Clinical Data Repository and Clinical Predictive Modeling

Harry Freitas da Cruz

Data Management for Digital Health

Winter 2018
Agenda

Real-world Use Cases
- Oncology
- Nephrology
- Intensive Care
- Additional Topics

Data Management & Foundations
- Biology Recap
- Data Sources
- Data Formats
- Business Processes
- Processing and Analysis

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
Agenda

- Clinical Data Repository (CDR)
- Establishing Clinical Prediction Models (CPM)
- Pitfalls when developing CPMs
Recap

- Nephrology primer
- Clinical decision support systems
- Machine learning for Nephrology

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
5
Predictive Analytics in Healthcare

Clinical Data Repository and Clinical Predictive Modeling

Data Management for Digital Health, Winter 2018

https://i.ytimg.com/vi/xEemneA_qwE/maxresdefault.jpg
Predictive Analytics in Healthcare

https://i.ytimg.com/vi/xEemneA_qwE/maxresdefault.jpg
Predictive Analytics in Healthcare Opportunities and Challenges

**OPPORTUNITIES**

- **Patient safety**
- **Precision medicine**
- **Risk stratification**

**CHALLENGES**

- **Physician and patient trust**
- **Data standardization**
- **Novel algorithms and tools**

Clincial Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018

https://www.healthcatalyst.com/three-approaches-to-predictive-analytics-in-healthcare
Clinical Data Repository
Heterogeneous Landscape


Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018 9

Clinical Data Repository Challenges and Opportunities

- Difficult and costly to develop a general purpose CDR
- Integration of different data sources is a daunting task
- Data inconsistency, redundancies are frequent
- Difficult to adapt to changing needs of users

- 360° patient view
- Explore patterns in disease progression and management
- Discover unknown patterns in the data
- Faster hypothesis testing -> Clinical studies

Clinical Data Repository
Nephrology Use Case (Acute Kidney Injury)

1. Acute patients (renal)
   - Medical devices
   - Laboratory Information System (LIS)
   - Patient Management System
   - Clinical Information System (departments)

2. Clinical Data Repository
   - Patient base data
   - Vital parameters (PDMS)
   - Laboratory findings
   - Medications taken
   - Therapy administered
   - Symptoms (text mining)
   - HL7/LOINC/POCTA-1 Communication Server (e.g., Cloverleaf)
   - Glomerular filtration rate (GFR)
   - Blood urea nitrogen (BUN)
   - Cystatin C
   - Creatinine (Cr)
   - Other parameters?

3. Predictive analytics engine
   - Validated Machine-learning models
   - Risk stratification
   - Severity scoring
   - Early warning

4. Notification system
   - Care teams
   - Specialist doctors

5. Clinical Data Repository and Clinical Predictive Modeling
   - Data Management for Digital Health, Winter 2018
   - 11

H. F. Cruz, Irish Nephrology Society Annual Scientific Meeting (2016)
Clinical Data Repository Nephrology Use Case: Patient Dashboard

Kidney dashboard

Patient dashboard

Age: 77
Sex: M
Ethnicity: WHITE

Acute Kidney Injury Risk

Serum creatinine (mg/dL)

Gromerular Filtration Rate (mL/min/1.73 m2)

Clinical Data Repository and Clinical Predictive Modeling

Data Management for Digital Health, Winter 2018
12
How to Design a Clinical Data Repository (1)?

- Establish an Enterprise Master Patient Index
- Identify source systems and databases
- Start with the known and available
- Develop ETL services for integration

Source: Chelico et al. (2016)

https://discoverbiblog.wordpress.com/2016/11/08/introduction-for-etl/

How to Design a Clinical Data Repository (2)?

- Build specific datamarts, e.g. intensive care
- Define strategy to ensure data is up-to-date
- Important: barrier between transactional and analytical is becoming blurred!

Clinical Prediction Models
Occam’s Razor

- Models are an abstraction of reality
- May still be useful depending on the purpose
- Start simple: robust and difficult to break
- As per William of Occam (1287): ontological parsimony
- In diagnostics, not always the case:
  - Hickam’s dictum
  - “Patients can have as many diseases as they damn well please”

Clinical Prediction Models
Supervised or Unsupervised?

