Animal Health Research Reviews

cambridge.org/ahr

A scoping review of 'big data', 'informatics', and 'bioinformatics' in the animal health and veterinary medical literature

Zenhwa Ouyang[1], Jan Sargeant[1,2,3], Alison Thomas[1], Kate Wycherley[1], Rebecca Ma[1], Rosa Esmaeilbeigi[1], Ali Versluis[4], Deborah Stacey[5], Elizabeth Stone[6], Zvonimir Poljak[1] and Theresa M. Bernardo[1]

[1]Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada; [2]Centre for Public Health and Zoonoses, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada; [3]Arrell Food Institute, University of Guelph, Guelph, Ontario, Canada; [4]Research and Scholarship Team, University of Guelph Library, Guelph, Ontario, Canada; [5]Department of Computer and Information Science, University of Guelph, Guelph, Ontario, Canada and [6]Department of Clinical Studies, University of Guelph, Guelph, Ontario, Canada

Review

Cite this article: Ouyang Z *et al* (2019). A scoping review of 'big data', 'informatics', and 'bioinformatics' in the animal health and veterinary medical literature. *Animal Health Research Reviews* **20**, 1–18. https://doi.org/10.1017/S1466252319000136

Received: 26 February 2019
Revised: 8 August 2019
Accepted: 13 August 2019

Key words:
Animal health; big data; bioinformatics; informatics; veterinary medicine

Author for correspondence:
Zenhwa Ouyang, Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada.
E-mail: zouyang@uoguelph.ca

Abstract

Research in big data, informatics, and bioinformatics has grown dramatically (Andreu-Perez J, *et al.*, 2015, *IEEE Journal of Biomedical and Health Informatics* 19, 1193–1208). Advances in gene sequencing technologies, surveillance systems, and electronic medical records have increased the amount of health data available. Unconventional data sources such as social media, wearable sensors, and internet search engine activity have also contributed to the influx of health data. The purpose of this study was to describe how 'big data', 'informatics', and 'bioinformatics' have been used in the animal health and veterinary medical literature and to map and chart publications using these terms through time. A scoping review methodology was used. A literature search of the terms 'big data', 'informatics', and 'bioinformatics' was conducted in the context of animal health and veterinary medicine. Relevance screening on abstract and full-text was conducted sequentially. In order for articles to be relevant, they must have used the words 'big data', 'informatics', or 'bioinformatics' in the title or abstract and full-text and have dealt with one of the major animal species encountered in veterinary medicine. Data items collected for all relevant articles included species, geographic region, first author affiliation, and journal of publication. The study level, study type, and data sources were collected for primary studies. After relevance screening, 1093 were classified. While there was a steady increase in 'bioinformatics' articles between 1995 and the end of the study period, 'informatics' articles reached their peak in 2012, then declined. The first 'big data' publication in animal health and veterinary medicine was in 2012. While few articles used the term 'big data' ($n = 14$), recent growth in 'big data' articles was observed. All geographic regions produced publications in 'informatics' and 'bioinformatics' while only North America, Europe, Asia, and Australia/Oceania produced publications about 'big data'. 'Bioinformatics' primary studies tended to use genetic data and tended to be conducted at the genetic level. In contrast, 'informatics' primary studies tended to use non-genetic data sources and conducted at an organismal level. The rapidly evolving definition of 'big data' may lead to avoidance of the term.

Introduction

Rationale

Society today produces more data in two days than it had cumulatively produced prior to 2003 (Sagiroglu and Sinanc, 2013). In human healthcare, data come from a variety of sources at a rapid pace. Data sources include social media, wearable sensors, surveillance systems, electronic medical records, and laboratory databases. Publications indexed in Google scholar that referenced 'big data' grew dramatically since 2008 (Andreu-Perez *et al.*, 2015). The top two health research areas were 'bioinformatics' and 'health informatics'.

In animal health, data also come from multiple sources at a rapid pace. Pet owners post photos and updates of their pets on social media. Wearables and other sensors have been developed for pets (https://www.whistle.com), horses (Peacock, 2012; Thompson *et al.*, 2018), and production animals (Andersson *et al.*, 2016; Haladjian *et al.*, 2018). Other sources of animal health data include government surveillance on animal diseases, veterinary electronic medical records, farm production records, and species-specific databases. These trends suggest that 'big data', 'informatics', and 'bioinformatics' might be growing in a similar fashion

© The Author(s) 2019. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

CAMBRIDGE UNIVERSITY PRESS

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

to that of human health. However, no one has evaluated how these terms are used in the veterinary medical and animal health literature.

Big data is frequently described in terms of three 'V's: volume, velocity, and variety (Schroeck *et al*., 2012). Volume refers to a large amount of data, velocity means that the data are generated quickly, and variety infers that the data come from different data sources and/or consist of different types of data (Schroeck *et al*., 2012). Veracity, or data reliability, is often considered a fourth characteristic of big data. Big data may also require non-traditional storage methods and analytical techniques (Elgendy and Elragal, 2014). Sources of big data in human healthcare include electronic medical records, genomics, imaging data, and data from social networks and sensors (Gaitanou *et al*., 2014).

Definitions of 'informatics' and 'bioinformatics' are broad and overlap with each other. The American Medical Informatics Association defines 'informatics' as 'the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health' (Kulikowski *et al*., 2012). The National Institutes of Health defines 'bioinformatics' as 'research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data' (Huerta *et al*., 2000).

Examining the use of these terms in the literature will provide insight into the type of research being conducted in each of these fields and may improve our understanding of big data, informatics, and bioinformatics and their relationships to (and how to distinguish them from) each other. Additionally, such examination will illuminate how research in these fields is conducted, who the leaders in the field are, the expertise needed to conduct such research and where the research is published.

For the remainder of this manuscript, we refer specifically to the terms big data, informatics, and bioinformatics with quotes (e.g. 'big data', 'informatics', and 'informatics'). When an article or group of articles is described using one of these terms in quotes (e.g. "'big data' article", "'big data' articles", and "articles about 'big data'"), we mean that the article or articles contain the quoted term.

Objectives

The purpose of this scoping review was to describe how 'big data', 'informatics', and 'bioinformatics' have been used in the animal health and veterinary medical literature by mapping the literature and describing the publications using these terms.

Materials and methods

Protocol

The authors used a scoping review approach as described by Arksey and O'Malley (Arksey and O'Malley, 2005). Study objectives and eligibility criteria were stated *a priori*. Most sections of the protocol were developed *a priori* with sections of the data charting tool and training tool modified after the review process started. The data synthesis plan was modified based on the findings of data charting.

Eligibility criteria

Smith and Williams (Smith and Williams, 2000) conducted a literature review of informatics in veterinary medicine from 1966 through 1995. Therefore, articles published in 1995 and later were selected for inclusion in the current study.

Information sources

The literature search covered the dates 1 January 1995 to 19 June 2017 in the following databases: Agricola (via ProQuest), ProQuest Dissertations and Theses, Medline (via PubMed), Web of Science, and IEEE Xplorer. The literature searches were conducted from 6 June 2017 to 19 June 2017. There were no language restrictions at this stage. Agricola, ProQuest, Medline, and Web of Science were chosen to capture scientific research in the animal health and veterinary medical literature. IEEE Xplorer was chosen to capture relevant engineering research in animal health and veterinary medicine.

Search

The search strategy was developed by a team of animal health and veterinary medical professionals, veterinary epidemiologists, a computer scientist and a library scientist (Table 1). The search strategy included conceptual and contextual terms (Peters *et al*., 2015). The conceptual terms were chosen to represent the topics of interest, which were 'big data', 'informatics' (lines 1 and 2 of Table 1a), and 'bioinformatics' (line 1 of Table 1b). Synonyms for 'informatics', 'information systems', and 'information technology', were also included as conceptual terms in the search strategy. The contextual terms were chosen to represent animal health and veterinary medicine. Contextual terms were limited to major small and large companion animals and food animals. Contextual terms included singular and plural variations (as well as scientific species names, e.g. canine, feline) of the following words: dog, cat, horse, dairy cattle, beef cattle, goat, sheep, layer poultry, broiler poultry, zoonoses, and foodborne (lines 3–17 of Table 1a and lines 2–16 of Table 1b). 'Zoonoses' and 'foodborne' were included to capture articles from a public health and food safety veterinary medical perspective, respectively.

Citations from Medline (via PubMed) were uploaded to Microsoft EndNote and then imported into DistillerSR (Evidence Partners, Ottawa, Canada). RIS files were downloaded from the other databases and uploaded directly to DistillerSR and deduplicated.

