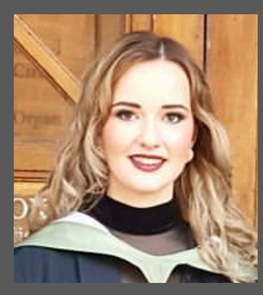
Corresponding Author:
Charlotte. S. C. Woolley
PhD Student
The Roslin Institute,
University of Edinburgh, UK
Email: s1576473@sms.ed.ac.uk



Acknowledgements
The Roslin Institute is supported by an Institute Strategic Programme Grant from the BBSRC (BB/J004235/1)

A novel, broadly applicable data cleaning algorithm for growth

Charlotte S. C. Woolley, Ian G. Handel, B. Mark Broonsvoort, Jeffrey J. Schoenebeck, Dylan N. Clements
The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK

Background

- Data cleaning is vital to preserve data quality and statistical power
- Few studies document the data cleaning process reproducibly
- Data cleaning Labrador retriever (LR) growth data is particularly problematic as it is very heterogeneous

Alfie; The UK's fattest Labrador, weighing 12st 5lb (80kg)[1]

Objectives

- To develop a novel data cleaning algorithm for growth, applied to LRs

Methods

- Four different LR growth datasets tested
- Non-linear mixed effects models applied to data in R
- Prediction intervals estimated to identify 'outliers'
- Algorithm uses rule-based decisions to clean

Step 1
- Remove complete duplications (where the animal ID, date of data entry and measurement are identical) by deleting the most recent duplicate(s)

Step 2
- Remove other duplications (where the animal ID and date of data entry are identical) by deleting the duplicate(s) that has the largest difference from the predicted measurement

Step 3
- Replace outliers with the closest correction (dividing or multiplying by 10/100/1000, subtracting or adding 100/1000, multiplying by the imperial/metric unit or transposing) to the predicted measurement

Step 4
- Remove outliers that jump in size in comparison with the largest predicted size change between the previous or the following measurements

Step 5
- Remove biologically implausible entries based on reported values for this species/breed

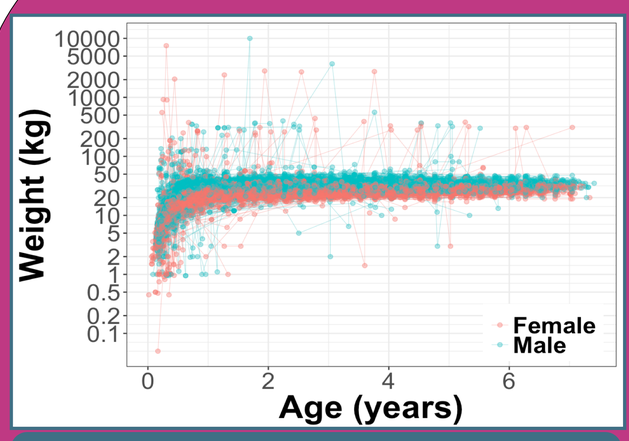
Conclusions

This algorithm:
- Identifies duplicate entries, typing, decimal point, unit and measurement errors
- Avoids modifying unusual but biologically plausible values
- Allows individuals to differ from the population
- Prioritises data repair over removal - fewer deletions
- Performs well in comparison with other data cleaning protocols
- Could be easily adapted for use in other breeds, species and fields

