equine/index.htm, accessed August 27 2009. SDA:APHIS:VS, CEAH (Ed.) Fort Collins, USA:

Bailey, C.J., Reid, S.W., Hodgson, D.R. & Rose, R.J. (1999). Impact of injuries and disease on a cohort of two- and three-year-old thoroughbreds in training. *Veterinary Record* 145(17), 487-93.

Bartlett, P.C., Kaneene, J.B., Kirk, J.H., Wilke, M.A. & Martenuik, J.V. (1986). Development of a Computerized Dairy-Herd Health Database for Epidemiologic Research. *Preventive Veterinary Medicine* 4(1), 3-14.

Boden, L.A., Anderson, G.A., Charles, J.A., Morgan, K.L., Morton, J.M., Parkin, T.D., Slocombe, R.F. & Clarke, A.F. (2006). Risk of fatality and causes of death of Thoroughbred horses associated with racing in Victoria, Australia: 1989-2004. *Equine Veterinary Journal* 38(4), 312-8.

Bonnett, B.N., Egenvall, A., Hedhammar, A. & Olson, P. (2005). Mortality in over 350,000 insured Swedish dogs from 1995-2000: I. Breed-, gender-, age- and cause-specific rates. *Acta Veterinaria Scandinavica* 46(3), 105-20.

Brouwer, H., Schouten, E.G., Noordhuizen, J.P.T.M. & Vanvoorthuysen, P.F. (1995). Potential of Computerized Medical Records for Epidemiologic Research. *Tijdschrift Voor Diergeneeskunde* 120(10), 296-299.

Bruun, J., Ersboll, A.K. & Alban, L. (2002). Risk factors for metritis in Danish dairy cows. *Preventive Veterinary Medicine* 54(2), 179-90.

Canner, P.L., Krol, W.F. & Forman, S.A. (1983). External Quality-Control Programs. *Controlled Clinical Trials* 4(4), 441-466.

Cogger, N., Evans, D.L., Hodgson, D.R., Reid, S.W. & Perkins, N. (2008). Incidence rate of musculoskeletal injuries and determinants of time to recovery in young Australian Thoroughbred racehorses. *Australian Veterinary Journal* 86(12), 473-480.

Cole, F.L., Hodgson, D.R., Reid, S.W. & Mellor, D.J. (2005). Owner-reported equine health disorders: results of an Australia-wide postal survey. *Australian Veterinary Journal* 83(8), 490-5.

Couetil, L.L. & Ward, M.P. (2003). Analysis of risk factors for recurrent airway obstruction in North American horses: 1,444 cases (1990-1999). *Journal of the American Veterinary Medical Association* 223(11), 1645-50.

Dambro, M.R. & Weiss, B.D. (1988). Assessing the quality of data entry in a computerized medical records system. *J Med Syst* 12(3), 181-7.

Dohoo, I., Martin, W. & Stryhn, H. (2003a). Validity in observational studies. In: *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC inc. pp. 207-232.

Dohoo, I., Martin, W. & Stryhn, H. (2003b). Introduction to observational studies. In: *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC inc.

Dohoo, I., Martin, W. & Stryhn, H. (2003c). Mixed models for discrete data. In: *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC Inc. pp. 473-498.

Egenvall, A., Bonnett, B.N., Hedhammar, A. & Olson, P. (2005a). Mortality in over 350,000 insured Swedish dogs from 1995-2000: II. Breed-specific age and survival patterns and relative risk for causes of death. *Acta Veterinaria Scandinavica* 46(3), 121-36.

Egenvall, A., Bonnett, B.N., Larsdotter, S. & Emanuelson, U. (2008a). Cost of veterinary care in insured Swedish horses 1997-2004 In: *Proceedings of Society for Veterinary Epidemiology and Preventive Medicine* Liverpool, UK pp. 27-42.

Egenvall, A., Bonnett, B.N., Ohagen, P., Olson, P., Hedhammar, A. & von Euler, H. (2005b). Incidence of and survival after mammary tumors in a population of over 80,000 insured female dogs in Sweden from 1995 to 2002. *Preventive Veterinary Medicine* 69(1-2), 109-27.

Egenvall, A., Bonnett, B.N., Olson, P. & Hedhammar, A. (1998). Validation of computerized Swedish dog and cat insurance data against veterinary practice records. *Preventive Veterinary Medicine* 36(1), 51-65.

