Data to Knowledge: A Personal Journey

11 July 2024

Jing Ma

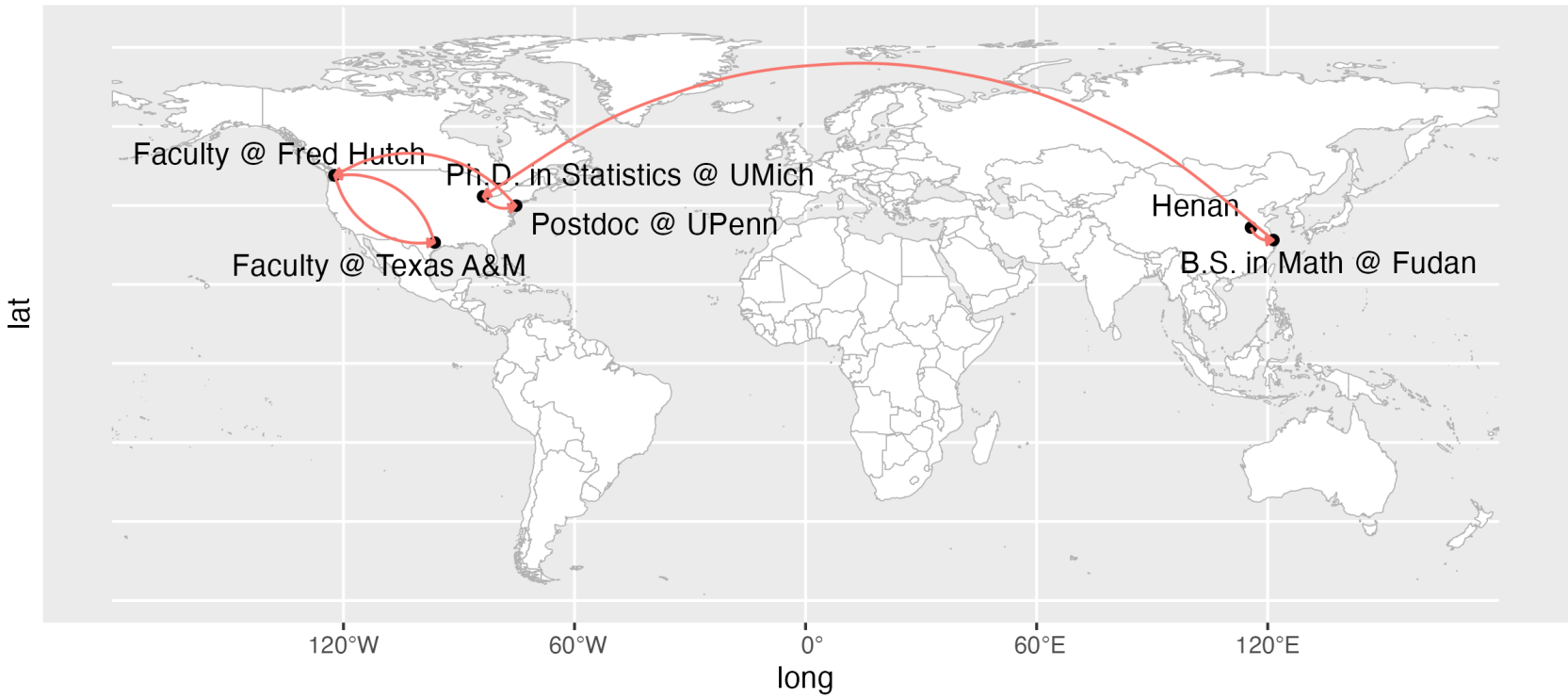
Associate Professor of Biostatistics
Division of Public Health Sciences

Outline

Fred Hutch Cancer Center

- My Journey
- Intro to Dog Aging Project
- Comorbidity Networks in Companion Dogs
- Closing Remarks

My Journey


Faculty @ Fred Hutch
Ph.D. in Statistics @ UMich
Postdoc @ UPenn
Faculty @ Texas A&M
Henan
B.S. in Math @ Fudan
lat
120°W
60°W
0°
60°E
120°E
long

Where I Grew Up

- Henan is the cradle of Chinese civilization
- 2 UNESCO World Heritage Sites (Longmen Grottoes and Anyang Yinxu)

Where I Grew Up

- Henan is the cradle of Chinese civilization
- 2 UNESCO World Heritage Sites (Longmen Grottoes and Anyang Yinxu)
- Population size >99m (2020)

Where I Grew Up

- Henan is the cradle of Chinese civilization
- 2 UNESCO World Heritage Sites (Longmen Grottoes and Anyang Yinxu)
- Population size >99m (2020)
- Breadbasket of China

From Farmer to Scientist

Fred Hutch Cancer Center

From Farmer to Scientist

Pursuit of Math & Statistics

Fred Hutch Cancer Center

Gaokao (China's National College Entrance Exam)

- 8.8m examinees in 2006 (780k in Henan)
- Had to choose school and declare major shortly after the exam

Pursuit of Math & Statistics

Fred Hutch Cancer Center

Gaokao (China's National College Entrance Exam)

- 8.8m examinees in 2006 (780k in Henan)
- Had to choose school and declare major shortly after the exam

Interest

- I have always been good at and enjoyed math.

Opportunity

- Fudan has one of the best mathematics programs.
- Career opportunities are unlimited for math majors...
- U of Michigan has one of the best statistics programs.
- Career opportunities for stat PhDs are better.

From Statistics to Biostatistics

Fred Hutch Cancer Center

Both emphasize on fundamental statistical principles and have the prospect of working in academia & industry.

Statistics

- More theoretical
- More independent
- Broader application areas

Biostatistics

- More time on data exploration
- More opportunities to collaborate
- Applications in public health

From Statistics to Biostatistics

Fred Hutch Cancer Center

Both emphasize on fundamental statistical principles and have the prospect of working in academia & industry.

Statistics

- More theoretical
- More independent
- Broader application areas

Biostatistics

- More time on data exploration
- More opportunities to collaborate
- Applications in public health

My transition: communication is key!

Academia vs Industry

Fred Hutch Cancer Center

Academia

- More competitive
- More time demanding, tenure pressure, & less pay
- Higher flexibility
- Your own 'startup'

Industry

- Higher pay
- Clear work schedule, but less flexibility
- Flexible locations

My Job

Fred Hutch Cancer Center

Fred Hutch is an independent, nonprofit organization, that also serves as UW Medicine's cancer program.

Research Institutes

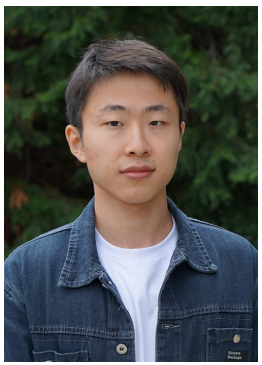
- Independent research
- Collaborative research
- Mentoring students

Universities

- Independent research
- Collaborative research
- Mentoring students
- Teaching

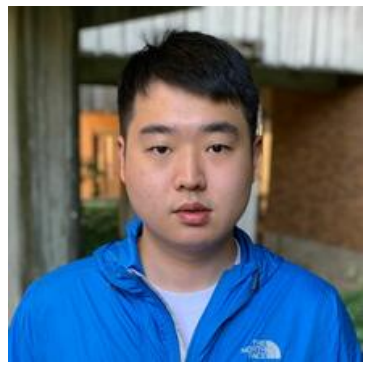
Ma Lab

We develop statistical methods to study the human microbiome and aging.



