Synthetic Data for Teaching Data Science Concepts

AMIA Education Forum

Session 16: Patient Simulation for Improved Learning

Ted Laderas and David Dorr

Assistant Professor, Medical Informatics & Clinical Epidemiology Oregon Health Science University

6-21-2018

Synthetic Data for Teaching Data Science Concepts

AMIA Education Forum

Session 16: Patient Simulation for Improved Learning

Ted Laderas and David Dorr

Assistant Professor, Medical Informatics & Clinical Epidemiology Oregon Health Science University

6-21-2018

Disclosure

- I disclose the following relevant relationship with commercial interests:
  - Instructor at DataCamp

About Me

- Assistant Professor, Bioinformatics & Computational Biomedicine
- Immune informatics and Data Visualization
- Data Nerd/R Enthusiast
- Teach Data Analytics/Statistical Methods/
- Lifelong Learner and plays well with others

Learning Objectives

- After participating in this session I hope you will be able to:
  - Explain why we need synthetic patient data
  - Explore the dataset for associations
  - Use this data set to teach data science

Cardiovascular Risk?

Say you were given a dataset

With the following covariates:

Say you were given a dataset

With the following covariates:

- BMI
- Type 2 Diabetes status
- Age
- Gender
- Race

Say you were given a dataset

With the following covariates:

- BMI
- Type 2 Diabetes status
- Age
- Gender
- Race
- Total Cholesterol
- Systolic Blood Pressure
- Hypertension status
- Genetics

Say you were given a dataset

With the following covariates:

- BMI
- Type 2 Diabetes status
- Age
- Gender
- Race
- Total Cholesterol
- Systolic Blood Pressure
- Hypertension status
- Genetics

- Could you predict whether someone is at risk for a cardiovascular disease?

Say you were given a dataset

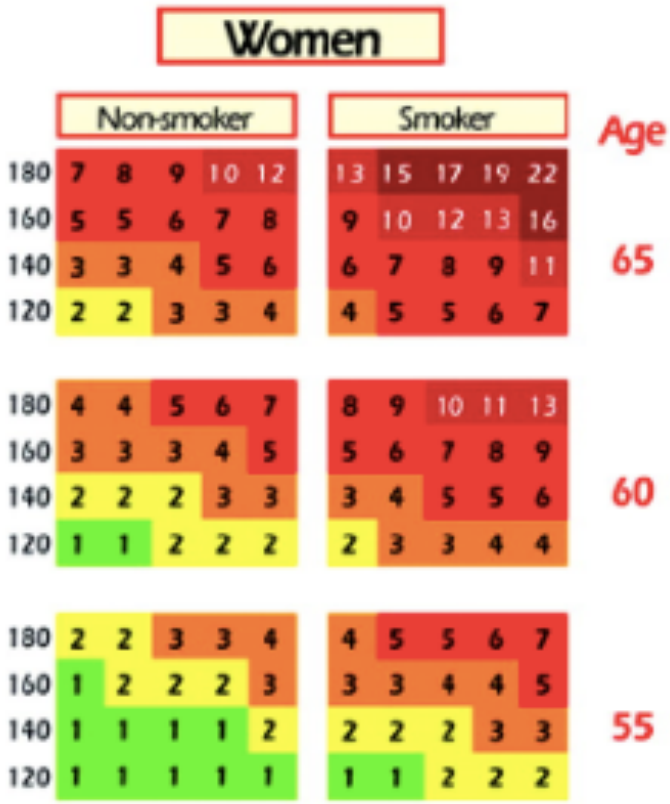
With the following covariates:

- BMI
- Type 2 Diabetes status
- Age
- Gender
- Race
- Total Cholesterol
- Systolic Blood Pressure
- Hypertension status
- Genetics

- Could you predict whether someone is at risk for a cardiovascular disease? - How do you know which of these variables is important?

CVD Risk Scores

- Risk Calculations relatively easy to understand
  - Framingham, MESA, Jackson Heart Study
- Hard to estimate for certain cohorts
  - People under 40 (low prevalence)
- How were scores estimated?
  - And for what population?



http://circ.ahajournals.org/content/122/3/300

Let's think about some relationships

Let's think about some relationships

- How are Age and Cardiovascular Risk related?

