Risk Stratification for Health-Care Associated *C. diff*

Learning Evolving Patient Risk Processes for *C. diff* Colonization

Jenna Wiens, John V. Guttag and Eric Horvitz*

Massachusetts Institute of Technology

* Microsoft Research

Clostridium difficile (*C. diff*)

- Bacteria takes over the gut when normal flora gets wiped out
- Transmitted through the mouth
- Causes severe diarrhea, intestinal diseases
- Treatment: metronidazole, oral vancomycin
- 20% of cases relapse within 60-days (Pepin J et al., 2005)

Source: http://www.helladelicious.com/wp-content/uploads/2012/05/Gut-Flora-Bacteria.jpg
http://newsimg.bbc.co.uk/media/images/41223000/jpg/_41223637_clostridium203spl.jpg

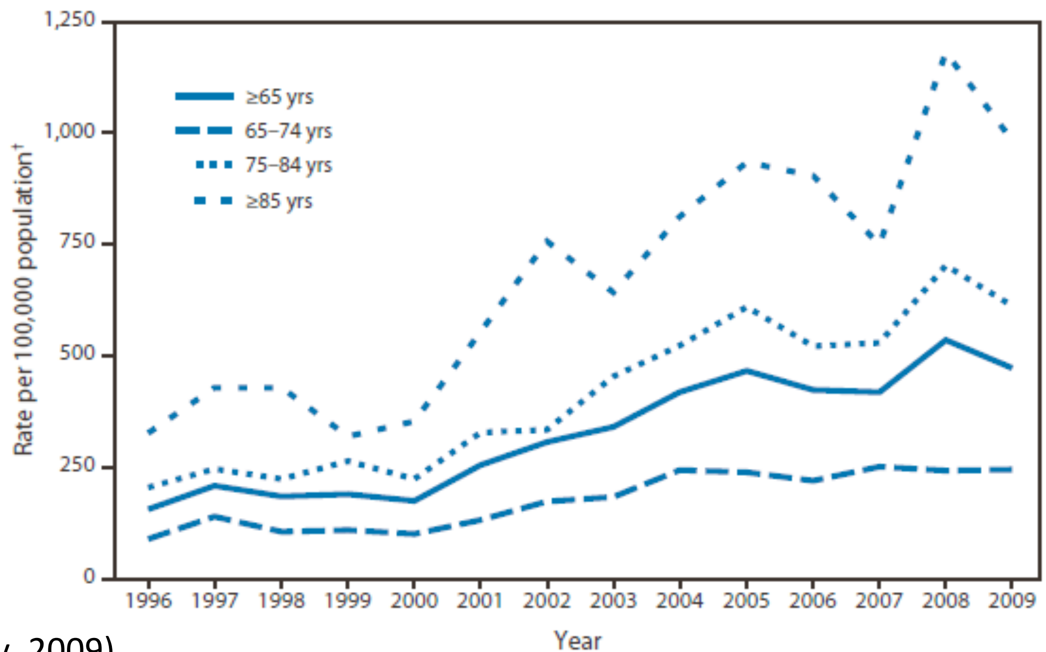
Prevalence

- Hospital-acquired: 178,000/year

  (McDonald et al., 2006)

- On par with number of new cases of invasive breast cancer in the US each year

  (American Cancer Society, 2009)



(CDC, 2011)