Egenvall, A., Hagman, R., Bonnett, B.N., Hedhammar, A., Olson, P. & Lagerstedt, A.S. (2001). Breed risk of pyometra in insured dogs in Sweden. *Journal of Veterinary Internal Medicine* 15(6), 530-538.

Egenvall, A., Lonnell, C. & Roepstorff, L. (2009). Analysis of morbidity and mortality data in riding school horses, with special regard to locomotor problems. *Preventive Veterinary Medicine* 88(3), 193-204.

Egenvall, A., Nodtvedt, A., Penell, J., Gunnarsson, L. & Bonnett, B.N.Insurance data for research in companion animals: benefits and limitations. *Acta Veterinaria Scandinavica* In press

Egenvall, A., Nodtvedt, A. & von Euler, H. (2007). Bone tumors in a population of 400 000 insured Swedish dogs up to 10 y of age: incidence and survival. *Canadian Journal of Veterinary Research* 71(4), 292-9.

Egenvall, A., Penell, J., Bonnett, B.N., Blix, J. & Pringle, J. (2008b). Demographics and costs of colic in Swedish horses. *Journal of Veterinary Internal Medicine* 22(4), 1029-37.

Egenvall, A., Penell, J.C., Bonnett, B.N., Olson, P. & Pringle, J. (2005c). Morbidity of Swedish horses insured for veterinary care between 1997 and 2000: variations with age, sex, breed and location. *Veterinary Record* 157(15), 436-443.

Egenvall, A., Penell, J.C., Bonnett, B.N., Olson, P. & Pringle, J. (2005d). Morbidity of Swedish horses insured for veterinary care between 1997 and 2000: variations with age, sex, breed and location. *Vet.Rec.* 157(15), 436-443.

Egenvall, A., Penell, J.C., Bonnett, B.N., Olson, P. & Pringle, J. (2006). Mortality of Swedish horses with complete life insurance between 1997 and 2000: variations with sex, age, breed and diagnosis. *The Veterinary Record* 158(12), 397-406.

Fall, T., Hamlin, H.H., Hedhammar, A., Kampe, O. & Egenvall, A. (2007). Diabetes mellitus in a population of 180,000 insured dogs: incidence, survival, and breed distribution. *Journal of Veterinary Internal Medicine* 21(6), 1209-16.

Fombonne, E., Heavey, L., Smeeth, L., Rodrigues, L.C., Cook, C., Smith, P.G., Meng, L. & Hall, A.J. (2004). Validation of the diagnosis of autism in general practitioner records. *BMC Public Health* 4, 5.

Gelatt, K.N. & MacKay, E.O. (2004). Secondary glaucomas in the dog in North America. *Veterinary Ophthalmology* 7(4), 245-259.

Gelatt, K.N., Wallace, M.R., Andrew, S.E., MacKay, E.O. & Samuelson, D.A. (2003). Cataracts in the Bichon Frise. *Veterinary Ophthalmology* 6(1), 3-9.

Gormley, G., Connolly, D., Catney, D., Freeman, L., Murray, L.J. & Gavin, A. (2008). Reporting of research data by GPs: a cautionary tale for primary care researchers. *Family Practice* 25(3), 209-212.

Gower, S.B., Weisse, C.W. & Brown, D.C. (2009). Major abdominal evisceration injuries in dogs and cats: 12 cases (1998-2008). *Journal of the American Veterinary Medical Association* 234(12), 1566-72.

Guptill, L., Glickman, L. & Glickman, N. (2003). Time trends and risk factors for diabetes mellitus in dogs: analysis of veterinary medical data base records (1970-1999). *Veterinary Journal* 165(3), 240-7.

Hamilton, W., Lancashire, R., Sharp, D., Peters, T.J., Cheng, K.K. & Marshall, T. (2009). The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. *Bmc Medicine* 7

Hassey, A., Gerrett, D. & Wilson, A. (2001). A survey of validity and utility of electronic patient records in a general practice. *British Medical Journal* 322(7299), 1401-1405.

Hennessy, S., Bilker, W.B., Weber, A. & Strom, B.L. (2003). Descriptive analyses of the integrity of a US Medicaid claims database. *Pharmacoepidemiology and Drug Safety* 12(2), 103-111.

Hogan, W.R. & Wagner, M.M. (1997). Accuracy of data in computer-based patient records. *Journal of the American Medical Informatics Association* 4(5), 342-55.

Jordan, K., Porcheret, M. & Croft, P. (2004). Quality of morbidity coding in general practice computerized medical records: a systematic review. *Family Practice* 21(4), 396-412.