Dwight Xu

Xinyi Xie



Wenjie Guan

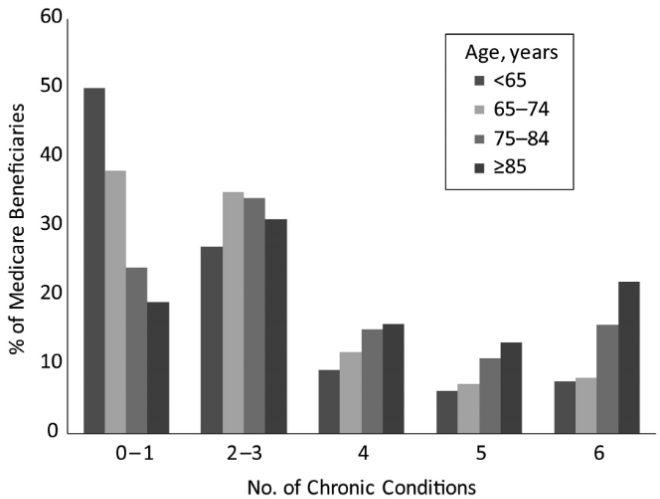
Antoinette Fang

Paizhe Xie

Aging & Comorbidities

Age is the single biggest risk factor of all diseases!

Figure 1. Percentage of the US population enrolled in the Medicare program, by number of chronic conditions and age ...

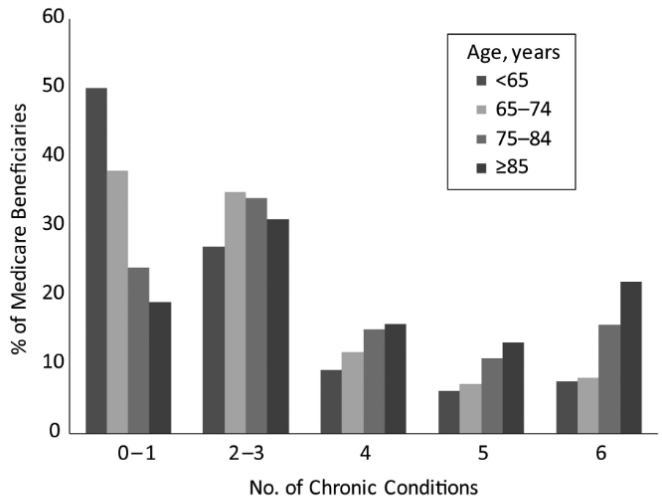


Epidemiol Rev, Volume 35, Issue 1, 2013, Pages 75–83, https://doi.org/10.1093/epirev/mxs009

Aging & Comorbidities

Age is the single biggest risk factor of all diseases!

Figure 1. Percentage of the US population enrolled in the Medicare program, by number of chronic conditions and age ...



Epidemiol Rev, Volume 35, Issue 1, 2013, Pages 75–83, https://doi.org/10.1093/epirev/mxs009

Scientific Questions

How does age impact comorbidity? Comorbidity refers to the presence of two health conditions in one subject.

Why Studying Comorbidities

Fred Hutch Cancer Center

- Comorbidities can offer insights into disease associations and progression, which can in turn improve predictive and diagnostic tools in the clinic.
- Comorbidity networks can help us better understand the underlying mechanism connecting various diseases.

Model Organism for Human Aging

Fred Hutch Cancer Center

The companion dogs are a good model organism in which to study aging and comorbidity.

- Dogs share many morbidities and environment with humans.
- Dogs have an almost equally sophisticated healthcare system.
- Dogs have a shorter lifespan.

Dog Aging Project

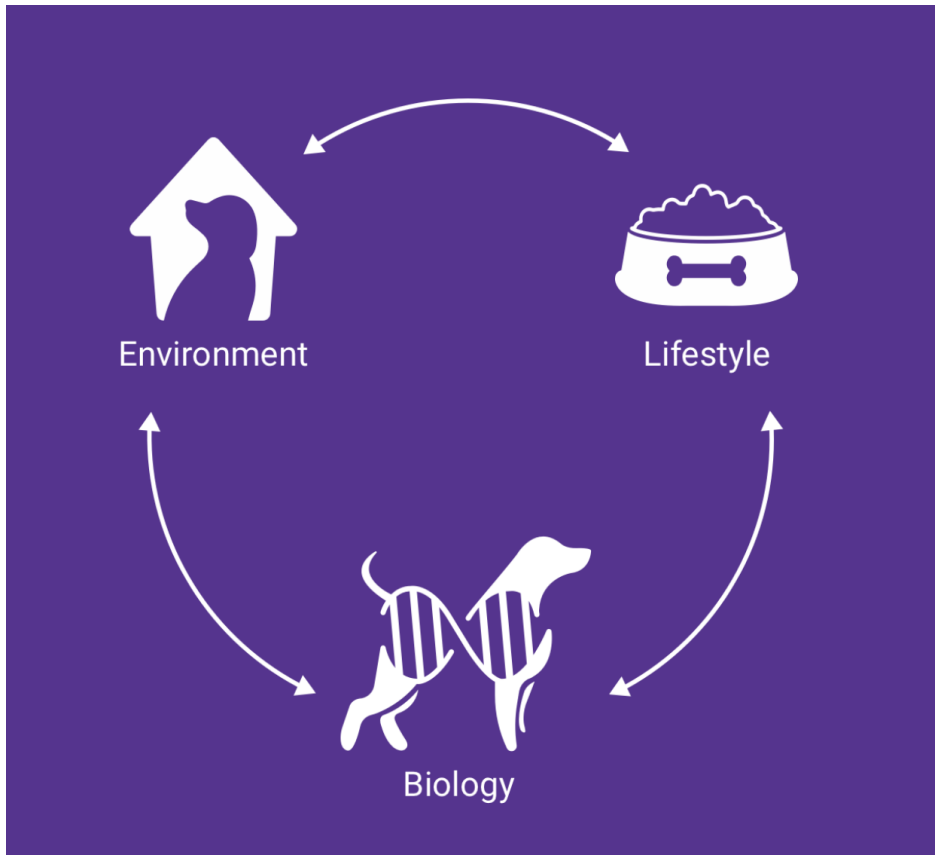


Figure 6: Dog Aging Project is a nationwide, long-term, longitudinal study on the biological, environmental, and lifestyle determinants of healthy aging in companion dogs.