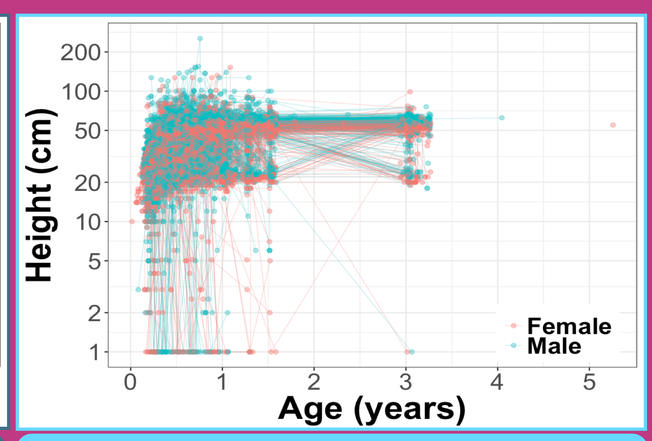
References

1. Mail Online. A too-many-sausages dog! Giant Alfie the Labrador is put on a crash diet after he tipped the scales at 12.5 STONE. (2012) Available from: http://www.dailymail.co.uk/news/article-2198235/Fattett-dog-Giant-Alfie-Labrador-crash-diet-tipped-scales-12-5-STONE.html. [Accessed 27/02/2018].
2. Clements, D. N. *et al.* Dogslife: a web-based longitudinal study of Labrador Retriever health in the UK. *BMC Vet. Res.* **9,** 13 (2013).
3. Jones, P. H. *et al.* Surveillance of diarrhoea in small animal practice through the Small Animal Veterinary Surveillance Network (SAVSNET). *Vet. J.* **201,** 412–418 (2014).

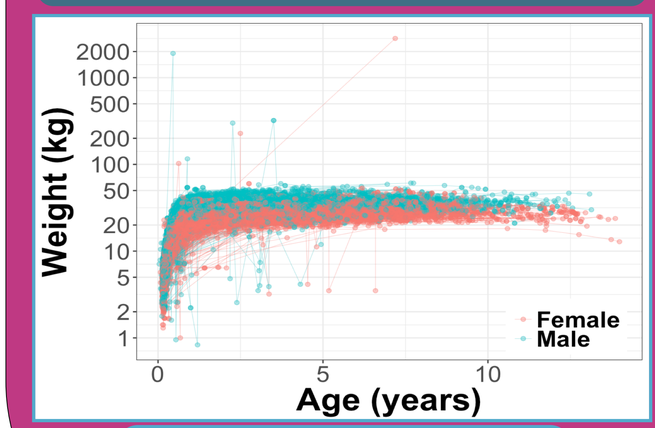
Uncleaned data



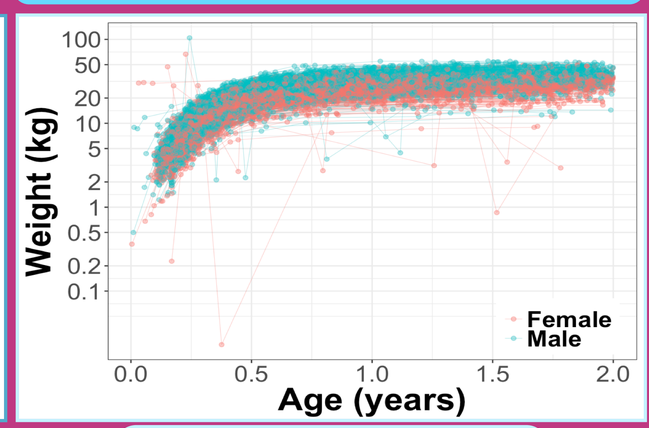
Uncleaned owner-reported weights of LRs by age from Dogslife[2] questionnaires



Uncleaned owner-reported heights of LRs by age from Dogslife[2] questionnaires

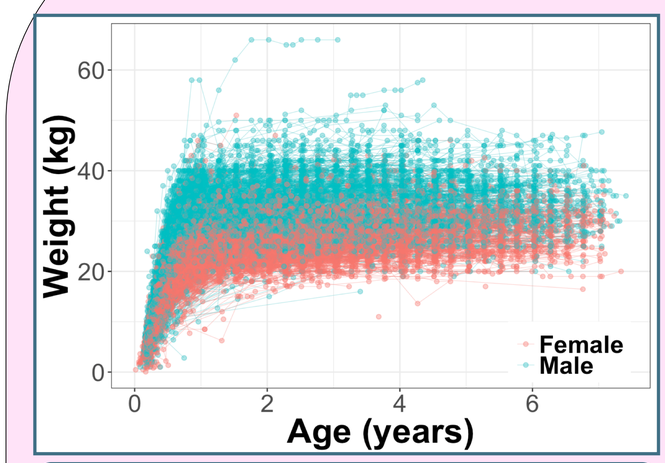


Uncleaned consultation weight records of LRs by age from SAVSNET[3]