Risk Factors

Time Invariant

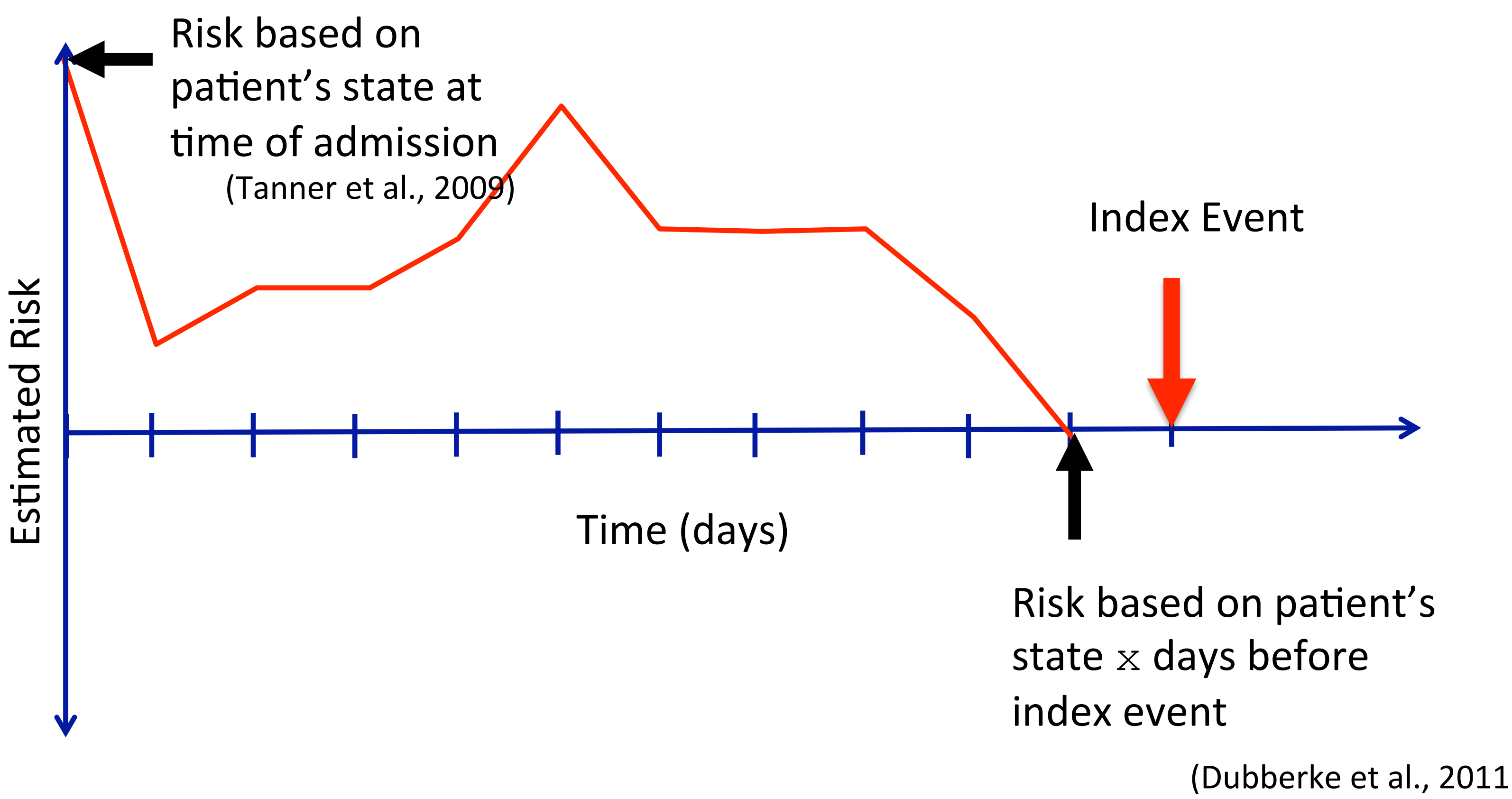
- Collected at the time of admission
- *e.g., admission complaint, previous admissions, home meds*

Time Varying

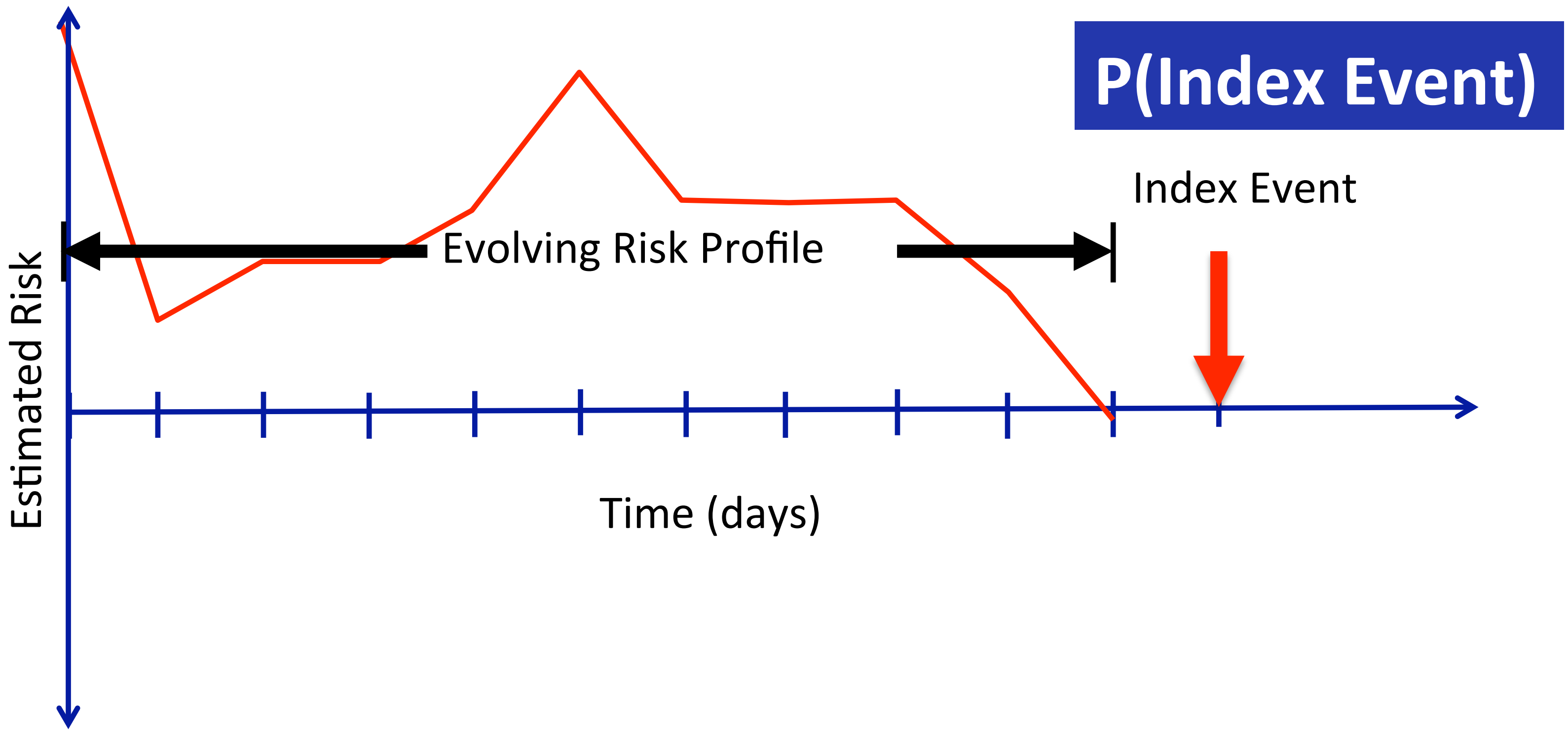
- Changes during the hospitalization
- *e.g., current meds, current procedures, current location, hospital conditions*