Selection of sources of evidence

Relevance screening was performed on title, abstract, and keyword (TAK) followed by full-text screening. The TAK relevance screening tool was piloted on randomly selected articles. Cohen's kappa was used to measure agreement between the primary reviewer (ZBO) and secondary reviewers. Cohen's kappa was used as a guide to help the research team train reviewers and refine questions in the relevance screening tool. Reviewer feedback on the relevance screening tool and/or a Cohen's kappa of 0.7 or more was used to determine sufficient agreement. For both TAK and full-text screening, agreement between two reviewers was required for articles to be included or excluded. Disagreements were resolved by consensus between the dissenting reviewers. If

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 1. Example of search strategy performed in Medline via PubMed to identify articles that use the terms (a) 'big data' or 'informatics' and (b) 'bioinformatics' in the animal health and veterinary medical literature

Number	Search String
(a)	
1	((informatic*[Title/Abstract] OR 'information system'[Title/Abstract] OR 'information systems'[Title/Abstract] OR 'information technology' [Title/Abstract] OR 'information technologies' [Title/Abstract]) OR informatic*[Other Term] OR 'information system'[Other Term] OR 'information systems'[Other Term] OR 'information technology'[Other Term] OR 'information technologies'[Other Term])
2	'big data'[Title/Abstract] OR 'big data'[Other Term]
3	((dog[Title/Abstract] OR dogs[Title/Abstract] OR canine[Title/Abstract] OR canines[Title/Abstract])) OR (dog[Other Term] OR dogs[Other Term] OR canine[Other Term] OR canines[Other Term])
4	((cat[Title/Abstract] OR cats[Title/Abstract] OR feline[Title/Abstract] OR feline[Title/Abstract])) OR (cat[Other Term] OR cats[Other Term] OR feline [Other Term] OR feline[Other Term])
5	((horse[Title/Abstract] OR horses[Title/Abstract] OR equine[Title/Abstract] OR equines[Title/Abstract])) OR (horse[Other Term] OR horses[Other Term] OR equine[Other Term] OR equines[Other Term])
6	(('dairy cattle'[Title/Abstract] OR 'dairy cow'[Title/Abstract] OR 'dairy cows'[Title/Abstract] OR 'dairy bovine'[Title/Abstract] OR 'dairy bovines'[Title/Abstract])) OR ('dairy cattle'[Other Term] OR 'dairy cow'[Other Term] OR 'dairy cows'[Other Term] OR 'dairy bovine'[Other Term] OR 'dairy bovines'[Other Term])
7	(((dairy[Title/Abstract]) AND (cattle[Title/Abstract] OR cow[Title/Abstract] OR cows[Title/Abstract] OR bovine[Title/Abstract] OR bovines[Title/Abstract])) OR dairy[Other Term]) AND (cattle[Other Term] OR cow[Other Term] OR cows[Other Term] OR bovine[Other Term] OR bovines[Other Term])
8	'beef cattle' [Title/Abstract] OR 'beef cow' [Title/Abstract] OR 'beef cows' [Title/Abstract] OR 'beef bovine' [Title/Abstract] OR 'beef bovines' [Title/Abstract] OR 'beef cattle' OR 'beef cow' OR 'beef cows' OR 'beef bovine' OR 'beef bovines'
9	(((beef[Title/Abstract]) AND (cattle[Title/Abstract] OR cow[Title/Abstract] OR cows[Title/Abstract] OR bovine[Title/Abstract] OR bovines[Title/Abstract])) OR beef[Other Term]) AND (cattle[Other Term] OR cow[Other Term] OR cows[Other Term] OR bovine[Other Term] OR bovines[Other Term])
10	((sheep[Title/Abstract] OR ovine[Title/Abstract] OR ovines[Title/Abstract])) OR (sheep[Other Term] OR ovine[Other Term] OR ovines[Other Term])
11	((goat[Title/Abstract] OR goats[Title/Abstract] OR caprine[Title/Abstract] OR caprines[Title/Abstract])) OR (goat[Other Term] OR goats[Other Term] OR caprine[Other Term] OR caprines[Other Term])
12	((swine[Title/Abstract] OR pig[Title/Abstract] OR pigs[Title/Abstract] OR porcine[Title/Abstract] OR porcines[Title/Abstract])) OR (swine[Other Term] OR pig[Other Term] OR pigs[Other Term] OR porcine[Other Term] OR porcines[Other Term])
13	(('layer poultry'[Title/Abstract] OR 'layer chicken'[Title/Abstract] OR 'layer chickens'[Title/Abstract] OR 'layer turkey'[Title/Abstract] OR 'layer turkeys'[Title/Abstract])) OR ('layer poultry'[Other Term] OR 'layer chicken'[Other Term] OR 'layer chickens'[Other Term] OR 'layer turkey'[Other Term] OR 'layer turkeys'[Other Term])
14	(('broiler poultry'[Title/Abstract] OR 'broiler chicken'[Title/Abstract] OR 'broiler chickens'[Title/Abstract] OR 'broiler turkey'[Title/Abstract] OR 'broiler turkeys'[Title/Abstract])) OR ('broiler poultry'[Other Term] OR 'broiler chicken'[Other Term] OR 'broiler chickens'[Other Term] OR 'broiler turkey'[Other Term] OR 'broiler turkeys'[Other Term])
15	((((broiler[Title/Abstract] OR layer[Title/Abstract])) AND (chicken[Title/Abstract] OR chickens[Title/Abstract] OR turkey[Title/Abstract] OR turkeys [Title/Abstract] OR poultry[Title/Abstract])) OR (broiler[Other Term] OR layer[Other Term])) AND (chicken[Other Term] OR chickens[Other Term] OR turkey[Other Term] OR turkeys[Other Term] OR poultry[Other Term])
16	((zoonosis[Title/Abstract] OR zoonoses[Title/Abstract] OR zoonotic[Title/Abstract])) OR (zoonosis[Other Term] OR zoonoses[Other Term] OR zoonotic[Other Term])
17	('food borne'[Title/Abstract]) OR 'food borne'[Other Term]
18	1 OR 2
19	3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13 OR 14 OR 15 OR 16
20	18 AND 19
21	20 AND ('1995/01/01'[PDat] : '2017/12/31'[PDat])
(b)	
1	(bioinformatic*[Title/Abstract]) OR bioinformatics*[Other Term]
2	((dog[Title/Abstract] OR dogs[Title/Abstract] OR canine[Title/Abstract] OR canines[Title/Abstract])) OR (dog[Other Term] OR dogs[Other Term] OR canine[Other Term] OR canines[Other Term])
3	((cat[Title/Abstract] OR cats[Title/Abstract] OR feline[Title/Abstract] OR felines[Title/Abstract])) OR (cat[Other Term] OR cats[Other Term] OR feline[Other Term] OR felines[Other Term])
4	((horse[Title/Abstract] OR horses[Title/Abstract] OR equine[Title/Abstract] OR equines[Title/Abstract])) OR (horse[Other Term] OR horses[Other Term] OR equine[Other Term] OR equines[Other Term])

(*Continued*)

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 1. (*Continued.*)

