PRE CONFERENCE RESEARCH METHODS WORKSHOP
Opening the Black Box: Ethically responsible use of big data

Society for Social Work and Research Annual Conference: Achieving Equal Opportunity, Equity, & Justice; Washington DC, January 11, 2018



Presenters

- **Brian Chor, Ph.D.,** Senior Researcher at the Chapin Hall Center for Children at the University of Chicago
- **Teresa De Candia, Ph.D.,** Deputy Director of Provider Outcomes, NYC Administration for Children Services
- **Diane DePanfilis, Ph.D., M.S.W.,** Professor at the Silberman School of Social Work, Hunter College, City University of New York
- **Maria Rodriguez, Ph.D., M.S.W.,** Assistant Professor at the Silberman School of Social Work, Hunter College, City University of New York
- **Ravi Shroff, Ph.D.,** Assistant Professor, Steinhardt School of Culture, Education, Human Development/Center for Urban Science and Progress, New York University
- **Dana Weiner, Ph.D.,** Policy Fellow at the Chapin Hall Center for Children at the University of Chicago
- **Allon Yaroni, Ph.D.,** Assistant Commissioner for Data Analytics, NYC Administration for Children's Services

ACKNOWLEDGEMENTS

NYC-ACS Predictive Analytics Technical Team

ACS:
- Allon Yaroni
- Teresa De Candia
- Leila Pree
- Synia Wong

CUNY NYU Team
- Diane DePanfilis
- Ravi Shroff
- Maria Rodriguez

Chapin Hall Team
- Dana Weiner
- Brian Chor
- Mike Stiehl
- Kiljoong Kim

Introductions: Create your Team Resume

In small groups, get to know each other by discussing the following…prepare to report back:

- The disciplines you represent (e.g. psychology, social work, public health)
- Total years of experience using big data
- Total years of Policy, Program, or Practice experience
- Something unique about your team/individual members
- 2-3 outcomes your group wants to achieve in this workshop

Report Out on Your Team Resume

- Present the results of team's answers to the intro questions.

Introductions: The Presenter Team

- Disciplines: Genetics, Mathematics, Public Policy, Psychology, Social Work
- # Years Big Data = 61
- # years policy, program, practice = 109

- Unique talents:
  - Black belt in Tae Kwon Do
  - Eat core of apple when eating an apple
  - Love snow storms
  - Can deadlift body weight
  - Former aerobic dance instructor
  - A strong British accent
  - Can bake a *mean* cheesecake

Outcomes Identified by Presenters

- Each participant takes home one new idea or skill
- We develop new connections
- Each team achieves outcomes

Ground Rules for this Session

- Stay “present” and fully participate (cell phones on silent and out of view)
- All of us are teachers and all of us are learners
- Each person is responsible for their own comfort
- Everything that is said in this room, stays in this room
- Questions are welcome any time

Workshop Objectives/Plan

- Participants will gain an understanding of:
  - Strategies for developing and managing inter-agency collaboration to address complex analytic questions with direct implications for policy, programs, and/or practice
  - Methods for using predictive and/or explanatory analytics to leverage administrative data to enhance understanding of the needs of populations served.
  - Essential conditions to enhance equity in models & results
  - The application of predictive and/or explanatory analytics for economizing resources, developing specific targets, and matching target groups to necessary resources

Essential Ingredients for Building Models to be Useful to the Field

DEFINITIONS

What is Predictive Analytics?

- *Predictive analytics (PA)* is the practice of extracting information from existing data sets in order to identify patterns and predict the likelihood of future outcomes.
- *Predictive Risk Modeling* (PRM) is an approach to predictive analytics that uses routinely collected administrative data to identify individuals at risk of an adverse event or to inform prevention efforts.

Predictive Analytics: Goals

Overall: Apply administrative data to understanding the level of care, attention, and service a target population may need. Examples may include:

- Assist decision-making by providing more information to supervisors and additional resources to front-line staff
- Adjust Quality Assurance reviews to account for the distribution of challenging/high-need cases
- Identify appropriate services that may mitigate propensity for negative outcomes and strengthen protective factors

Essential Ingredient for Building Models to be Useful to the Field

COLLABORATIONS

Why Collaborate?

Internally, to:

- Build consensus around common questions or problems that exist that can be informed by these approaches
- Promote buy in among internal stakeholders
- Make meaning of findings
- Ensure the validity of the data used for modeling

Externally, to:

- Enhance the rigor of analytic approaches by engaging academic partners who can explore the convergent validity of different methods
- Facilitate broad understanding among advocates
- Build public awareness of the ethical considerations of predictive analytics

Developing Inter-Agency Collaboration: Inclusive Process *(NY Example)*

Internal Stakeholders

- Child Welfare Programs
- Division of Child Protection
- Division of Preventive Services
- Family Permanency Services
- Division of Policy, Planning, and Measurement

External Stakeholders

- Chapin Hall
- City University of New York
- New York University
- Contracted Provider Agencies Quality Assurance
- Casey Family Programs

Connecting to an Agency Priority: Frequently Encountered Families *(NYC Example)**

- Families that are the subject of several child protective investigations where safety and risk remain a concern
  - Families with two or more reports within the prior six months, or four or more within the prior two years
- Families that have been receiving multiple preventive spells for years without achieving their goals.
  - Families involved with Preventive Services who are experiencing elevated risk factors.
  - Families involved with Preventive Services with long length of services, as measured by 18 months.
- Children who are in and out of foster care and have yet to achieve permanency.
  - Children who achieve permanency and later re-enter into placement.
  - Children who experience foster care placement and later are involved in a case as a case parent.

**-Example of focusing on a unifying theme*

Bridging Practice & Data

Practice

- What outcomes do we wish to affect?
  - How can we use data to help you meet your goals?
- What are some potential predictive variables?
- What is the application?
  - What processes do we want to put into place to increase positive outcomes?

Data

- What data are available?
  - Enough quantity?
  - Enough quality?
- What analytic approach do we wish to use?
  - Exploratory?
  - Machine Learning?
- Internal or external?
  - Internal capacity building?
  - Contracted partners?

What assets (partnerships) do you have? What road blocks do you anticipate?

DISCUSSION/ACTIVITY

In your Teams

- Discuss the degree to which you are already partnering with agencies interested in PRM.

- What roadblocks have you experienced or do you anticipate?
- What options do you have to address these barriers?

Report out: Partnerships

- Types of partnerships:
  - Formal
  - Informal

- What roadblocks have you experienced or do you anticipate?
  - Report Out

- What options do you have to address these barriers?
  - Report Out

Selecting the right methods to answer questions

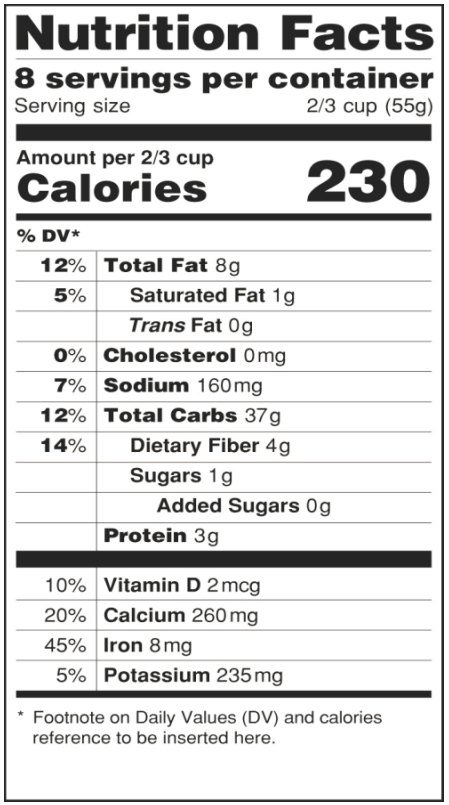
METHODS

Outline

Background
Key Concepts
NYC Example
Pitfalls

Methods and Ethics: Why it Matters?