Representing and reasoning about temporal processes promises to enhance the accuracy of inferences about risk.

Typical Approach in Clinical Literature



Our Approach

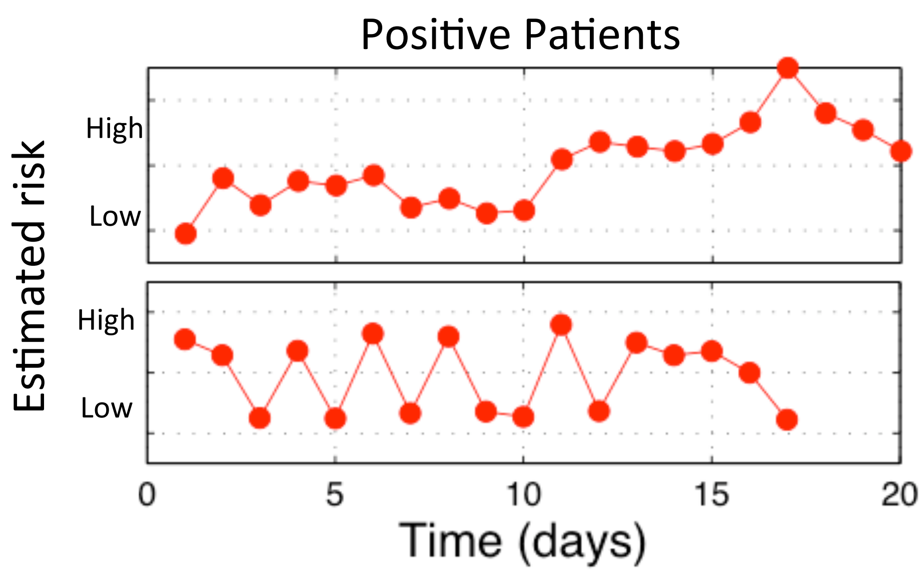


Risk Processes

Hypothesis: extracting and analyzing evolving patient risk can lead to a more accurate model for predicting infections

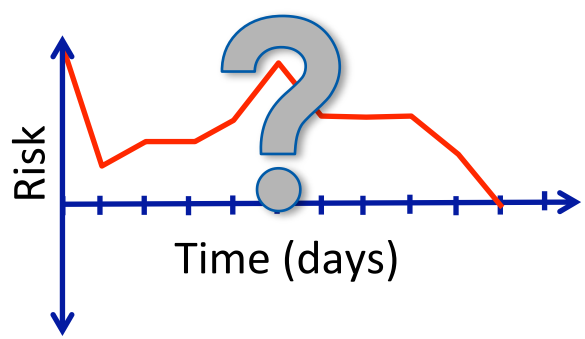
Risk Process:

describes the evolution of risk over the course of a hospital admission



Inferring Risk Processes

- Challenges:
  - No ground truth about risk
    - Retrospective data → not all patients get tested
    - Actual risk on any day is unknowable
  - Thousands of correlated variables

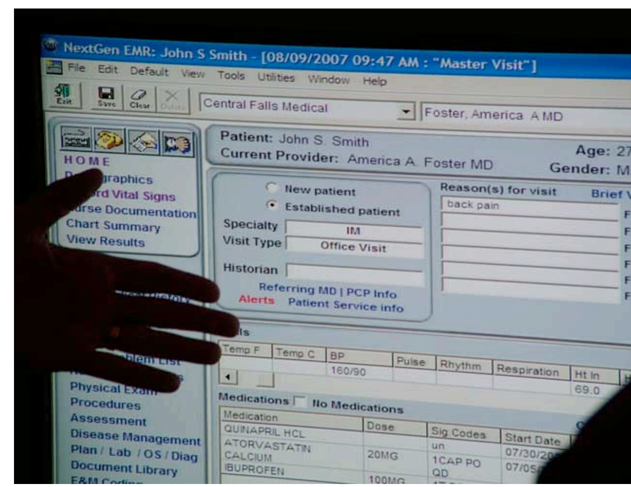


The Data

- Database from a large urban hospital in the US
- In-patient stays from a single year
- Inclusion criteria (see paper for details)
  - Eliminate easily identifiable cases

