Leveraging unified big data and AI: insights from VetCompass

Noel Kennedy nkennedy@rvc.ac.uk

Senior Research Fellow in AI

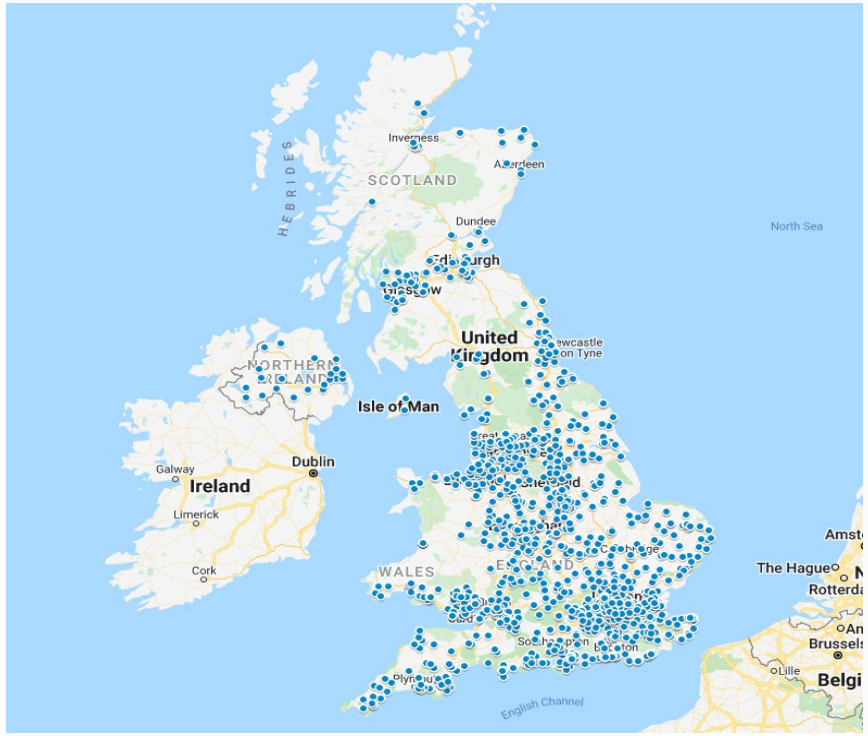
The Royal Veterinary College

Veterinary Healthcare Landscape

- First opinion
- Charity
- Referral centres
- Emergency centres
- University teaching hospitals
- Farm / large animal

VetCompass UL

Clinic locations



- Epidemiological animal welfare research programme
- > 100 peer reviewed papers
- Analogous to CPRD for veterinary
- Around 30% of vet clinics in UK contribute data
  - > 1,800 practices
  - Primary care, referral, charity, out of hours, emergency
- >21 million animals
- Multiple practice systems

RVC

What is the data like

- Nearly all free text notes
- Typed by clinician or receptionist
- Data comprises
  - Patient metadata – DoB, DoD, breed/species
  - Clinical notes
  - Billed items – drugs, treatments, food etc
  - Structured fields – body weights, BCS, test results

- Almost no clinical coding is performed in consult room, or even during visit

Data feed overview

- Export side
  - Nightly data feeds
  - File format
  - SFTP / HTTPS
- Import side
  - Pseudonymised
  - Cleaned / normalized
  - Indexed via ElasticSearch

File format – Flat file .csv

- Pros
  - Simple to export
  - All systems can use
- Cons
  - Nightmare to import!
  - Ambiguous string delimiters
  - Ambiguous row delimiters
  - No ability to validate file integrity
  - No referential integrity
  - No data typing integrity
  - Hard to debug

File format – XML

- Pros
  - Many systems can use
  - File integrity
  - Data type integrity
  - Some referential integrity
- Cons
  - Nightmare to import!
  - Many developers find xml very confusing
  - Big overheads
  - Hard to debug

File format – .bak

- Pros
  - Most groups use SQL Server
  - File integrity
  - Data type integrity
  - Full referential integrity
  - Space efficient
- Cons
  - Restricted to SQL server