- Stewardship of administrative data
- Repercussion of prediction quality on target populations
- Fairness/Unfairness of prediction
- Transparency of prediction
  - *Food for thought: Much like a nutrition value label, knowing what goes into and out of a predictive analytic model, the expected impact of the model on target populations, should be a requirement*
- Other reasons why it is important to understand methods in the context of ethics of using big data?



Nutrition Facts

8 servings per container

Serving size 2/3 cup (55g)

Amount per 2/3 cup

Calories 230

% DV*	
12%	**Total Fat** 8g
5%	**Saturated Fat** 1g
	***Trans* Fat** 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	**Dietary Fiber** 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	**Vitamin D** 2mcg
20%	**Calcium** 260mg
45%	**Iron** 8mg
5%	**Potassium** 235mg

* Footnote on Daily Values (DV) and calories reference to be inserted here.

Predictive Analytic Methods

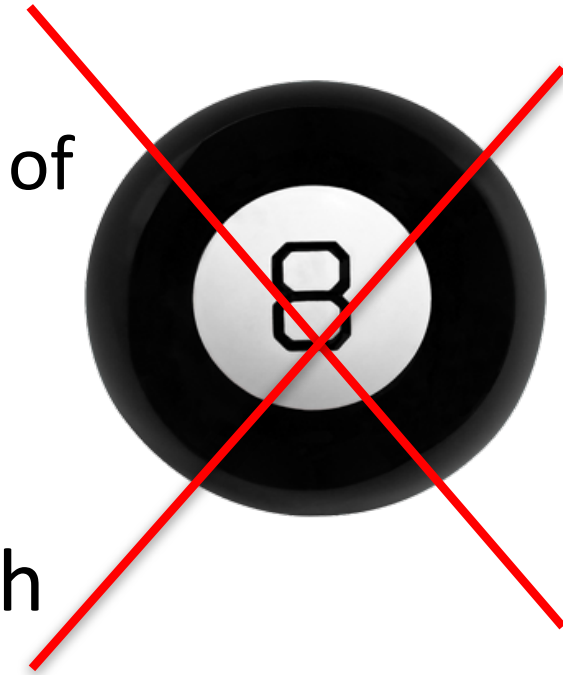
Goal: Empirically predict/estimate the likelihood/probability of an event/outcome of interest

Prediction ≠ Causality

Prediction ≠ Crystal ball

Prediction ≠ Absolute truth

Prediction ≠ Error free



Food for thought: Other common preconceived notions about predictive analytic methods?

Predictive Analytic Methods: Building Blocks

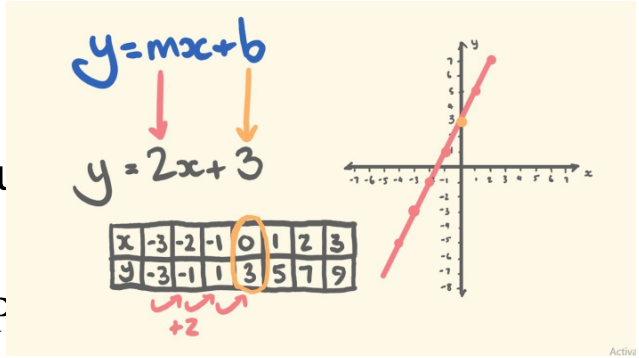
- Administrative data: What are existing data sources?
  - Longevity (for training set vs. tessting set)
  - Volume
  - Accessibility
  - Replicability
- Predictors: What associations are you hoping to find?
  - Kitchen sink
  - Theoretically and/or empirically based
  - Well-defined and specific
- Outcome: What prediction are you hoping to intervene?
  - Availability of appropriate interventions
  - Well-defined and specific
- Target population: Who are you hoping to identify?
  - Well-defined and specific

<u>Ethics Checkpoints:</u>

- ❑ Data entry bias
- ❑ Missing data
- ❑ Operationalization
- ❑ Relevance
- ❑ Feasibility
- ❑ Representativeness

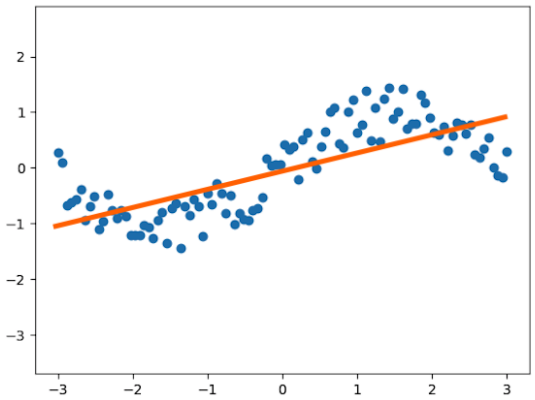
Predictive Analytic Methods: Regression

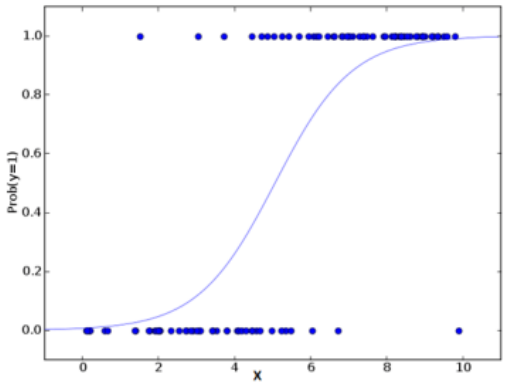
- Estimates relationship among variables
- Recall 6th grade math: y = mx + b
- Dependent variable (Y) = Outcome (categorical or continu
- Independent variables (X) = Predictors
- Relationship = Regress Y on X's = Association between I and Outcome
  - Each X carries a "weight" that indicates the strength of its relationship with the outcome
- Prediction = Best-fitted regression line = "How closely do the Predicted Outcomes align with the Observed Outcomes?"



Regression

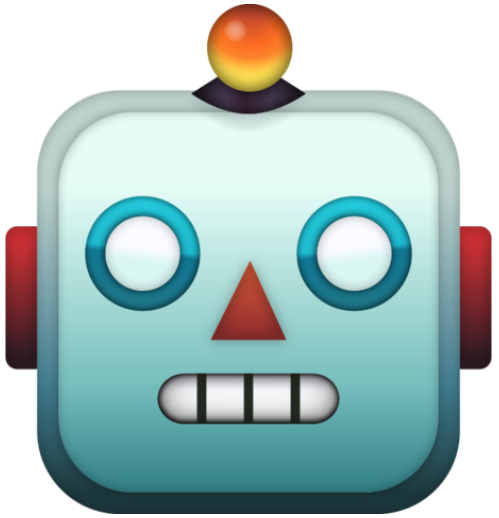
- Estimates relationship among variables
  - Allows you to describe the strength ("weight") of a predictor's relationship with an outcome
  - e.g., Youth with a prior investigation is 3 times more likely to re-enter foster care than you with no prior investigation
  - Prediction: Minimized difference between a "predicted" outcome and an "observed" outcome





