Machine Learning for Healthcare

Dataset Shift

David Sontag

CSAIL

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

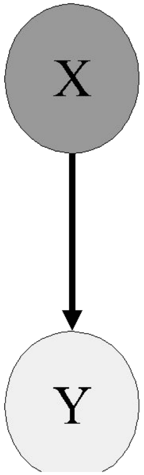
HEALTH SCIENCES & TECHNOLOGY

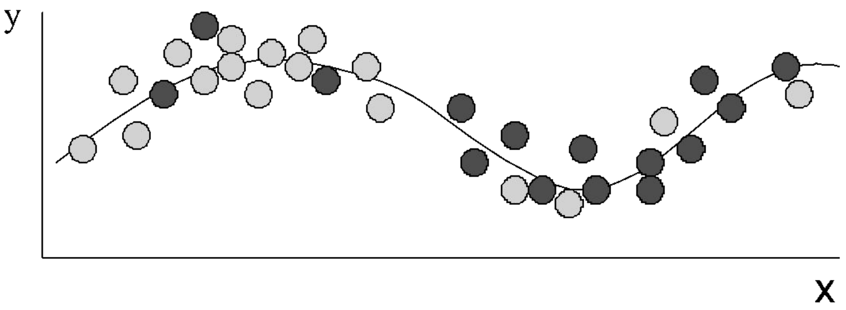
Outline for today's class

- **Examples & formalization of dataset shift**
- Testing for dataset shift
- Mitigating dataset shift
- Case studies

Types of dataset shift

- $Pr_{old}(x,y)$ versus $Pr_{new}(x,y)$, where X are the features / covariates and Y is the label / outcome
- (Simple) covariate shift: $Pr_{old}(x) \neq Pr_{new}(x)$ but $Pr_{old}(y|x) = Pr_{new}(y|x)$

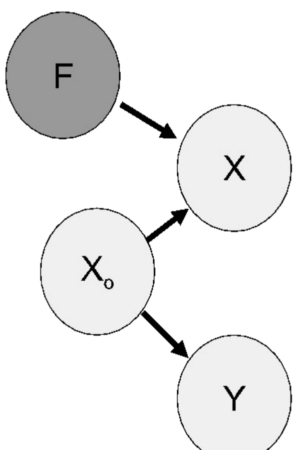


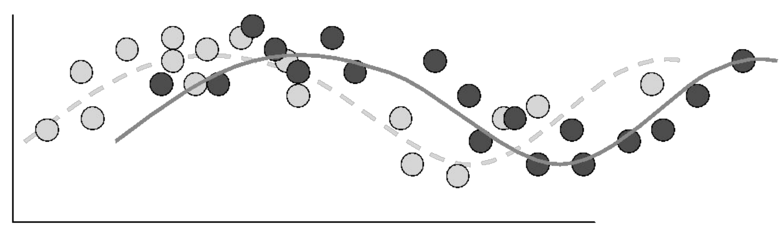


(Quiñonero-Candela et al., Dataset Shift in Machine Learning, MIT Press 2008)

Types of dataset shift

- $Pr_{old}(x,y)$ versus $Pr_{new}(x,y)$, where X are the features / covariates and Y is the label / outcome
- (Simple) covariate shift: $Pr_{old}(x) \neq Pr_{new}(x)$ but $Pr_{old}(y|x) = Pr_{new}(y|x)$
- Domain shift: $Pr_{old}(y|x) \neq Pr_{new}(y|x)$ due to data transformation





(Quiñonero-Candela et al., Dataset Shift in Machine Learning, MIT Press 2008)

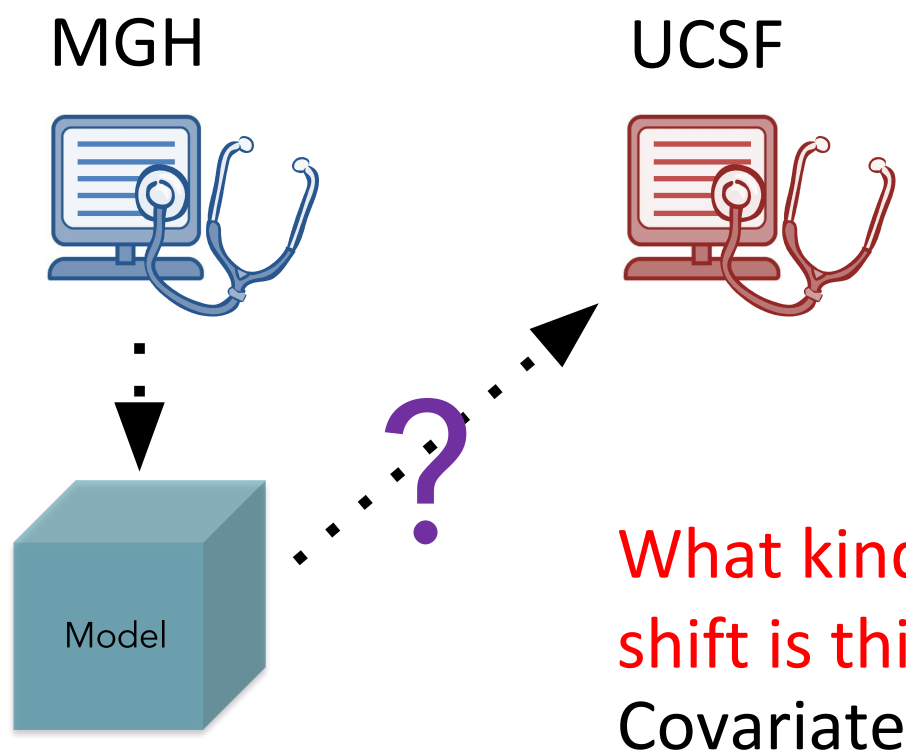
Types of dataset shift

- $Pr_{old}(x,y)$ versus $Pr_{new}(x,y)$, where X are the features / covariates and Y is the label / outcome
- (Simple) covariate shift: $Pr_{old}(x) \neq Pr_{new}(x)$ but $Pr_{old}(y|x) = Pr_{new}(y|x)$
- Domain shift: $Pr_{old}(y|x) \neq Pr_{new}(y|x)$ due to feature transformation
- Label shift: $Pr_{old}(y|x) \neq Pr_{new}(y|x)$ due to labels taking on a new meaning

(Quiñonero-Candela et al., Dataset Shift in Machine Learning, MIT Press 2008)

Dataset shift / non-stationarity:

Models often do not generalize

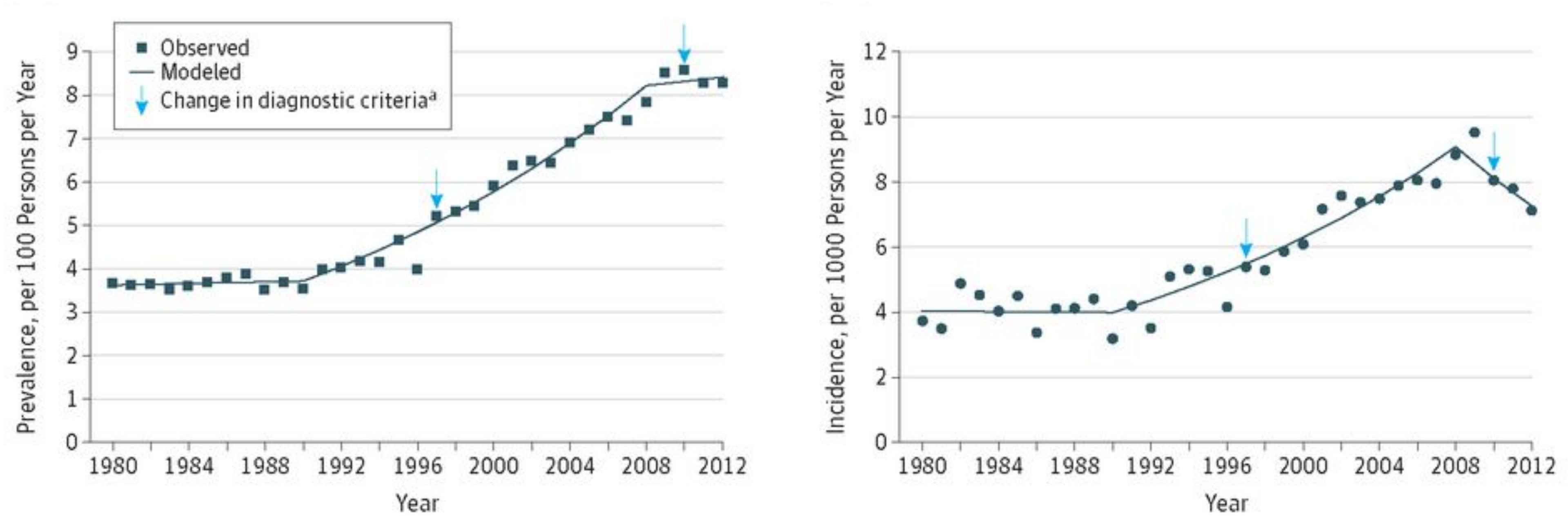


What kind of dataset shift is this?

Covariate and domain shift

[Figure adopted from Jen Gong and Tristan Naumann]

Dataset shift / non-stationarity:
Diabetes Onset After 2009



→ Automatically derived labels may change meaning

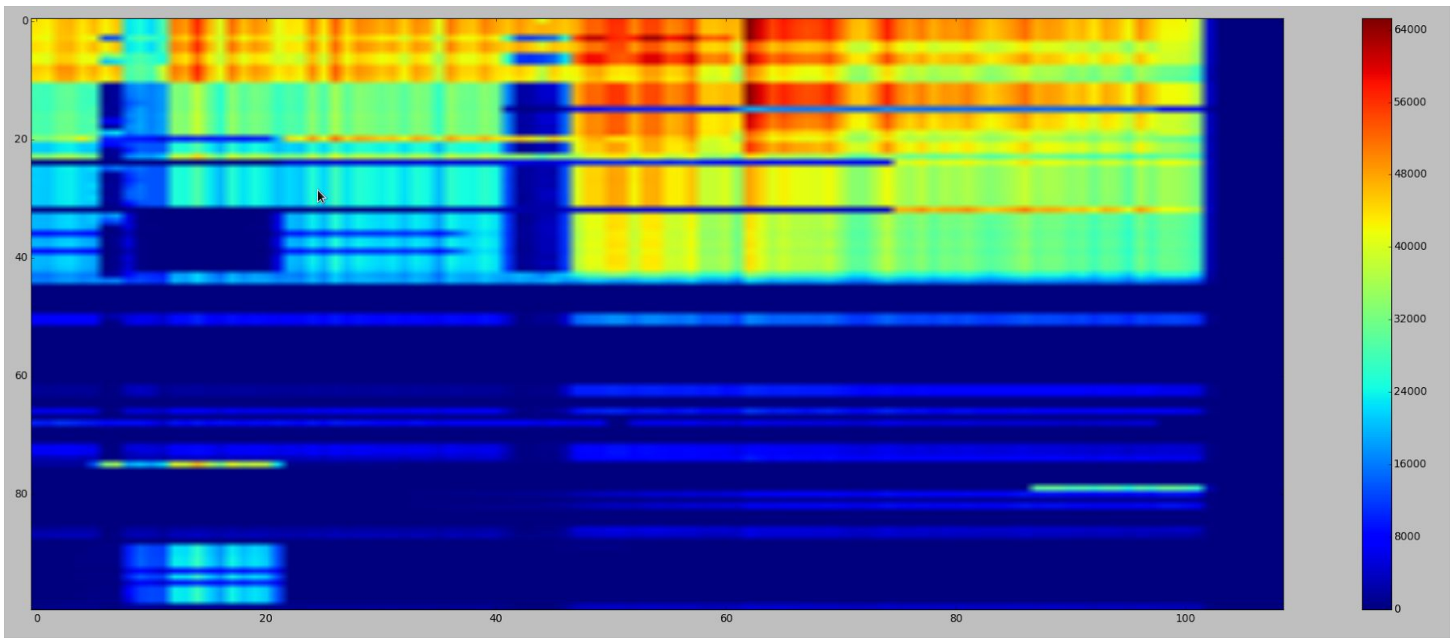
Label shift

[Geiss LS, Wang J, Cheng YJ, et al. Prevalence and Incidence Trends for Diagnosed Diabetes Among Adults Aged 20 to 79 Years, United States, 1980-2012. JAMA, 2014.]

Dataset shift / non-stationarity:

Top 100 lab measurements over time

Labs



Time (in months, from 1/2005 up to 1/2014)

