Standardized Syndromic Classification of Animal Health Data (SSynCAHD)

FC Dórea¹, C Dupuy², F Vial³, CW Revie⁴, A Lindberg¹

¹Swedish Zoonoses Centre, National Veterinary Institute, Sweden.
²Epidemiology Unit, French Agency for Food, Environmental and Occupational Health & Safety, France
³Veterinary Public Health Institute, Universitáti Bern, Switzerland
⁴Atlantic Veterinary College, University of Prince Edward Island, Canada
*fernanda.dorea@sva.se

**Keywords:** syndromic surveillance, classification, syndromes, animal health, standards

**Abstract**
As the field of veterinary syndromic surveillance continues to develop, the adoption of syndromic classification standards becomes essential in order to promote maturity of the field and allow comparability of outputs from systems using different animal health data sources from different origins. We propose to review a number of animal health data sources, combining machine learning methods with expert opinion, in order to define a robust syndromic classification. Particular focus will be given to setting translational rules that allow classification of data sources with different levels of specificity, and coded according to different institution’s vocabularies.

**Introduction**
Veterinary syndromic surveillance (VSS) has been developing fast in the past 5 years (1). An inventory made by the Triple-S project (syndromic surveillance systems in Europe) in 2013 documented 27 VSS systems or initiatives in 12 European countries (2).

The development of syndromic surveillance using animal health data has at times been limited by the lack of standards for recording animal health events and thus to categorize these into syndromes. In practice this means that classification structures must be developed individually for each VSS system, and that they are specific to each data source. The step of data classification has been identified by our group as the least portable component of the VSS systems we have developed (3).

The lack of standards also renders outputs largely meaningless outside the system in which they are generated. That is, results are not comparable among systems using different data sources, and especially hard to compare among countries. This constitutes an important obstacle to the early detection of emerging diseases at an international scale as well as to the development and application of multivariate surveillance methods at any scale.

The lack of standardized data is not the result of a lack of nomenclature systems, but of a lack of incentives to comply with them. Standardized nomenclature systems, such as SNOMED (Systematized Nomenclature of Medicine) exist, but animal health data sources are rarely documented using these standards (4). This is in contrast to human clinical data, which is generally coded in compliance with the most current International Classification of Diseases (ICD-10) in order, *inter alia*, to bill insurance companies.

The difficulty in creating common vocabularies for VSS systems also results from the great variety of data sources explored. The inventory of VSS systems by the Triple-S project (2) documented the use of data sources such as drug sales, website hits or help line calls, clinical data, laboratory test requests, mortality reported (or rendering plant activity records) and meat inspection data. These data sources capture health events in different stages of the diagnostic continuum, and for that reason vary in specificity (Figure 1).

**Figure 1:** Sources of animal health data commonly explored for syndromic surveillance.

Considering the variety of animal health data sources available, and the use of specific vocabularies by each institution (even within the same country), we propose that the solution is not the development of (more) nomenclatures, but the development of classification methods based on translational rules. That is, rules would be designed to translate different types of data, offering varying degrees of specificity, into a standardized set of syndromes. The advantages would include: an ability to achieve syndromic classification from different sources of data which are (and will continue to be) recorded using the institution’s own vocabulary; the ability to compare VSS system outputs; and the more timely development of VSS systems.

**Materials and methods**
Comprehensive nomenclatures based on the manual development of taxonomies by experts, such as ICD-10, take years to develop, and require frequent manual revision to accommodate new knowledge or new language terms (5). In contrast modern machine learning methods allow for the automated extraction of structures and relationships among terms found in data sources. In this data-driven process the availability of more and more datasets favors the quality and speed of creation of a classification system,
rather than adding to the daunting amount of data that must be reviewed manually by experts.

Animal health data sources already have varied degrees of medical knowledge built into them: the relationship between the clinical signs seen by a veterinarian, and the samples she takes for diagnostic purposes; the signs or organ systems commonly observed and described in the same health event; the carcass alterations and reason for condemnation; etc. Machine learning methods allow this medical knowledge to be extracted and structured into classification rules, which can then be reviewed by experts.

As this cycle “learning from data → review by experts” repeats, the accuracy of classification increases, as does the system’s ability to deal with new terms. As the total amount of medical knowledge input into these classification systems increases, in time they are able to learn almost solely from data, based on predicted relationships between new and existing terms. To take an oversimplified example: once experts associate “avian influenza” to the classification term “respiratory”, a machine learning algorithm will be able to determine that the meaningful medical word here is “influenza”, since the term “avian” is not restricted to respiratory cases. When the system comes across “swine influenza” it may well be able to also classify it as respiratory, and associate to this group any new terms frequently found in the swine influenza instances. However, experts could always revise the rules.

While it is not our intention to create a comprehensive nomenclature, the work would nevertheless be informed by the latest thinking relating to the Semantic Web, which supports the development of frameworks to maximize the potential of data sharing and reuse (6). In particular, the use of ontology languages, such as OWL (Web Ontology Language) will constitute the main framework of development. In such frameworks the use of synonyms for created classes would allow the handling of multiple languages, stimulating cooperation among multiple countries.

Animal species, and many other characteristics of the health events, such as the step in the disease continuum at which they occurred (Figure 1), will influence the decision to associate specific health terms to a given syndrome. The use of ontologies allows a classification system to store not only the hierarchy among classification levels, but also the relationships among the objects classified.

Machine learning methods can be supervised or unsupervised. In supervised machine learning the algorithms learn classification methods by example, provided that classified datasets already exist. We have applied this method to laboratory test request data (7), creating classification rules which were then reviewed and altered based on expert knowledge. For less structured datasets, such as meat inspection data, unsupervised methods proved efficient (8).