Kaneene, J.B., Ross, W.A. & Miller, R. (1997a). The Michigan equine monitoring system .2. Frequencies and impact of selected health problems. *Preventive Veterinary Medicine* 29(4), 277-292.

Kaneene, J.B., Saffell, M., Fedewa, D.J., Gallagher, K. & Chaddock, H.M. (1997b). The Michigan equine monitoring system .1. Design, implementation and population estimates. *Preventive Veterinary Medicine* 29(4), 263-275.

Lam, K.H., Parkin, T.D., Riggs, C.M. & Morgan, K.L. (2007). Descriptive analysis of retirement of Thoroughbred racehorses due to tendon injuries at the Hong Kong Jockey Club (1992-2004). *Equine Veterinary Journal* 39(2), 143-8.

Lawrenson, R., Todd, J.C., Leydon, G.M., Williams, T.J. & Farmer, R.D.T. (2000). Validation of the diagnosis of venous thromboembolism in general practice database studies. *British Journal of Clinical Pharmacology* 49(6), 591-596.

Lawrenson, R., Williams, T. & Farmer, R. (1999). Clinical information for research; the use of general practice databases. *Journal of Public Health Medicine* 21(3), 299-304.

Leblond, A., Villard, I., Leblond, L., Sabatier, P. & Sasco, A.J. (2000). A retrospective evaluation of the causes of death of 448 insured French horses in 1995. *Veterinary Research Communications* 24(2), 85-102.

Levine, J.M., Ngheim, P.P., Levine, G.J. & Cohen, N.D. (2008). Associations of sex, breed, and age with cervical vertebral compressive myelopathy in horses: 811 cases (1974-2007). *Journal of the American Veterinary Medical Association* 233(9), 1453-8.

Lewis, J.D., Brensinger, C., Bilker, W.B. & Strom, B.L. (2002). Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease. *Pharmacoepidemiology and Drug Safety* 11(3), 211-218.

Lloyd, S.S. & Rissing, J.P. (1985). Physician and coding errors in patient records. *JAMA* 254(10), 1330-6.

Maca, S.M., Scharitzer, M. & Barisani-Asenbauer, T. (2006). Uveitis and neurologic diseases: an often overlooked relationship. *Wiener Klinische Wochenschrift* 118(9-10), 273-9.

Malm, S., Fikse, W.F., Danell, B. & Strandberg, E. (2008). Genetic variation and genetic trends in hip and elbow dysplasia in Swedish Rottweiler and Bernese Mountain Dog. *Journal of Animal Breeding and Genetics* 125(6), 403-412.

Mann, R. & Williams, J. (2003). Standards in medical record keeping. *Clinical Medicine* 3(4), 329-332.

Mellor, D.J., Love, S., Gettinby, G. & Reid, S.W.J. (1999). Demographic characteristics of the equine population of northern Britain. *Veterinary Record* 145(11), 299-304.

Mellor, D.J., Love, S., Walker, R., Gettinby, G. & Reid, S.W.J. (2001). Sentinel practice-based survey of the management and health of horses in northern Britain. *The Veterinary Record* 149(14), 417-423.

Menendez, S., Steiner, A., Witschi, U., Danuser, J., Weber, U. & Regula, G. (2008). Data quality of animal health records on Swiss dairy farms. *Veterinary Record* 163(8), 241-6.

Mohammed, H.O., Rebhun, W.C. & Antczak, D.F. (1992). Factors Associated with the Risk of Developing Sarcoid Tumors in Horses. *Equine Veterinary Journal* 24(3), 165-168.

Mulder, C.A.T., Bonnett, B.N., Martin, S.W., Lissemore, K. & Page, P.D. (1994). The Usefulness of the Computerized Medical Records of One Practice for Research Into Pregnancy Loss in Dairy-Cows. *Preventive Veterinary Medicine* 21(1), 43-63.

Mullooly, J.P. (1990). The Effects of Data-Entry Error - an Analysis of Partial Verification. *Computers and Biomedical Research* 23(3), 259-267.

Mörk, M., Lindberg, A., Alenius, S., Vagsholm, I. & Egenvall, A. (2009). Comparison between dairy cow disease incidence in data registered by farmers and in data from

a disease-recording system based on veterinary reporting. *Preventive Veterinary Medicine* 88(4), 298-307.