Population:

- ~10,000 hospital admissions
- ~200 Positive *C. diff* cases



Experimental Setup

- Training & Testing
  - Randomly subsampled the negative class
  - Split data into stratified training and test sets 70/30.
  - Training set 1,251 admissions (127 positive)
  - Testing set 532 admissions (50 positive)

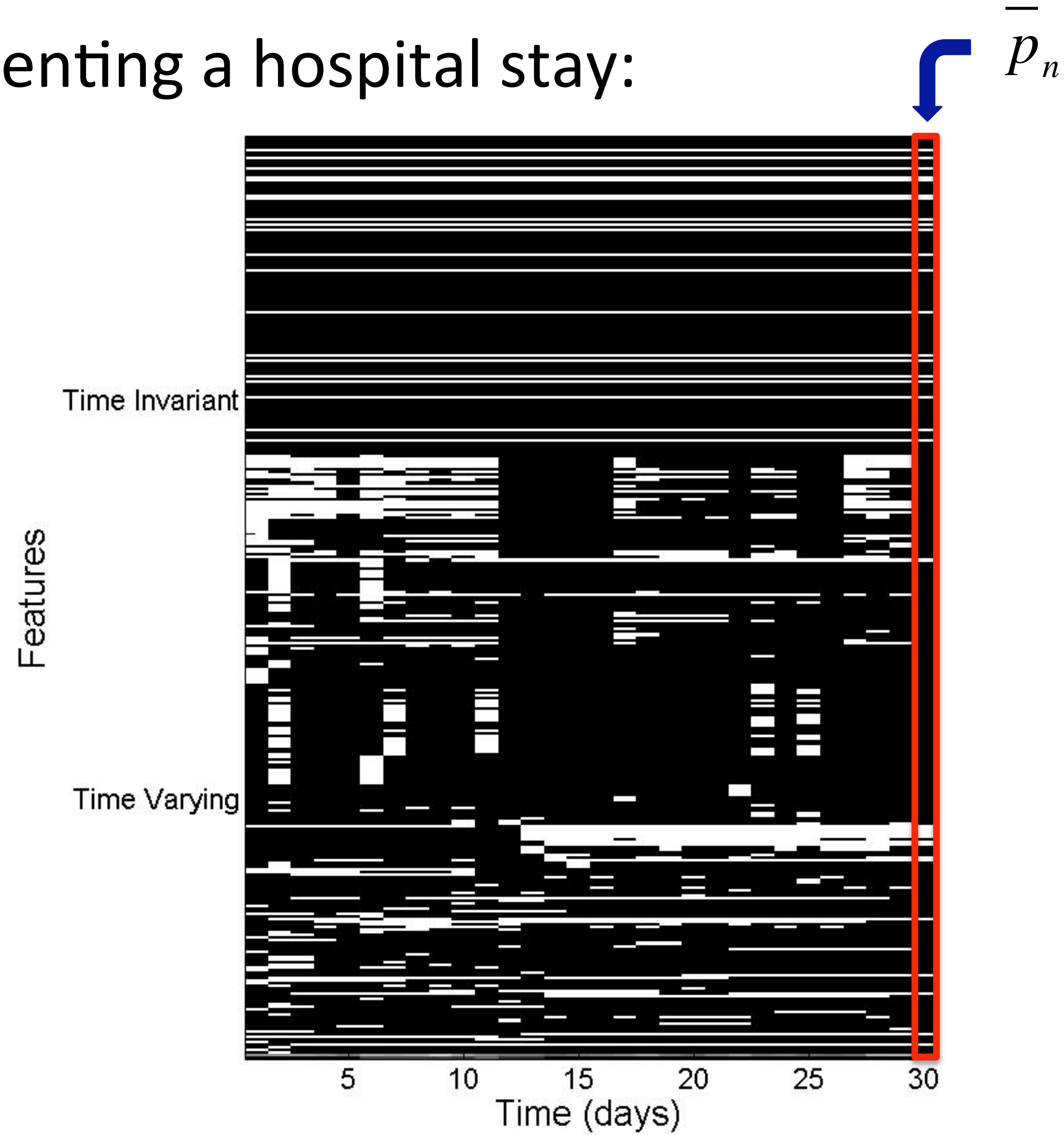
Features

Time Invariant		Time Varying
• prev. ICD 9 codes	• patient's age	• lab results
• home medications	• patient's marital status	• procedures
• prev. admission medications	• patient's sex	• location room
• patient's city	• expected surgery	• location unit
• attending MD	• ER admission	• medications
• Hospital service	• dialysis	• vitals
• admission source	• diabetic	• day of admission
• financial class code	• history of C. diff	• unit CP
• admission complaint	• num. hospital visits (90 days)	• hospital wide CP
• admission procedure	• avg., max., total los (90 days)	
• patient's race		