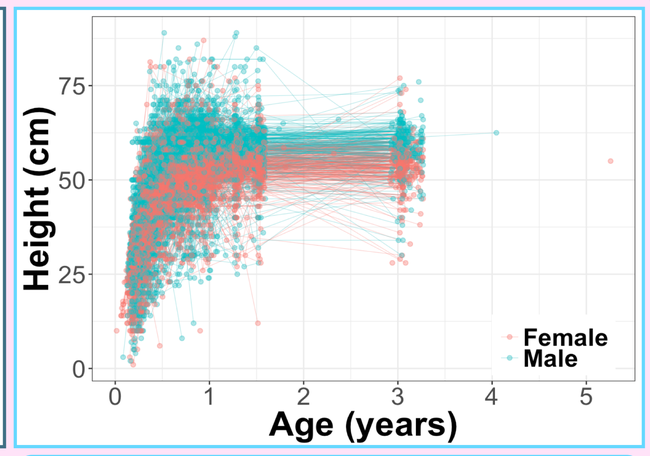


Uncleaned clinical weight records of LRs by age from a veterinary hospital network

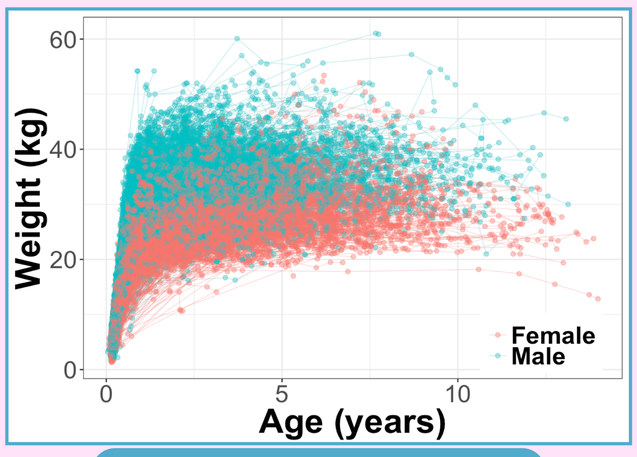
Cleaned data



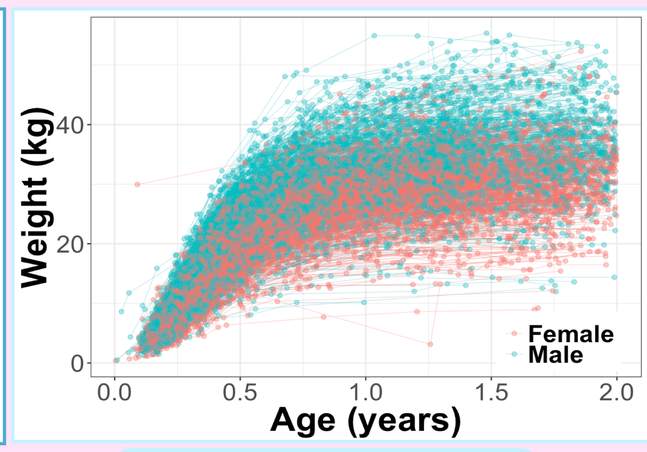
Cleaned owner-reported weights of LRs by age from Dogslife questionnaires



Cleaned owner-reported heights of LRs by age from Dogslife questionnaires



Cleaned consultation weight records of LRs by age from SAVSNET



Cleaned clinical weight records of LRs by age from a veterinary hospital network