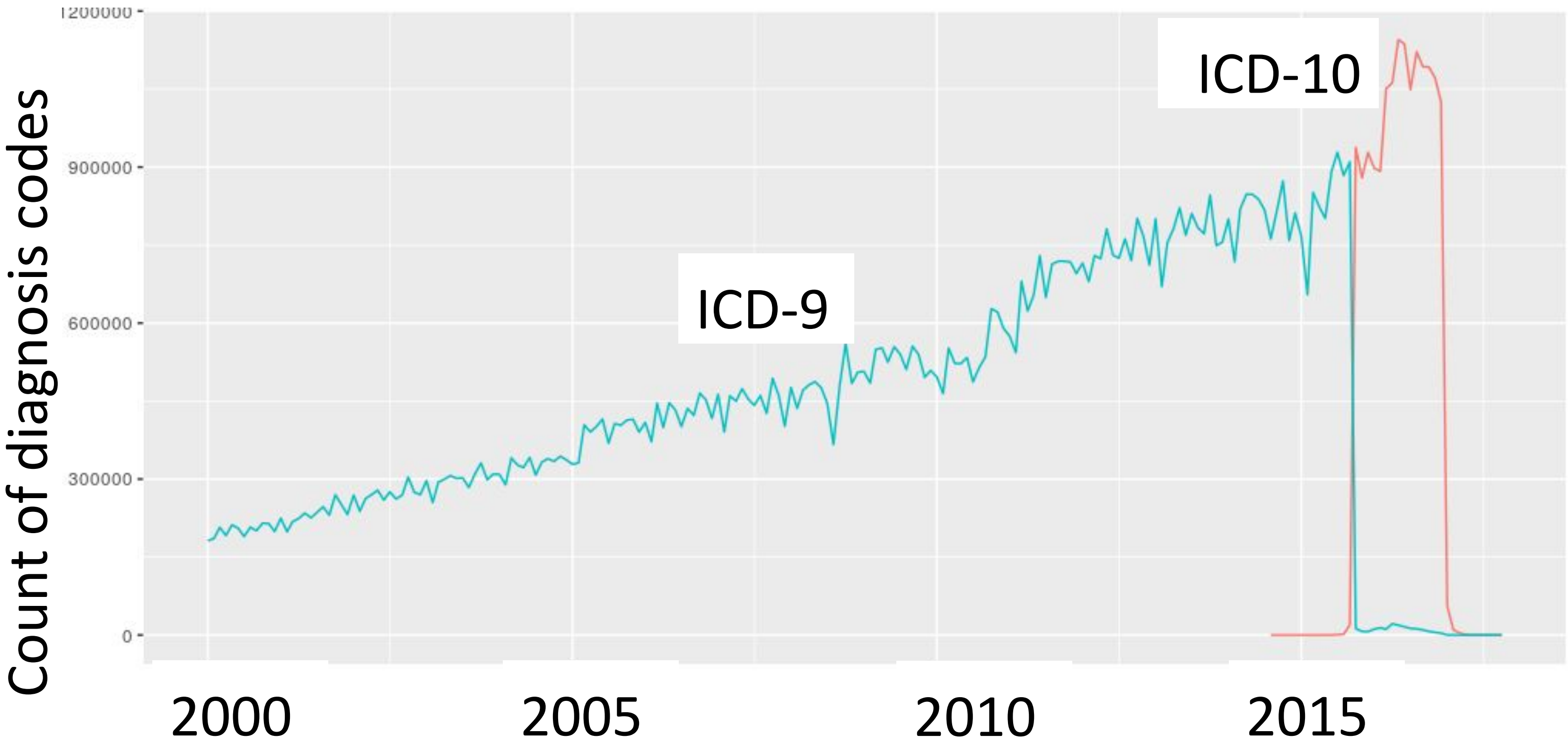
→ Significance of features may change over time

Covariate shift

[Figure credit: Narges Razavian]

Dataset shift / non-stationarity:

ICD-9 to ICD-10 shift



→ Significance of features may change over time

Covariate shift (domain shift if mapping ICD10 to ICD9)

[Figure credit: Mike Oberst]

Outline for today's class

- Examples & formalization of dataset shift
- **Testing for dataset shift**
- Mitigating dataset shift
- Case studies

Testing for dataset shift

- Shift in p(y):
  - Plot distributions
- Shift in p(x) or p(x|y):
  - Compare feature means
  - Use kernel two-sample test (Gretton et al., JMLR '12)

Integral probability metric:
(Muller, 1997)

$$\mathrm{IPM}_{\mathcal{L}}(p,q) := \sup_{\ell \in \mathcal{L}} \left|\mathbb{E}_p[\ell(x)] - \mathbb{E}_q[\ell(x)]\right|$$

Maximum mean discrepancy (MMD): L are functions with norm 1 in a RKHS:
(Gretton et al., 2012)

samples $x_1, \ldots, x_m \sim p,\ x'_1, \ldots, x'_n \sim q$

$$\hat{\mathrm{MMD}}_k^2(p,q) := \frac{1}{m-1}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, x'_j) + \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x'_i, x'_j)$$

Testing for dataset shift

- Shift in p(y):
  - Plot distributions
- Shift in p(x) or p(x|y):
  - Compare feature means
  - Use kernel two-sample test such as maximum mean discrepancy/MMD (Gretton et al., JMLR ‘12)
  - (Attempt to) learn a classifier to distinguish one dataset from the other

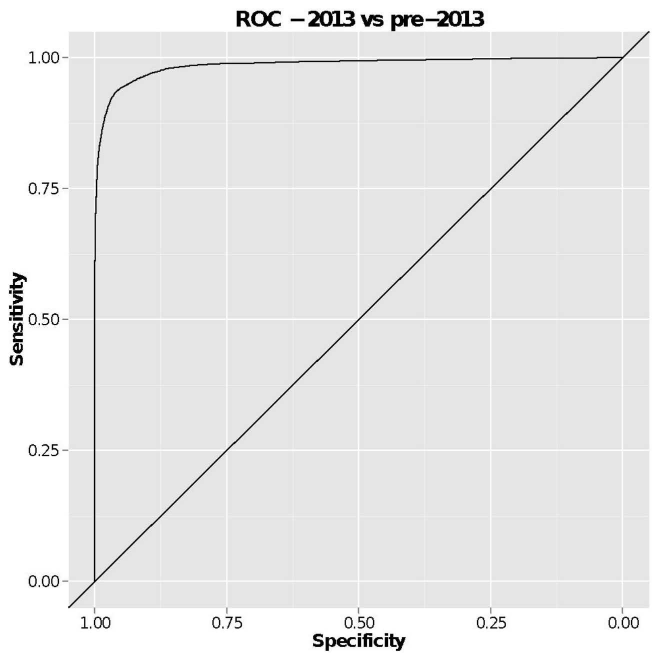
$$\text{samples } x_1, ..., x_m \sim p,\ x'_1, ..., x'_n \sim q$$

Binary classification (0 vs. 1)

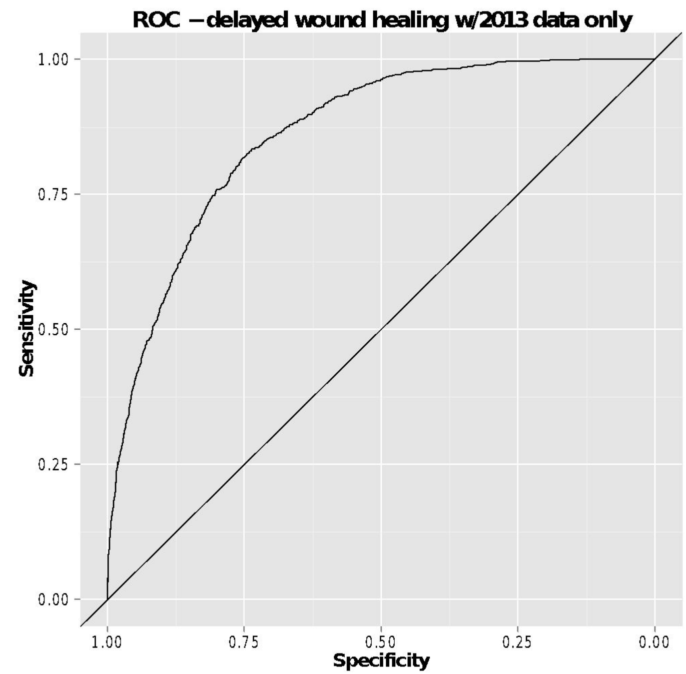
$$\mathcal{D} = \{(x_1, 1), \ldots, (x_m, 1), (x'_1, 0), \ldots, (x'_n, 0)\}$$