Number	Search String
5	(('dairy cattle'[Title/Abstract] OR 'dairy cow'[Title/Abstract] OR 'dairy cows'[Title/Abstract] OR 'dairy bovine'[Title/Abstract] OR 'dairy bovines'[Title/Abstract])) OR ('dairy cattle'[Other Term] OR 'dairy cow'[Other Term] OR 'dairy cows'[Other Term] OR 'dairy bovine'[Other Term] OR 'dairy bovines'[Other Term])
6	(((dairy[Title/Abstract]) AND (cattle[Title/Abstract] OR cow[Title/Abstract] OR cows[Title/Abstract] OR bovine[Title/Abstract] OR bovines[Title/Abstract])) OR dairy[Other Term]) AND (cattle[Other Term] OR cow[Other Term] OR cows[Other Term] OR bovine[Other Term] OR bovines[Other Term])
7	(('beef cattle'[Title/Abstract] OR 'beef cow'[Title/Abstract] OR 'beef cows'[Title/Abstract] OR 'beef bovine'[Title/Abstract] OR 'beef bovines'[Title/Abstract])) OR ('beef cattle'[Other Term] OR 'beef cow'[Other Term] OR 'beef cows'[Other Term] OR 'beef bovine'[Other Term] OR 'beef bovines'[Other Term])
8	(((beef[Title/Abstract]) AND (cattle[Title/Abstract] OR cow[Title/Abstract] OR cows[Title/Abstract] OR bovine[Title/Abstract] OR bovines[Title/Abstract])) OR beef[Other Term]) AND (cattle[Other Term] OR cow[Other Term] OR cows[Other Term] OR bovine[Other Term] OR bovines[Other Term])
9	((sheep[Title/Abstract] OR ovine[Title/Abstract] OR ovines[Title/Abstract])) OR (sheep[Other Term] OR ovine[Other Term] OR ovines[Other Term])
10	((goat[Title/Abstract] OR goats[Title/Abstract] OR caprine[Title/Abstract] OR caprines[Title/Abstract])) OR (goat[Other Term] OR goats[Other Term] OR caprine[Other Term] OR caprines[Other Term])
11	((swine[Title/Abstract] OR pig[Title/Abstract] OR pigs[Title/Abstract] OR porcine[Title/Abstract] OR porcines[Title/Abstract])) OR (swine[Other Term] OR pig[Other Term] OR pigs[Other Term] OR porcine[Other Term] OR porcines[Other Term])
12	(('layer poultry'[Title/Abstract] OR 'layer chicken'[Title/Abstract] OR 'layer chickens'[Title/Abstract] OR 'layer turkey'[Title/Abstract] OR 'layer turkeys'[Title/Abstract])) OR ('layer poultry'[Other Term] OR 'layer chicken'[Other Term] OR 'layer chickens'[Other Term] OR 'layer turkey'[Other Term] OR 'layer turkeys'[Other Term])
13	(('broiler poultry'[Title/Abstract] OR 'broiler chicken'[Title/Abstract] OR 'broiler chickens'[Title/Abstract] OR 'broiler turkey'[Title/Abstract] OR 'broiler turkeys'[Title/Abstract])) OR ('broiler poultry'[Other Term] OR 'broiler chicken'[Other Term] OR 'broiler chickens'[Other Term] OR 'broiler turkey'[Other Term] OR 'broiler turkeys'[Other Term])
14	((((broiler[Title/Abstract] OR layer[Title/Abstract])) AND (chicken[Title/Abstract] OR chickens[Title/Abstract] OR turkey[Title/Abstract] OR turkeys[Title/Abstract] OR poultry[Title/Abstract])) OR (broiler[Other Term] OR layer[Other Term])) AND (chicken[Other Term] OR chickens[Other Term] OR turkey[Other Term] OR turkeys[Other Term] OR poultry[Other Term])
15	((zoonosis[Title/Abstract] OR zoonoses[Title/Abstract] OR zoonotic[Title/Abstract])) OR (zoonosis[Other Term] OR zoonoses[Other Term] OR zoonotic[Other Term])
16	('food borne'[Title/Abstract]) OR 'food borne'[Other Term]
17	2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13 OR 14 OR 15 OR 16
18	1 AND 17
19	18 AND ('1995/01/01'[PDat] : '2017/12/31'[PDat])

consensus was not achieved between two reviewers, a third reviewer was consulted.

Articles with TAKs containing at least one contextual term and at least one conceptual term proceeded to full-text relevance screening. Reviewers could select 'unsure' during TAK relevance screening. These articles also proceeded to full-text screening.

Searches for full-text articles were conducted on the University of Guelph library website. If not available, an interlibrary loan request was placed. Full-text articles that were not acquired via interlibrary loan were then searched for in the Google search engine and in Google Scholar using titles and first author. Any full-text articles that were not found on Google or Google Scholar were excluded.

In full-text relevance screening, reviewers determined whether at least one contextual term present in the article referred to an animal (e.g. 'cat' versus 'CAT scan') and whether the contextual terms implied that the study was relevant to animals (e.g. a study that utilized an *equine* virus in the development of a human vaccine for use in humans with no mention of animal health implications would be excluded; a study that utilized an *equine* virus in the development of a human vaccine that has implications for both human and animal health would be included). If the contextual terms satisfied these conditions, the article proceeded to the final stage of relevance screening. In the final stage, reviewers determined whether the conceptual terms were used to describe the study or if the conceptual terms described a study referenced by the article (e.g. an article that stated 'The current study utilizes *big data*' would be included; a study that stated 'Previous studies utilizing *big data* suggested an association', but 'big data' did not apply to the study itself would be excluded). If the conceptual terms were used to describe the study in the article, the article proceeded to full-text screening. Non-English articles were excluded at this stage of the study.

Data charting process

We developed a data collection form which went through two iterations of review by the entire research team and was piloted among ZBO, RE, RM, AT, and KW before being finalized.

Data collection was performed by eight members of the review team (ZBO, AT, EM, RE, VS, KW, JS, and IS). Reviewers were given a set of articles and initially met with ZBO for consensus after 10–50 articles were complete. Questions about the review protocol were addressed and disagreements in data collection were resolved.

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Data items

Articles were identified as either describing: (1) primary studies (studies where the research team collected original data, conducted an original analysis or performed simulation-modeling); or (2) reviews (systematic, scoping, narrative), commentaries/editorials, letters-to-the-editor or conference proceedings. Although conference proceedings may have described primary studies, due to variations in the format of conference proceedings (i.e. some were abstracts only while others resembled complete scientific papers), conference proceedings were not grouped with primary studies.

Species that the articles were describing were identified. Species were limited to those described in Table 8. The search and subsequent data collection was limited to the major domestic species encountered in veterinary medicine and animal health. Inclusion of other species (e.g. exotics, wildlife) was beyond the scope of this review.

Data were collected on the geographic region of the study. If it was not provided, the first author location was used. Geographic regions were based on the Standard Country or Area Codes for Statistical Use published by the United Nations (https://unstats.un.org/unsd/methodology/m49/).

The first author affiliation was collected to provide an understanding of the fields of study involved in producing research in big data, informatics or bioinformatics in veterinary medicine and animal health. Journal of publication was collected to provide an understanding of who is interested in this research. The classification scheme for the first author affiliation and journal of publication is presented in Table 2.

The data items shown in Table 10 were collected for articles that described primary studies. Primary studies were classified into types (Table 10) and study levels (Table 3). Studies classified as having study levels at the 'genes, proteins, molecules and metabolites of animals' or 'genes, proteins, molecules and metabolites of organisms found on/in animals' investigated genetic material will be referred to as 'genetic studies', and may include, but not limited to, gene sequencing, genomic, metagenomic, and microbiome studies. Data sources used in primary studies were also categorized (Table 4).

Initially, no distinction was made between genetic databases and non-genetic databases in the government-sourced category. After data classification was completed, it was decided post-hoc to estimate the number of government genetic databases. The number of articles classified as using government data sources that had the terms 'NCBI' (National Center for Biotechnology Information), 'GenBank' or 'DAVID' (Database for Annotation, Visualization and Integrated Discovery) were counted. GenBank and DAVID are nucleotide and protein sequence databases. GenBank is hosted by NCBI, which is an organization that hosts search engines of several databases, including GenBank. Genetic data from non-government databases were classified under 'genetic databases'.

Reviewers were given the option of selecting multiple answers for each data item. For the study level and study type, each selection must have been stated in the study objectives. Thus, an article with a study objective that states that only prevalence of a bacterium was measured may have reported the results of a hypothesis test; however, the reviewer could not select 'hypothesis test' under study type because it was not reflected in the study objectives.

Synthesis of results

The number of articles per year that used the conceptual terms 'big data', 'informatics', and 'bioinformatics' was compiled into a timeline (Fig. 2). The frequency of articles that used the conceptual terms was compared to publication type (Table 7). Data regarding species, geographic region, first author affiliation, and journal of publication for each conceptual term were extracted for all articles and compiled in Table 8. A layered barplot (Fig. 4) (post-hoc) was created to illustrate the number of articles about each species by the geographic region. Most studies about pigs used the term 'bioinformatics' (Table 8), so it was decided post-*hoc* to determine if this was true for each geographic region (Table 9). The study level, study type, and data sources for each conceptual term were collected and were presented in Table 10.

Results

Selection of sources of evidence

The literature search yielded 8602 articles. There were 1093 articles included in data characterization after de-duplication, TAK relevance screening, and full-text screening. Of these, 918 were full-text articles that described a primary research study and 175 articles were conference proceedings or were not primary research studies (e.g. narrative reviews, scoping reviews, letter-to-the-editor, conference proceedings, and commentaries). Of the 578 articles that were excluded on full-text screening, 147 articles were not found, 93 articles were not in English, and 338 articles did not pass full-text relevance screening (Fig. 1).

Results of individual sources of evidence and synthesis of results

Figure 2 shows that the use of the term 'bioinformatics' increased rapidly since 1995. The use of 'informatics' increased until 2012, then began to decline. The term 'big data' was first used in 2012 in one publication and was used in one publication in 2013 and 2014. The use of the term increased to four articles in 2015 and five articles in 2016. Data for 2017 are for a partial year, as the search period ended June 19, 2017

The majority of articles used 'bioinformatics' (Fig. 3). Articles about 'informatics' were the second most common, of which 57% (250/438) described using geographic information systems (GIS). Only 14 articles in the veterinary medical and animal health literature used the term 'big data', and half of them were narrative reviews, commentaries, editorials or letters-to-the-editor (Table 7). 'Informatics' and 'bioinformatics' articles were most frequently primary studies. The characterization for the 'big data' articles is shown below (Tables 5 and 6).