Dog Aging Project

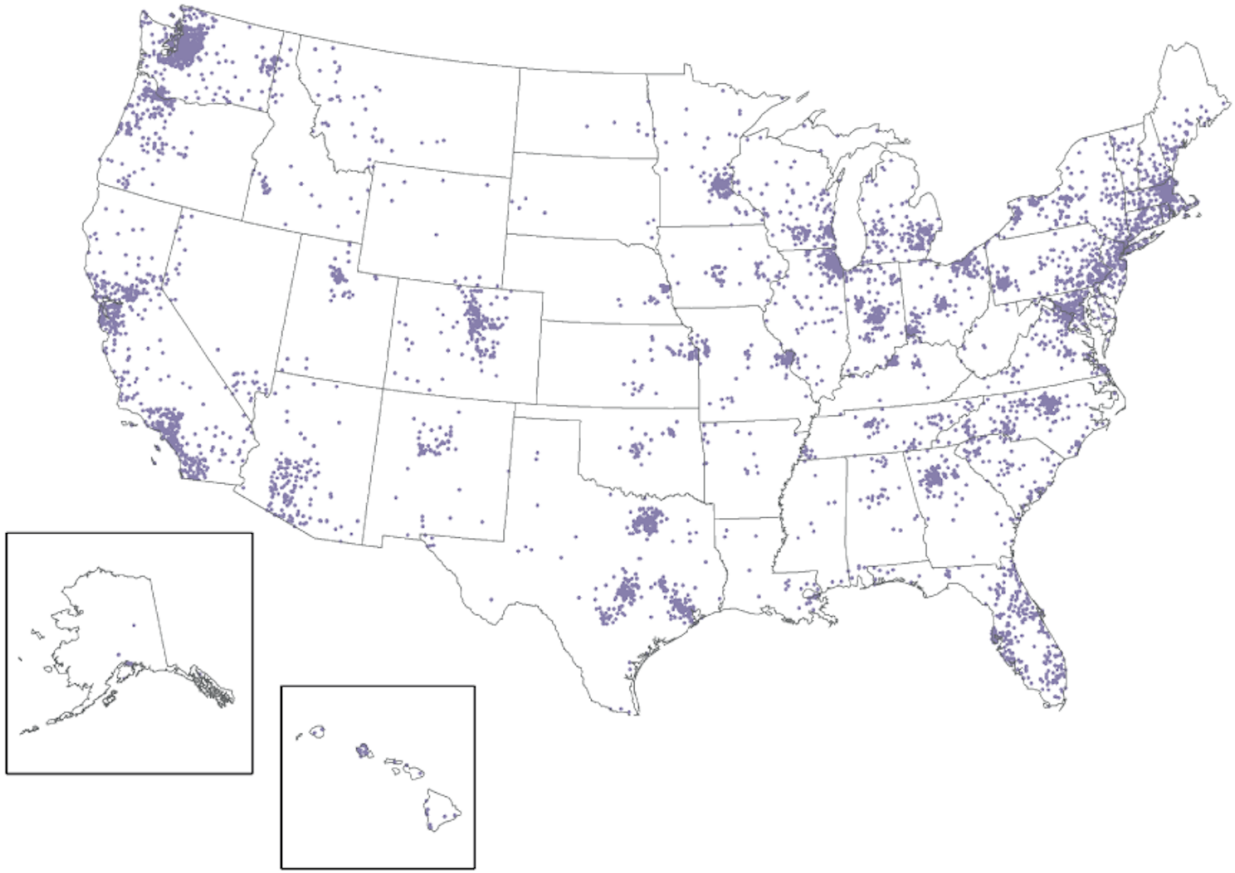


Figure 7: DAP has participants all over the US.

Current and Planned Data

Fred Hutch Cancer Center

Owner-reported surveys and measures

- Health and Life Experience Survey
- Canine Social and Learned Behavior Survey
- End-of-Life Survey
- Morphometrics and mobility
- Cognitive task performance

Biological and physical data

- Cheek swab
- Biospecimen kit
- Activity monitors
- Cardiology examination
- Banked samples

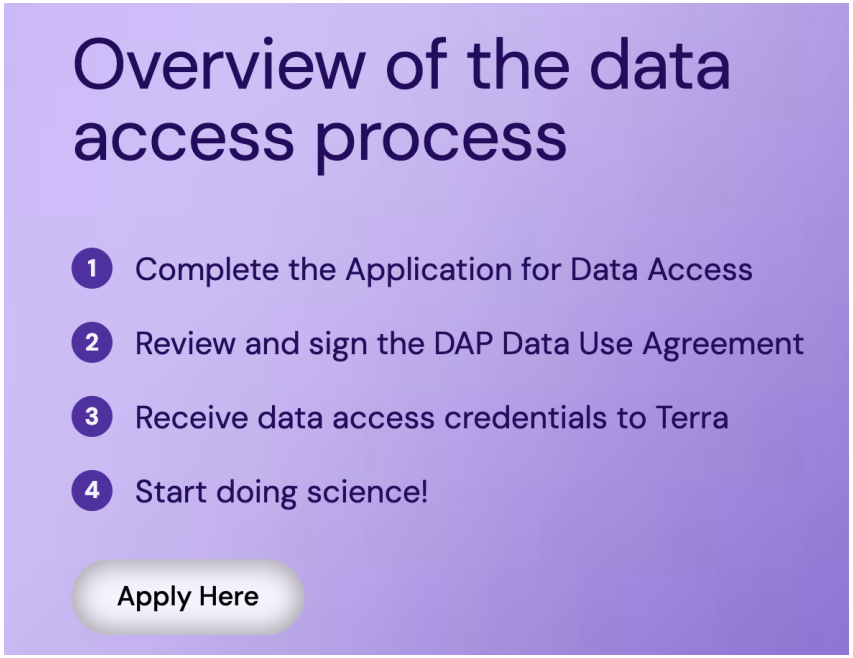
Veterinary data

- Veterinary electronic medical records

Environmental data

- Geospatial data

DAP Data Availability



GitHub: https://github.com/dogagingproject
Shiny App: https://data.dogagingproject.org:8000/

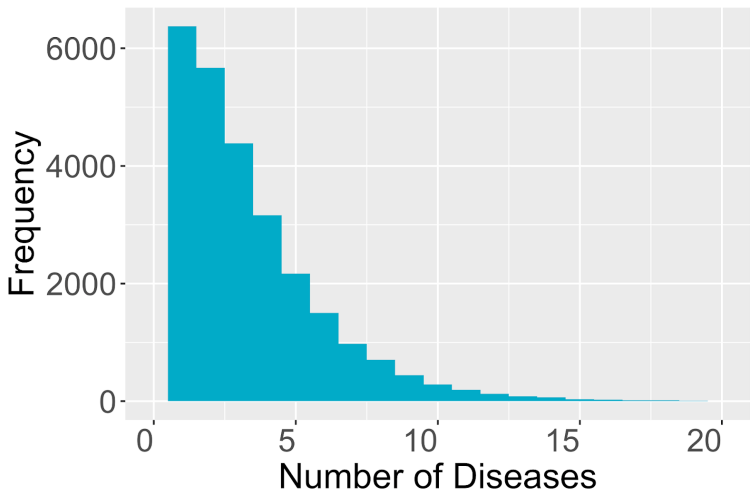
Study Cohort

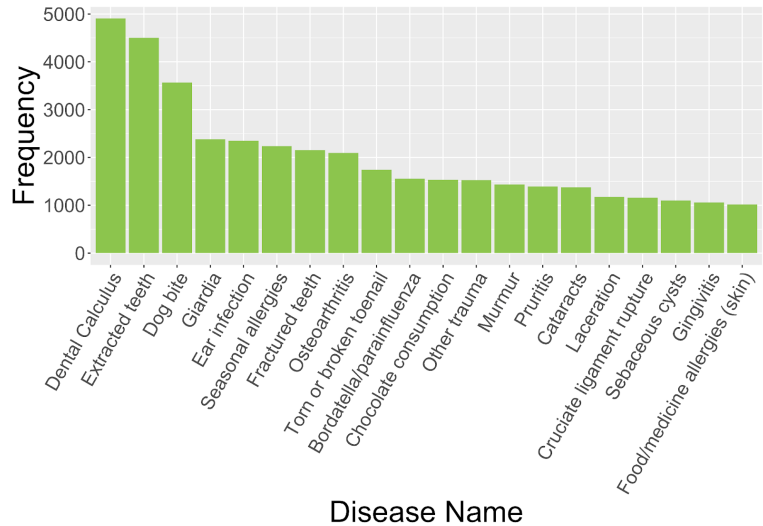
Fred Hutch Cancer Center

Demographics of the canine cohort

Variable	No. (%) of subjects (N=26526)
age	
Mean (SD)	7.8 (± 4.2)
weight	
Mean (SD)	51 (± 29)
breed	
Purebred	13152 (50 %)
Mixed breed	13374 (50 %)
sex	
Male intact	1264 (5 %)
Male neutered	12166 (46 %)
Female intact	679 (3 %)
Female spayed	12417 (47 %)