Testing for dataset shift

- Testing for covariate shift (wound healing):



Distinguish 2013 from pre-2013



Distinguish first 2/3 of 2013 from last 1/3 of 2013

(Slide credit: Ken Jung)

Outline for today's class

- Examples & formalization of dataset shift
- Testing for dataset shift
- **Mitigating dataset shift**
  - *Covariate shift* Do nothing. Regression just "works"
  - *Covariate shift* Importance sampling
  - *Domain shift* Causal invariances
- Case studies

Covariate shift: nonparametric regression just "works"

-

When can we expect training on p(x,y) and testing on q(x,y) to give good results, for $\mathrm{p} \neq q$?

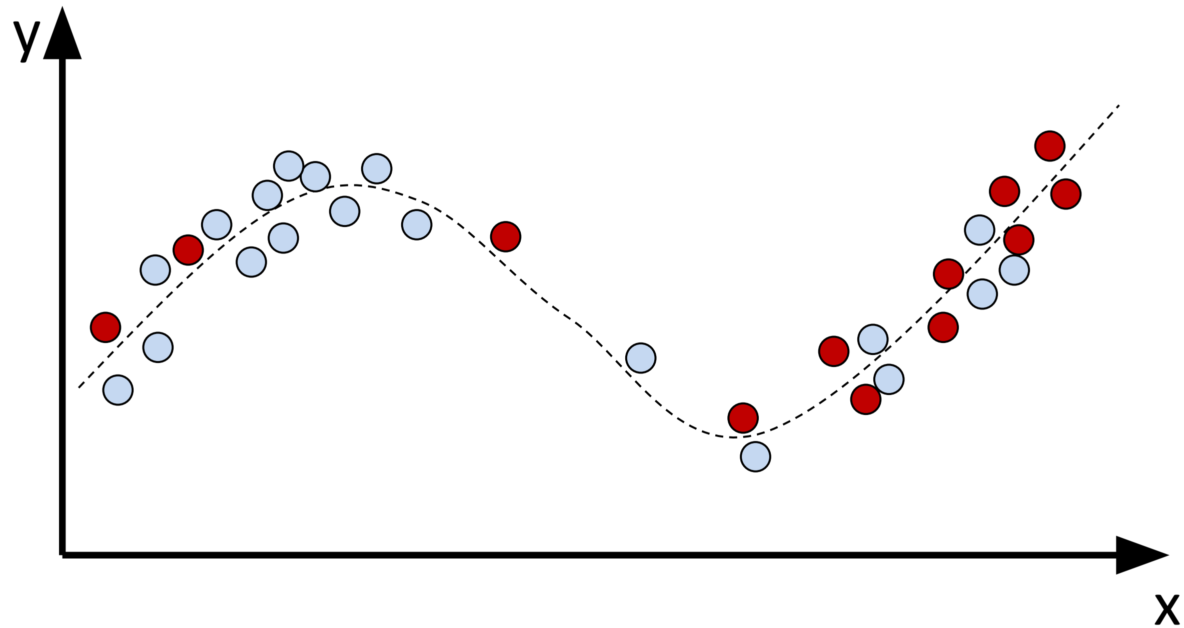
<u>Theorem:</u> If $\mathrm{p(x)} > 0$ whenever $\mathrm{q(x)} > 0$ and $p(y \mid x) = q(y \mid x)$, then in the limit of infinite data from *p*, can achieve Bayes' error on *q*

But we might not have infinite data!

We may have to use a more restricted model (e.g. a linear model despite true one being non-linear)

Effect of covariate shift when (naively) learning with misspecified models

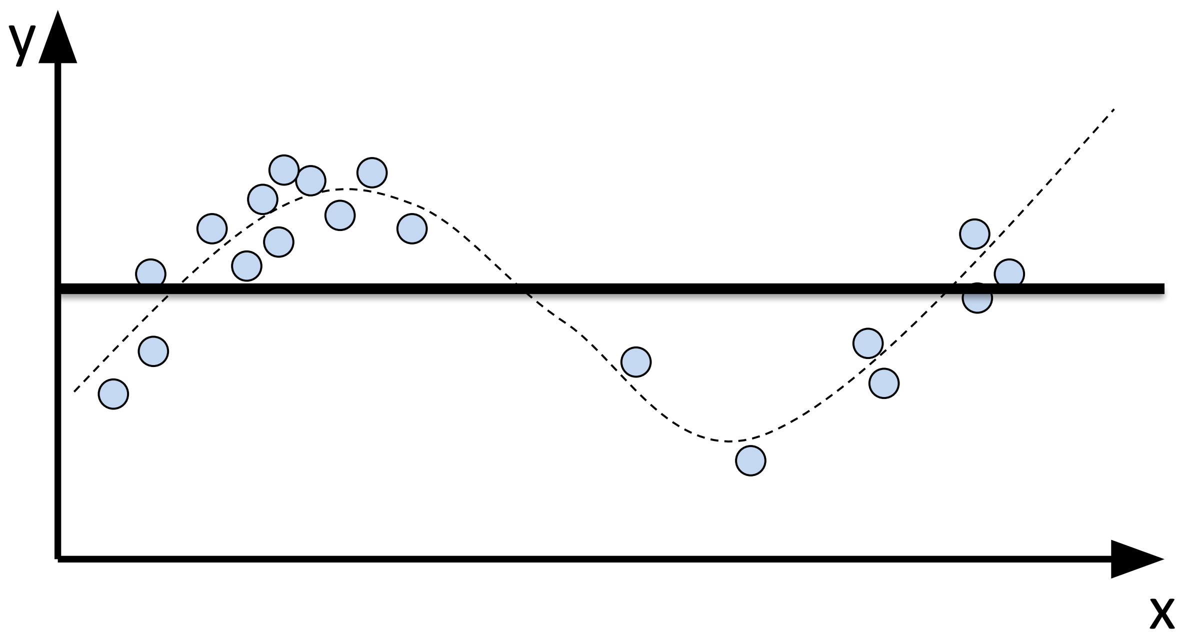
- Training data p(x,y)=● and test data q(x,y)=○



[Storkey, "When Training and Test Sets are Different", Dataset in Machine Learning, MIT Press 2009]