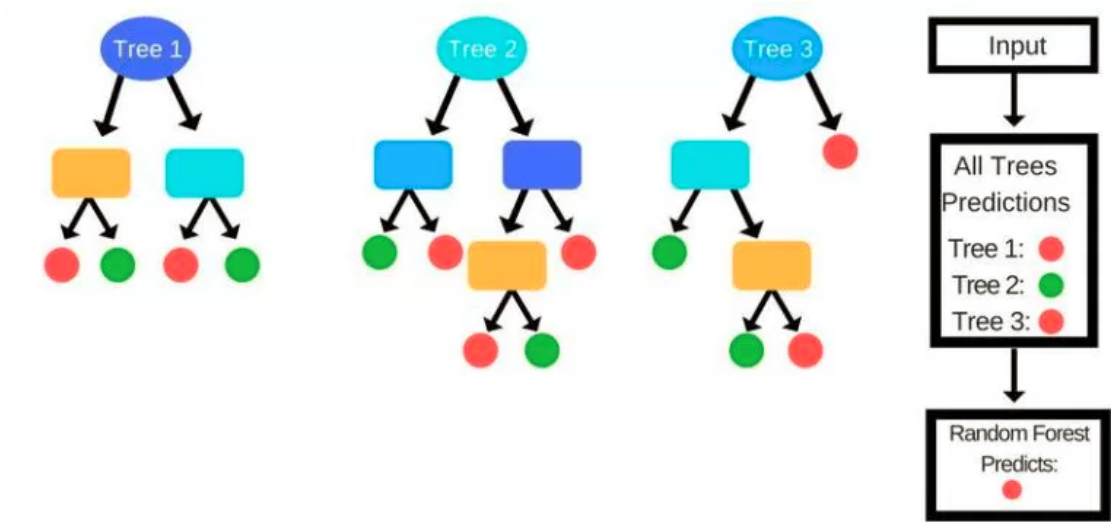
Predictive Analytic Methods: Machine Learning

- A form of artificial intelligence (i.e., “learning”) that automates analytic model building using mass data by searching for patterns and creating algorithms to make prediction
- Examples of machine learning methods
  - Decision tree learning (e.g., random forest)
  - Deep learning (e.g., neural network)
- Inputs = ~Predictors that are to be learned and studied
- Output = Outcome prediction
  - Categorical outcome: Classification algorithm
  - Continuous outcome: Regression algorithm

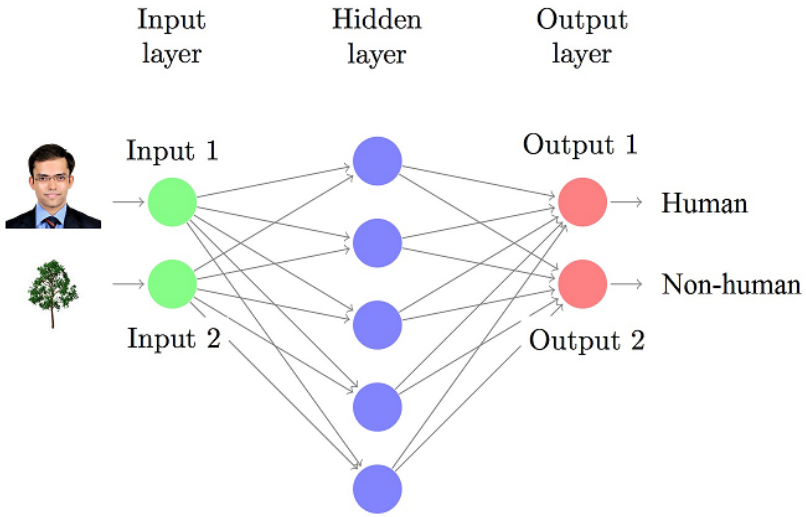


Predictive Analytic Methods: Machine Learning

Random Forest

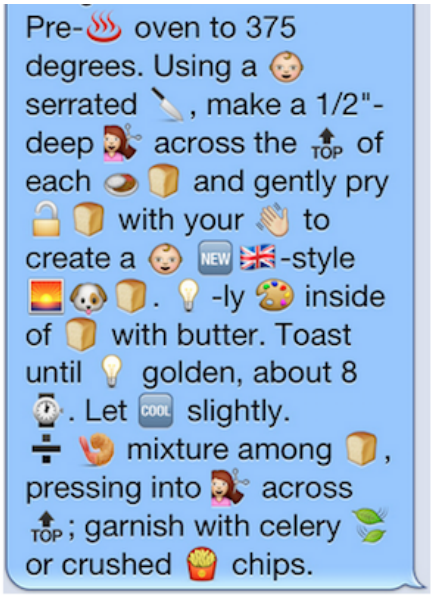


Neural Network



Learning from the Past to Predict the Future

- Develop a recipe, test your recipe yourself (does it taste good?), and have someone else use your recipe (does it taste good?)
- "Training" dataset – To develop a model
  - Includes predictors and outcomes for a population similar to your target population
  - To develop and fit the parameters that produce a predictive model (i.e., tinkering with your recipe)
- "Testing" dataset – To evaluate a model
  - As similar to the "training" dataset as possible
  - To provide an unbiased evaluation of the final model from the "training" dataset (i.e., an independent judge validating and tasting your recipe)



Concepts 1 - Predictions, Errors, and Error Rates

True Positive (TP)

True Negative (TN)

False Negative (Error) or FN

False Positive (Error) or FP

Refers to whether prediction is correct

Refers to prediction

ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

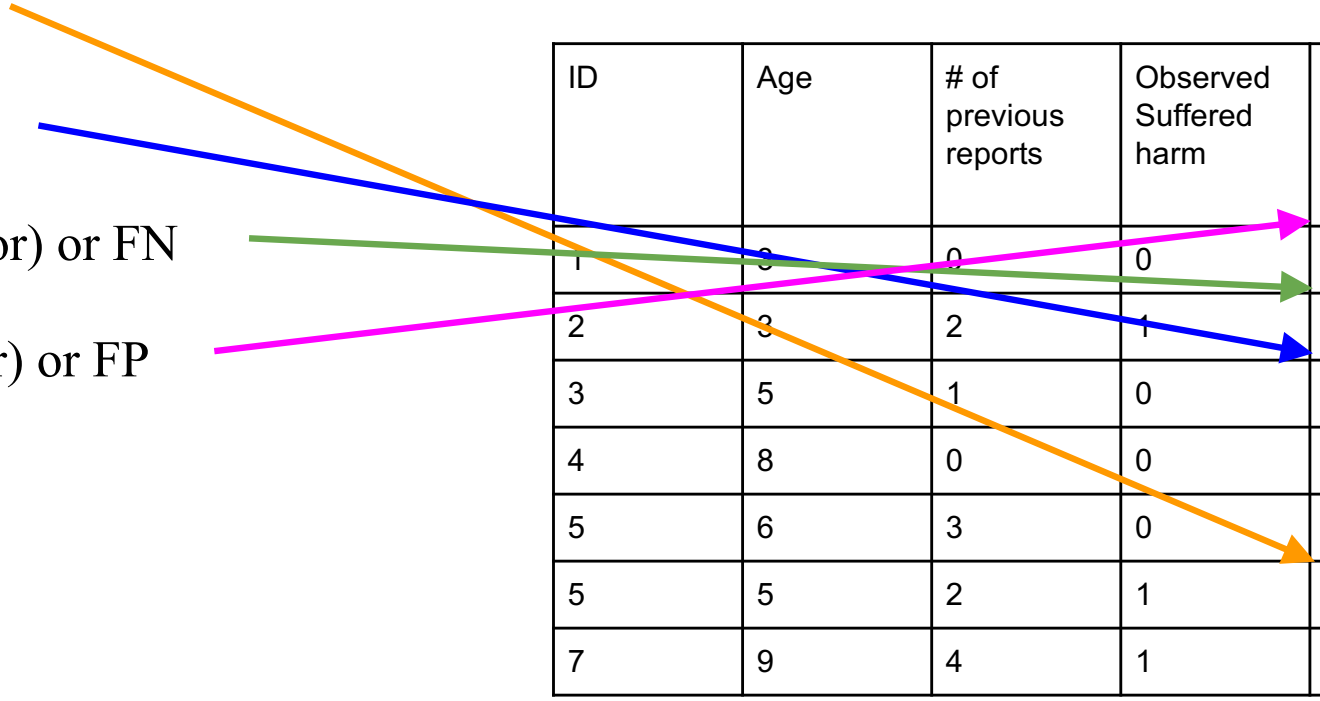
Concepts 1 - Predictions, Errors, and Error Rates

True Positive (TP)

True Negative (TN)

False Negative (Error) or FN

False Positive (Error) or FP



ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

Concepts 1 - Predictions, Errors, and Error Rates

Accuracy - proportion of correct predictions; often too general and NOT the most important thing to us

We can frame our questions in two ways:

1. What proportion of our observations are classified correctly?
2. What proportion of our predictions are correct?

ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

Concepts 1 - Predictions, Errors, and Error Rates

1. What proportion of our observations are classified correctly?

True Positive Rate (ie sensitivity, recall): What proportion of positive observations are correctly predicted?

