Digital Transformation of Healthcare

Data Cleaning

Michoel Snow, M.D. Ph.D., Glen Ferguson, Ph.D.

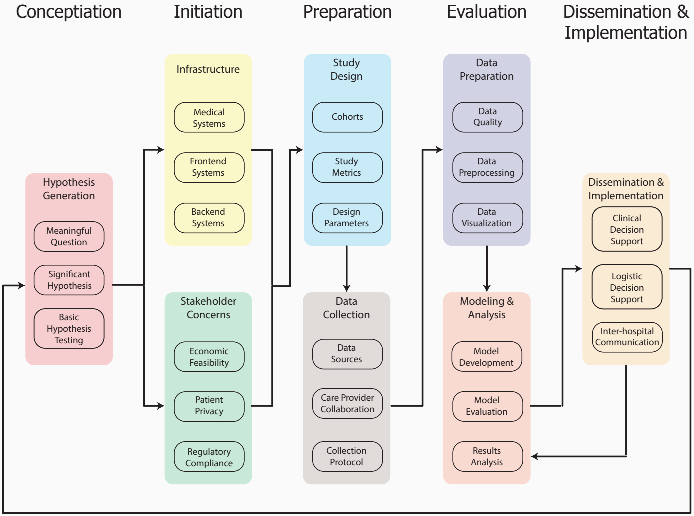
Center for Health Data Innovations

Data Cleaning

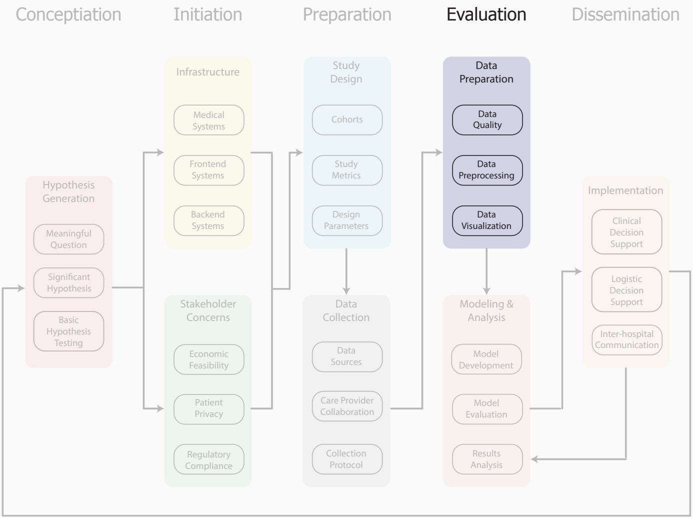
After this lecture students will be able to

- Discuss and apply the steps involved in cleaning data for modeling
- Design a process for the imputation of missing data
- Build a bioinformatics pipeline starting from given data

Bioinformatics Pipeline

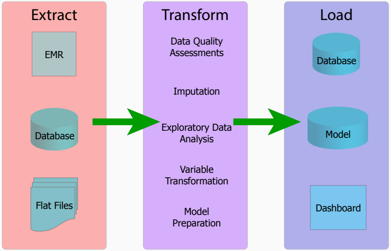


Data Cleaning



Scenario

- You are trying to optimize the use of pain medication regimens for pediatric sickle cell patients
- You have just collected all the data on all peds hem-onc patients on the floor for the past month.
- How can you systematically assess, organize and prepare the data for modeling and analysis?
    - What are the steps to transform raw data into a usable format


Extract
EMR
Database
Flat Files

Transform
Data Quality Assessments
Imputation
Exploratory Data Analysis
Variable Transformation
Model Preparation

Load
Database
Model
Dashboard

Imputation and Extrapolation

- what are the different reasons why data might be missing
- What are the different ways that data could be missing
- Can we develop a systematic way to deal with missing data
  - pain score
  - pain medication usages
  - retic count
  - infection status
  - imaging results
- How do you evaluate imputation

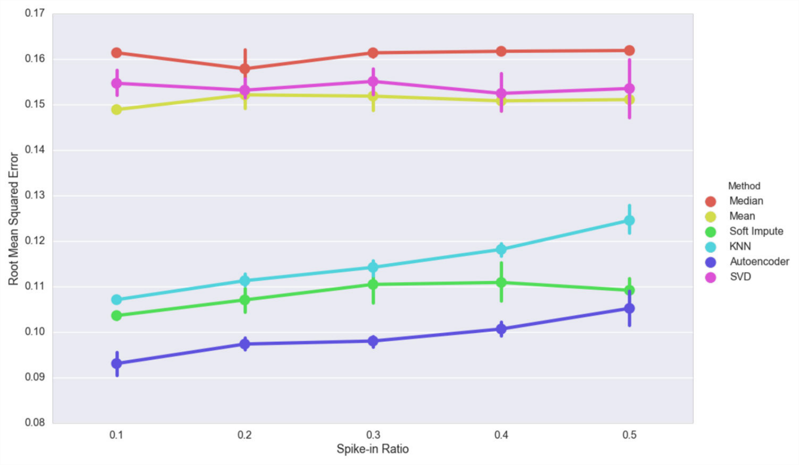
Imputation Example

Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders

- Researchers started from an ALS clinical trials database of 10,723 patients
- The dataset includes patient demographic data, family history, concomitant medications, vital sign measurements, laboratory results, and patient clinical history
- They removed data using either an MCAR or MNAR approach
- For end metrics they considered both accuracy in imputation and ALS functional rating scale