Effect of covariate shift when (naively) learning with misspecified models

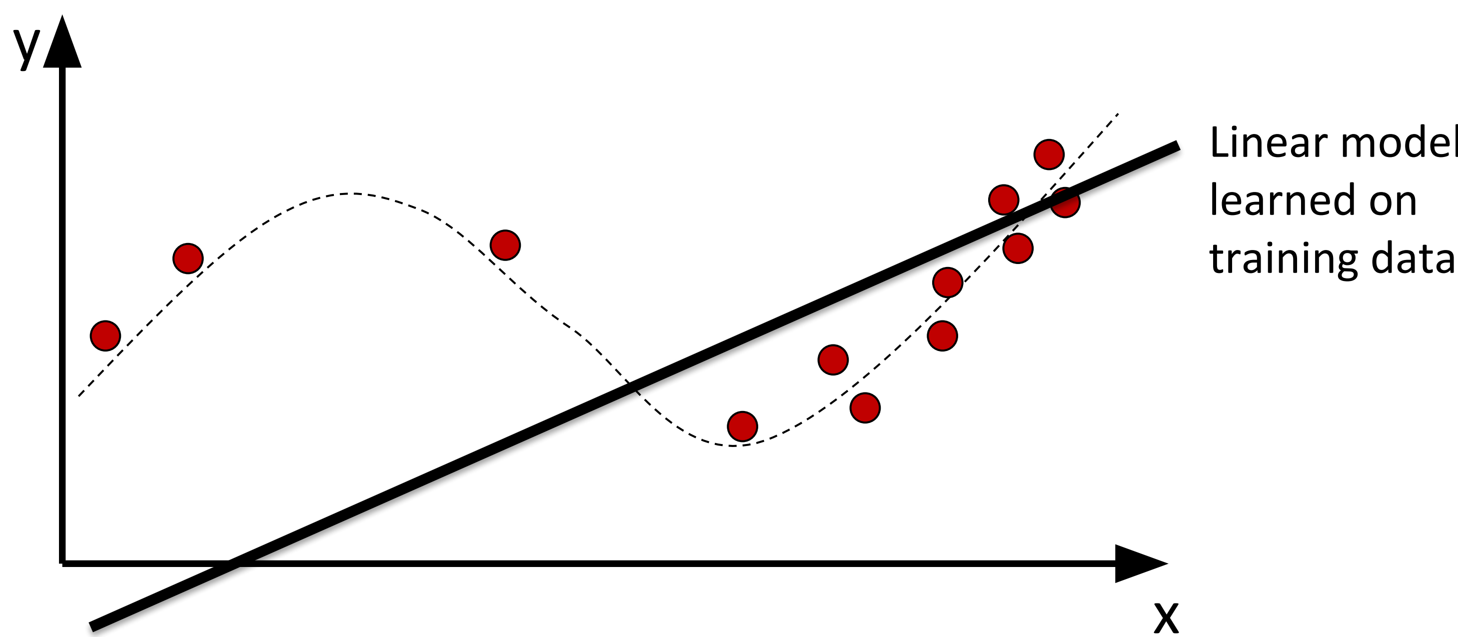
- Training data p(x,y)= and test data q(x,y)=



[Storkey, "When Training and Test Sets are Different", Dataset in Machine Learning, MIT Press 2009]

Effect of covariate shift when (naively) learning with misspecified models

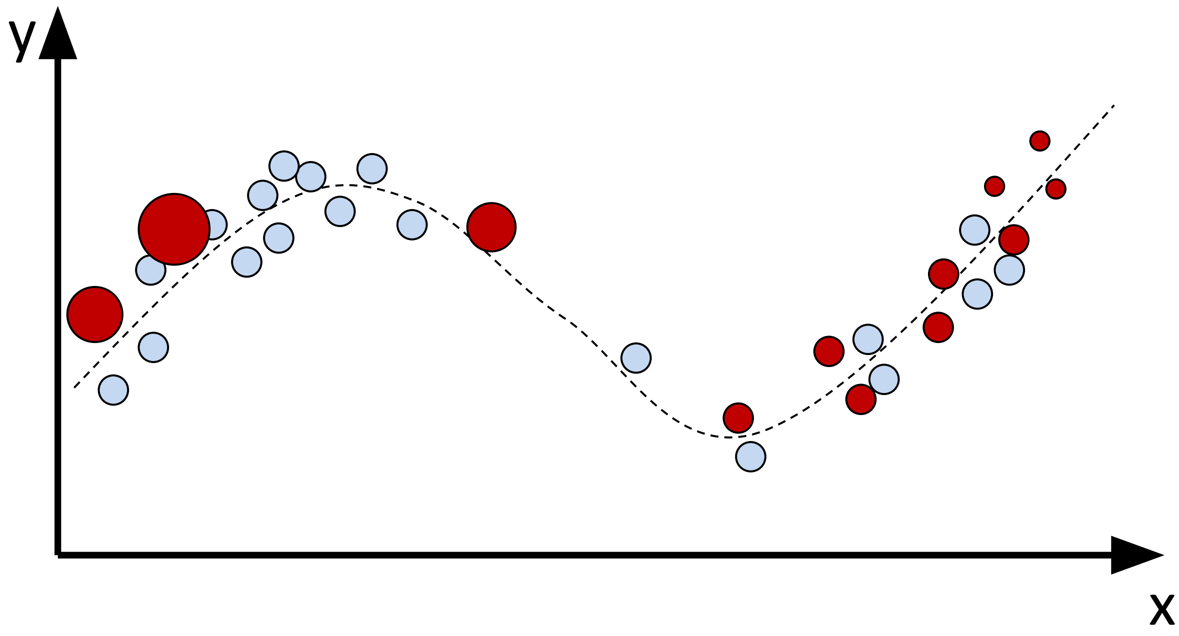
- Training data p(x,y)= and test data q(x,y)=



[Storkey, "When Training and Test Sets are Different", Dataset in Machine Learning, MIT Press 2009]