False Positive Rate (inverse is called specificity): What proportion of negative observations are incorrectly predicted as positive?

ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

Concepts 1 - Predictions, Errors, and Error Rates

2. What proportion of our predictions are correct?

Positive Predictive Value (ie precision): What proportion of positive predictions are correct?

..aargh so much jargon!! 😱😱😱

That's why these terms make up what is called a "confusion table" 😉

ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

Concepts 1 - Predictions, Errors, and Error Rates

Exercise - Try calculating:

a) Accuracy =
b) True Positive Rate =
c) False Positive Rate =
d) Positive Predictive Value =

ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

Concepts 1 - Predictions, Errors, and Error Rates

Exercise - Try calculating:

a) Accuracy = 4/7
b) True Positive Rate = 2/3
c) False Positive Rate = 1/2
d) Positive Predictive Value = 1/2

ID	Age	# of previous reports	Observed Suffered harm	Predicted Suffered Harm
1	3	0	0	1
2	3	2	1	0
3	5	1	0	0
4	8	0	0	0
5	6	3	0	1
5	5	2	1	1
7	9	4	1	1

Concepts 2 - Thresholds

Our rule: an observation is predicted to have a positive outcome if and only if it has a risk score ≥ 10

Wait a sec! Predictions are usually continuous "risk scores". How do we come up with categorical predictions?

Oops, we didn't show you the extra step before, which involves picking a threshold (which is basically a kind of rule). Tsk tsk.

In this case risk scores range between 1-20, and we had decided to pick the threshold of 10. Stay with me for why one threshold is picked over another.

ID	Age	# of previous reports	Observed Suffered harm	Risk Score (1-20)	Predicted Suffered Harm ≥ 10
1	3	0	0	12	1
2	3	2	1	7	0
3	5	1	0	2	0
4	8	0	0	6	0
5	6	3	0	11	1
5	5	2	1	17	1
7	9	4	1	11	1

Concepts 2 - Thresholds

a) Threshold of >=10 gives you:
   i) TPR of 2/3 and FPR of 1/2
b) Threshold of >=13
   i) We will catch fewer kids who have outcome, so it should lower our true positive rate, which is bad
   ii) Should we expect fewer or more false positives? Fewer, which is good
   iii) Calculate new predictions with new threshold ≥ 14 and compute TPR and FPR using new threshold. TRP = ? and FPR = ?

ID	Age	# of previous reports	Observed Suffered harm	Risk Score (1-20)	Predicted Suffered Harm ≥ 10
1	3	0	0	12	1
2	3	2	1	7	0
3	5	1	0	2	0
4	8	0	0	6	0
5	6	3	0	11	1
5	5	2	1	17	1
7	9	4	1	11	1

Concepts 2 - Thresholds

a) Threshold of >=10 gives you:
   i) TPR of 2/3 and FPR of 1/2

b) Threshold of >=13
   i) We will catch fewer kids who have outcome, so it should lower our true positive rate, which is bad
   ii) Should we expect fewer or more false positives? Fewer, which is good
   iii) Calculate new predictions with new threshold ≥ 14 and compute TPR and FPR using new threshold. TPR = ⅓ and FPR = 0

ID	Age	# of previous reports	Observed Suffered harm	Risk Score (1-20)	Predicted Suffered Harm ≥ 14
1	3	0	0	12	0
2	3	2	1	7	0
3	5	1	0	2	0
4	8	0	0	6	0
5	6	3	0	11	0
5	5	2	1	17	1
7	9	4	1	11	0

Concepts 3 - Thresholds

The freedom to set thresholds is a good thing. It allows us more control to maximize our objectives.

Setting thresholds is done by weighting the (often totally abstract) costs of false positive errors and false negative errors (e.g., reduced risk of child abuse vs. increased risk of over-monitoring family).

If one of these is small compared to the other, thresholds are often set according to available resources (e.g., service will be offered to 400 families with highest risk scores).

Concepts 3 - ROCs and their AUCs

ROCs are graphical representations of the tradeoff between True Positive Rate and False Positive Rate afforded by the full range of thresholds for a given model.

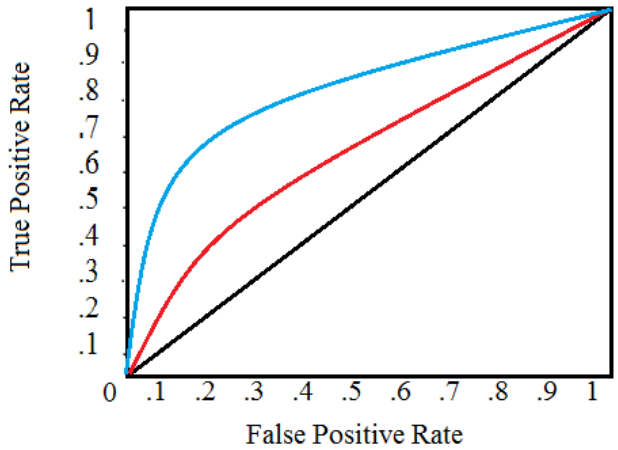
We can catch more positives

-at the cost of mislabeling negatives as positive

-medical tests do this A LOT and we are mostly ok with that

Alternatively, we can keep false positives low

-at the cost of missing warning signs



-natural disaster prediction does this and we are mostly ok with it (evacuating people is expensive)

Concepts 3 - ROCs and their AUCs

Area under the curve (AUC):

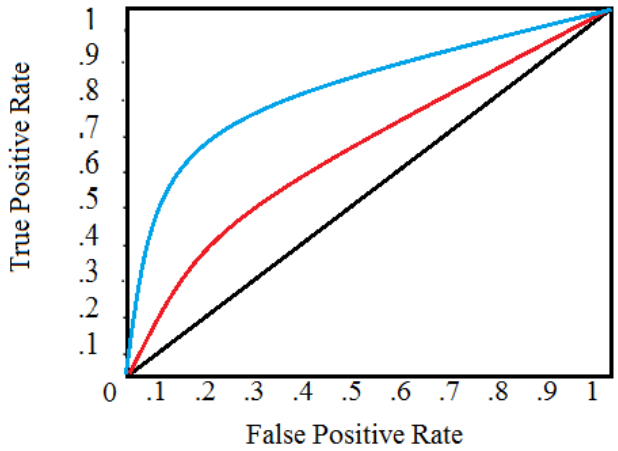
1 is perfect, 0.5 is guessing

How high is decent depends on:

-particular domain

-absolute cost of errors

-what decision making alternatives exist



Severe harm example

- **Outcome:** Severe maltreatment, defined as one or more future Substantiated Severe Harm Allegations against the child and occurring within 2 years of investigation start date, 5.7% prevalence (more on next slide)
- **Training and Test Samples:** ~200k children in investigations ending between Jan. 1, 2013 and Dec. 31, 2014 (2 years)
- **Time of prediction:** day 7 of investigation
- **Predictors:** ~200, collected from data prior to time of prediction, including demographic data, past and current investigation data