Data transmission

- Two options
  - SFTP
  - HTTPS
- Both encrypted in transit and at rest

Data import

- Data validated and imported into staging DB
- Data enriched – breed/species normalized
- Further anonymization

What is the data like - example

NOTE

Reason: Booster Vaccination and Health Exam

Appointment Notes:
Tartar build up on upper molars, rest nad, BAR

Passed very diluite urine in the consult room

Booster and KC given

Expressed cncern over very diluite urine

Offered bloods - O will consider

Line items

Doc:Vaccination Certificate

Nobivac Lepto 4 (50 box)

Nobivac KC (25 box)

Vacc 1yr KC/ Bordetella / Kennel Cough

VAR-Vaccinations & Health Check

Vacc 1yr Lepto

VAR-Vaccinations & Health Check

Doc:Patient Report Card Dog

VetCompass animal welfare research

- Disease studies
  - Hypoadrenocortism, skin fold dermatitis, seizures etc
  - Risk factors for diagnosis
  - Characterising presentation
  - Describing treatments
  - Survival

- Breed/species studies
  - Prevalence/incidence
  - ‘What kinds of diseases do breeds get compared with other dogs/cats’

- Healthcare/treatment focused
  - Antimicrobial usage
  - Reasons for prescribing

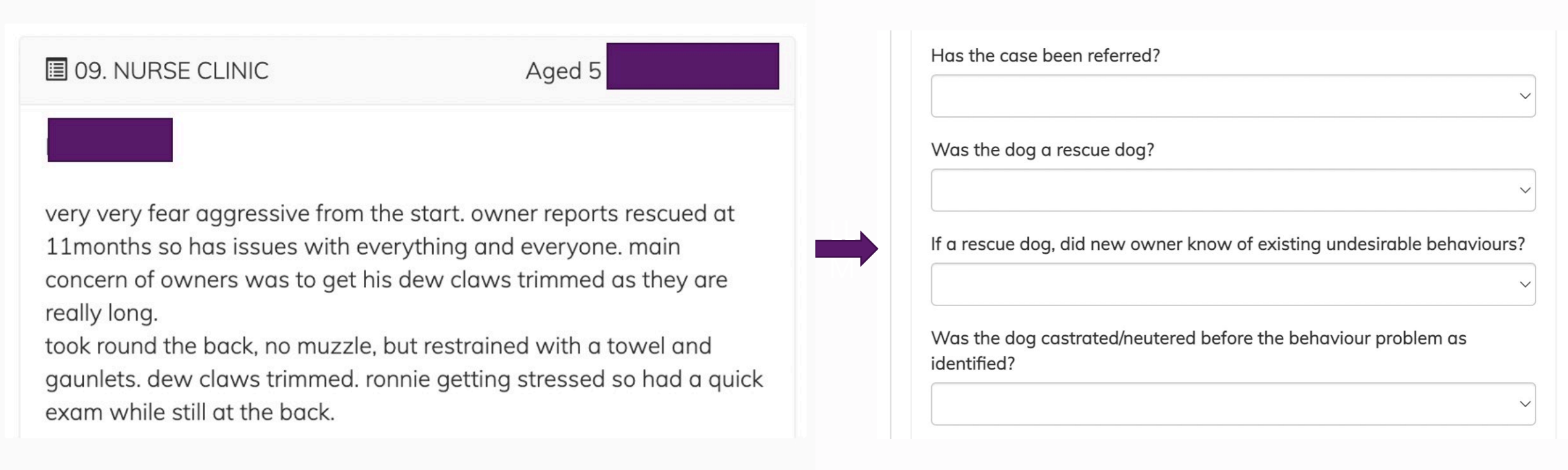
- 100+ research papers:

https://www.rvc.ac.uk/vetcompass/papers-and-data/original-publications

How do researchers use the data?