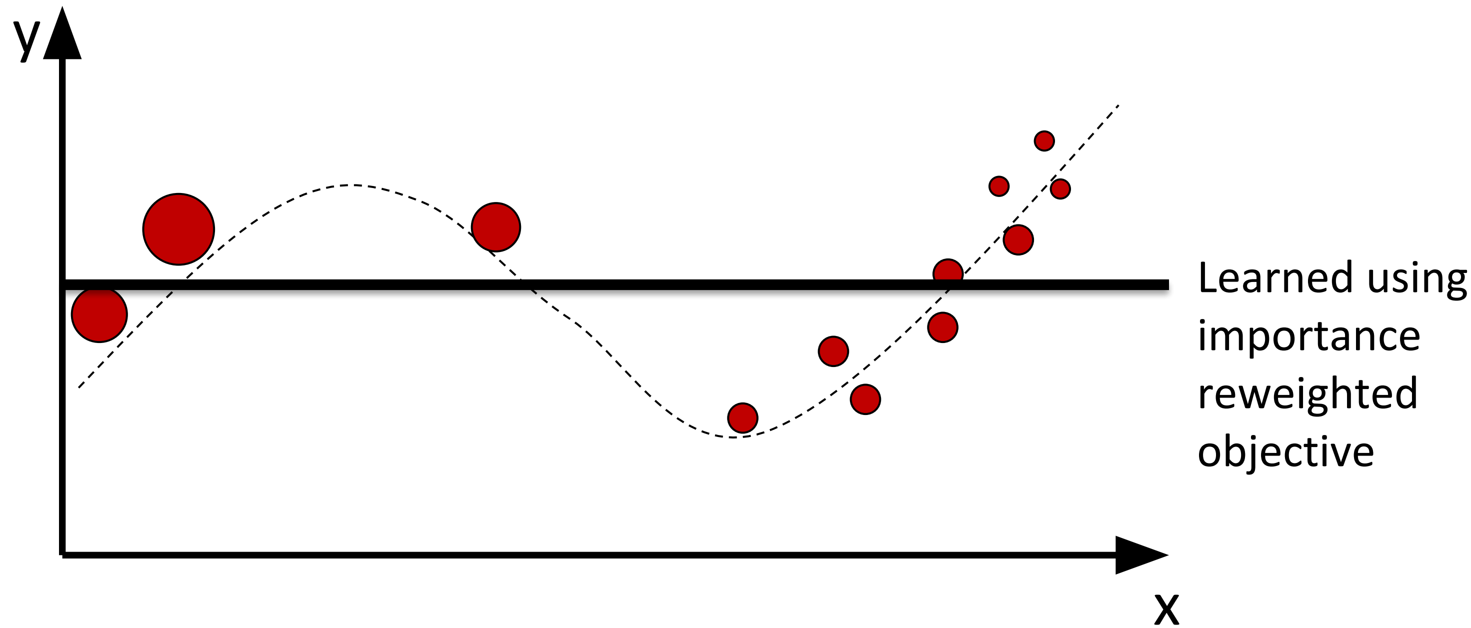
Learning using importance reweighting

- Training data p(x,y)=● and test data q(x,y)=○



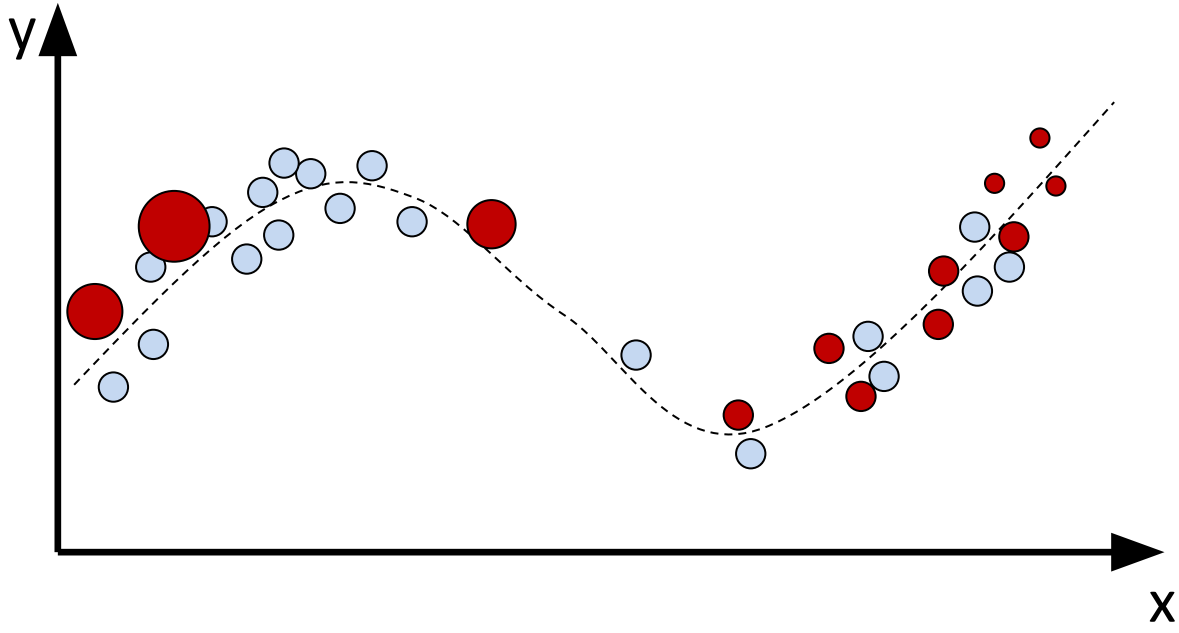
Learning using importance reweighting

- Training data p(x,y)= and test data q(x,y)=



Learning using importance reweighting

- Training data p(x,y)= and test data q(x,y)=



We only needed to know q(x) to figure out how to reweight the training data! Example of *unsupervised* domain adaptation

When importance reweighting is not enough

- Importance reweighted estimator can be high variance
- If there is no *overlap*, then unsupervised domain adaptation is in general impossible – even with infinite data
  - E.g., ICD9 to ICD10

Learning under domain shift

- Must make additional assumptions, e.g.
  - Covariate shift assumption holds for a *subset* of features (Rojas-Carulla '18)
  - Can disentangle factors of variation so as to learn models robust to them (Heinze-Deml & Meinshausen '19):

(a) Example 3, training set. (b) Example 3, test set.

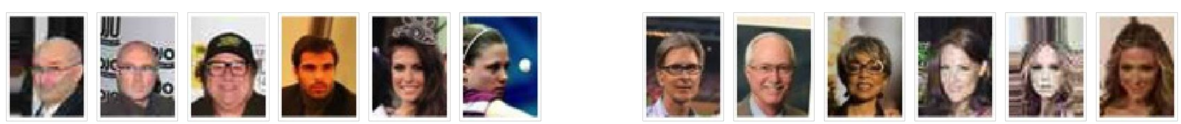
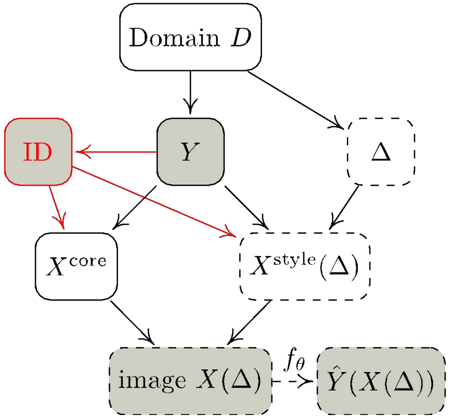


Figure 2: Motivating example 3: The goal is to predict whether a person is wearing glasses. The distributions are shifted in test data by style interventions where style is the image quality. A 5-layer CNN achieves 0% training error and 2% test error for images that are sampled from the same distribution as the training images (a), but a 65% error rate on images where the confounding between image quality and glasses is changed (b). See §5.3 for more details.