General characteristics of the articles are included in Table 8. Articles about small animals (dogs and cats) used 'informatics' more than 'bioinformatics'. 'Informatics' and 'bioinformatics' were relatively balanced between articles about cattle where the production system (dairy, beef) was specified. Articles where the production systems were unspecified were more often about 'informatics'. Articles about pigs, on the other hand, tended to be about 'bioinformatics' (Table 8). For articles that used the term 'informatics', there were ~2.1 species mentioned per article. For articles that used the term 'bioinformatics', there were ~1.4 species mentioned per article.

Five of six geographic regions produced articles about 'big data'. Articles about 'informatics' and 'bioinformatics' have been

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 2. Classification scheme of first authors and journal types

Classification	Description	Examples
Veterinary medicine and animal health	Author affiliation or journal title must explicitly indicate relevance to animals. Includes, but not limited to veterinary medicine, animal science, and animal agriculture and food science.	Author affiliations • School of Veterinary Medicine • Department of Surgery, School of Veterinary Medicine • Department of Statistics, School of Veterinary Medicine • Department of Dairy Sciences • Department of Animal Biology • Department of Animal Genetics Journal titles • Journal of Veterinary Medicine • Journal of Veterinary Surgery • Journal of Animal Sciences • Journal of Dairy Sciences • Journal of Animal Biology • Journal of Animal Genetics
Human medicine and health.	Author affiliations or journal titles that contain the words 'medicine' or 'health' or words that pertain to any medical specialty (e.g. surgery, opthalmology, dermatology, nutrition, pediatrics, and geriatric). Does not contain words that indicate relevance to animals, e.g. 'veterinary', 'animal' or 'dairy'.	Author affiliations • Department of Medicine • Department of Surgery, School of Medicine • Department of Statistics, School of Medicine • Department of Public Health • Department of Pediatrics • Department of Environmental Sciences, School of Public Health Journal titles • Journal of Medicine • Journal of Surgery • Journal of Public Health • Journal of Geriatrics • Journal of Psychiatry • Journal of Environmental Medicine
Biological sciences	Author affiliations or journal titles that pertain to biology, microbiology, biochemistry, genetics, zoology, environmental sciences or engineering, entomology, parasitology, bioengineering, and biomedical engineering. Terms such as 'biostatistics' and 'biological mathematics' would be excluded from this classification and placed in the 'statistics, data science, mathematics' classification.	Author affiliation • Department of Biology/Biological Sciences/Biosciences • Department of Biological Sciences • Department of Genetics • Department of Zoology • Department of Parasitology • Department of Environmental Sciences/Environmental Engineering Journal titles • Journal of Biological Sciences • Journal of Genetics • Journal of Zoology • Journal of Parasitology • Journal of Environmental Sciences
Bioinformatics	Author affiliations or journal titles that explicitly reference the terms (or variations of the terms) 'bioinformatics', 'genomics', 'proteomics', 'metabolomics' or any other type of OMIC.	Author affiliations • Department of Bioinformatics • Department of Genomics • Department of Metabolomics • Department of Foodomics Journal titles • Journal of Bioinformatics • Journal of Genomics • Journal of Metabolomics • Journal of Foodomics

(*Continued*)

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 2. (*Continued.*)

Classification	Description	Examples
Physical sciences	Author affiliations or journal titles with words that indicate relevance to a science without indicating relevance to an animal or biological science. Includes, but not limited to, geography, physics, chemistry and engineering (e.g. mechanical, electrical). 'Biological geography', 'biophysics', 'biochemistry' and 'biomedical engineering' would be excluded from this classification and placed in the 'biological sciences' classification.	Author affiliations • Department of Physics • Department of Geography • Department of Chemistry • Department of Materials Engineering Journal titles • Journal of Physics • Journal of Geography • Journal of Chemistry • Journal of Materials Engineering
Statistics and mathematics	Author affiliations or journal titles containing the words (or variations of) 'statistics', 'data science' or 'mathematics'. 'Biostatistics' and 'mathematical biology' would be placed in this category.	Author affiliations • Department of Statistics • Department of Statistical Analysis • Department of Data Science • Department of Data Analysis • Department of Mathematics Journal titles • Journal of Statistics • Journal of Statistical Analysis • Journal of Data Science • Journal of Data Analysis • Journal of Mathematics
Computer science and information technology	Author affiliations or journal titles that use the words 'computer science', 'computer programming' or 'information technology or some type of variation or abbreviation.	Author affiliations • Department of Computer science • Department of Computer Programming • Department of Information Technology Journal titles • Journal of Computer Science • Journal of Computer Programming • Journal of Information Technology
Social sciences	Author affiliations or journal titles that use the words 'economics, 'social sciences' or 'business' or variations.	Author affiliations • Department of Economics • Department of Social Sciences • Department of Sociology • Department of Psychology • Department of Marketing • Department of Business Journal titles • Journal of Economics • Journal of Social Sciences • Journal of Sociology • Journal of Psychology • Journal of Marketing • Journal of Business

published in all geographic regions. North America and Europe had similar numbers of publications for 'informatics' and 'bioinformatics'; however, most publications from Asia were about 'bioinformatics'.

Articles about cattle were most common across all geographic regions except Asia. Articles about pigs were the most common in Asia (Fig. 4). To determine whether studies about pigs conducted in Asia contributed significantly to the counts for articles that used the term 'bioinformatics', we present data specific to pigs in Table 9. Articles describing studies performed in Asia or with the first authors based in Asia overwhelmingly used the term 'bioinformatics' more than 'big data' and 'informatics'. Articles describing studies performed in North America or with first authors based in North America also used the term

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 3. Study level classification (organized by subject area domain) for data charting of primary studies using the terms 'big data', 'informatics', and 'bioinformatics'

Domain	Study level	Examples
Methodology		
	Lab techniques	• Development of a new method to isolate DNA from bacteria. • Comparison of bacterial culture techniques. • Validation of a new bacterial culture technique.
	Analytical techniques	• Development of a new statistical method. • Comparison of various statistical methods. • Validation of a new simulation-model. • Development, comparison and/or validation of analytical techniques that will be packaged into software, but not at the time of the study.
	Software	• Development of software. • Comparison of various software products. • Validation of analytical techniques within a software product.
Environment		
	Effects of animals on the environment	• A study that investigates how cattle manure affects local water sources. • A study that investigates how ambient air pollution from a swine farm affects local residents. • A study that investigates how feral cats affect the wild bird population.
Animal product or by-product		
	Animal product or by-product	• A study that measures milk production to determine whether the presence of a certain protein is associated with increased milk production in dairy cattle. • A study that investigates factors that promote wool quality in sheep. • A study that investigates the efficacy of pig feces as crop fertilizer. • A study that investigates best practices in the handling of cattle carcasses in the abattoir to improve hide quality.
Bacteria, viruses, parasites or fungi found on/in animals		
	Bacteria, viruses, parasites or fungi found on/in animals	• A study that estimates the prevalence of a specific bacteria on the skin of dogs visiting a veterinary clinic. • A study that investigates the association between specific bacteria found in feces of sick dogs and a specific dog food. • A study that investigates the control of avian influenza in poultry. • A study that measures the efficacy of an anthelmintic in cattle.
	Genes, proteins, molecules, and metabolites of organisms found on/in animals	• A DNA sequencing study of cattle liver flukes. • A study that investigates the genetic relationship between Staphylococci found on the skin of humans and dogs. • A study that attempts to trace the spread of avian influenza in poultry in an outbreak by analyzing genetic sequences. • A study that characterizes genes and proteins of an antimicrobial resistant bacteria in horses to inform development of pharmaceuticals.
Animal		
	Animal	• A study that investigates how certain feeds can improve average daily gain in cattle. • A study that investigates risk factors for bone fractures in horses. • A study that investigates whether dogs can be used to detect wild turtles in the desert. • A study that investigates and reports the biological development of certain cancers in dogs. • A study that investigates the efficacy of a cancer treatment for cats. • A study that compares the effects of open-range and traditional poultry production systems on welfare.

(*Continued*)

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 3. (*Continued.*)

Domain	Study level	Examples
	Genes, proteins, molecules, and metabolites of animals	• A study that describes the similarities between a certain gene of domesticated dogs and wolves. • A study that identifies a gene responsible for immunity to certain diseases in pigs. • A study that sequences a gene responsible for milk production in cattle. • A study that describes the amino acid sequence of a certain protein associated with laminitis in horses.

'bioinformatics' more often, however, the difference was not as pronounced.

Most of the articles had first authors with affiliations in 'veterinary medicine and animal health' (Table 8). 'Informatics' articles more frequently had first authors from 'physical sciences' (29 versus 6), 'computer science and information technology' (15 versus 0), and 'social sciences' (24 versus 1) than 'bioinformatics' articles.

The two most common types of journals of publication were 'biological' (484) and 'veterinary medicine and animal health' (355) (Table 8). 'Veterinary medicine and animal health' was the most common journal of publication for 'big data' and 'informatics' articles. 'Biological' was the most common journal of publication for 'bioinformatics' articles. 'Informatics' articles were more frequently published in 'physical sciences' (25 versus 3) and 'computer science and information technology' (49 versus 2) journals than 'bioinformatics'. 'Bioinformatics' articles more frequently published to 'bioinformatics' journals (81 versus 6) than 'informatics' journals.