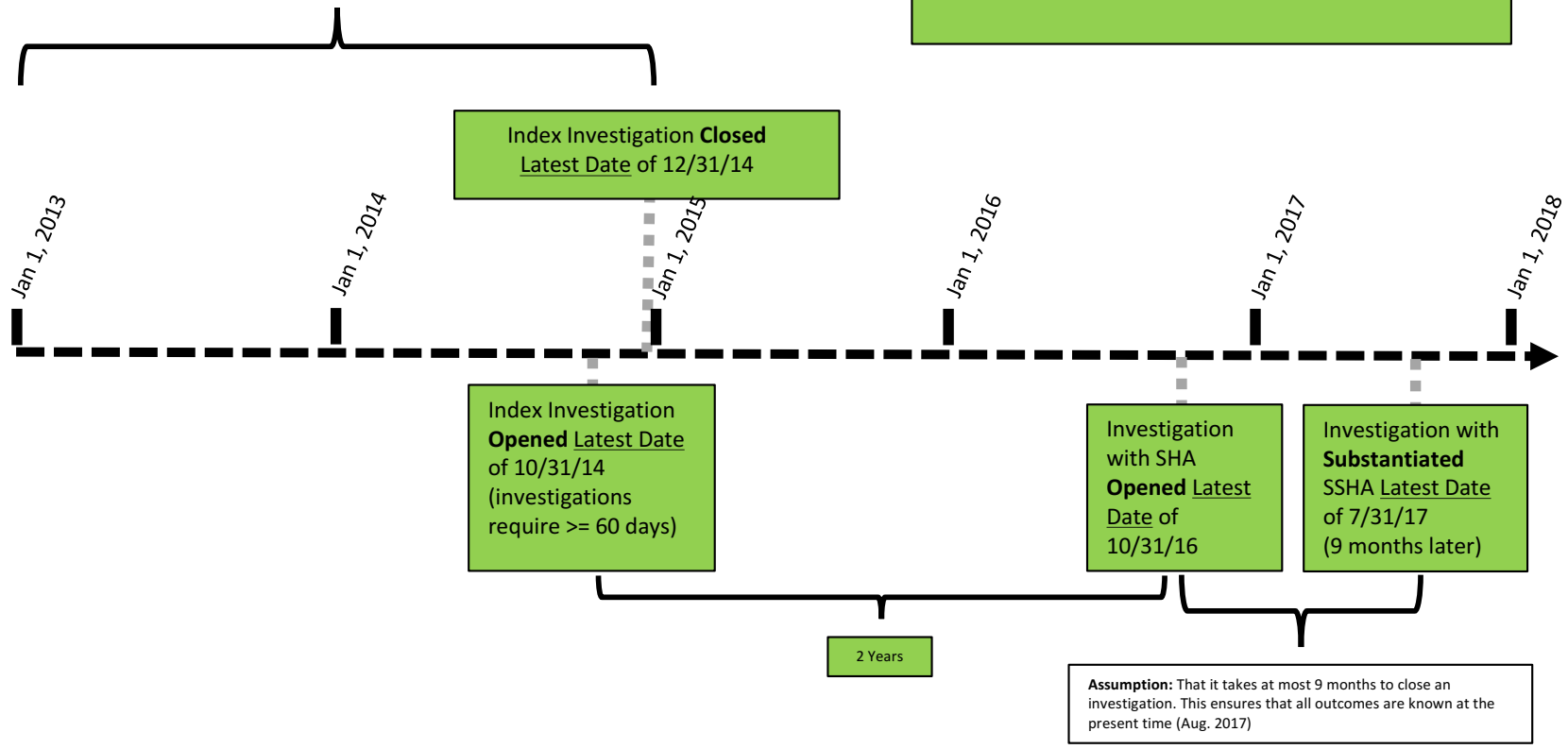
Severe Harm example: Outcome

Allegation	Include
Abandonment	If child is under 3
Burns/Scalding	Yes
Child Drugs/alcohol Use	No
Choking/Twisting/Shaking	Yes
Education Neglect	If child is under 3
Emotional Neglect	No
Excessive Corporal Punishment	If child is under 7
DOA/Fatality	Yes
Fractures	Yes
Inadequate Food/Clothing/ Shelter	No
Internal Injuries	Yes
Inappropriate Custodial Conduct	No
Inadequate Guardianship	If child is under 3
Inappropriate Isolation/Restraint	If child is under 7
Lacerations/Bruises/Welts	Yes
Lack of Medical Care	If child is under 7
Lack of Supervision	If child is under 3
Malnutrition/ Failure to Thrive	Yes
Parent Drug/ Alcohol Misuse	If child is under 3
Poisoning/ Noxious Substances	Yes
Swelling/ Dislocation/Sprains	Yes
Sexual Abuse	Yes
Other	No

Severe Harm example: Timeline

Training and Test Samples for Inclusion:
Children in investigations that closed between 1/1/13 and 12/31/14