[Rojas-Carulla, Schölkopf, Turner, Peters. Invariant Models for Causal Transfer Learning, JMLR '18]
[Heinze-Deml, Meinshausen. Conditional Variance Penalties and Domain Shift Robustness, '19]

Learning under domain shift

- Must make additional assumptions, e.g.
  - Covariate shift assumption holds for a *subset* of features (Rojas-Carulla ‘18)
  - Can disentangle factors of variation so as to learn models robust to them (Heinze-Deml & Meinshausen ‘19):



Learning algorithm assumes we have (some) training data with *grouped* observations (e.g. pictures of the same person with different image quality)

[Rojas-Carulla, Schölkopf, Turner, Peters. Invariant Models for Causal Transfer Learning, JMLR ‘18]
[Heinze-Deml, Meinshausen. Conditional Variance Penalties and Domain Shift Robustness, ‘19]

Outline for today's class

- Examples & formalization of dataset shift
- Testing for dataset shift
- Mitigating dataset shift
- **Case studies**
  - Framingham risk score
  - Antibiotic resistance

Case study: Framingham risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - Model based on 6 major risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

[Wilson et al., Circulation, 1998]

CHD score sheet for men using TC or LDL-C categories.

Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-1	[-1]
35-39	0	[0]
40-44	1	[1]
45-49	2	[2]
50-54	3	[3]
55-59	4	[4]
60-64	5	[5]
65-69	6	[6]
70-74	7	[7]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-3
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	1
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-3]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[2]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	2	[2]
35-44	0.91-1.16	1	[1]
45-49	1.17-1.29	0	[0]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-1	[-2]

Step 4

Blood Pressure					
Systolic	Diastolic (mm Hg)				
(mm Hg)	<80	80-84	85-89	90-99	≥100
<120	0 [0] pts				
120-129		0 [0] pts			
130-139			1 [1] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 7 (sum from steps 1-6)

Adding up the points	
Age	______
LDL-C or Chol	______
HDL - C	
Blood Pressure	______
Diabetes	______
Smoker	______
Point total	______

Step 8 (determine CHD risk from point total)

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
<-3	1%		
-2	2%		
-1	2%	[<-1]	[2%]
0	3%	[0]	[3%]
1	4%	[1]	[3%]
2	4%	[2]	[4%]
3	6%	[3]	[5%]
4	7%	[4]	[7%]
5	9%	[5]	[8%]
6	11%	[6]	[10%]
7	14%	[7]	[13%]
8	18%	[8]	[16%]
9	22%	[9]	[20%]
10	27%	[10]	[25%]
11	33%	[11]	[31%]
12	40%	[12]	[37%]
13	47%	[13]	[45%]
≥14	≥56%	[≥14]	[≥53%]

Step 9 (compare to average person your age)

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	3%	1%	2%
35-39	5%	4%	3%
40-44	7%	4%	4%
45-49	11%	8%	4%
50-54	14%	10%	6%
55-59	16%	13%	7%
60-64	21%	20%	9%
65-69	25%	22%	11%
70-74	30%	25%	14%

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Peter W. F. Wilson et al. Circulation. 1998;97:1837-1847

American Heart Association.

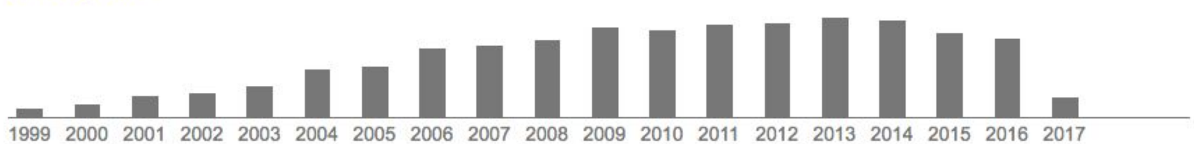
Case study: Framingham risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score

Prediction of coronary heart disease using risk factor categories

[HTML] from ahajournals.org
Full text - MIT Libraries

Authors	Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, William B Kannel
Publication date	1998/5/1
Journal	Circulation
Volume	97
Issue	18
Pages	1837-1847
Publisher	Lippincott Williams & Wilkins
Description	Background—The objective of this study was to examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk, to incorporate them into coronary prediction algorithms, and to compare the discrimination properties of this approach with other noncategorical prediction functions. Methods and Results—This work was designed as a prospective, single-center study in the setting of a community-based ...
Total citations	Cited by 8422



Case study: Framingham risk score

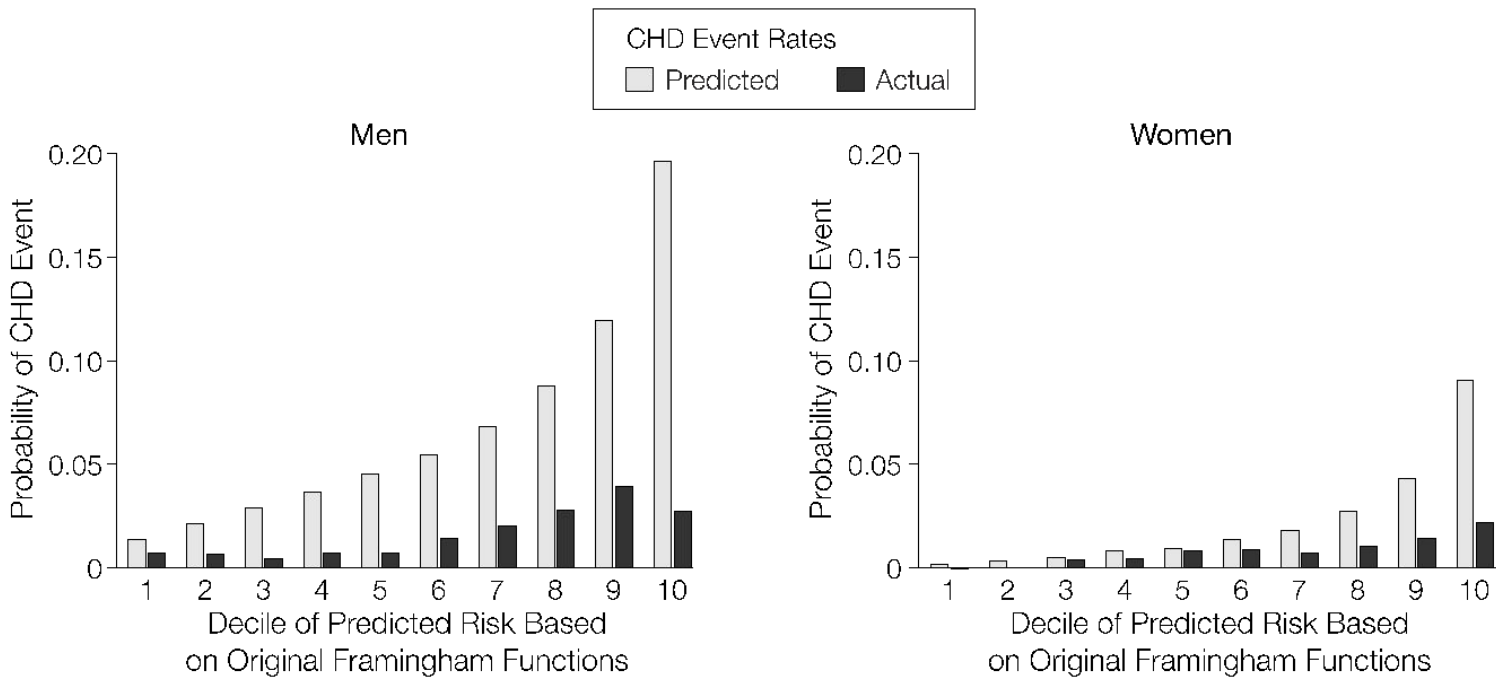
- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?
- C-statistic (=AUC on censored data) on Chinese population is 0.705/0.742 (M/F)
- What else should we look at?