Primary studies described in 'bioinformatics' articles tended to be conducted at the 'animal genes, proteins, metabolites' level (354/589; 60%) (Table 10). 'Informatics' articles describing primary studies tended to be conducted at the 'animal bacteria, virus, parasite, fungus' level (121/326; 37%) and 'animal' level (67/326; 21%) or were 'software, analytical technique, lab technique development/validation studies' (87/326; 27%). Primary studies described by 'informatics' articles focused more on the 'effects of animals on environment' (35/326; 11%) than those described by 'bioinformatics' articles (2/589; 0.3%).

'Bioinformatics' articles also described 'software, analytical technique, lab technique development/validation studies' (45/589; 8%) (Table 10). Of these articles, 'bioinformatics' articles were largely about laboratory techniques while 'informatics' articles were about analytical techniques and software.

Primary studies classified as 'hypothesis testing (observational)' were more frequently in 'informatics' articles (168/326; 52%) than in 'bioinformatics' articles (166/589; 28%) (Table 10). Primary studies classified as 'hypothesis testing (experimental)' were more frequently in 'bioinformatics' articles (136/589; 23%) than 'informatics' articles (5/326; 2%). 'Bioinformatics' studies (258/589; 44%) were also more often classified as 'descriptive' than 'informatics' studies (41/326; 13%).

'Bioinformatics' primary studies tended to use genetic databases (234/589; 40%) and government-sourced databases (301/589; 51%) (Table 10). Of the 301 'bioinformatics' primary studies that used government-sourced data, 89% (269/301) of those databases were NCBI (National Center for Biotechnology Information), GenBank or DAVID (Database for Annotation, Visualization and Integrated Discovery). 'Informatics' primary studies tended to use non-genetic sources of data. Although 'informatics' primary studies used biologic samples, they also used other data sources, e.g. electronic medical records, farm production records, internet search engines, climate data, questionnaires, and wearables/sensors. Forty-seven percent (154/326) of 'informatics' primary studies used government-sourced data; however, only seven of these data sources were NCBI, GenBank or DAVID.

Discussion

Summary of evidence

Although research in 'big data', 'informatics', and 'bioinformatics' has been growing in human medicine, with the exception of 'bioinformatics', we currently do not see a similar growth in the animal health and veterinary medical research literature. There appears to be a lag in the production of 'big data' articles in veterinary medicine and animal health compared to human health (Andreu-Perez *et al.*, 2015).

The use of the term 'big data' is relatively recent and uncommon, perhaps due to the rapidly evolving definition of what big data is (Natarajan *et al.*, 2017). The greater number of reviews compared to primary studies would suggest that the potential of big data in veterinary medicine and animal health is still being explored (see Table 6). Researchers interested in learning about 'big data' in veterinary medicine and animal health may need to search other bodies of literature.

An effective definition needs to address what characteristics are necessary for a study to be considered a big data study. The development of such a definition could be addressed by a systematic review. Big data is often characterized by the Vs, e.g. volume, velocity and variety (Laney, 2001; Schroeck *et al.*, 2012). Although data volume remains a necessary component for the approach to be considered a big data approach, the latter two components are becoming equally or more important (Natarajan *et al.*, 2017), a trend which has been attributed to more widespread availability of large volumes of data. It has also been argued that the relationships between the three Vs of big data should be examined in order to declare data as 'big' (Natarajan *et al.*, 2017). This complexity, when coupled with the relatively stringent initial definition of big data, and the definition's now evolving nature (Ylijoki and Porras, 2016) could have influenced, in different ways, the low number of studies declared as using a big data approach in the veterinary medical and animal health literature. First, it is possible that research conducted in this area did not fit the contemporary definition, even if loosely defined, of big data. Second, it is possible that published literature addresses

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 4. Descriptions and examples of data sources used in primary studies using the terms 'big data', 'informatics', and 'bioinformatics'

Data sources	Description	Examples
Biologic samples	• Any biologic sample taken from an animal. • Any direct observation by a researcher made about the animal by the researcher.	• Blood, hair, skin samples • Biopsies • Visual examination of an animal by a researcher • Visual examination of a video of an animal by a researcher
Genetic databases	• Any database containing genetic data not owned by the government. • Includes genetic, genomic, metagenomic, microbiomic and any other database that contains nucleic acid, amino acid or protein sequence data.	• Gene sequencing data owned by a cattle breeding association.
Electronic medical records	• Any electronic medical record used and maintained by health professionals.	• Electronic medical record of a veterinary hospital.
Farm production records	• Any production record used and maintained by agricultural producers.	• Dairy production records of a farm.
Internet search engines, social media	• Any data produced by analyzing internet searches (e.g. text entered by user into a search engine), internet search results (e.g. webpages resulting from an internet search), or by mining data from social media.	• Webpages returned from an internet search. • Frequency of keywords used in internet searches. • Posts on Twitter that would subsequently be analyzed to assess public opinion.
Scientific literature databases	• Data based on the capturing of search results or search behaviors in scientific literature databases. • Results reported in scientific literature.	• Frequency of scientific publications in a variety of scientific literature databases about a certain topic. • Data collected from various publications from searches in scientific literature databases to estimate parameters for simulation-modeling.
Geographic	• Geographic data collected by the researchers.	• Researchers travel from household-to-household recording geographic coordinates produced by a GPS (global positioning system).
Environment	• Data collected by researchers on the climate, weather, plant life or soil. • Does not include data collected on animals.	• Researchers travel to various locations to collect plant samples to estimate plant density in a certain area. • Images of plant life which researchers use to estimate plant density via image analysis.
Government-sourced	• Any data that was taken from a government database.	• Data from government agricultural databases. • Genetic databases from the government.
Non-government-sourced	• Data from a database that was not from the government and cannot be classified into any of the other categories.	• Health data collected by a private company given to researchers for research different from the original purpose.
Wearable sensors	• Researchers utilized a device that was either attached to or carried within the animal's body to collect data.	• Activity monitors on a dog collar to measure activity and record location. • GPS devices placed on cattle. • Chips implanted in the skin of dogs to record identity and location.
Questionnaires	• Data collected from questions administered to another person or people. Questions may be administered orally, on paper or electronically.	• Paper or electronic surveys. • Interviews or focus groups.
No data used	• Any study that did not use recorded or observed data as input.	• Mathematical simulation studies that explore hypothetical parameter values.

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

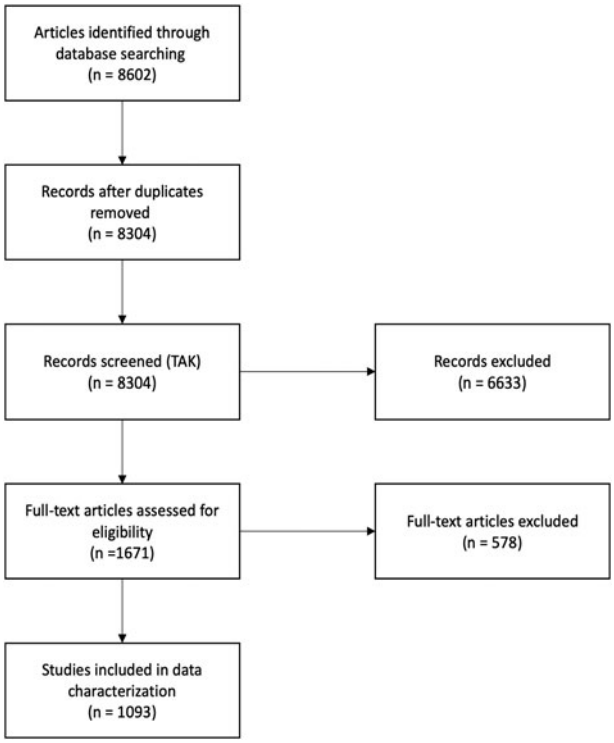

Fig. 1. Flow of articles and citation from literature search through data characterization.