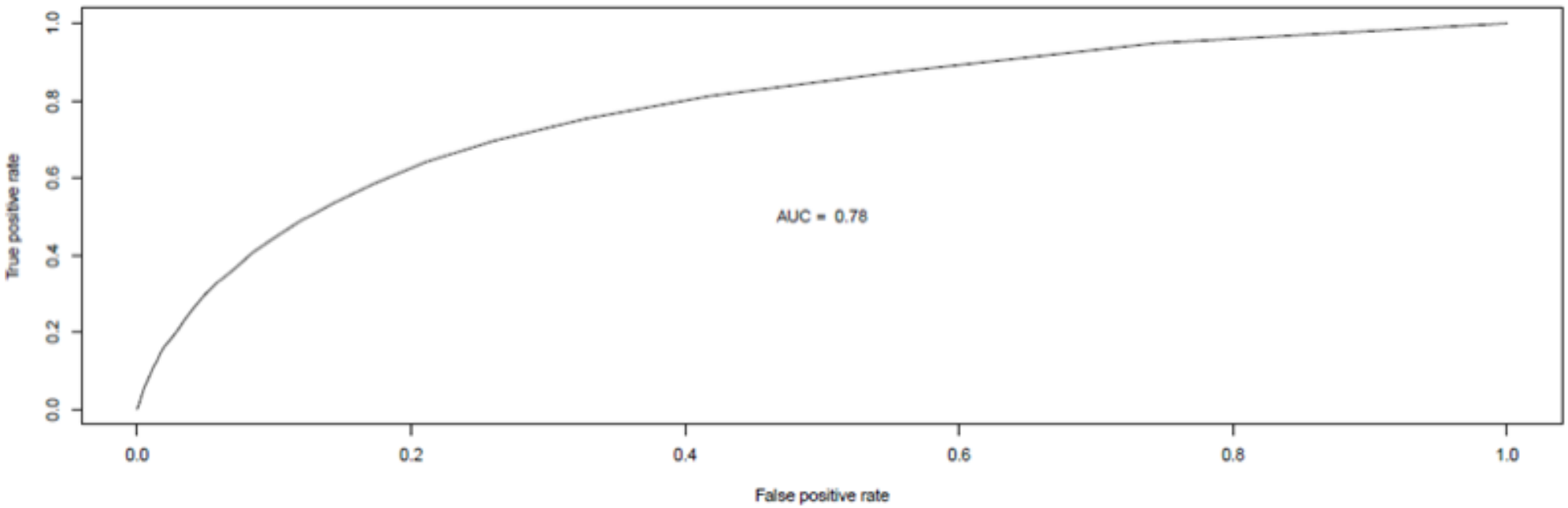
Example:
Latest Case Scenario



Severe harm example: Predictors

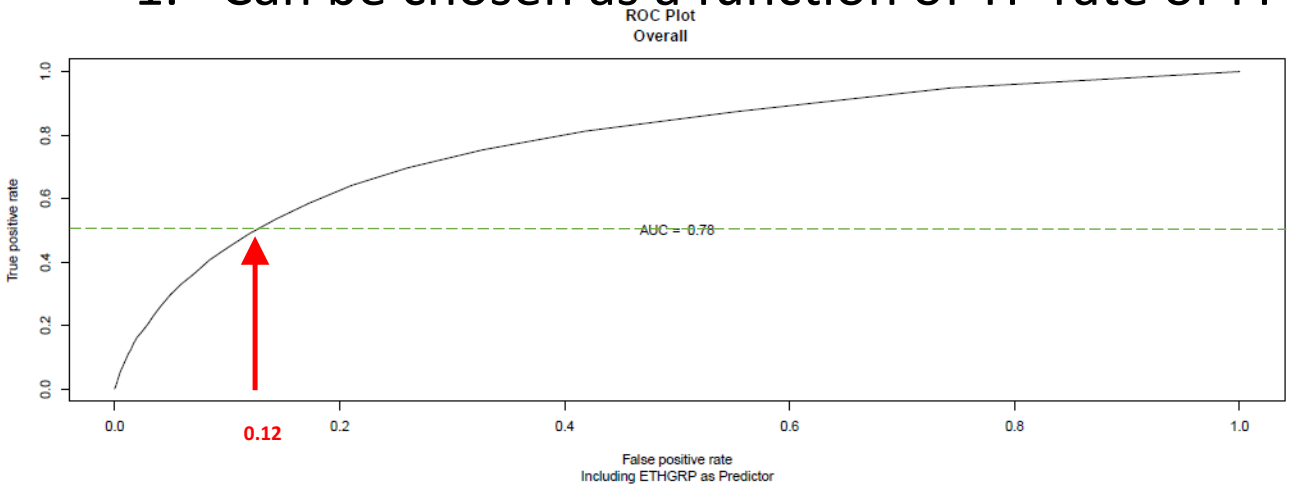
- **Current and past investigations**
  - Number of investigations
    - Total and indicated
    - Recent total and indicated
    - Child has role
    - Perpetrator (confirmed + non-confirmed)
  - Time known to DCP
  - 13 High Priority Codes
  - 23 Allegation types
  - 19 Safety Factors
  - Risk Assessment Profile (RAP) scores
- **Demographics**
  - Ages of child and mother; sibling counts by age (e.g., 1 sibling between 11-18)
  - Child's race ('Hispanic', 'Afr Am', 'White', 'Asian/Pacific Island', 'Other', 'Unknown')
  - Child's gender
  - Community district and county (from current stage)
- **Model Refinements**
  - FASP/RAP Questions
  - Foster care history

Severe harm example: ROC



Severe harm example: Thresholds

1. Can be chosen as a function of TP-rate or FP-rate (not both)



TP rate of 0.5 corresponds here to a FP rate of 0.12

2. Can be chosen as a function of available resources

The above threshold makes positive predictions for the top 15% of children at risk (some of whom will be FPs)

Potential pitfalls: exercise 1

Training Data I

ID	Age	# of previous reports	Suffered harm Observed
1	3	0	0
2	3	2	1
3	5	1	0
4	8	0	0
5	6	3	0
5	5	2	1
7	9	4	1

Algorithm

If age is under 6 AND no reports, **no harm**

If age is 6 or over AND < 3 reports, **no harm**

Otherwise, **harm**

Exercise: Calculate the accuracy of this algorithm on the training data

Training Data II

Id	Age	# previous reports	Gender	# siblings	Single parent household	Suffered harm Observed
1	3	0	M	2	1	0
2	3	2	M	3	0	1
3	5	1	M	1	1	0
4	8	0	F	1	1	0
5	6	3	F	3	0	0
5	5	2	M	0	1	1
7	9	4	F	2	0	1

Algorithm

If age is 3 AND no previous reports AND male AND > 1 siblings, **no harm**

If age is 5 AND one previous report AND male AND single parent household, **no harm**

If age > 7 AND no previous reports AND female AND single parent household, **no harm**

If age is 6 AND has > two previous reports AND female AND > two siblings, **no harm**

OTHERWISE, **harm**

Exercise: Calculate the accuracy of this algorithm on the training data

Question: Which algorithm performs better on the training data?

Question: Which algorithm do you think will generalize better to new data?
[How would you test this?]

Question: Which algorithm do you think you should use in practice?

Overfitting

“The production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.”
[from Oxford Dictionaries]

How can we tell if we’re overfitting [our algorithm to our training data]?

What can we do about it?

Solution: check performance on “test” data

Use more data [NOT the same data used to train the algorithm] to check performance.

Key idea: compare the *actual* outcomes in a “test” dataset to the predictions of the algorithm on that “test” dataset.

Solution: check performance on “test” data

Use more data [NOT the same data used to train the algorithm] to check performance.

Key idea: compare the *actual* outcomes in a “test” dataset to the predictions of the algorithm on that “test” dataset.

Question: Where should we get “test” data from?

Question: How much “test” data do we need?

Where should we get “test” data from?

Usually reserve a subset of original training data

Where should we get “test” data from?

Usually reserve a subset of original training data

Question: what about data from a different jurisdiction?

Where should we get “test” data from?

Usually reserve a subset of original training data

Question: what about data from a different jurisdiction?

Generally, we want test data to be as similar to real data (that you will apply model to) as possible.

How much “test” data do we need?

Possible options: ½ or ⅓ of training data, chosen randomly. This really depends on how much data you need to train your model
[*cross-validation* is a strategy that conserves training data]

Questions:
Does population change over time (i.e., do we need to regularly check performance on a new test set?)

When should we retrain the model?

Potential pitfalls: exercise 2

Exercise: Which of the following predictors should we definitely **not** include when training the model which predicts (on day 7 of investigation) the outcome of severe harm (within 2 years of investigation start)?

- Age of child
- Age of mother
- Post-investigation survey results
- Sibling count by age (e.g., “1 sibling between 11-18”)
- Foster care history (e.g., “3 previous spells in foster care”)
- Time known to DCP
- 23 allegation types
- Case outcome/disposition (e.g., “indicated” or “unsubstantiated”)
- ZIP code at start of investigation
- ZIP code at end of investigation

Exercise: Which of the following predictors should we definitely **not** include when training the model which predicts (on day 7 of investigation) the outcome of severe harm (within 2 years of investigation start)?

- Age of child
- Age of mother
- Post-investigation survey results
- Sibling count by age (e.g., "1 sibling between 11-18")
- Foster care history (e.g., "3 previous spells in foster care")
- Time known to DCP
- 23 allegation types
- Case outcome/disposition (e.g., "indicated" or "unsubstantiated")
- ZIP code at start of investigation
- ZIP code at end of investigation

Upshot: when training a model, make sure you don't use features that wouldn't be available at the time of prediction.

Doing so can yield misleading measures of model accuracy

Also, pay attention to **variable operationalization**: make sure variables are constructed the same way in the training data, testing data, and the "real" implementation data.

Validation

Principles of model validation:

Compare to best possible alternative (could be another model, or human decisions made prior to model implementation). Usually not “ideal world with no errors.”

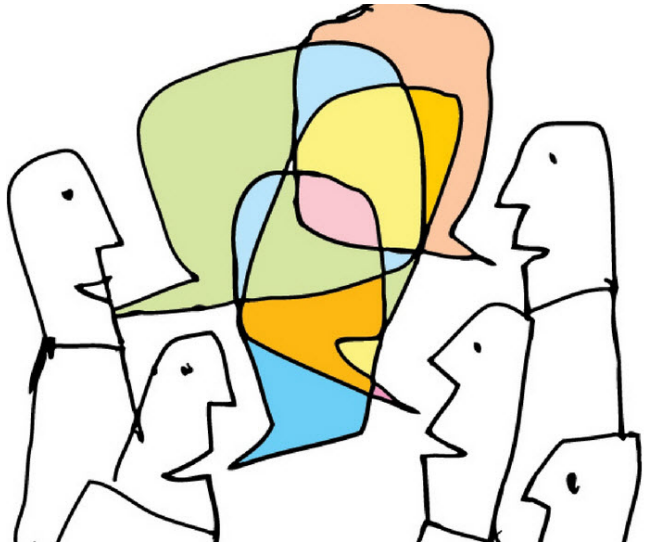
Think about model evaluation when designing and implementing predictive analytics.