- Data is free text
- Researchers require structured (eg categorical) data
- Structured data used to produced descriptive stats:
  - ‘Among the 464/3,308 (14.03%) Yorkshire Terriers that died during the study period, the median age at death was 13.56 years’ [1]
  - ‘Dogs prescribed pimobendan lived longer (adjusted mean survival time 1051 days, 95% CI 967–1125) than dogs not prescribed pimobendan (905 days, 95% CI 871–940 days)’ [2]

Researcher manual annotation



Big data = big problems

- Denominator populations have millions of animals
- Researchers can only read a few thousand
- And that takes a few weeks
- Most patient records unread - rarer outcomes less understood
- Some factors are too time-consuming to examine manually eg co-morbidities

Can we use an LLM?

- LLM as research assistant
- Machine can read faster
- Not a chatbot
- Batch mode
- Conceptually:
  - Input = Researcher question + EHR of a patient
  - Output = Answer for that patient

Can we use an LLM?

- Researcher asks LLM the questions
- *LLM* answers the questions:

Canine Chronic kidney disease study

- First VetCompass study with two sites
  - VetCompass Australia – University of Sydney, University of Queensland
  - VetCompass UK – Royal Veterinary College London
- Retrospective cohort
- Same methods + case definition
- Researchers at both sites study equivalent cohorts
- Aim: “facilitate diagnosis and case management of dogs with CKD to improve animal welfare”
  - Incidence/prevalence
  - Disease progression
  - Case management
  - Survival factors

How well does LLM work?

- Model [3] : deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
- Qwen2.5 14B parameter base model
- Data: Questions + Answers binarized
- Gold standard is researcher’s answer

CKD - Case finding

Question	Total Patients	True CKD Patients	Error	F1	PPV	Sensitivity
Was the dog diagnosed with chronic kidney disease (CKD) or synonym? + {list of synonyms}	242	49	3	0.6	0.6	0.59
Was the dog diagnosed with chronic kidney disease (CKD) or synonym? + {list of synonyms} + {inclusion criteria} + {exclusion criteria}	242	49	2	0.68	0.65	0.71

Periodontal disease - Case categorization

Question	R^2	Error	F1	PPV	Sensitivity
What was the age of this patient when it first met the inclusion criteria for periodontal disease?	0.99	33			
Is there evidence recorded of diagnostic methods used to contribute to diagnosis of this periodontal disease?		1	0.3	1.0	0.17
Was there visual inspection of the patient's mouth whilst the patient was conscious?		12	0.82	0.99	0.7
Was there evidence of gingival recession in the patient's mouth?		0	0.61	0.47	0.86
After the initial diagnosis of periodontal disease, was the patient re-examined for periodontal disease?		0	0.32	0.25	0.43
Was a referral to a dental specialist discussed or carried out?		0	0.33	0.5	0.25

Future directions

- Further prompt engineering
- Waiting for annotations to complete to get a training set
- AI will label around 30k cases

Acknowledgements

- VetCompass Australia
  - Aruna Shrestha, Mary Thompson, Ricardo Magalhaes,
- VetCompass UK
  - Maddie Matin, Rebecca Edwards, Camilla Pegram, Rosanne Jepson, Jonathan Elliot, Dan O'Neill, Dave Brodbelt
- Zoetis
  - Andrea Wright
- James Hoontrakul and IT colleagues
- All practices participating in VetCompass

References

[1] O'Neill, Dan G., Sara D. Witkowska, Dave C. Brodbelt, David B. Church, and Karolina S. Engdahl. 2025. 'Yorkshire Terriers under Primary Veterinary Care in the UK – Demography and Disorders'. *Companion Animal Health and Genetics* 12 (1): 6. https://doi.org/10.1186/s40575-025-00145-y.

[2] Pegram, Camilla, Karla Diaz-Ordaz, Dave C. Brodbelt, et al. 2025. 'Emulating the EPIC Trial Using VetCompass Primary-Care Data: Causal Effects of Pimobendan in UK Dogs with Grade IV/VI Heart Murmurs'. *PLOS ONE* 20 (6): e0325695. https://doi.org/10.1371/journal.pone.0325695.

References

[3] Guo, Daya, Dejian Yang, Haowei Zhang, et al. 2025. ‘DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning’. *Nature* 645 (8081): 633–38. https://doi.org/10.1038/s41586-025-09422-z.