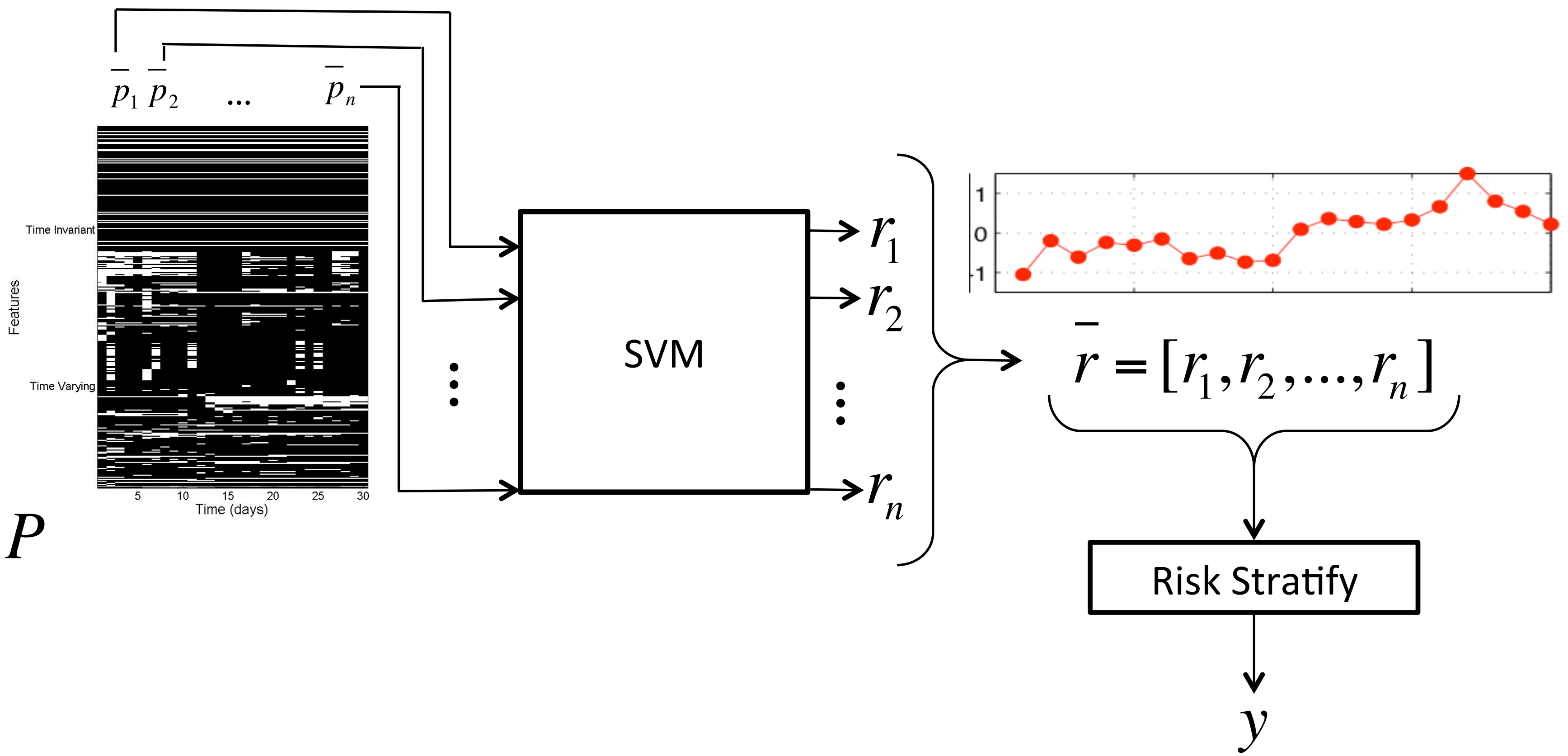
Features: >10,000 variables for each day of every hospital admission

Time Invariant		Time Varying
• prev. ICD 9 codes	• patient's age	• lab results
• home medications	• patient's marital status	• procedures
• prev. admission medications	• patient's sex	• location room
• patient's city	• expected surgery	• location unit
• attending MD	• ER admission	• medications
• Hospital service	• dialysis	• vitals
• admission source	• diabetic	• day of admission
• financial class code	• history of C. diff	• unit CP
• admission complaint	• num. hospital visits (90 days)	• hospital wide CP
• admission procedure	• avg., max., total los (90 days)	
• patient's race		