Discussion Questions

- What questions do you hope to answer using predictive analytics?

- What challenges and barriers do you see in developing, translating, and applying predictive analytic models?

- What additional precautions should we consider when using predictive analytics?



Applying an equity lens when planning analysis and interpreting results

ETHICS & EQUITY

Outline

- **Background -** Defining Ethical PRM
- **Conditions for Equity -** Model Predictions
- **Conditions for Equity -** Beyond Model Predictions

Outline

- **Background -** Defining Ethical PRM
- **Conditions for Equity** - Model Predictions
- **Conditions for Equity** - Beyond Model Predictions

Background – *Key Ethical Concerns*

“As for the poisonous effect of ideology on the debate over public assistance: Big data promises something closer to an unbiased, ideology-free evaluation of the effectiveness of these social programs. We could come closer to the vision of a meritocratic, technocratic society that politicians from both parties at state and local levels — those closest to the practical problems their constituents face — have begun to embrace” (Mason, 2018).

Background - *Key Ethical Concerns*

- Data (Capotasto, 2017; Teixeira & Boyas, 2017)
    - incomplete, hard to get from other pertinent systems, overall quality
- Resources (Capotasto, 2017; Teixeira & Boyas, 2017)
    - jurisdictions lack expertise
    - outsourcing model development also problematic
    - upfront costs pricey either way
    - long term sustainability unknown

Background - *Key Ethical Concerns*

- Impacts (Capotasto, 2017; Teixeira & Boyas, 2017: O'Neill 2016):

    - classifying families based on individual-level risk profiles

    - historical bias encoded in data

Background - *Defining Ethical Predictive Modeling*

Ethics at ACS

PAAC (Predictive Analytics Advisory Committee) was established as a **representative** body that focuses on **ethics** of its **predictive models** and their **applications**, and helps:

- develop guidelines around practice
- provide oversight (accountability, training, etc) in implementation
- ensure core principles are met

Background - *Defining Ethical Predictive Modeling*

PAAC Core Values

Validity

Transparency

Equity

Relevance

Application

Background - *Defining Ethical Predictive Modeling*

PAAC Core Values

- **Validity** - Good predictive power; appropriate and sufficient data; sound technical analyses.
- **Transparency** - Technical documents that are accessible to internal staff and external stakeholders, and access to the following information: model predictors and outcomes; model performance; origin of ideas; data collection mechanisms; intended applications.
- **Equity** - Provision of resources consistent with individual levels of need, regardless of group membership (**we refer to the lack of this as a type of bias**). Impact analyses to ensure that new practices counter implicit biases and mitigate disproportionality.

Background - *Defining Ethical Predictive Modeling*

PAAC Core Values

- **Relevance** - Models should predict outcomes we are trying to have a positive impact on (or proxies for them), and that we know how to to counter with the use of specific interventions.

- **Application -** Interventions should be: effective and achievable in the context of current practice; in line with ACS priorities; effective; associated with costs (e.g., risk to family) or in limited supply.

Outline

- **Background -** Defining Ethical PRM
- **Conditions for Equity -** Model Predictions
- **Conditions for Equity -** Beyond Model Predictions

Conditions for Equity - *Model Predictions*

Several proposed definitions of fairness/equity

Parity	• Impact is the same on all race groups. For example, if more blacks than whites are rated high-risk by the algorithm, **people in this camp would call that unfair**.
Same Error Rates	• Related to the idea of impact. **Several often mutually exclusive versions exist**: • similar AUC across groups • equally risky cases treated the same (equal precision, e.g., COMPAS bail algorithm) • same error rates between groups (TPR and FPR)
No sensitive attributes, like race or gender are included	• But even if you don't explicitly consider such attributes, that information is usually baked into other factors, like place of residence or income. Some people argue that it can be unfair **not to** consider this information when making decisions (e.g., gaps in resume for men vs women).

Fairness often comes at the cost of overall performance, and even then, many of these definitions of fairness are mutually exclusive.

Conditions for Equity - *Model Predictions*

In practice, some suggestions to consider (with a grain of salt):

Increasing parity

Preferring similar prediction quality

Noting effects of sensitive predictors

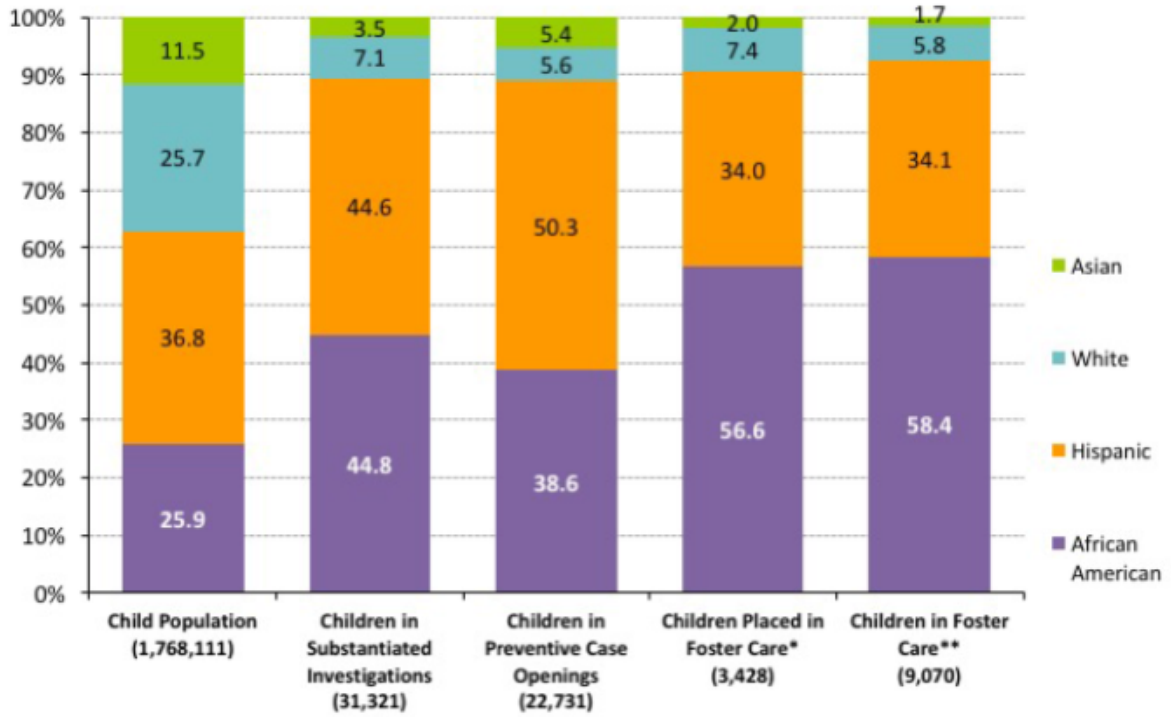
Conditions for Equity - *Model Predictions*

In practice, some suggestions to consider (with a grain of salt):

Increasing parity - Whereas absolute parity would not be a good condition, if we believe that SOME disparity is due to discrimination or implicit biases, we might feel safe to aim for more parity, compared with other available alternatives (human decision making, other models, structured decision making tools, etc.).

Conditions for Equity - *Model Predictions*

Race/Ethnicity and Path through the Child Welfare System, 2016



Note: Missing values and other race are excluded from percent calculations.
*Excludes youth placed in Close to Home.
**Excludes youth in Close to Home placements.

Conditions for Equity - *Model Predictions*

In practice, some suggestions to consider (with a grain of salt):

Preferring similar prediction quality - Look at performance measures within groups (AUC, TPR, FPR). No strong concensus exists as to which error rates should be made equal. Overall poorer performance in one group can be due to variables that are more noisy or otherwise less predictive; potential solution is to rebuild model with better data, when possible.