Let's think about some relationships

- How are Age and Cardiovascular Risk related?
- How is Smoking and Cardiovascular Risk related?

Let's think about some relationships

- How are Age and Cardiovascular Risk related?
- How is Smoking and Cardiovascular Risk related?
- How are Race and Total Cholesterol related?

Let's think about some relationships

- How are Age and Cardiovascular Risk related?
- How is Smoking and Cardiovascular Risk related?
- How are Race and Total Cholesterol related?
    - What variables are best to predict risk?
    - Are particular variables more predictive for a cohort?

Synthetic Datasets as puzzles for increasing curiosity

- Need data that is "safe" to learn on
  - Understand the difficulty of predicting risk in different cohorts
- Increase curiosity about health data
- Lower barriers for learning in a safe environment
- Reduce the fear of data by looking at it and talking about it

You be the Driver

What associations should we look at?

- BMI
- Type 2 Diabetes status
- Age
- Gender
- Race
- Total Cholesterol
- Systolic Blood Pressure
- Hypertension status
-

Patient 1

Patient 2

Sample Patients

Which of these patients is more at Risk for Cardiovascular Disease?

Magic Happens

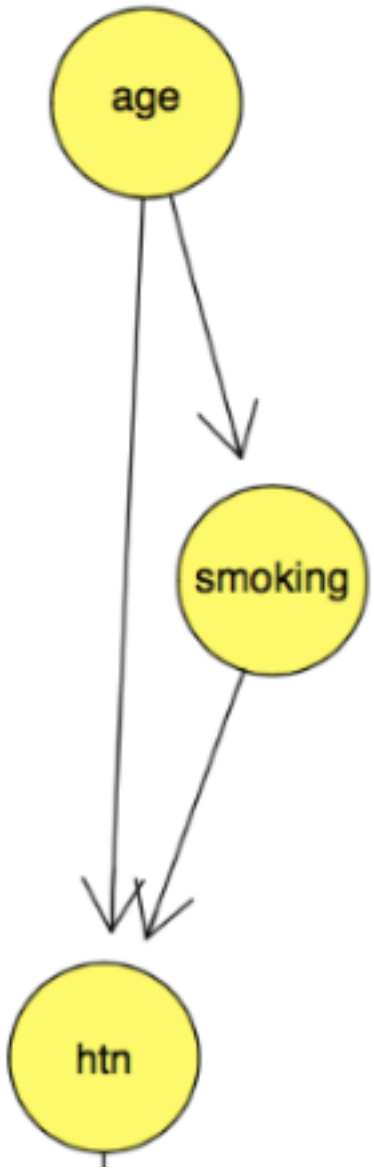
- When we look at data together!

About the Dataset

Model associations in the dataset as a causal network:

- Race/Total Cholesterol
- Hypertension/Hypertensio Treatment
- Age/Smoking

Data was generated as a Bayesian network - designed iteratively with Clinician (David Dorr)



Additional Problem: Genotype of patients

1. For more advanced students
2. Four SNPs are included for a smaller subset of patients
    - Frequencies are race dependent
3. Only one SNP increases risk
4. Other SNPs are associated with total cholesterol