- **Supervised learning**
  - Labeled data is available
  - Categorical or numerical responses
  - Decision trees, Bayesian nets, ridge regression, etc.

- **Unsupervised learning**
  - No labeled data
  - Finding hidden patterns in data
  - Hierarchical clustering, k-means, etc.
Clinical Prediction Models
Supervised Learning: Types of Models

- Classification-based
  - Binary or multi-class
  - Diagnosis, risk stratification, hospital readmission

- Regression-based
  - Optimal drug dosage
  - Treatment plan adjustment
  - Survival analysis / survival curve
  - Time-to-event models, e.g. cancer mortality

https://blog.statsbot.co/machine-learning-algorithms-183cc73197c

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018

Clinical Prediction Models
Examples for Classification-based Models

Task: find the best discriminants for known outcomes

- Using logistic regression to predict stroke outcome
- Applying Deep Learning to diagnose cancer patients
- Analyzing electrocardiograms to detect atrial fibrillation
- Predict incidence of heart disease with life-style data
- Drawing a line is not always useful -> in non-linear contexts, machine learning may provide better outcomes

https://blog.statsbot.co/machine-learning-algorithms-183cc73197c

Data Management for Digital Health, Winter 2018
18
Clinical Prediction Models
Examples for Regression-based Models

Task: fit the best curve to predict a continuous variable

- Predicting cancer survival time using a Cox model
- Forecasting with Support Vector Regression (SVR) reduction of viral load after treatment
- Predicting length of stay (LOS) of ICU patients using local polynomial regression
- Never enough said: correlation is not causation
What does a Predictive Model look like?

- Statistical (or ML) analysis of predictor variables
- Contains thorough description of methods and derivation/validation cohort
- Outcomes usually entails a score calculation

A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure

- Navdeep Tangri, MD, FRCPC
- Lesley A. Stevens, MD, MS, FRCPC
- John Griffith, PhD
- Hocine Tighiouart, MS
- Ognjenka Djurdjev, MSc
- David Naimark, MD, FRCPC
- Adeera Levin, MD, FRCPC
- Andrew S. Levey, MD

**Context:** Chronic kidney disease (CKD) is common. Kidney disease severity can be classified by estimated glomerular filtration rate (GFR) and albuminuria, but more accurate information regarding risk for progression to kidney failure is required for clinical decisions about testing, treatment, and referral.

**Objective:** To develop and validate predictive models for progression of CKD.

**Design, Setting, and Participants:** Development and validation of prediction models using demographic, clinical, and laboratory data from 2 independent Canadian cohorts of patients with CKD stages 3 to 5 (estimated GFR, 10-59 mL/min/1.73 m²) who were referred to nephrologists between April 1, 2001, and December 31, 2008. Models were developed using Cox proportional hazards regression methods and evaluated.
Kidney Failure Risk Equation (4 Variable)

Estimate risk of progression to end-stage renal disease in CKD patients using age, sex, eGFR and proteinuria with KFRE

Age?
Unanswered

Years

eGFR?
Unanswered

mL/min/1.73m²

Urine Albumin Creatinine Ratio? (Note units carefully)
Unanswered

ng/mmol

Patient location?

North America

Non-North America

Results

Please answer all questions. The results will be computed once all questions are answered.
Predictive Models in the ICU
Acute Physiology And Chronic Health Evaluation (APACHE)

Source: Armed Forces Institute of Cardiology & National Institute of Heart Diseases (Pakistan)
What does a Predictive Model look like? Acute Physiology And Chronic Health Evaluation (APACHE)

- Risk classification for severely ill patients (e.g. ICU)
- Even though newer scores exist, still widely used
- Based on routine physiological measurements e.g. temperature, blood pressure, creatinine, white blood cell count, etc.

clinical investigations in critical care

The APACHE III Prognostic System*
Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults


What does a Predictive Model look like?
Acute Physiology And Chronic Health Evaluation (APACHE)

MD+Calc, https://www.mdcalc.com/apache-ii-score
Clinical Prediction Models
Establishing a CPM: Preparation and Dataset Selection

■ Step 1: Preparation
  □ What is the target outcome?
  □ What is the target patient?
  □ What is the target user?