Nguhiu-Mwangi, J., Mbithi, P.M.F., Wabacha, J.K. & Mbuthia, P.G. (2008). Retrospective study of foot conditions in dairy cows in urban and periurban areas of Kenya. *Israel Journal of Veterinary Medicine* 63(2), 40-45.

Nodtvedt, A., Bergvall, K., Emanuelson, U. & Egenvall, A. (2006a). Canine atopic dermatitis: validation of recorded diagnosis against practice records in 335 insured Swedish dogs. *Acta Veterinaria Scandinavica* 48, 8.

Nodtvedt, A., Egenvall, A., Bergvall, K. & Hedhammar, A. (2006b). Incidence of and risk factors for atopic dermatitis in a Swedish population of insured dogs. *The Veterinary Record* 159(8), 241-6.

Nyman, A.K., Ekman, T., Emanuelson, U., Gustafsson, A.H., Holtenius, K., Waller, K.P. & Sandgren, C.H. (2007). Risk factors associated with the incidence of veterinary-treated clinical mastitis in Swedish dairy herds with a high milk yield and a low prevalence of subclinical mastitis. *Preventive Veterinary Medicine* 78(2), 142-60.

Perkins, E., Stephens, J., Xiang, H. & Lo, W. (2009). The cost of pediatric stroke acute care in the United States. *Stroke* 40(8), 2820-7.

Physick-Sheard, P.W. (1986). Career profile of the Canadian Standardbred. I. Influence of age, gait and sex upon chances of racing 4. *Can.J.Vet.Res.* 50(4), 449-456.

Pilpel, D., Fraser, G.M., Kosecoff, J. & Brook, R.H. (1993). Validation of A Centrally Maintained Computerized Hospital Database - Comparison with Operating-Room Logbooks. *Israel Journal of Medical Sciences* 29(5), 287-291.

Piscitelli, P., Santoriello, A., Buonaguro, F.M., Di Maio, M., Iolascon, G., Gimigliano, F., Marinelli, A., Distante, A., Serravezza, G., Sordi, E., Cagossi, K., Artioli, F., Santangelo, M., Fucito, A., Gimigliano, R., Brandi, M.L., Crespi, M. & Giordano, A. (2009). Incidence of breast cancer in Italy: mastectomies and quadrantectomies performed between 2000 and 2005. *Journal of Experimental and Clinical Cancer Research* 28, 86.

Pollari, F.L., Bonnett, B.N., Allen, D.G., Bamsey, S.C. & Martin, S.W. (1996a). Quality of computerized medical record abstract data at a veterinary teaching hospital. *Preventive Veterinary Medicine* 27(3-4), 141-154.

Pollari, F.L., Bonnett, B.N., Bamsey, S.C., Meek, A.H. & Allen, D.G. (1996b). Postoperative complications of elective surgeries in dogs and cats determined by examining electronic and paper medical records. *Journal of the American Veterinary Medical Association* 208(11), 1882-1886.

Priester, W.A. (1970). A summary of diagnoses in the ox, horse, dog and cat from 12 veterinary school clinics in the U. S. and Canada. *Vet.Rec.* 86(22), 654-658.

Priester, W.A. (1974). Data from eleven United States and Canadian colleges of veterinary medicine on pancreatic carcinoma in domestic animals. *Cancer Res.* 34(6), 1372-1375.

Pringle, M., Ward, P. & Chilvers, C. (1995). Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. *British Journal of General Practice* 45(399), 537-41.

Ray, W.A., Griffin, M.R., Fought, R.L. & Adams, M.L. (1992). Identification of Fractures from Computerized Medicare Files. *Journal of Clinical Epidemiology* 45(7), 703-714.

Report of the Commission on Equine Policy, S. (2000). A Swedish Equine Policy (En svensk hästpolitik). SOU 2000:109 Stockholm:

Rohn, M., Tenhagen, B.A. & Hofmann, W. (2004). Survival of dairy cows after surgery to correct abomasal displacement: 1. Clinical and laboratory parameters and overall survival. *Journal of Veterinary Medicine. A, Physiology, Pathology, Clinical Medicine* 51(6), 294-9.

Roos, L.L., Jr., Roos, N.P., Cageorge, S.M. & Nicol, J.P. (1982). How good are the data? Reliability of one health care data bank. *Med Care* 20(3), 266-76.

Roos, L.L., Sharp, S.M. & Cohen, M.M. (1991). Comparing Clinical Information with Claims Data - Some Similarities and Differences. *Journal of Clinical Epidemiology* 44(9), 881-888.