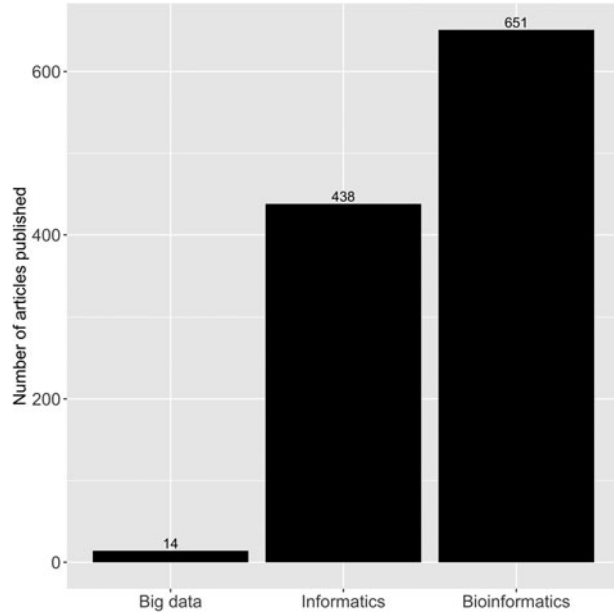

Fig. 3. Number of articles that used the words 'big data', 'informatics' or 'bioinformatics'.

only one component of big data (e.g predictive analytics) in isolation from other components and therefore cannot be, and was not considered, an approach to research consistent with big data. Only when combined with other components, do these isolated parts form an approach to big data. This integration may be beyond the scope of individual research contributions. Finally, it is also possible that the big data research has been conducted, but has not been communicated under the name 'big data', or the approach has been utilized not for the purposes of publication but for product or process development within specific organizations, e.g. livestock commodity groups that are used by industry and researchers. Research about one component of big data and big data research used within specific organizations, if published at all, may only be found within specialized literature.

Another possible explanation for why 'big data' was uncommon is that existing big datasets in veterinary medicine and animal health, like human health, may have been extracted from data sources that were not designed to answer questions currently held by researchers (Lazer *et al.*, 2014; Chen and Asch, 2017), making it difficult to conduct studies that use big data. This supports the notion that pipelines must be created to 'turn big data into "smart data"' (VanderWaal *et al.*, 2017). Further, large datasets that do

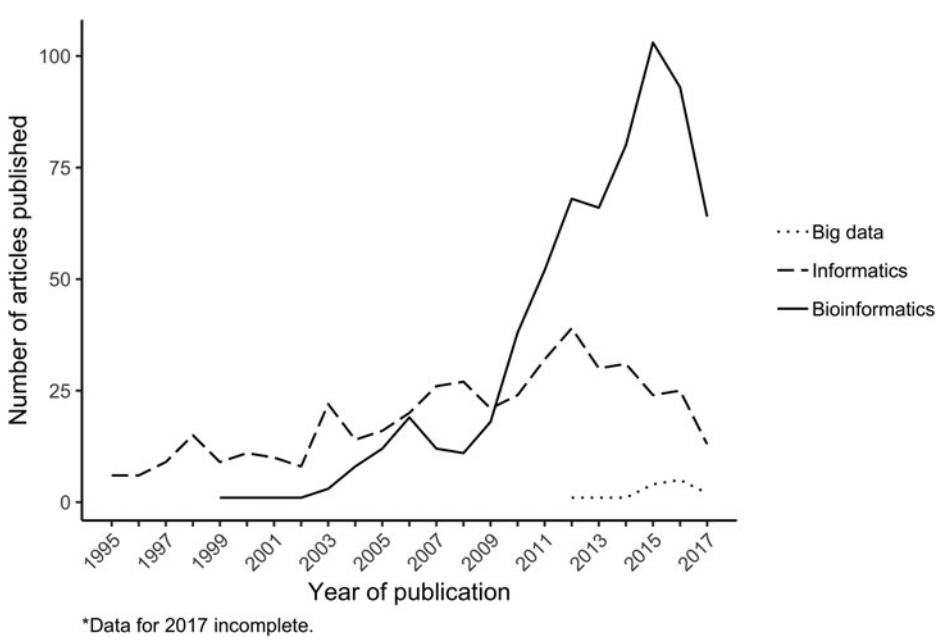

Fig. 2. Frequency of the use of 'big data', 'informatics', and 'bioinformatics' per year.

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 5. List of five primary studies that contain the term 'big data'

Year	Title	Species	Geographic region	First author affiliation	Journal of publication	Study level	Study type	Data sources
2016	Applications of Bayesian phylodynamic methods in a recent U.S. porcine reproductive and respiratory syndrome virus outbreak. (Alkhamis *et al*., 2016)	Pigs	North America	Veterinary medicine and animal health	Biological sciences	• Methodology	• Development or validation of analytical methods.	• Genetic databases
2016	Use of big data in the surveillance of veterinary diseases: early detection of tick paralysis in companion animals. (Guernier *et al*., 2016)	Dogs, cats	Australia/ Oceania	Veterinary medicine and animal health	Biological sciences	• Animal bacteria, virus, parasite, fungus • Methodology	• Hypothesis testing (observational) • Description, development or validation of software product.	• Internet search engines, social media • Non-government organizations
2015	Big data analytics for empowering milk yield prediction in dairy supply chains. (Yan *et al*., 2015)	Dairy cattle	Asia	Social sciences	Statistics and mathematics	• Methodology	• Development or validation of analytical methods.	• No data used.
2015	Big data and the dairy cow: factors affecting fertility in UK herds. (Hudson, 2015)	Dairy cattle	Europe	Veterinary medicine and animal health	Biological sciences	• Methodology	• Hypothesis testing (observational) • Theoretical study (simulation modeling, SIR/mathematical modeling, predictive) • Development or validation of analytical methods.	• Electronic medical records. • No data used.
2016	Evidence in practice – a pilot study leveraging companion animal and equine health data from primary care veterinary clinics in New Zealand. (Muellner *et al*., 2016)	Dogs, cats, horses	Australia/ Oceania	Veterinary medicine and animal health	Veterinary medicine and animal health	• Methodology	• Description, development or validation of software product.	• Electronic medical records

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 6. List of nine reviews, commentaries, editorials, letters-to-the-editor, and conference proceedings that contain the term 'big data'

First author	Year	Title	Other conceptual terms	Species	Geographic region	First author affiliation	Journal of publication	Publication type
Cole	2012	Breeding and genetics symposium: Really big data: Processing and analysis of very large data sets.	Informatics	Dairy cattle, beef cattle	North America	Veterinary medicine and animal health.	Veterinary medicine and animal health.	Conference proceedings.
Greenwood	2014	Consequences of nutrition during gestation, and the challenge to better understand and enhance livestock productivity and efficiency in pastoral ecosystems.		Beef cattle	Australia/ Oceania	Veterinary medicine and animal health.	Veterinary medicine and animal health.	Narrative review.
Hirata	2013	Development of quality control and breeding management system of goats based on information and communication technology.	Informatics	Goats	Asia	Physical sciences.	Computer science and information technology.	Commentary, editorial, letter-to-the-editor.
Hostens	2016	Bovi-analytics: A platform to educate veterinary students. Big data in dairy cows. An initiative to create the veterinary stethoscope version 3.0?		Dairy cattle	Europe	Veterinary medicine and animal health.	Veterinary medicine and animal health.	Conference proceedings.
Kulatunga	2017	Opportunistic wireless networking for smart dairy farming.		Dairy cattle	Europe	Computer science and information technology.	Computer science and information technology.	Commentary, editorial, letter-to-the-editor.
Pang	2016	Veterinary oncology: Biology, big data and precision medicine.	Bioinformatics	Dogs, cats	Europe	Veterinary medicine and animal health.	Veterinary medicine and animal health.	Narrative review.
Tan	2017	Environmental sustainability analysis and nutritional strategies of animal production in China.		Cattle (unspecified), pigs, layer poultry, broiler poultry	Asia	Veterinary medicine and animal health.	Veterinary medicine and animal health.	Narrative review.
Asokan	2015	Leveraging 'big data' to enhance the effectiveness of 'one health' in an era of health informatics.	Informatics	Dogs, cats, horses, cattle (unspecified), sheep, goats, pigs, poultry (unspecified)	Asia	Human.	Human.	Commentary, editorial, letter-to-the-editor.
Deusch	2015	News in livestock research — use of Omics-technologies to study the microbiota in the gastrointestinal tract of farm animals.	Bioinformatics	Cattle (unspecified); sheep, goats, pigs, poultry (unspecified)	Europe	Veterinary medicine and animal health.	Computer science and information technology.	Narrative review.

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 7. Frequency of 'big data', 'informatics', and 'bioinformatics' in 1093 publications in the animal health and veterinary medical literature

	Big data	Informatics	Bioinformatics	Total counts
Primary studies (not including conference proceedings)	5	326	589	920
Systematic review	0	1	0	1
Scoping review	0	1	0	1
Narrative review	4	24	57	85
Commentary, editorial, letter-to-the-editor	3	57	5	65
Conference proceeding	2	29	0	31
Total counts	14	438	651	1103[a]

[a]Exceeds 1093 because articles may contain multiple conceptual terms.