■ Step 2: Dataset selection
  □ Is the data needed available?
  □ Is the data representative?
  □ What is the validation strategy?


Data Management for Digital Health, Winter 2018
25
Clinical Prediction Models
Establishing a CPM: Preparation and Dataset Selection

- Step 1: Preparation
  - What is the target outcome?
  - What is the target patient?
  - What is the target user?

- Step 2: Dataset selection
  - Is the data needed available?
  - Is the data representative?
  - What is the validation strategy?


Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
26
Clinical Prediction Models
Establishing a CPM: Variable Handling

- Expert judgement often necessary
- Previous significant factors should be used
- Avoid predictors that are possible correlated
- Merging categorical variables should be considered
- Scale matters in continuous variables: consistency
- Nominalization often needed for continuous variables
- Consider scale transformation (e.g. log)
- Normalization of values (e.g. from 0-1)
- Identify and handle outliers

---


Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
27
Clinical Prediction Models
Establishing a CPM: Missing Data

- Leaving out the missing ones: complete case analysis
- Single imputation x multiple imputation
  - Using “other” or “unknown”
  - Averaging occurrences, median or mean
- MICE (Multiple Imputation using Chained Equations)
- Regression model from the existing variables

---

Clinical Prediction Models
Establishing a CPM: Model Generation

- Selecting the proper algorithm
- Performing feature selection
  - Backward elimination
  - Stepwise selection
- Trade-offs between goodness of fit and complexity
  - Akeike Information Criterion
  - Bayes Information Criterion
- Perform parameter tuning
  - Optimization of hyperparameters

Clinical Prediction Models
Establishing a CPM: Model Generation - Decision Trees

- Decision rules inferred from data
- Advantages:
  - Interpretability ("white box")
  - Can often be combined with other algorithms
  - Requires little data preparation
- Disadvantages
  - As dimensions increase, so does complexity
  - May lack generalization, prone to overfitting
  - Creates bias if classes are unbalanced


http://web.eecs.umich.edu/~cscott/research/decision_tree.jpg

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
30
Clinical Prediction Models
Establishing a CPM: Model Generation - Decision Trees

- Given training set of size $N$, with feature vectors $x_i \in \mathbb{R}^d$, $i = 1, \ldots, N$ and corresponding class assignments $c_1, \ldots, c_N$.

- Calculate for every new node in the tree the best split of the w.r.t. one of the $d$ features at some threshold.

- Best split is determined using some impurity measure that should be minimal for the resulting split “populations” of training data, e.g. Gini-impurity: $Gini(E) = 1 - \sum_{i=1}^{\text{classes}} P_i^2$

Clinical Prediction Models
Establishing a CPM: Model Generation - Decision Trees

- Split criterion
  - Gini impurity: $Gini(E) = 1 - \sum_{i=1}^{\text{classes}} p_i^2$
  - Information gain (entropy)

- Unbalanced training data
  - Class weight: $Gini(E) = 1 - \sum_{i=1}^{\text{classes}} w_i \cdot p_i^2$

- Regularization hyperparameters:
  - Minimum samples for split
  - Minimum impurity decrease
  - Maximum tree depth

Restrict growth of decision → Averts overfitting

Clinical Prediction Models
Establishing a CPM: Evaluation and Validation

- Internal validation
  - Cross-validation
  - Bootstrapping

- External validation
  - Using a different data source
  - Ensure transportability and generalizability

- Measures of performance
  - ROC Curve
  - $R^2$, p-values

- True positive rate (TPR) / true negative rate
Clinical Prediction Models
Establishing a CPM: Measures of Performance

Measure
Sensitivity and specificity
Discrimination (ROC/AUC)
Predictive values: positive, negative
Likelihood ratio: positive, negative
Accuracy: Youden index, Brier score
Number needed to treat or screen
Calibration: Calibration plot, Hosmer-Lemeshow test
R² statistical significance: p-value (e.g., likelihood ratio test)
Magnitude of association, e.g., β coefficients, odds ratio
Model quality: Akeike IC/ Bayes IC
Net reclassification index and integrated discrimination improvement
Net benefit
Cost-effectiveness