Teaching Data Science Skills in R

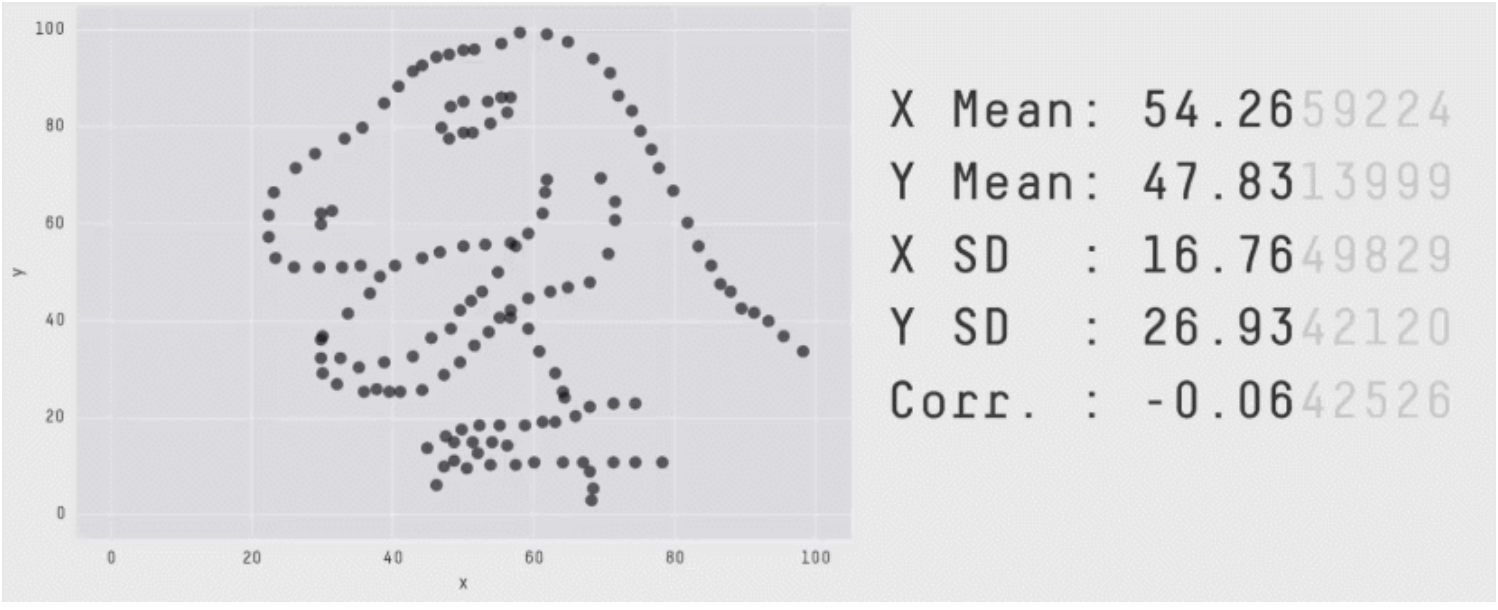
1. Cardiovascular Risk Prediction Workshop over two nights
    - Exploring cohorts and their risk prevalence
    - Predicting cardiovascular risk in a cohort using machine learning
2. Big Data to Knowledge Skills Course
    - Held for Portland State University Students/Staff/Faculty
    - Students must know some R
    - Must have some math/statistics background
3. 11 attended

Day 1: Understanding risk prediction

- Learning Objective: understand how we currently predict cvd risk
  - Calculate risk for patients using current score calculators
- Learning Objective: which variables are associated with CVD risk?
  - How are they related to each other?
- Learning Objective: select cohort of data to predict risk in for Day 2
  - Assess prevalence of cardiovascular disease in cohort using Shiny dashboard