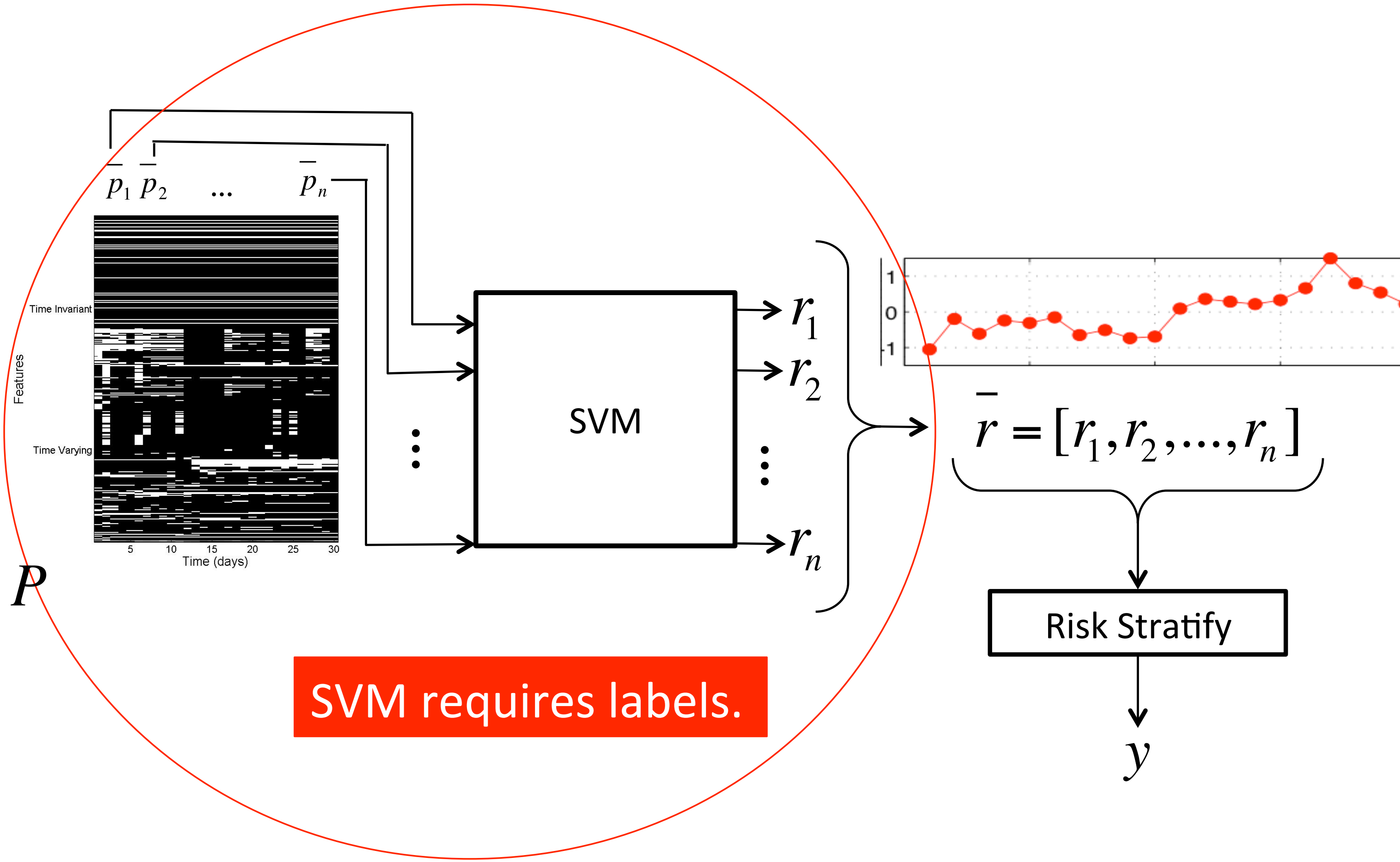
Representing a hospital stay:



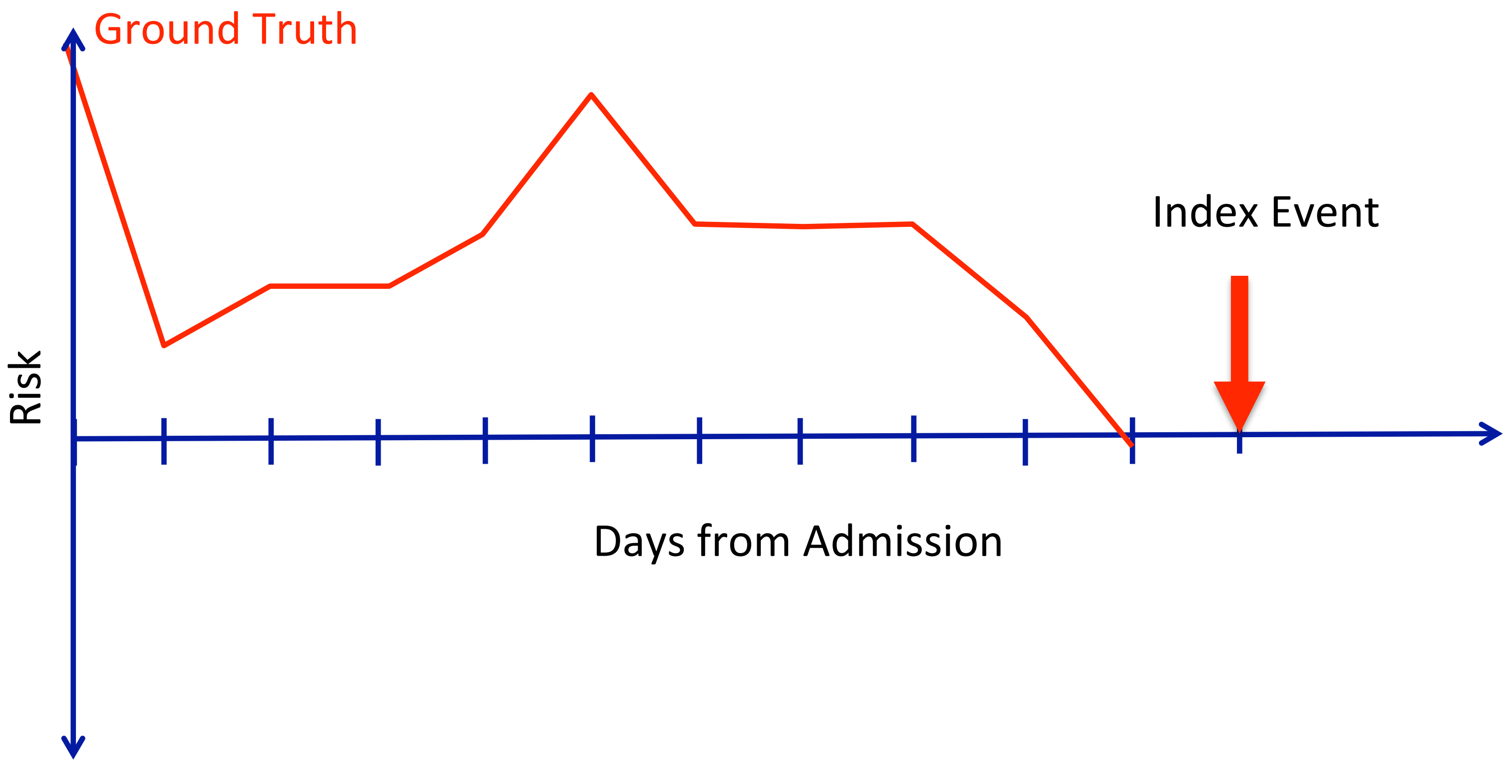
Our Approach to Risk Stratification



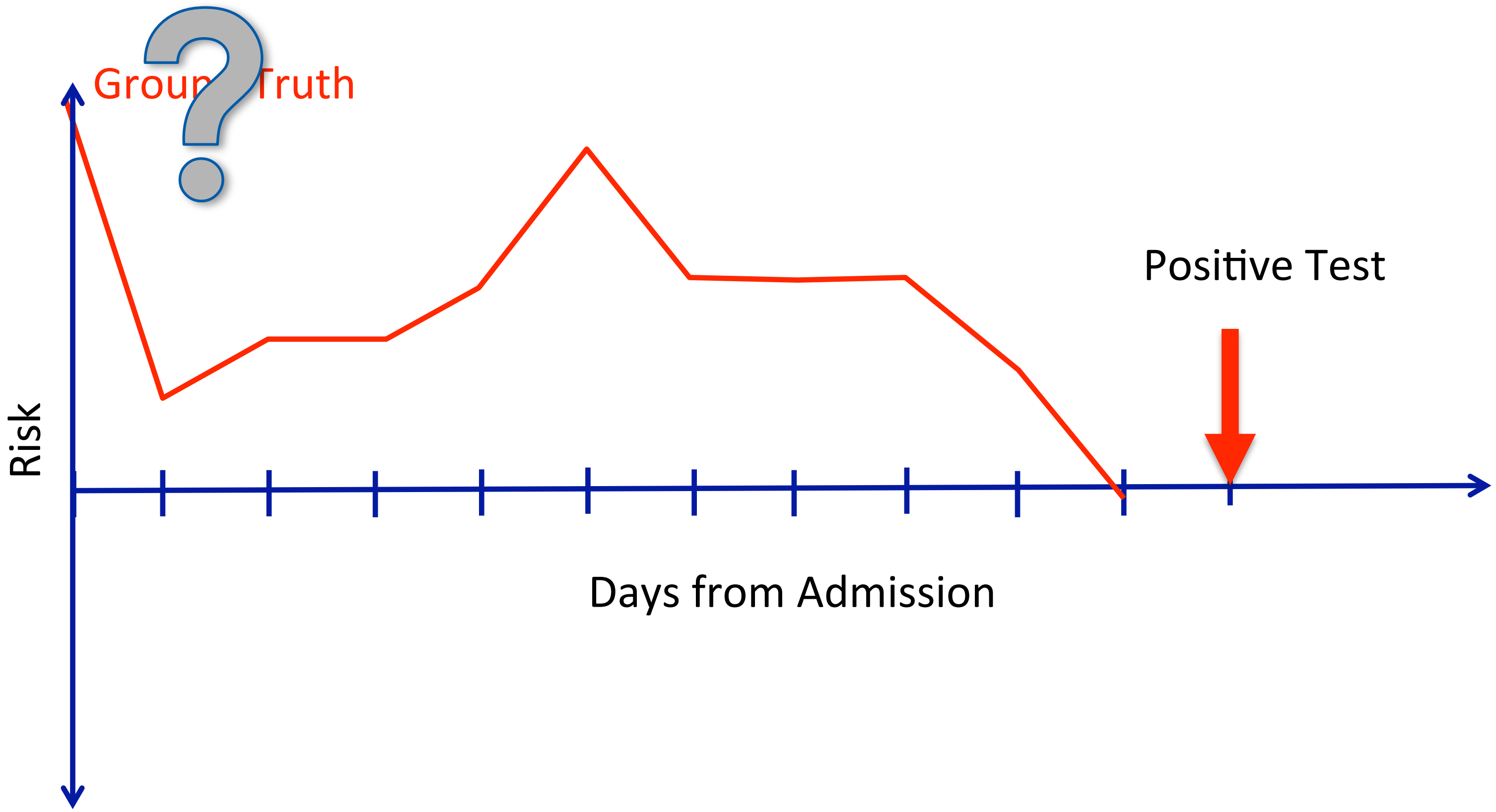