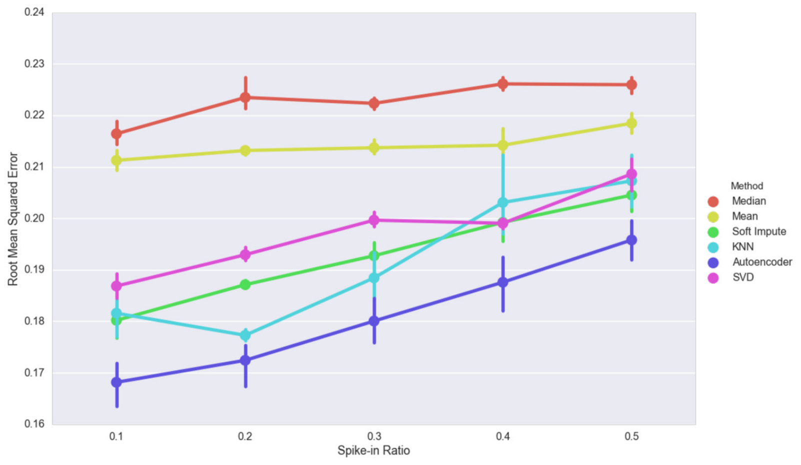
Imputation Example

Missing Completely at Random

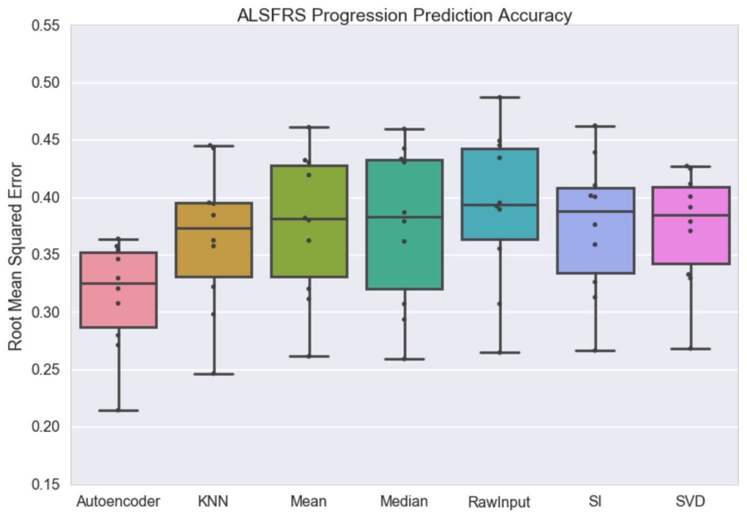


Imputation Example

Missing Not at Random

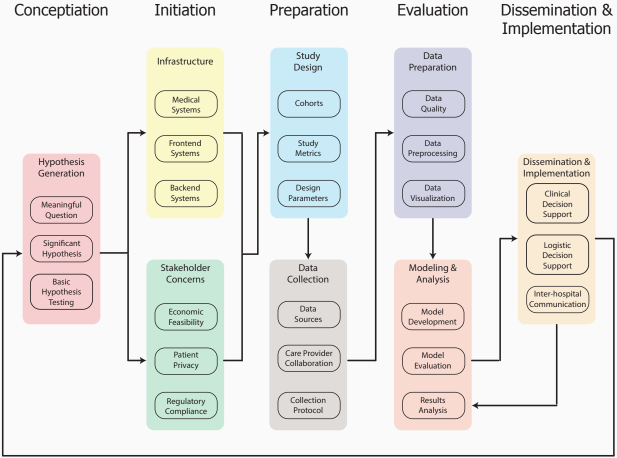


Imputation Example



Pipeline

- Using the pediatric sickle cell example let's walk through building the second half of the bioinformatics pipeline


Conception
Initiation
Preparation
Evaluation
Dissemination & Implementation
Hypothesis Generation
Meaningful Question
Significant Hypothesis
Basic Hypothesis Testing
Infrastructure
Medical Systems
Frontend Systems
Backend Systems
Stakeholder Concerns
Economic Feasibility
Patient Privacy
Regulatory Compliance
Study Design
Cohorts
Study Metrics
Design Parameters
Data Collection
Data Sources
Care Provider Collaboration
Collection Protocol
Data Preparation
Data Quality
Data Preprocessing
Data Visualization
Modeling & Analysis
Model Development
Model Evaluation
Results Analysis
Dissemination & Implementation
Clinical Decision Support
Logistic Decision Support
Inter-hospital Communication