Measures of model performance (Lee 2016)

Clinical Prediction Models

Measures of Performance: Precision and Recall

- Both are relevance measures
- Precision := % retrieved instances that are relevant
- Recall := % relevant documents in the result set

\[
\text{Precision} = \frac{3}{5} \\
\text{Recall} = \frac{3}{6}
\]
Clinical Prediction Models
Measures of Performance: Confusion Matrix

- Precision := \( \frac{TP}{TP + FP} \)
- Recall := \( \frac{TP}{TP + FN} \)
- F-measure := \( 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \)
- Sensitivity := True Positive Rate := Recall
- Specificity := True Negative Rate

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>True</td>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td>Predicted</td>
<td>False</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
AKI prediction (classifier accuracy)
Receiver Operating Characteristic (ROC) curve

- **TPR (Sensitivity)**
- **FPR = (1 - Specificity)**
- **Area Under the Curve (AUC) = 0.87**

**Recall (sensitivity)**
- \[ \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \]

**Precision**
- \[ \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \]

**Specificity**
- \[ \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \]
Clinical Prediction Models
Receiver Operating Characteristic (ROC) Curve

- Performance of a binary classifier
- Plot showing TPR and FPR
- Varying classification thresholds
- Allows comparison between classifiers
- The higher the AUC (Area Under the Curve) the better

\[ TPR = \frac{TP}{TP + FN} \]

\[ FPR = \frac{FP}{FP + TN} \]

TPR = True Positive Rate
FPR = False Positive Rate
TP = True Positives
FP = False Positives
TN = True Negatives
FN = False Negatives

AUC = Area Under the Curve

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
Developing a Clinical Prediction Model
Cross-validation (k-fold)

- How general is the model?
- Important to avoid overfitting
- Useful for internal validation of a model
- Training and validation sets
- Random split into k subsamples

Final Accuracy = Average(Round 1, Round 2, ...)

Clinical Data Repository and Clinical Predictive Modeling
Data Management for Digital Health, Winter 2018
39
Con: Most clinical risk scores are useless

Friedo W. Dekker, Chava L. Ramspek and Merel van Diepen

Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

Correspondence and offprint requests to: Friedo W. Dekker; E-mail: F.W.Dekker@lumc.nl
Clinical Prediction Models
Most Common Pitfalls and Flaws

Flaws and pitfalls in model development

Flaws and pitfalls in model validation

Lack of impact studies

Clinical Prediction Models
Most Common Pitfalls and Flaws

- Too many predictors for the number of events
- Binarizing continuous variables
- Assuming predictors are linear
- Low predictive performance
- Poor calibration (or not reported at all)

Clinical Prediction Models
Most Common Pitfalls and Flaws

- Incomplete or poorly conducted
- Research focused on new predictive models
- External validation rarely performed
- Models tend to perform poorly

Clinical Prediction Models
Most Common Pitfalls and Flaws

- Impact is different from development and validation
- Risk scores do not necessarily improve care
- Randomized controlled trials are necessary
- Costly and time-intensive undertaking

Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations

Luke Eliot Hodgson,1 Alexander Sarnowski,2 Paul J Roderick,1 Borislav D Dimitrov,1 Richard M Venn,3 Lui G Forni2,4
Clinical Prediction Models
Prediction Models for Hospital-Acquired Acute Kidney Injury

- 53 HA-AKI prediction models
- Methodological shortcomings
- Incomplete reporting (e.g. calibration)
- Only 3 employing multiple imputation
- Rare external validation
- Little consideration for use in practice
- No impact analysis studies

---

Con: Most Clinical Risk Scores are Useless
... If not Properly Developed, Validated and Assessed!

Flaws and pitfalls in model development

Flaws and pitfalls in model validation

Lack of impact studies

What to Take Home?

- Aspects regarding Clinical Data Repository
- Step-by-step process to establishing clinical prediction models (CPMs)
- Measures of performance for CPMs: precision, recall, ROC curve, etc.
- Common pitfalls with CPMs: developing is only the beginning