Study Cohort





Comorbidity

Fred Hutch Cancer Center

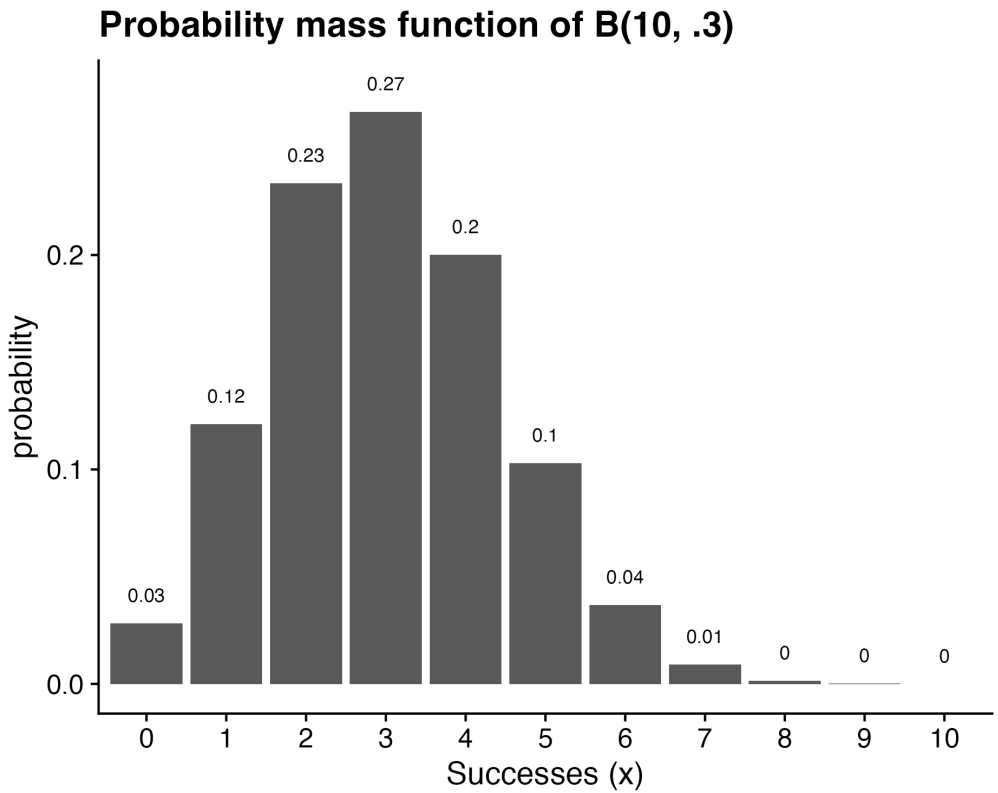
Comorbidity refers to the presence of two health conditions in one subject.

Table 1: Example data in first 10 subjects. Here 1 indicates presence.

id	atopic dermatitis	allergies	Cushing's disease
1	0	0	0
2	1	0	0
3	0	0	0
4	1	1	0
5	1	1	1
6	0	0	1
7	0	0	0
8	1	1	0
9	1	1	0
10	1	0	0
⋮	⋮	⋮	⋮

Binomial Random Variables

A binomial random variable is defined as X where X=# of successes in n trials of a binomial experiment.



Normal Approximation

The distribution of $B(n, p)$ can be approximated by a normal distribution if $np > 10$ and $n(1 - p) > 10$.

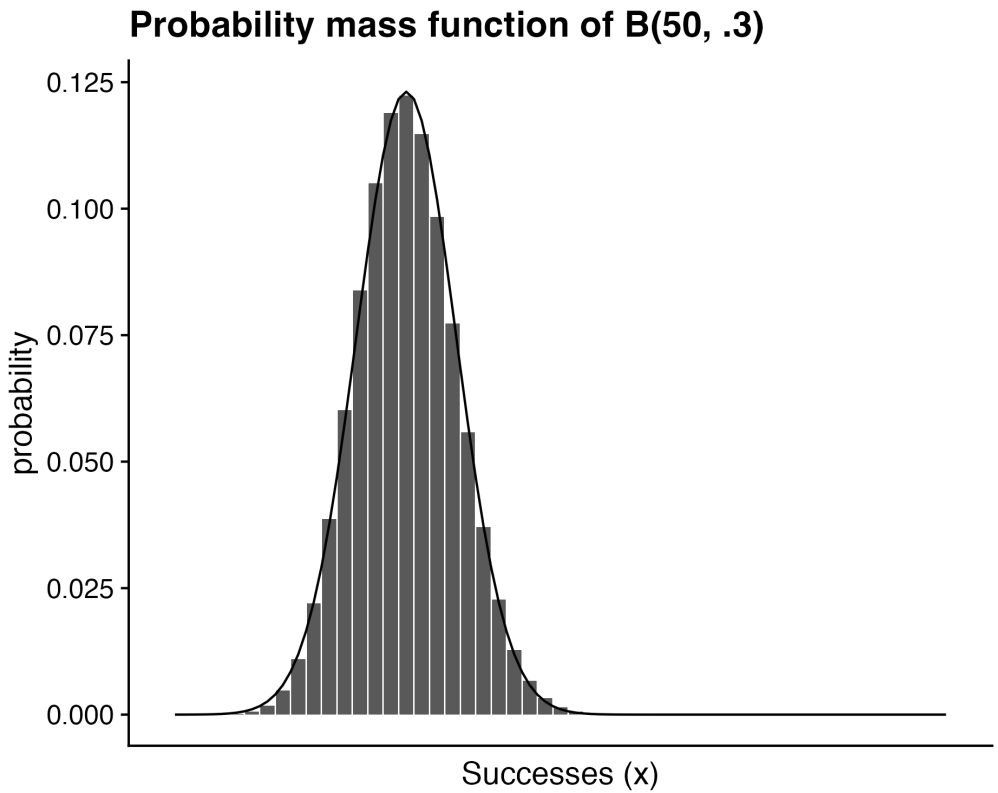


Figure 8: Overlaid line is the density curve for $N(np, np(1 - p))$

Hypothesis Testing

Fred Hutch Cancer Center

The null hypothesis is that a pair of diseases y and z are independently distributed:

$$P(\{y, z\} \subset d) = P(y \in d)P(z \in d)$$

Hypothesis Testing

Fred Hutch Cancer Center

The null hypothesis is that a pair of diseases y and z are independently distributed:

$$P(\{y, z\} \subset d) = P(y \in d)P(z \in d)$$

Conventional test assumes that the disease risk is constant for all dogs

$$P(y \in d) = \frac{\#\text{ dogs with disease } y}{\text{total number of dogs}},$$

$$P(z \in d) = \frac{\#\text{ dogs with disease } z}{\text{total number of dogs}}.$$

Under the null, # of dogs with both diseases $\{y, z\}$ follows a binomial distribution with $P(y \in d)P(z \in d)$.

An Illustration

Take $y =$ atopic dermatitis & $z =$ allergies. The total number of dogs is $n = 26526$. The number of dogs with diseases y and z is 919 and 1015, respectively.

An Illustration

Take $y =$ atopic dermatitis & $z =$ allergies. The total number of dogs is $n = 26526$. The number of dogs with diseases y and z is 919 and 1015, respectively.