Conditions for Equity - *Model Predictions*

In practice, some suggestions to consider (with a grain of salt):

Noting effects of sensitive predictors - If overall performance does not suffer, we may decide to include/exclude race and other variables based on which increases parity and similarity in prediction quality (e.g. Allegheny). However, including these might trigger “strict scrutiny”: https://en.wikipedia.org/wiki/Strict_scrutiny.

Conditions for Equity - *Model Predictions*

Question: What do these have in common : “Increasing parity”, “Preferring similar prediction quality”, “Noting effects of sensitive predictors”?

Proposed answer: These are all relativist strategies. The best we can do is to treat equity as relative and to compare it with alternative decision making tools.

“Corbett-Davies’s conclusion: the only way to achieve a fair algorithm is to try to avoid obvious errors that scientists call miscalibration, redlining, sample and label bias. Once those sources of bias are excluded, apply a common threshold to everybody – and live with the numerical oddities that might arise.”
https://simons.berkeley.edu/news/algorithms-discrimination **Do you agree?**

Outline

- **Background -** Defining Ethical PRM
- **Conditions for Equity** - Model Predictions
- **Conditions for Equity -** Beyond Model Predictions

Conditions for Equity - *Beyond Model Predictions*

Select unbiased outcomes to predict (model can only be as fair as outcome)

for which there are good applications (otherwise, what's the point)

Create good practice around implementation (training etc.)

Evaluate evaluate evaluate (does model work and does impact match intent?)

and keep invested parties in the loop (e.g., stakeholders, community)

Let's examine these in more detail as they pertain to equity

Conditions for Equity - *Beyond Model Predictions*

"The fairest of them all" - **Select unbiased outcomes to predict** in order to mitigate perpetuating unfairness (hiring decision example). Outcomes should be:

- Easy to measure
- Least likely to be biased
- Directly related to actionable goals

Conditions for Equity - *Beyond Model Predictions*

"The fairest of them all" - **Select unbiased outcomes to predict** in order to mitigate perpetuating unfairness (hiring decision example). Outcomes should be:

- Easy to measure - less noise and potentially more likely to be objective
- Least likely to be biased (e.g., fatality vs indicated investigation vs any investigation)
- Directly related to actionable goals - arguably unethical to predict outcomes we don't have ethical interventions for (e.g., World Health Organization screening guidelines). Outcomes should be related to the risk the intervention is meant to mitigate. Idea: how about predicting potential gain from intervention rather than risk?

Ethics and Equity - *Exercise 1*

Using the criteria above as well, brainstorm and discuss what might be some examples of appropriate and inappropriate outcomes to predict in your own fields

Conditions for Equity - *Beyond Model Predictions*

Predict outcomes **for which there are good applications**

When possible, interventions should not be punitive:

- Except, perhaps, when they also offer the most benefit?
- Potentially punitive interventions can require higher standards, such as additional human review

Comparing costs and benefits only makes sense in the context of applications (free money example). Even if they are unquantifiable, we should still ask:

- If we get it wrong (false positive) more often for one group than another, do we: Simply waste taxpayer dollars? Overmonitor those families? Risk harming children and communities?
- Is it worth the potential benefits of: Reducing (rather than eliminating) danger to child?

Interventions, and not simply algorithms, should be rigorously tested. Let's suppose we have an awesome model, highly predictive and equitable, and our outcome isn't biased. Then what? How often do we know the effectiveness and costs of an intervention:

- On average
- Within particular groups

Conditions for Equity - *Beyond Model Predictions*

Create good practice around implementation

Workers should be trained:

- Confidence in model as well as confidence in over-riding (and documenting rationale) are both important

Business process should be designed to:

- Take into consideration when, with whom, and to what extent model predictions are shared (example of resources offered to high risk families after rather than before reunification decision has been made)

Such practices can reduce the effect of implicit biases

Conditions for Equity - *Beyond Model Predictions*

Evaluate, evaluate, evaluate!

Validate model and validate alternative decision-making tools (human, SDM, other models, etc.).

- Predictive models make it easier to track things, including bias. Human decisions are a black box!
- Question shouldn't simply be "are these error rates acceptable?" Rather, we should ask, "which among various decision making tools is the most effective across the board, and which is the most equitable?"

Validate through time, even on data for which there aren't outcomes yet.

- Remember that patterns in training set may be outdated (e.g., change in business process or variable definitions). Check for validity as new outcomes are observed and update models as necessary.

Impact evaluations should be done in addition to model validations.

- Are we able to mitigate outcomes? Often difficult to measure, given confounding between level of risk and intervention.

Conditions for Equity - *Beyond Model Predictions*

Keep ‘em in the loop

Stakeholder engagement, active discussion, system oversight, and transparency help us to be responsive and accountable.

This is a multidisciplinary undertaking so we need many voices, in order to understand, for example:

- How practice translates into predictors going into model
- How integration of predictive analytics is felt throughout agency and communities

In addition, we need to find better ways to detect bias and distinguish it from true differences in our outcomes of interest (which needless to say, are themselves driven by societal differences in access, opportunity, risk, etc.)

Conditions for Equity - *Beyond Model Predictions*

Select unbiased outcomes to predict (model can only be as fair as outcome) EQUITY

for which there are good applications (otherwise, what's the point) RELEVANCE

Create good practice around implementation (training etc.) APPLICATION

Evaluate evaluate evaluate (does model work and does impact match intent?) VALIDITY

and keep invested parties in the loop (e.g., stakeholders, community) TRANSPARENCY

Conclusion: Each core value needed to ensure maximum equity

Ethics and Equity - *Exercise 2*

What are some of the institutional biases in your own fields and what might be some of the strategies and pitfalls in overcoming them?

Looking for opportunities to use data to support policies, programs, practices

APPLICATIONS OF RESULTS

NYC Application Area Examples

Repeat Reports	• Clinical Consultation program development • DCP Quality Assurance • Safe Measures Dashboard
Children at a high likelihood of experiencing elevated risk	• Court Ordered Supervision Program Development • Preventive Services program development • Scorecard Risk Adjustment
Re-entry	• Trial Discharge program development • Aftercare program development • Linkage to service referral
Intergenerational Involvement	• Service Development • Linkage to service referral

NYC Application Area Examples

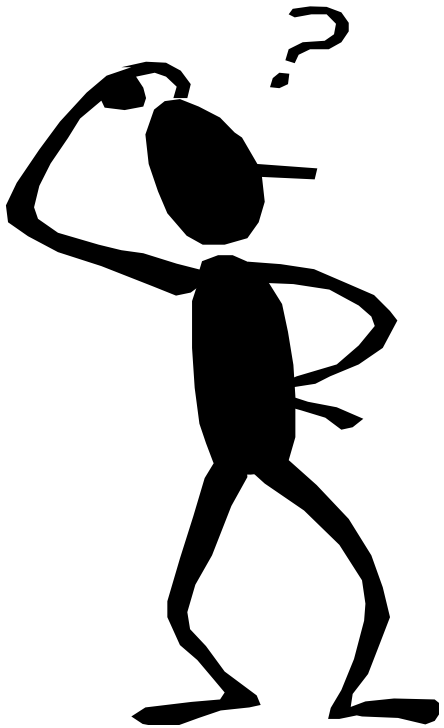
Repeat Reports	• Clinical Consultation program development • DCP Quality Assurance • Safe Measures Dashboard
Children at a high likelihood of experiencing elevated risk	• Court Ordered Supervision Program Development • Preventive Services program development • Scorecard Risk Cohorts
Re-entry	• Trial Discharge program development • Aftercare program development • Linkage to service referral
Intergenerational Involvement	• Service Development • Linkage to service referral

Group Work

- Matching potential applications to your use of PRM.
  - What would you do, if you knew (fill in the blank)
    - What would you do to prevent this adverse event?

Reflections:

- How well did your team accomplish your outcomes?
- What other strategies will you employ to further your capacity in this area?



Questions/ Reflections?