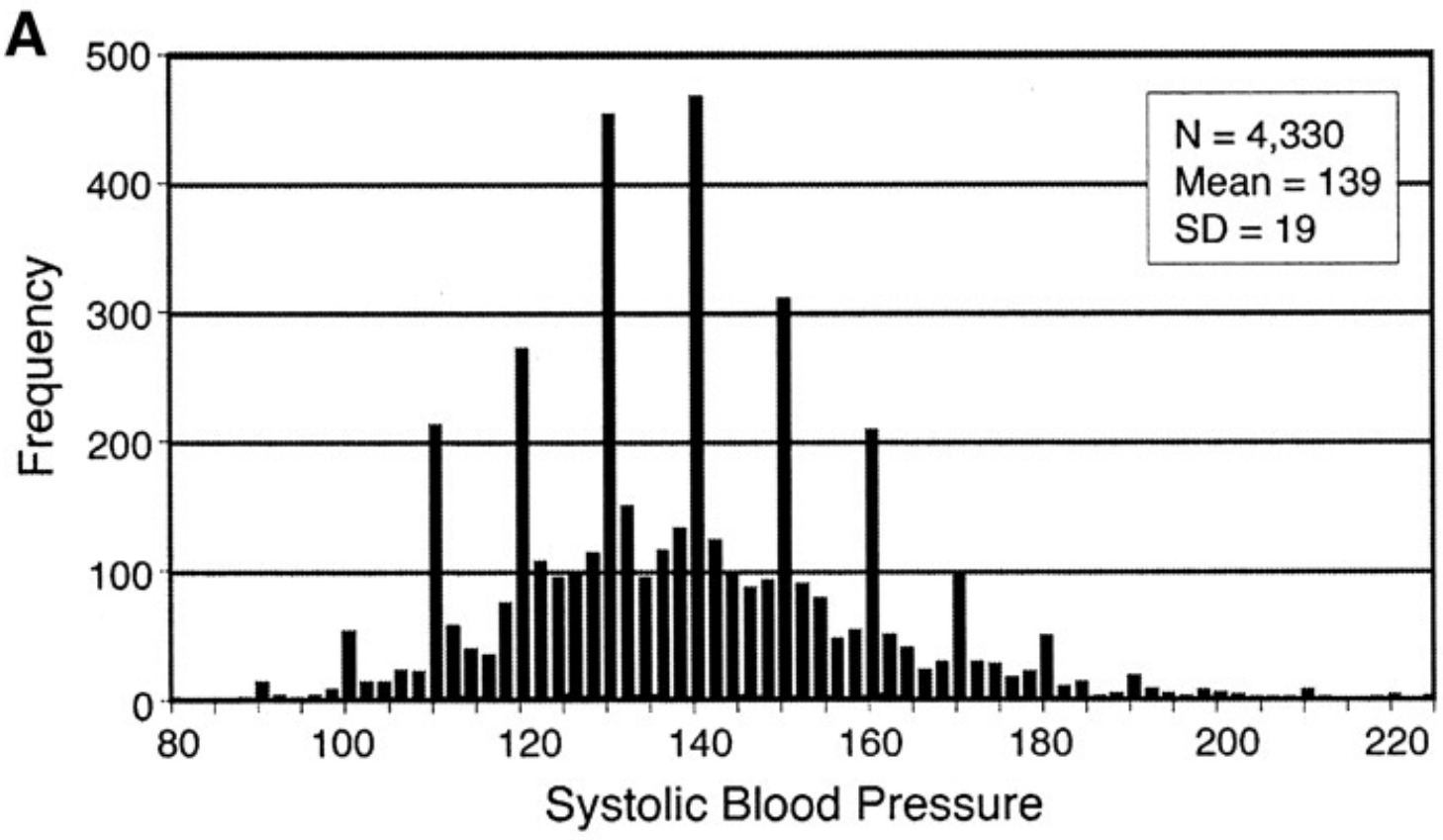
Always Look at Your Data

- Need to understand the value of Exploratory Data Analysis



https://github.com/stephlocke/datasauRus

Uh...Why is this?



http://care.diabetesjournals.org/content/30/8/1959

How to not be afraid of data

How to not be afraid of data

- Students were given a Shiny dashboard with the dataset
  - Reduce cognitive load in exploring data
  - Make data exploration more accessible

How to not be afraid of data

- Students were given a Shiny dashboard with the dataset
  - Reduce cognitive load in exploring data
  - Make data exploration more accessible- Asked to select a cohort of interest to subset data

How to not be afraid of data

- Students were given a Shiny dashboard with the dataset
  - Reduce cognitive load in exploring data
  - Make data exploration more accessible- Asked to select a cohort of interest to subset data- Asked to assess within cohort
  - Prevalence
  - Association between covariates in the dataset

Data Explorer

Check it out here: http://bit.ly/cvdDash

1. Summary Tables (What categories exist in the data?)
2. Cross-tables (How does age influence CVD risk?)
3. Bar Graphs (How is CVD risk associated with T2D?)
4. Histograms (How is age distibuted in our population?)
5. Boxplots (How is CVD related to Systolic Blood Pressure?)