Table 8. General characteristics of 1093 included articles containing terms related to 'big data', 'informatics', and 'bioinformatics' in the animal health and veterinary medicine literature

Category (*n* = number of articles)	Big data (*n* = 14)	Informatics (*n* = 438)	Bioinformatics (*n* = 651)	Total counts[a]
Species				
Dogs (*n* = 185)	4	116	69	189
Cats (*n* = 85)	4	51	34	89
Horses (*n* = 117)	2	58	59	119
Dairy cattle (*n* = 152)	5	74	74	153
Beef cattle (*n* = 116)	2	61	54	117
Cattle (*n* = 192)	3	108	85	196
Sheep (*n* = 227)	2	122	107	231
Goats (*n* = 180)	3	93	88	184
Pigs (*n* = 382)	4	138	244	386
Layer poultry (*n* = 14)	1	2	11	14
Broiler poultry (*n* = 36)	1	10	25	36
Poultry (*n* = 101)	2	65	37	104
Total counts[b]	33	898	887	1818
Geographic region				
North America (*n* = 271)	2	124	148	274
South America (*n* = 57)	0	42	15	57
Europe (*n* = 331)	5	171	157	333
Africa (*n* = 24)	0	19	6	25
Asia (*n* = 362)	4	69	291	364
Australia/Oceania (*n* = 57)	3	20	36	59
Total counts[b]	14	445	653	1112
First author affiliation				
Veterinary medicine and animal health (*n* = 720)	10	245	471	726
Human medicine and health (*n* = 102)	1	58	45	104
Biological sciences (*n* = 176)	0	59	117	176
Bioinformatics (*n* = 16)	0	6	10	16
Physical sciences (*n* = 35)	1	29	6	36
Statistics and mathematics (*n* = 4)	0	2	3	5
Computer science and information technology (*n* = 16)	1	15	0	16
Social sciences (*n* = 26)	1	24	1	26

(*Continued*)

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 8. (*Continued.*)

Category (*n* = number of articles)	Big data (*n* = 14)	Informatics (*n* = 438)	Bioinformatics (*n* = 651)	Total counts[a]
Total counts[b]	14	438	653	1105
Journal of publication				
Veterinary medicine and animal health (*n* = 351)	6	199	150	355
Human medicine and health (*n* = 79)	1	48	31	80
Biological (*n* = 481)	3	104	377	484
Bioinformatics (*n* = 87)	0	6	81	87
Physical sciences (*n* = 28)	0	25	3	28
Statistics and mathematics (*n* = 8)	1	2	5	8
Computer science and information technology (*n* = 52)	3	49	2	54
Social sciences (*n* = 3)	0	3	0	3
Total counts[b]	14	436	649	1099

[a]May exceed *n* because articles may contain multiple conceptual terms.
[b]May exceed 1093 because articles may have been classified into multiple categories.

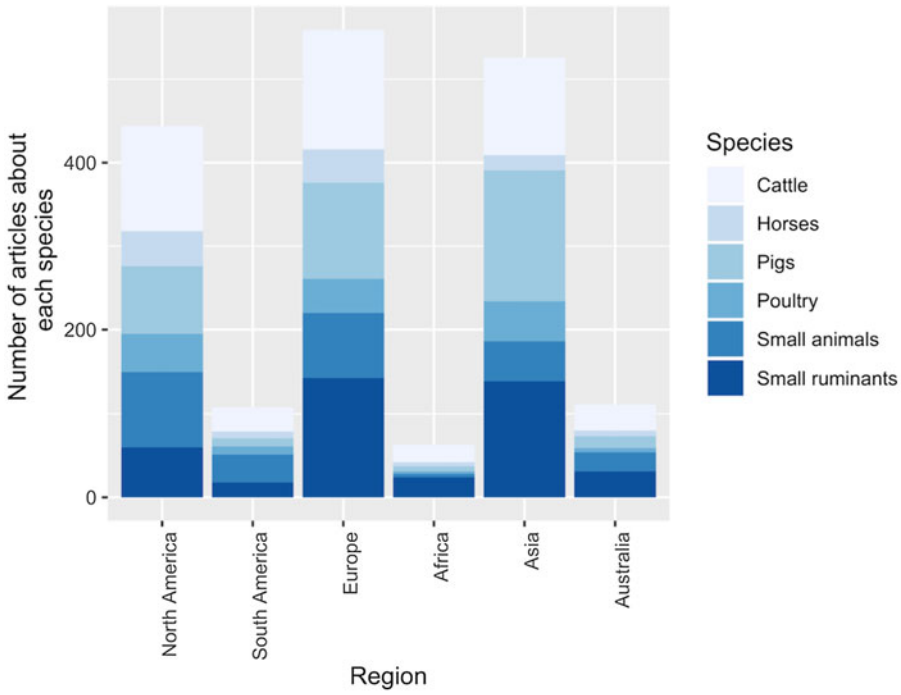

Fig. 4. Number of articles about each species, by geographic region.

Table 9. Number of 'big data' or 'informatics' articles versus 'bioinformatics' articles, by geographic region, for studies related to swine populations

Region	Big data or informatics	Bioinformatics	Total
North America	27	54	81
South America	8	2	10
Europe	57	58	115
Africa	5	1	6
Asia	34	123	157
Australia/ Oceania	7	7	14

exist may contain meaningful information that does not answer predefined research questions. Unsupervised machine learning and pattern recognition algorithms may shed light on what is hidden in these datasets, by revealing patterns that were not expected. Such methodologies may be relatively new to animal health and veterinary medicine. Finally, big datasets may simply be difficult for researchers to acquire.

'Informatics' studies tend to use a variety of data sources, such as 'geospatial information systems', government databases, scientific literature databases, and electronic medical/production records, that have been described as being or becoming big data (VanderWaal *et al.*, 2017). Remote sensing technologies have existed in dairies since the 1980s, which would explain the large

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Table 10. Data classification of 918 primary studies into study level, study type and data sources

Category (*n* = number of articles)	Big data (*n* = 5)	Informatics (*n* = 326)	Bioinformatics (*n* = 589)	Total counts[a]
Study level				
Animal (*n* = 82)	0	67	15	82
Animal genes, proteins, metabolites (*n* = 354)	0	9	345	354
Animal product or by-product (*n* = 49)	0	15	34	49
Animal bacteria, virus, parasite, fungus (*n* = 134)	1	121	13	135
Genes of animal bacteria, virus, parasite, fungus (*n* = 173)	0	5	168	173
Effects of animals on environment (*n* = 37)	0	35	2	37
Software, analytical technique, lab technique development/validation study (*n* = 136)	5	87	45	137
Total counts[b]	6	339	622	967
Study type				
Descriptive (*n* = 299)	0	41	258	299
Hypothesis testing (experimental) (*n* = 141)	0	5	136	141
Hypothesis testing (observational) (*n* = 336)	2	168	166	336
Theoretical study (simulation-modeling) (*n* = 24)	1	21	2	24
Development of validation of laboratory methods (*n* = 35)	0	0	35	35
Comparison of laboratory methods (*n* = 3)	0	0	3	3
Development or validation of analytical methods (*n* = 66)	3	53	11	67
Comparison of analytical methods (*n* = 8)	0	4	4	8
Description, development or validation of software product (*n* = 48)	2	42	5	49
Comparison of software product (*n* = 5)	0	5	0	5
Total counts[b]	8	339	620	967
Data sources				
Biologic samples (*n* = 662)	0	126	536	662
Genetic databases (*n* = 240)	1	5	234	240
Electronic medical records (*n* = 36)	2	35	0	37
Farm production records (*n* = 27)	0	23	4	27
Internet search engines, social media (*n* = 4)	1	3	0	4
Scientific literature databases (*n* = 28)	0	19	9	28
Geographic (measured by researchers) (*n* = 36)	0	36	0	36
Climate, weather, plant life, soil (*n* = 35)	0	35	0	35
Government-sourced (*n* = 454)	0	154	301	455
Non-government-sourced (*n* = 31)	1	28	2	31
Wearables, sensors, electronic identification (*n* = 14)	0	14	0	14
Questionnaire, surveys (*n* = 71)	0	68	3	71
No data used (*n* = 14)	2	12	0	14
Total counts[b]	7	558	1089	1654

[a]May exceed *n* because articles may contain multiple conceptual terms.
[b]Total may exceed 918 because articles may have been classified into multiple categories.

number of cattle studies classified as 'informatics' studies (Rutten *et al.*, 2013).

Despite an overlap in the definitions of 'informatics' and 'bioinformatics', there is a strong distinction in the literature. 'Bioinformatics' studies were about genes, amino acids, and proteins while 'informatics' studies were about an organism or pathogen (e.g. animal, bacteria, and virus). 'Bioinformatics' studies also tended to about laboratory techniques while 'informatics' studies tended to be about analytical techniques and software. Bioinformatic laboratory techniques may contain an analytical

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

component; however, if this was not stated explicitly, the study was not classified as being about analytical techniques. Genetic datasets (including genomic and metagenomic datasets) are often considered large, and multiple sources of data may be used (e.g. biological samples, government databases). However, once collected for a research study, the genetic dataset does not change. This lack of velocity may explain why most 'bioinformatics' articles do not use the term 'big data'.