Under the null, the probability of observing both diseases in one dog is

$$\frac{919}{26526}\frac{1015}{26526} \approx 0.0013$$

The null distribution can be approximated by a normal with mean 35 and standard deviation 5.93.

An Illustration

Take y = atopic dermatitis & z = allergies. The total number of dogs is $n = 26526$. The number of dogs with diseases y and z is 919 and 1015, respectively.

Under the null, the probability of observing both diseases in one dog is

$$\frac{919}{26526}\frac{1015}{26526} \approx 0.0013$$

The null distribution can be approximated by a normal with mean 35 and standard deviation 5.93.

What is the probability of observing 116 or more dogs with both diseases y and z? $1.1e^{-42}$!

Issues with Binomial Test

The binomial test assumes constant disease risks, which may not hold, because disease risks can depend on age, sex, sterilization status, and weight.

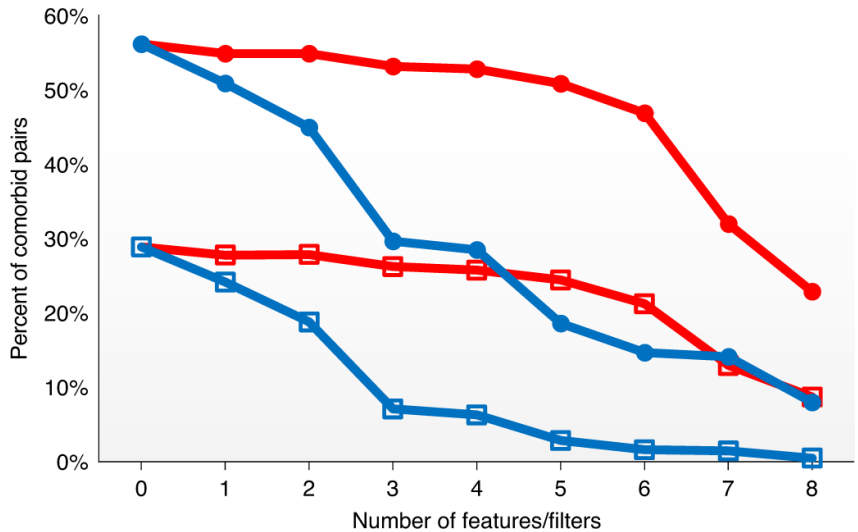


Figure 9: Blue line: Binomial test with stratification by confounding variables; Red line: Poisson Binomial test[1]. Shape indicates sample sizes.

[1] Lemmon et al. (2021). Nature Computational Science

Poisson Binomial Test

The Poisson Binomial distribution generalizes the binomial distribution to allow for N distinct disease risks

$$p_1, \ldots, p_N$$

where $p_d = P(y \in d)$.

Poisson Binomial Test

The Poisson Binomial distribution generalizes the binomial distribution to allow for N distinct disease risks

$$p_1, \ldots, p_N$$

where $p_d = P(y \in d)$.

The Binomial distribution is a special case where $p_1 = \ldots = p_N = p$.

Poisson Binomial Test

The Poisson Binomial distribution generalizes the binomial distribution to allow for N distinct disease risks

$$p_1, \ldots, p_N$$

where $p_d = P(y \in d)$.

The Binomial distribution is a special case where $p_1 = \ldots = p_N = p$.

The cumulative distribution function of a Poisson Binomial distribution is often approximated with a normal distribution

$$N\left(\sum_{d=1}^{N} p_d, \sum_{d=1}^{N} p_d(1-p_d)\right)$$

Poisson Binomial Test for Comorbidity

Need to estimate individualized disease risks for y and z.

Poisson Binomial Test for Comorbidity

Fred Hutch Cancer Center

Need to estimate individualized disease risks for y and z. This can be done by fitting two separate logistic regression models.

Poisson Binomial Test for Comorbidity

Need to estimate individualized disease risks for y and z. This can be done by fitting two separate logistic regression models.

Under the null of independence, the mean of the joint distribution is

$$\sum_{d=1}^{N} P(y \in d)P(z \in d)$$

and variance is

$$\sum_{d=1}^{N} P(y \in d)P(z \in d)\{1 - P(y \in d)P(\in d)\} + S^2,$$

where S^2 accounts for the uncertainty in estimating the individualized disease risks.

Logistic Regression

Fred Hutch Cancer Center

The disease risk for the d-th dog depends on demographic variables

$$P(y_d = 1) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{d1} + \beta_2 x_{d2} + \cdots + \beta_6 x_{d6})}$$

where x_{dj} is the j-th variable. Here we treat age and weight as continuous, and breed and sex as categorical variables.

Logistic Regression

Fred Hutch Cancer Center

The disease risk for the d-th dog depends on demographic variables

$$P(y_d = 1) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{d1} + \beta_2 x_{d2} + \cdots + \beta_6 x_{d6})}$$

where x_{dj} is the j-th variable. Here we treat age and weight as continuous, and breed and sex as categorical variables.

Logistic regression is sensitive to class imbalance. Here we only consider diseases that occur in at least 60 dogs.