We propose to extend these methods, previously employed for specific datasets, to include multiple datasets. The iteration between data-driven methods and expert knowledge will be carried out to include more data types (Figure 1), and multiple data sets of each type. Expert opinion will be elicited through a series of workshops and questionnaires. The proposed steps are listed below.

1. Gather documentation regarding VSS initiatives, in order to review the current syndromic classifications used. An inventory of initiatives has already been carried out by the Triple-S project (2). We will focus on documenting, in relation to the data types used, the syndromic groups defined and the classification methods employed.

2. Create a more extensive inventory of animal health sources with potential for syndromic surveillance. Again, part of this task has been carried out or is planned in other projects. Our focus will be on documenting the differences and similarities in vocabularies used, in order to define: the appropriate data types to start the process; the datasets with most potential for standardization across countries; and the biggest challenges to be addressed by expert opinion.

3. Gather a list of experts interested in the process. Initial contacts from previous inventories of VSS initiatives will be used, and these individuals will be encouraged to involve relevant contacts. A meeting with members of the International Society of Disease Surveillance (ISDS) is scheduled for December 2013, following an offer to share their expertise on standards development.

4. Initial compilation of a list of syndromic groups. Because many animal health data sources remain unclassified, or have been classified according to different criteria, the initial work will require expert input and manual review of datasets in order to create a substantive amount of training data to be used in supervised machine learning. Expert opinion will be elicited through a workshop planned for the ICAHS 2014. The results of steps 1 and 2 above will be presented, in order to seek agreement about the ideal list of syndromic groups to start working with. Additionally, points of divergence among different classification systems identified in steps 1 and 2 will be presented to try to reach consensus.

5. The main phase of the work will be conducted after the first expert elicitation. Starting from the list of syndromes agreed in the workshop, a few datasets will be used individually as case studies to derive rules that would allow classification of health events into the defined syndromes. Machine learning methods will be employed as described above. This will allow the generation of robust rules, taking advantage of the medical knowledge already input by diagnosticians into animal health data sets. The rules generated in this automated manner will then be reviewed manually, taking into consideration expert opinion gathered during the first workshop. Additional datasets will then be compared in two ways:

 — Comparing multiple datasets of the same type (i.e. more than one source of laboratory data, meat inspection data, clinical data, animal movement, etc.) in order to identify how rules can be generalized, within that data type, to be robust for different
datasets. Particular attention will be given to datasets expected to be available in several countries, which may follow similar data recording formats (e.g., animal movement data in Europe), or may be very varied (e.g., meat inspection data). Inconsistencies will again be documented and expert opinion elicited to consolidate rules.

— Comparing multiple data source types, to identify how the syndromic groups need to be broken down into progressively more specific disease groups, in order to accommodate data types with more specific etiology identification. Different data types, due to their different specificity, will fall into different levels of this hierarchy, and ultimately it may prove necessary to use fuzzy methods, such as faceted taxonomies. The ultimate goal is to achieve a taxonomy which will allow data from different types to be compared.

6. As we review datasets, it will become evident which data types are too unspecific to fit the initial syndromic groups used. This is an expected problem when using production data for instance. This new phase will constitute of iterative reviews of various data sources to identify health events that could not be classified into any of the previously listed syndromes, creation of new syndromes, and review of the classification rules and their coverage.

The work process in steps 5 and 6 is shown schematically in Figure 2.

Figure 2: Schematic view of the process of development of a standardized syndromic classification.

These steps are iterative. As more diseases and data sources are reviewed, the hierarchy will grow based on the development of practical rules to translate existing data into the syndromic classification. These rules will in turn be generalized to include different vocabularies as well as languages other than English.

A last stage will be the development of computational tools (freeware) to apply the translational rules to any data source.

Results
This work is still in its planning stages, and the most concrete result so far is the formation of an organizing group, listed as the authors of this document, who will be responsible for organizing expert elicitation, conducting the data sources review, applying the machine learning methods, and reviewing outputs based on expert opinion gathered. The group accumulates experience from syndromic surveillance development, machine learning methods for classification, and previous experience with the development of classification standards in veterinary medicine. Steps 1 to 3 will be performed before the ICAHS, the workshop listed in step 4 during the conference, and the main body of work will start after that.

Discussion
It has been our experience that attempts to establish joint surveillance and contingency plans will most likely fail if they rely on the sharing of data among institutions, even within the same country. Institutions and countries are prepared to share summary statistics and outputs of surveillance analysis, but these outputs are often hard to interpret and compare. Nomenclatures for animal disease classification exist but it cannot be expected that existing animal health data will be coded in compliance to these standards.

In this document we have outlined steps to initiate the construction of a standardized syndromic classification of animal health data based mainly on the creation of translational tools to classify existing data. The adoption of the system would not rely on changing the way data are recorded in their institution of origin, but would result in standardized input among veterinary syndromic surveillance systems using different data sources. This would be the first and most fundamental step towards generating outputs that could be shared and compared among countries, while protecting the privacy of animal health data.

Moreover, the creation of a system with a hierarchy that can accommodate data types offering different levels of specificity will allow comparison of outputs from multiple VSS systems, or combination of multiple data sources into a single system. In the case of emerging diseases, for instance, for which a taxonomy of clinical signs and etiology would not be available, the unspecific signs detected early in the process of disease emergence would be reflected in several datasets. Once these dataset were made comparable, their evidence could be pooled to enhance confidence and speed of disease detection, nationally and even internationally.

References
3) Dórea et al. (2013) SVEPM, Proc of the annual meeting.
5) ICD, WHO. http://www.who.int/classifications/icd/en/
6) W3 Semantic web: http://www.w3.org/2001/sw
78 Dupuy et al. BMC Vet Res 2013, 9:88