Limitations

The literature search was limited to the conceptual terms 'big data', 'informatics', and 'bioinformatics'. A more complete picture of the concepts of big data and informatics may require a search of a larger list of terms. For instance, articles describing studies that used big data may be better identified by the names of analytical techniques designed specifically for big data. Similarly, many articles about informatics or big data may have been excluded for not using those specific words. Research conducted using data sources such as animal industry datasets (e.g. performance, health, and breeding records) as well as data from animal (and human) health surveillance systems may be relevant to 'informatics' research. Further, searches using words such as 'robotic milkers', 'wearable sensors', and 'electronic medical records' may also have provided articles relevant to 'informatics'. Although the search yielded a large number of publications, it is possible that the search would have been more complete by including these terms in the search. The authors began with a literature search with a larger list of conceptual terms; however, the number of articles returned was extremely large (data not shown).

The literature search was limited to English abstracts. Articles with English abstracts but non-English full-text were excluded from the study. Articles that used the terms 'big data', 'informatics', and 'bioinformatics' in non-English languages would not have been captured potentially biasing the study.

Conclusions

'Big data' was an uncommon term. 'Bioinformatics' was the most common term. There were more 'informatics' articles about small animals and livestock with unspecified production systems (e.g. cattle, poultry) than 'bioinformatics' articles. A large number of 'pig' articles contributed to 'bioinformatics' studies.

All geographic regions produced literature using the terms 'informatics' or 'bioinformatics'. Two geographic regions (South America, Africa) did not produce literature using the term 'big data'. Asia produced the most literature using the term 'bioinformatics'. Articles about pigs contributed heavily to the 'bioinformatics' articles from Asia.

While most articles had first author affiliations in 'veterinary medicine and animal health', a higher proportion of 'informatics' articles had affiliations that were not veterinary/animal, medical/health or biologically related. 'Big data' and 'informatics articles' were more often published in 'veterinary medicine and animal health' journals. 'Bioinformatics' articles were more often published in 'biological' journals.

'Bioinformatics' studies tended to be conducted at the gene level. 'Informatics' studies tended to be conducted at the 'animal' or 'animal bacteria, virus, parasite, fungus' level. 'Informatics' studies also tended to examine analytical techniques and software. 'Bioinformatics' studies tended to examine laboratory techniques. 'Informatics' studies were often observational. Experiments were more common in 'bioinformatics' studies.

'Bioinformatics' studies used biologic samples, genetic databases, and government databases. 'Informatics' studies used a wider variety of data sources (e.g. 'electronic medical records', 'farm production records', 'scientific literature databases', 'geographic', 'wearables, sensors, electronic identification').

The definition of big data has evolved rapidly and should be taken into account when describing research. As big data research is more common in human medicine, it may serve as a model for researchers in animal health and veterinary medicine. Techniques such as unsupervised machine learning and pattern recognition algorithms may uncover unrecognized associations within big datasets.

Finally, as with any study, it is important to focus resources on collecting and analyzing data in a way that meets the research objectives.

Acknowledgements. This research was undertaken thanks in part to:

- IDEXX Laboratories,
- Funding from the Canada First Research Excellence Fund through the Food from Thought program at the University of Guelph,
- International Graduate Tuition Scholarships at the University of Guelph,
- International Doctoral Tuition Scholarships at the University of Guelph,
- The Natural Sciences and Engineering Research Council's Undergraduate Summer Research Awards, and
- Undergraduate Research Assistantships at the University of Guelph.

We thank Erin McGill, Vivienne Steele and Inthuja Selvaratnam for their assistance with data extraction.

References

Alkhamis MA, Perez AM, Murtaugh MP, Wang X and Morrison RB (2016) Applications of Bayesian phylodynamic methods in a recent U.S. Porcine reproductive and respiratory syndrome virus outbreak. *Frontiers in Microbiology* **7**, 67.

Andersson LM, Okada H, Miura R, Zhang Y, Yoshioka K, Aso H and Itoh T (2016) Wearable wireless estrus detection sensor for cows. *Computers and Electronics in Agriculture* **127**, 101–108.

Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC and Yang G-Z (2015) Big data for health. *Ieee Journal of Biomedical and Health Informatics* **19**, 1193–1208.

Arksey H and O'Malley L (2005) Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* **8**, 19–32.

Chen JH and Asch SM (2017) Machine learning and prediction in medicine – beyond the peak of inflated expectations. *New England Journal of Medicine* **376**, 2507–2509.

Elgendy N and Elragal A (2014) Big data analytics: a literature review paper. *Industrial Conference on Data Mining*, 214–227. doi: 10.1007/978-3-319-08976-8_16.

Gaitanou P, Garoufallou E and Balatsoukas P (2014) The effectiveness of big data in health care: a systematic review. *Research Conference on Metadata and Semantics Research*, 141–153. doi: 10.1007/978-3-319-13674-5_14.

Guernier V, Milinovich GJ, Bezerra Santos MA, Haworth M, Coleman G and Soares Magalhaes RJ (2016) Use of big data in the surveillance of veterinary diseases: early detection of tick paralysis in companion animals. *Parasites & Vectors* **9**, 303.

Haladjia J, Haug J, Nüske S, Bruegge B, Haladjian J, Haug J, Nüske S and Bruegge B (2018) A wearable sensor system for lameness detection in dairy cattle. *Multimodal Technologies and Interaction* **2**, 27.

Hudson C (2015) Big data and the dairy cow: factors affecting fertility in UK herds', *PQDT – UK & Ireland*. Ann Arbor: The University of Nottingham (United Kingdom).

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.

Huerta M, Downing G, Haseltine F, Seto B and Liu Y (2000) NIH working definition of bioinformatics and computational biology. *Biomedical Information Science & Technology Initiative and National Institutes of Health*.

Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, Kalet IJ, Lenert LA, Musen MA, Ozbolt JG, Smith JW, Tarczy-Hornoch PZ and Williamson JJ (2012) AMIA board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *Journal of the American Medical Informatics Association : JAMIA* **19**, 931–938. doi: 10.1136/amiajnl-2012-001053

Laney D (2001) 3D data management: controlling data volume, velocity and variety. *META group research note* **6**, 1.

Lazer D, Kennedy R, King G and Vespignani A (2014) The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205.

Muellner P, Muellner U, Gates MC, Pearce T, Ahlstrom C, O'Neill D, Brodbelt D and Cave NJ (2016) Evidence in practice – A pilot study leveraging companion animal and equine health data from primary care veterinary clinics in New Zealand. *Frontiers in Veterinary Science*. 2017/01/10. Epi-interactive Ltd., Wellington, New Zealand. Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand. Equine Health Association, Wellington, New Zealand. The Royal Veterinary College, Hatfield, U. 3, 116.

Natarajan P, Frenzel JC and Smaltz DH (2017) *Demystifying Big Data and Machine Learning for Healthcare*. Boca Raton: CRC Press.

Peacock L (2012) 'The Equine Distress Monitor Project', *All Graduate Plan B and other Reports*. (Accessed 26 September 2018).

Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D and Soares CB (2015) Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare* **13**, 141–146.

Rutten CJ, Velthuis AGJ, Steeneveld W and Hogeveen H (2013) Invited review: sensors to support health management on dairy farms. *Journal of Dairy Science* **96**, 1928–1952. doi: 10.3168/jds.2012-6107

Sagiroglu S. and Sinanc D. (2013) Big data: A review, in 2013 *International Conference on Collaboration Technologies and Systems* (*CTS*). IEEE, pp. 42–47. doi: 10.1109/CTS.2013.6567202.

Schroeck M, Shockley R, Smart J, Romero-Morales D and Tufano P (2012) Analytics: the real-world use of big data: how innovative enterprises extract value from uncertain data. Executive report. *IBM Institute for Business Value and Said Business School at the University of Oxford*.

Smith RD and Williams M (2000) Applications of informatics in veterinary medicine. *Bulletin of the Medical Library Association* **88**, 49–55.

Thompson CJ, Luck LM, Keshwani J, Pitla SK and Karr LK (2018) Location on the body of a wearable accelerometer affects accuracy of data for identifying equine gaits. *Journal of Equine Veterinary Science* **63**, 1–7.

VanderWaal K, Morrison RB, Neuhauser C, Vilalta C and Perez AM (2017) Translating big data into smart data for veterinary epidemiology. *Frontiers in Veterinary Science* **4**, 110.

Yan WJ, Chen X, Akcan O, Lim J and Yang D (2015) Big data analytics for empowering milk yield prediction in dairy supply chains. *Proceedings 2015 IEEE International Conference on Big Data*. IEEE, pp. 2132–2137. doi: 10.1109/BigData.2015.7363997

Ylijoki O and Porras J (2016) Perspectives to definition of big data: a mapping study and discussion. *Journal of Innovation Management* **4**, 69–91.

Downloaded from https://www.cambridge.org/core. 05 Mar 2022 at 03:52:22, subject to the Cambridge Core terms of use.