LRM Coefficients

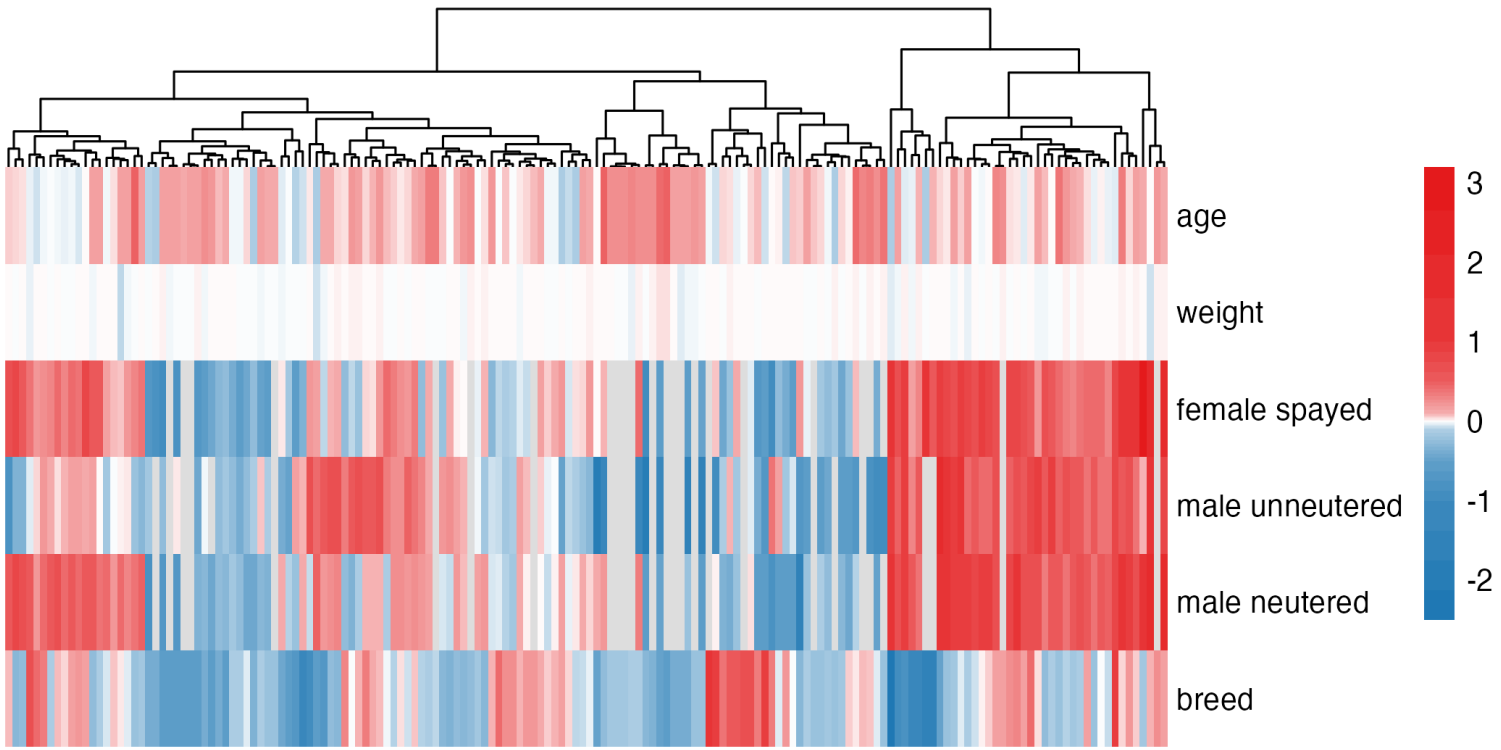


Figure 10: Age is an important risk factor for many diseases. Being purebred does not necessarily raise a dog's disease risk.

Undirected Comorbidity Network

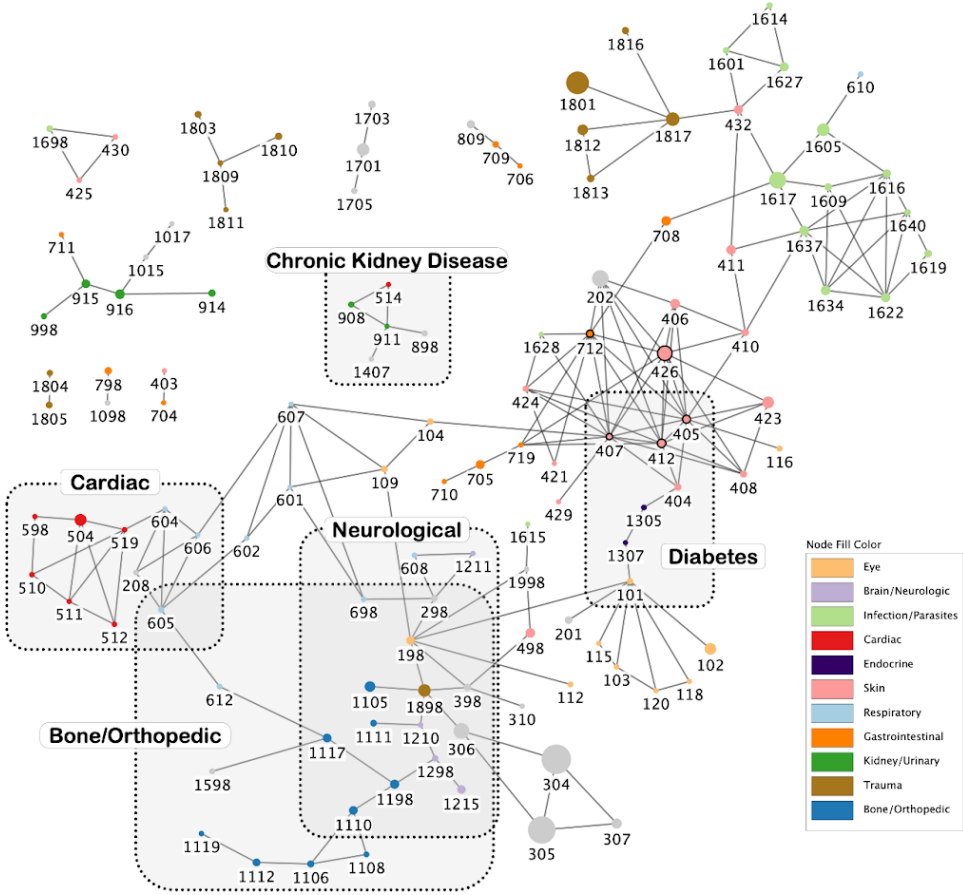


Figure 11: Nodes are diseases, edges are comorbidity P-values statistically significant at a Benjamini-Hochberg adjusted alpha level of 0.01. Node sizes represent prevalence, with more common diseases having larger nodes. Nodes with degree greater than 10 have a bold black outline.

Comorbidities

- Diabetes – Blindness ($p<0.001$);
- Atopic dermatitis – Allergies ($p<0.001$);
- Cushing's disease – Alopecia ($p=0.002$);
- Hypertension – Chronic kidney disease ($p<0.001$)[2]
- Anemia – Proteinuria – Chronic kidney disease[3]

[2]Syme, H. (2011). Veterinary Clinics: Small Animal Practice
[3]Sannamwong et al. (2023) Veterinary World.

Degree Distribution

The degree of a node in a network is the number of edges it has.

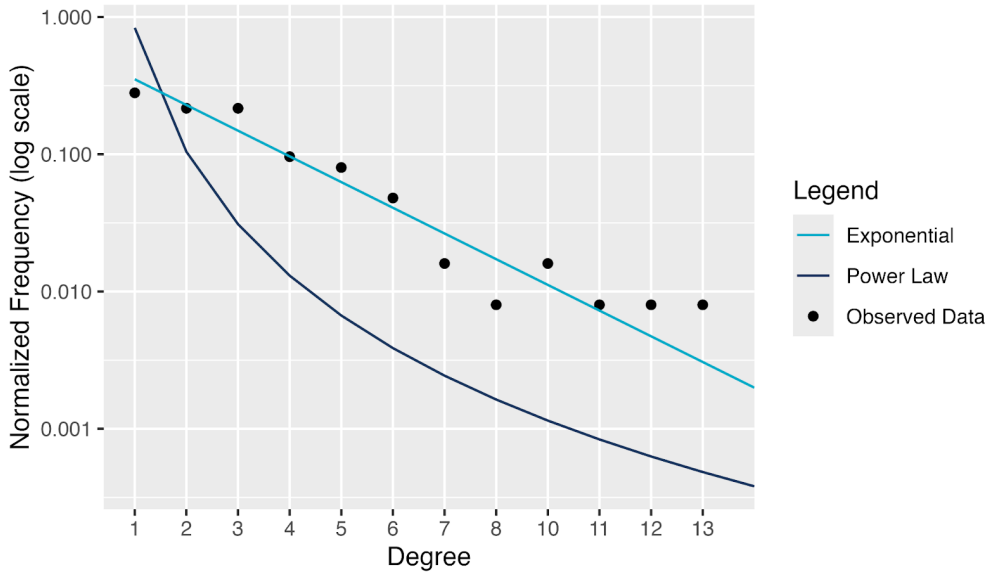


Figure 12: Log scale density distribution of node degrees in the comorbidity network.

The rate of accumulating diseases is constant, regardless of the number of diseases one already has.