Rosenlund, M., Bellander, T., Nordquist, T. & Alfredsson, L. (2009). Traffic-generated air pollution and myocardial infarction. *Epidemiology* 20(2), 265-71.

Rossdale, P.D., Hopes, R., Digby, N.J. & offord, K. (1985). Epidemiological study of wastage among racehorses 1982 and 1983. *Vet.Rec.* 116(3), 66-69.

Ru, G., Terracini, B. & Glickman, L.T. (1998). Host related risk factors for canine osteosarcoma. *Veterinary Journal* 156(1), 31-9.

Smith, L.J., Marr, C.M., Payne, R.J., Stoneham, S.J. & Reid, S.W.J. (2004). What is the likelihood that Thoroughbred foals treated for septic arthritis will race? *Equine Veterinary Journal* 36(5), 452-456.

Sultan, I., Rodriguez-Galindo, C., Saab, R., Yasir, S., Casanova, M. & Ferrari, A. (2009). Comparing children and adults with synovial sarcoma in the Surveillance, Epidemiology, and End Results program, 1983 to 2005: an analysis of 1268 patients. *Cancer* 115(15), 3537-47.

Swedish Animal Hospital Association, O. (1993). Diagnostic registry for the horse, the dog and the cat (Diagnosregister för häst, hund och katt). Tåberg, Tåbergs tryckeri, Sweden.

Thrusfield, M. (2003). *Veterinary Epidemiolgy*. Malden, USA: Blackwell Publishing, Inc.

Traub-Dargatz, J.L., Salman, M.D. & Voss, J.L. (1991). Medical problems of adult horses, as ranked by equine practitioners. *J.Am.Vet.Med.Assoc.* 198(10), 1745-1747.

Udomprasert, P. & Williamson, N.B. (1990). The Dairychamp Program - a Computerized Recording-System for Dairy Herds. *Veterinary Record* 127(10), 256-262.

USDA:APHIS:VS, C. (2006). *Equine 2005, Part I: Baseline Reference of Equine Health and Management, 2005.* Fort Collins: ISBN

USDA:APHIS:VS, C. (2007). *Trends in Equine Mortality, 1998-2005.* Fort Collins, USA:#N471-0307). ISBN

Wagner, M.M. & Hogan, W.R. (1996). The accuracy of medication data in an outpatient electronic medical record. *Journal of the American Medical Informatics Association* 3(3), 234-244.

Walker, C., Williams, H. & Phelan, J. (1998). Allergic rhinitis history as a predictor of other future disqualifying otorhinolaryngological defects. *Aviation Space and Environmental Medicine* 69(10), 952-6.

Wallin, L., Strandberg, E., Philipsson, J. & Dalin, G. (2000). Estimates of longevity and causes of culling and death in Swedish warmblood and coldblood horses 63(3), 275-289.

Ward, M.P. (2002). Seasonality of canine leptospirosis in the United States and Canada and its association with rainfall. *Preventive Veterinary Medicine* 56(3), 203-13.

Watts, R., Al-Taiar, A., Mooney, J., Scott, D. & Macgregor, A. (2009). The epidemiology of Takayasu arteritis in the UK. *Rheumatology* 48(8), 1008-11.

Whitelaw, F.G., Nevin, S.L., Milne, R.M., Taylor, R.J., Taylor, M.W. & Watt, A.H. (1996). Completeness and accuracy of morbidity and repeat prescribing records held on general practice computers in Scotland. *British Journal of General Practice* 46(404), 181-186.

Williams, P.C.W. & Ward, W.R. (1989). Development of a Coding System for Recording Clinical Findings in Farm Animal Practice. *Veterinary Record* 124(5), 118-122.

Wisten, A., Forsberg, H., Krantz, P. & Messner, T. (2002). Sudden cardiac death in 15-35-year olds in Sweden during 1992-99. *Journal of Internal Medicine* 252(6), 529-36.

(This is yet another empty page. Again, a white rectangle is drawn on top of the page number.)

Acknowledgements

The studies were carried out at the Division of Ruminant Medicine and Veterinary Epidemiology, Department of Clinical Sciences, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden. Financial support was provided by the Agria Research Fund and ATG.

I would like to thank all who in any way contributed to this work over the years, and especially:

Professor **Björn Ekesten**, Head of the Department of Clinical Sciences, for placing the facilities of the department at my disposal, and for your interest and support.