Day 2: Machine Learning and CVD Risk

1. Students used their cohort
2. Students were given an RMarkdown template and example code for modeling
3. Students attempted to predict risk in cohort using three different methods:
4. Scoreboard to compare results
5. Discussion

RMarkdown Documents

1. Similar to Jupyter Notebooks
2. Give students example code

Predicting CV risk: Which Variables?

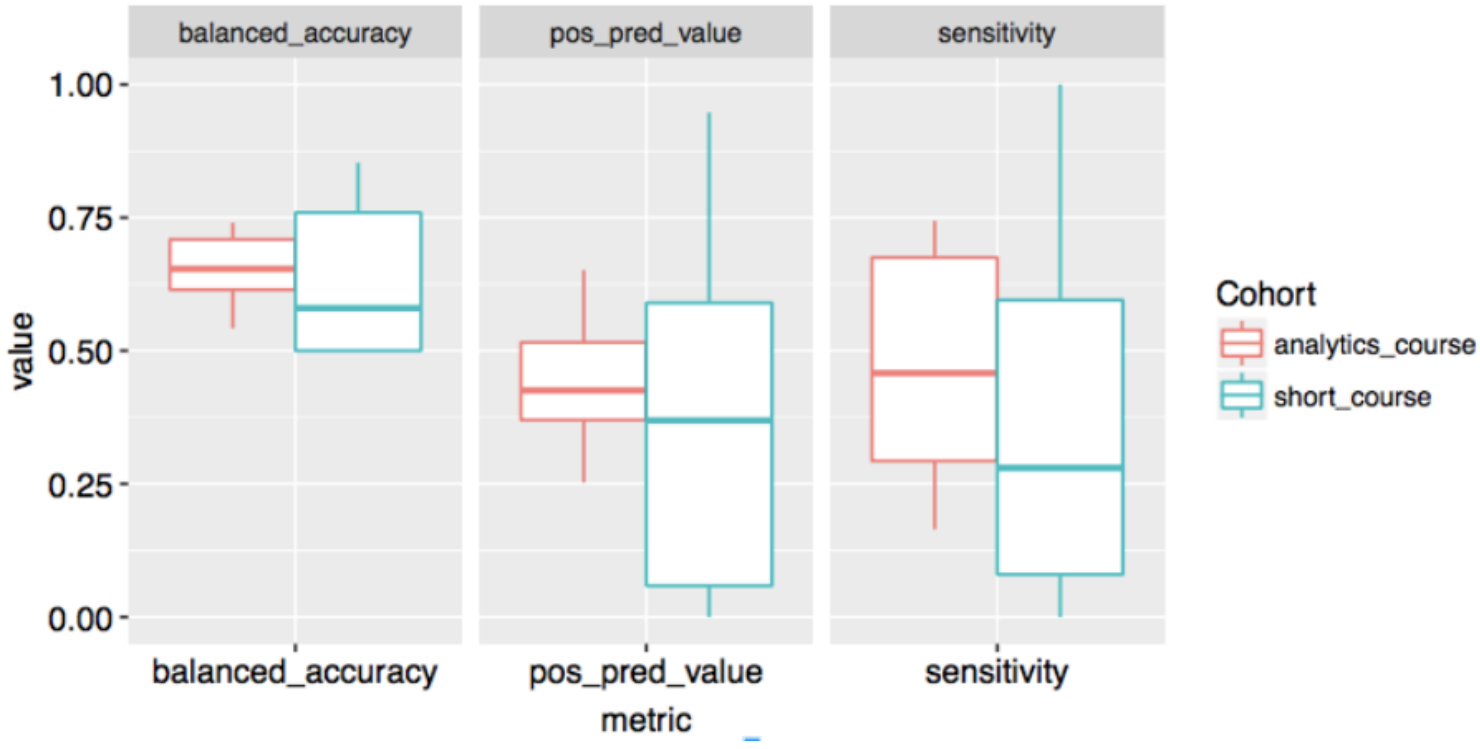
1. Variable selection is key to predicting cardiovascular risk
2. Students were asked to select variables based on their previous findings
3. Comparison of different machine learning methods

Assessing Models

Because overall prevalence is low, need to use balanced metrics to assess performance

1. Balanced Accuracy
2. Positive Predictive Value
3. Sensitivity

Student Results



Student Feedback

- Enjoyed learning about EDA
    - "hands-on data analysis"
    - "Great mix of background material combined with tools for exploring the data and analyzing predictive ability"
    - "Providing us the way to think about predictive analytics"
    - "I'm moving from basic research into public health, so talking about applicability of models to patient help was helpful."