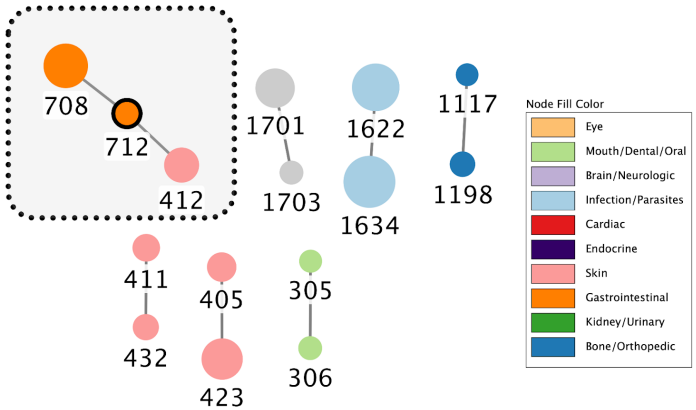
Stratification by Life Stage

Fred Hutch Cancer Center

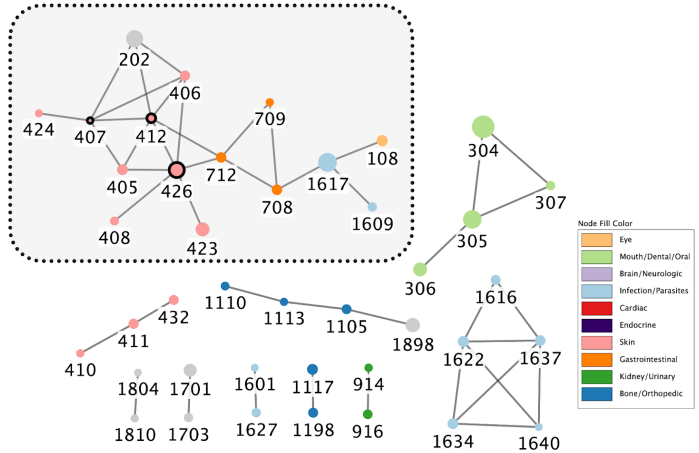
We performed stratified analysis to study age-dependent patterns in comorbidities.

Rather than stratifying dogs by chronological age, we use the concept of life stage. In DAP, life stage is defined using a combination of chronological age and dog size, because small dogs tend to live longer than large dogs, and often by a substantial margin.

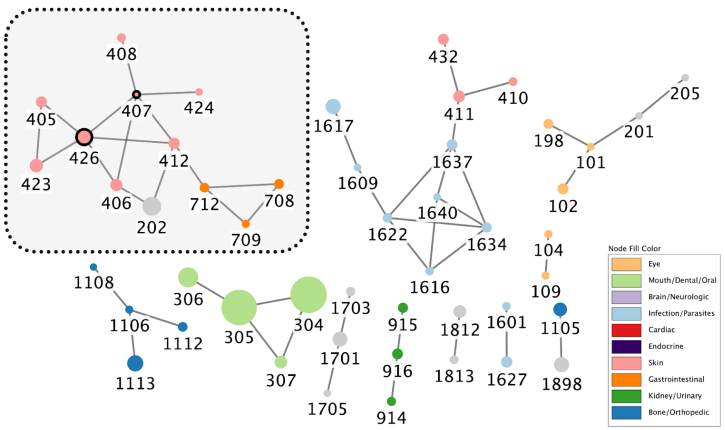
Age-dependent Comorbidity Networks



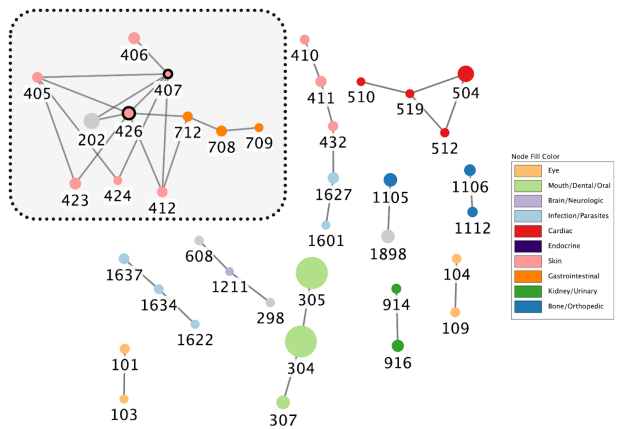
Young



New Mature



Mature



Senior

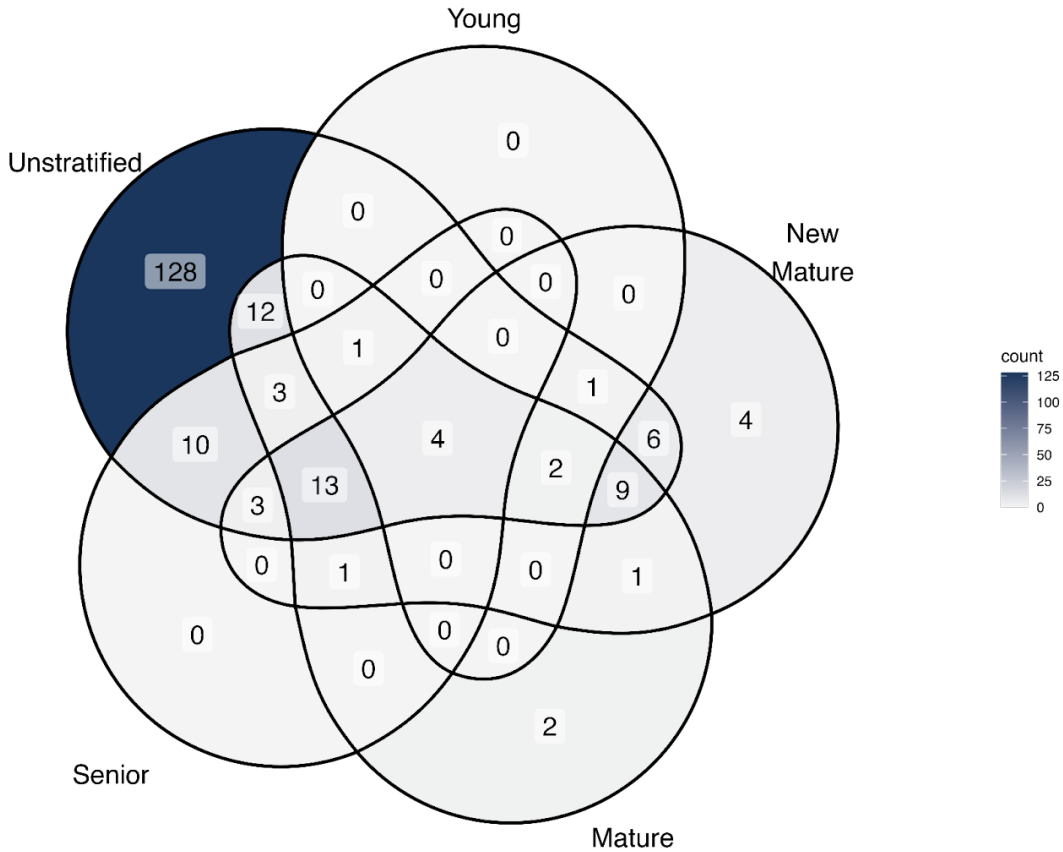
Age-dependent Comorbidity Networks

Fred Hutch Cancer Center

	Node count	Edge count	Edge density
Young	3	2	0.667
New Mature	15	22	0.210
Mature	12	16	0.242
Senior	11	16	0.291

Table 2: Edge densities for the largest subnetwork in each age-stratified network.

Shared Edges across Networks



Neighboring life stages share more edges.

Time-directed Comorbidity Network

Using the date of diagnosis as a proxy to when each dog obtained a disease, we can compute the probability of one disease occurring before another one. Furthermore, we account for the dog's medical history.