Associate professor **Agenta Egenvall**, my main supervisor, colleague and friend for sharing your genuine knowledge on epidemiologic research and data analyses, for taking so much of your time in assisting and explaining different issues over and over again. Thanks also for your efficiency, and for always keeping your door, email and telephone line open to me. I will remember all the laughs and great team work!

Associate professor **Brenda Bonnett**, my supervisor, for convincingly presenting the wonderful world of epidemiology to me while still in my undergraduate studies, for sharing your vast knowledge of epidemiology and encouraging me to see the big picture. Thanks for improving my research and writing skills by a master's eye for methodological details and presentations.

Professor **John Pringle**, my supervisor, for your calm, friendly, positive, and professional support of my work and especially for encouraging me

when I needed it. Thanks also for sharing your expertise in equine medicine and pointing out the clinician's situation. I always found our discussions very rewarding and inspiring.

Agria, and especially Dr **Lotta Gunnarsson**, **Pekka Olson** and **Johan Blix**, among others, for letting me access the insurance data, for support during the studies and being positive regarding my research. A special thanks to **Karin Ljunggren** for sharing your knowledge on the equine insurance processes and taking time to answer all my questions.

ATG and the **ATG Equine Clinics Ltd,** for providing the clinic data and for positive support during the studies. Thanks especially to **Görel Agri** and **Eva McLaren** at ATG for explaining and answering questions on the ATG database, making sure I got things right, and for sharing my interest in computerized information, and to **Jenny Ennerdal** at ATG Equine Clinics Ltd for friendly and invaluable help with reading manuscripts and answering all kinds of questions related to the ATG Equine Clinics. Thanks also to former company head veterinarian Dr **Peter Franzén** for support during the planning stage of the studies III and IV.

All the clinics and veterinarians in studies I and III for assisting me with copies of the veterinary journals, although very busy with clinical work. Your assistance was very much appreciated and needed!

Professor Ian Dohoo for hosting my visit to the Atlantic Veterinary College, Prince Edward Island, Canada, and for first class teaching in epidemiology, and Dr **Henrik Stryhn**, for professional teaching in biostatistics. Special thanks to everyone who made my visit on PEI so great, and in particular all my colleagues and friends: **Abu**, **Carol**, **Fabienne**, **Fortune**, **Isabel**, **Javier**, **Jill**, **Nick**, **Pascal**, **Richard**, **Vicky** and **Wendela**.

Kjell-Åke Ahlin, for making sure my computer worked, and for fast healing when it was not.

The **KC library staff** for all your help with references.

All former and present **colleagues at the division of ruminant medicine and veterinary epidemiology**, for sharing everyday working life with me and for providing a friendly, inspirational environment.
The former and present PhD-students at IME: **Ann L, Anna, Ann-Charlotte, Ann-Kristina, Aran, Carina, Cecilia L, Cecilia W, Charlotte, Emma, Helena, Helene, Jaruwan, Lena, Maria, Nils, Oscar** – it was great sharing the post-graduate time with you!

M**y large animal practice colleagues in the districts of Kalmar, Finsta and Norrköping** for offering me the opportunity to do clinical work in periods during the graduate studies, for sharing professional experience, and for your outstanding friendliness and encouragement. You are all animal heroes!

Drs **Ane Nødtvedt**, **Jenny Frössling** and **Ann Nyman**, and **Marie Mörk** (almost Dr now!) for being inspirational and fun and sharing an interest in epidemiology in general and many other relevant topics in particular.

All participants in the "**Epiforum**" group for interesting discussions. Thanks especially to Professor **Ulf Emanuelson** and Dr **Marie Engel** for your friendliness and engagement and interest in epidemiology.

My fiends in "**Bokis"**, especially **Anneli**, **Andrea**, **Bine**, **Linda**, **Johanna N**, **Vickan** and **Ylva** for sharing my interest in literature and fika. for over a decade of great times of discussion books, of spa trips, cruise trips, luxury afternoon teas and so much fun!

Jenny Lindley and **Eva Axelsson,** for being great friends for so many years.

Åsa Cajander, my dear friend for all the good times, for always being interested in discussing scientific theory and research and all other aspects of life.

My **family** and **relatives** in Sweden and Canada for being part of my life. You are a great mix of people!

My parents **Cia** and **Olle**, and brothers **David** and **Jacob** for your love, for believing in me and knowing me quite well by now. Your engagement, interest and support are priceless!

My very own wonderful family for being the most important part of my life and making it all worth while; **Neil**, **Nea** and **Hanna**.