[Liu et al., JAMA '04]

Case study: Framingham risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (directly applied to Chinese population)

Figure 2. Ten-Year Prediction of CHD Events in CMCS Men and Women Using the Original Framingham Functions



[Liu et al., JAMA '04]

Case study: Framingham risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?

- C-statistic (=AUC on censored data) 0.705/0.742 (M/F)
- Re-fit using local data only slightly improves C-statistic (=AUC on censored data), to 0.736/0.759 (M/F)

[Liu et al., JAMA ‘04]

Case study: Framingham risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (re-fit to Chinese population)

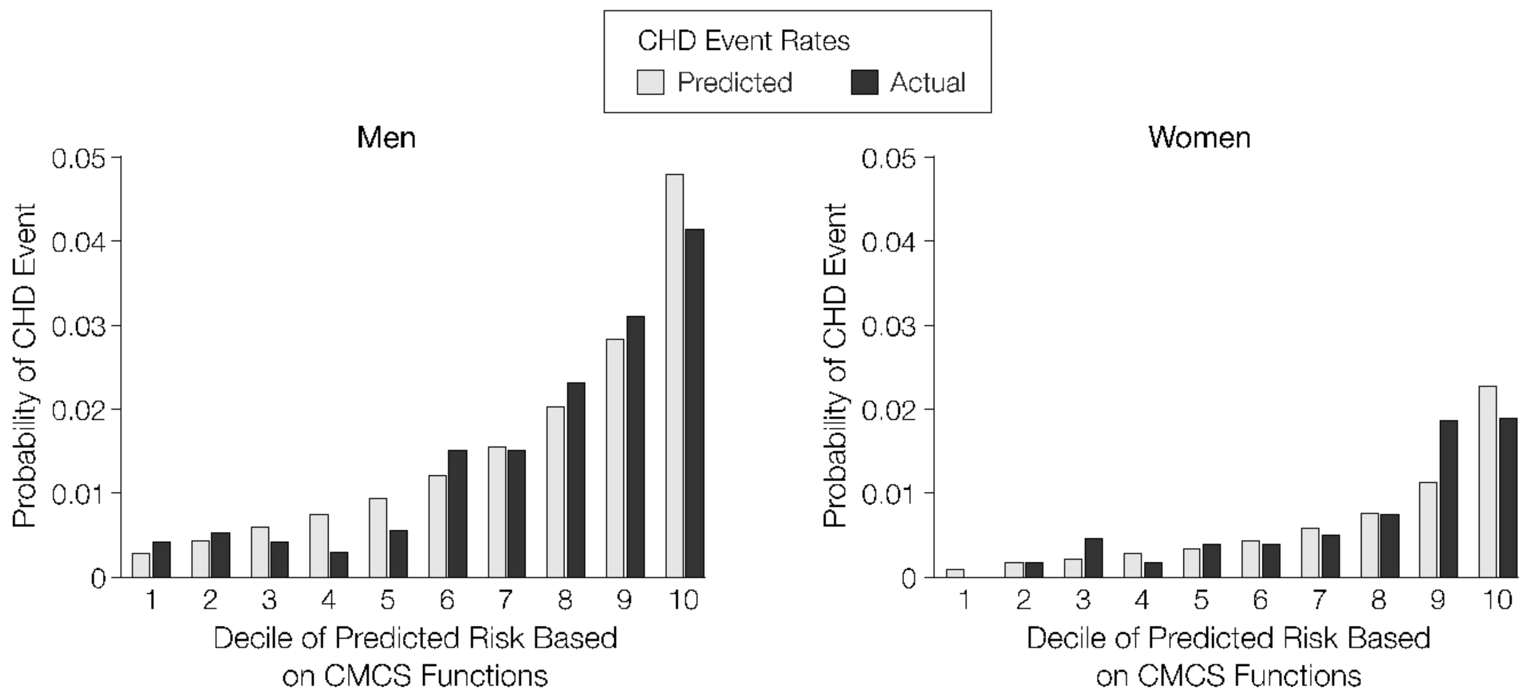
	CMCS	Framingham*
Risk Factors	β	β
Age	0.07	† 0.05
Age squared	NA	NA
Blood pressure		
Optimal	−0.51	0.09
Normal		
High normal	0.21	0.42
Stage 1 hypertension	0.33	0.66
Stage 2-4 hypertension	0.77	0.90
TC, mg/dL		
<160	−0.51	−0.38
160-199		
200-239	0.07	† 0.57
240-279	0.32	0.74
≥280	0.52	0.83
HDL-C, mg/dL		
<35	−0.25	† 0.61
35-44	0.01	0.37
45-49		
50-59	−0.07	0.00
≥60	−0.40	−0.46
Diabetes	0.09	0.53
Smoking	0.62	0.73

[Liu et al., JAMA '04]

Case study: Framingham risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (re-fit to Chinese population)

Figure 1. Ten-Year Prediction of CHD Events in CMCS Men and Women Using the CMCS Functions



[Liu et al., JAMA '04]

Case study: predicting antibiotic resistance

[Oberst, Boominathan, Zhou, Kanjilal, Sontag]

Case study: predicting antibiotic resistance

- Guide choice of antibiotic, even before culture results come back

- Data from MGH & BWH hospitals in Boston
- We show that we can nearly **eliminate** 2nd line antibiotic usage while **decreasing** the rate of inappropriate antibiotics prescribed
- Key tool: *predicting antibiotic resistance*

Case study: predicting antibiotic resistance

- In our early investigations, we included features derived from clinical notes
- We noticed that top predictors were ‘2010’, ‘2009’, ‘2014’, etc.
- We knew there was non-stationarity due to levels of resistance changing, but this was *much* more than we expected

Case study: predicting antibiotic resistance

What happened in 2006?

A new card was introduced to MIC testing with a lower range dilutions (more dynamic range)

As a result, cut points to decide difference between resistant/susceptible were moved down

This resulted in many more “positives” for pre-2006 years, but which were simply because these were the lowest possible values that could be recorded

Label shift detected by model introspection

[Figure from Helen Zhou]

Conclusion

- Dataset shift happens all the time with healthcare data
- It doesn’t always hurt performance
- Interpretability methods can help with detecting and mitigating dataset shift
- Safe deployments should include automated checks for dataset shift
- Active area of research in ML