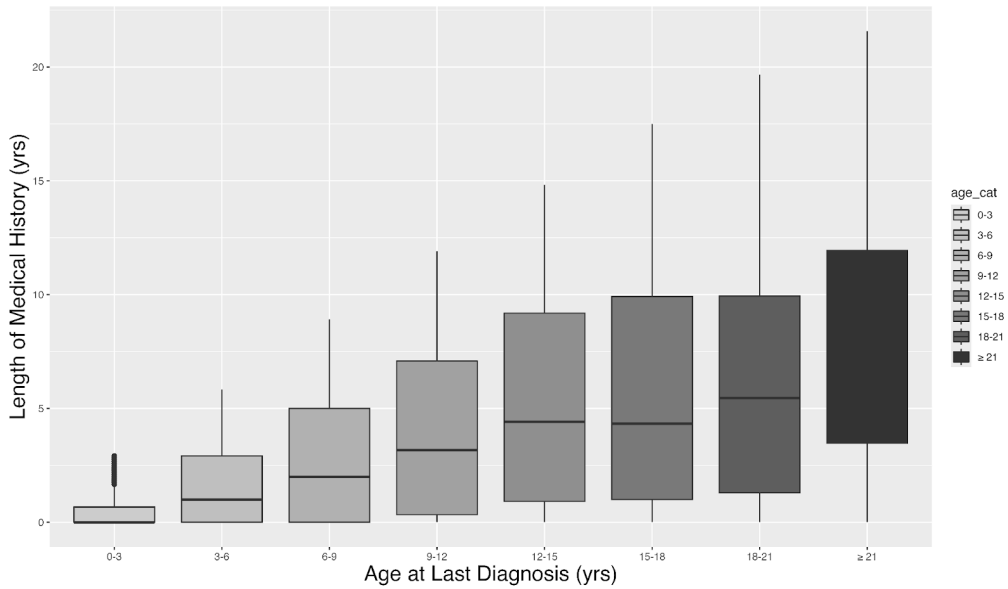


Figure 17: Dog's medical history is approximated using the difference between each dog's most recent date of diagnosis and earliest date of diagnosis.

Time-directed Comorbidity Network

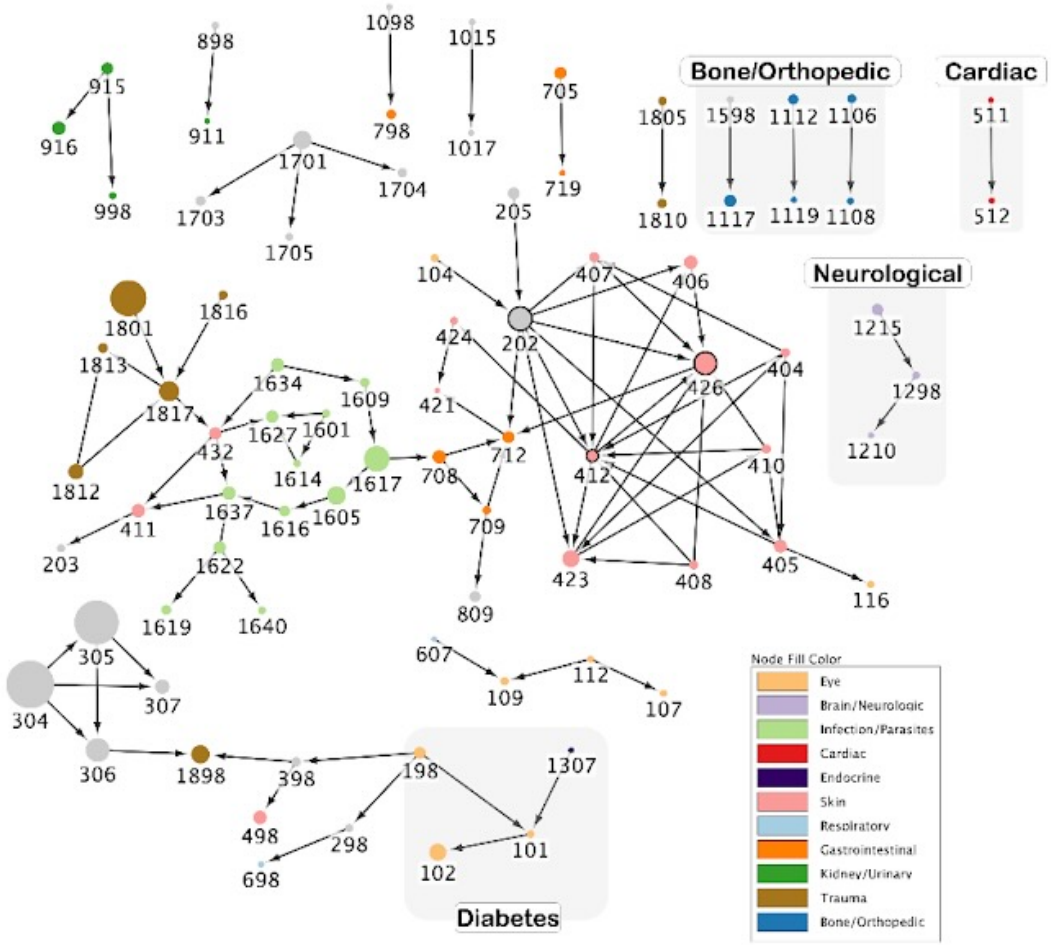


Figure 18: Nodes are diseases, edges are temporal P-values statistically significant at a Benjamini-Hochberg adjusted alpha level of 0.01. Node sizes represent prevalence, with more common diseases having larger nodes.

Comorbidities with Direction

- Diabetes $\rightarrow$ Blindness[4]
- Cluster of parasite and parasite-related diseases is worth further investigation.

[4] Williams, D. L. (2017). Veterinary Sciences.

Summary

We provide the largest analysis of canine comorbidities among 166 diseases.

- The degree distribution of the comorbidity network is exponential rather than power law.

Summary

We provide the largest analysis of canine comorbidities among 166 diseases.

- The degree distribution of the comorbidity network is exponential rather than power law.
- Stratified analysis revealed important age-specific patterns in comorbidities.

Summary

We provide the largest analysis of canine comorbidities among 166 diseases.

- The degree distribution of the comorbidity network is exponential rather than power law.
- Stratified analysis revealed important age-specific patterns in comorbidities.
- Time-directed comorbidity network provides insights into disease progression.

Limitations

- Data are cross-sectional, but longitudinal data are becoming available.

Limitations

- Data are cross-sectional, but longitudinal data are becoming available.
- Rare diseases are susceptible to overfitting by logistic regression.

Limitations

- Data are cross-sectional, but longitudinal data are becoming available.
- Rare diseases are susceptible to overfitting by logistic regression.
- Owner-reported surveys may be subject to bias.

Limitations

- Data are cross-sectional, but longitudinal data are becoming available.
- Rare diseases are susceptible to overfitting by logistic regression.
- Owner-reported surveys may be subject to bias.
- The owners in DAP are predominantly white, highly educated, and wealthier than the US population.

Opportunities

- Investigation into negatively associated disease pairs

Opportunities

Fred Hutch Cancer Center

- Investigation into negatively associated disease pairs
- Incorporate veterinary electronic medical records using NLP

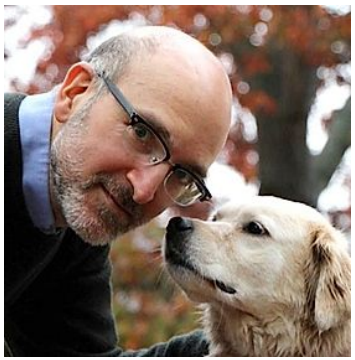
Opportunities

- Investigation into negatively associated disease pairs
- Incorporate veterinary electronic medical records using NLP
- Incorporate genetic profiles to more accurately adjust for breed differences.

Acknowledgement

Antoinette Fang

Lakshin Kumar



Daniel Promislow

Dog Aging Project Consortium