Where we could do better

- Students wanted more background reading and material
  - "More math and R background. Especially about Shiny."
  - "A little hard for a person without experience of coding"
  - "Example metric of self-assessment with regard to proficiency with R."

Lessons Learned

- Generating realistic synthetic data is difficult
- Need to tune data so that problem is solvable
- Students like the hands-on aspects of using it
- Collaborative process with Clinical and Biology sides of informatics

Future Directions

- Clinical Data Wrangling Bootcamp

Acknowledgements

- Materials developed under Big Data to Knowledge BD2K grant: 1R25EB020379-01
- Thank you to:
  - DMICE
  - Harold Lehman
  - Christopher Chute

Grab the Dataset!

We want you to use it! Break it, help us make it better!

http://github.com/laderast/cvdRiskData

Available in R as an installable dataset:

Read the preprint here:
https://www.biorxiv.org/content/early/2018/04/21/232611

Get the Course Material

- All course material is available online
  - http://github.com/laderast/cvdNight1
  - http://github.com/laderast/cvdNight2

Questions/Comments?

Email: tedladeras@gmail.com
Twitter: @tladeras
Github: http://github.com/laderast/
Web: http://laderast.github.io/

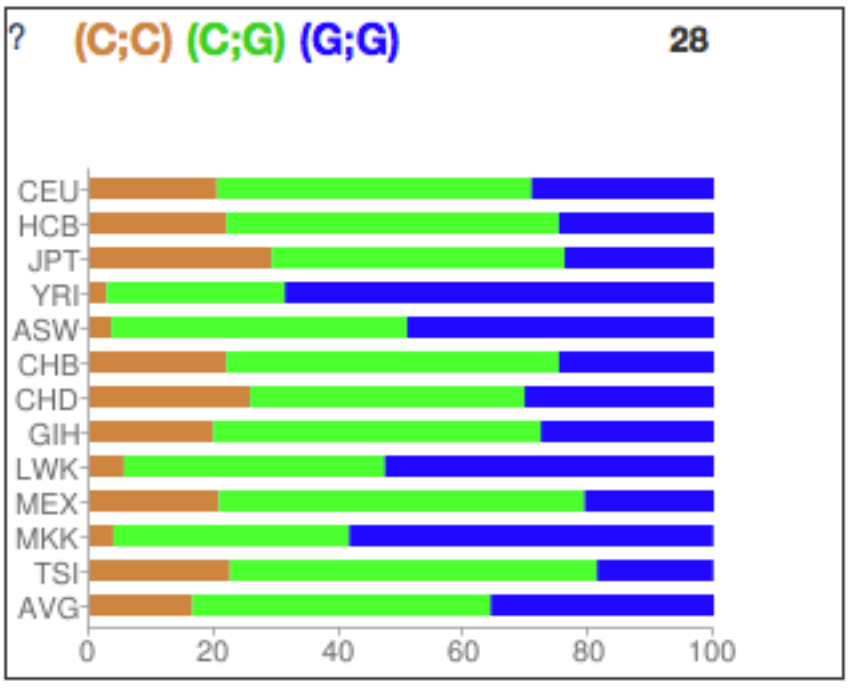
Backup Slides

Bayesian Networks

- Use to generate multivariate distribution
- Associate marginal probabilities
- Specify relationships between variables using conditional probability tables
  - / /
  - / / /
- Iterative process
  - Sanity Check by David (is this realistic?)

Genetic Covariates

- Four SNPs (from SNPedia)
  - Used real world distributions in Race from SNPedia
- Simplified dataset so each SNP had only two genotypes
- Three SNPs were associated with Total Cholesterol
- One SNP increased overall risk



Example Patients