Our Approach to Risk Stratification



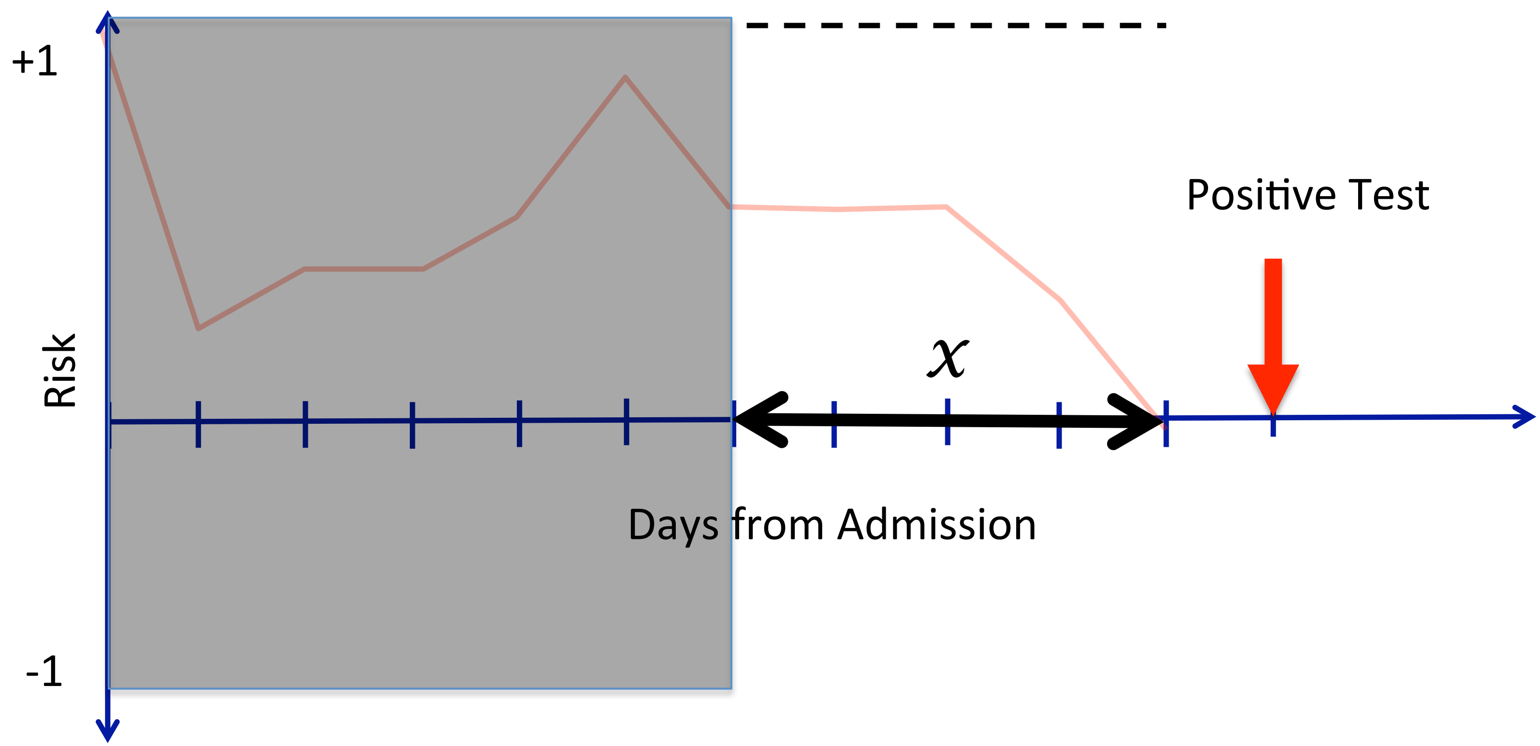
Labeling the Data



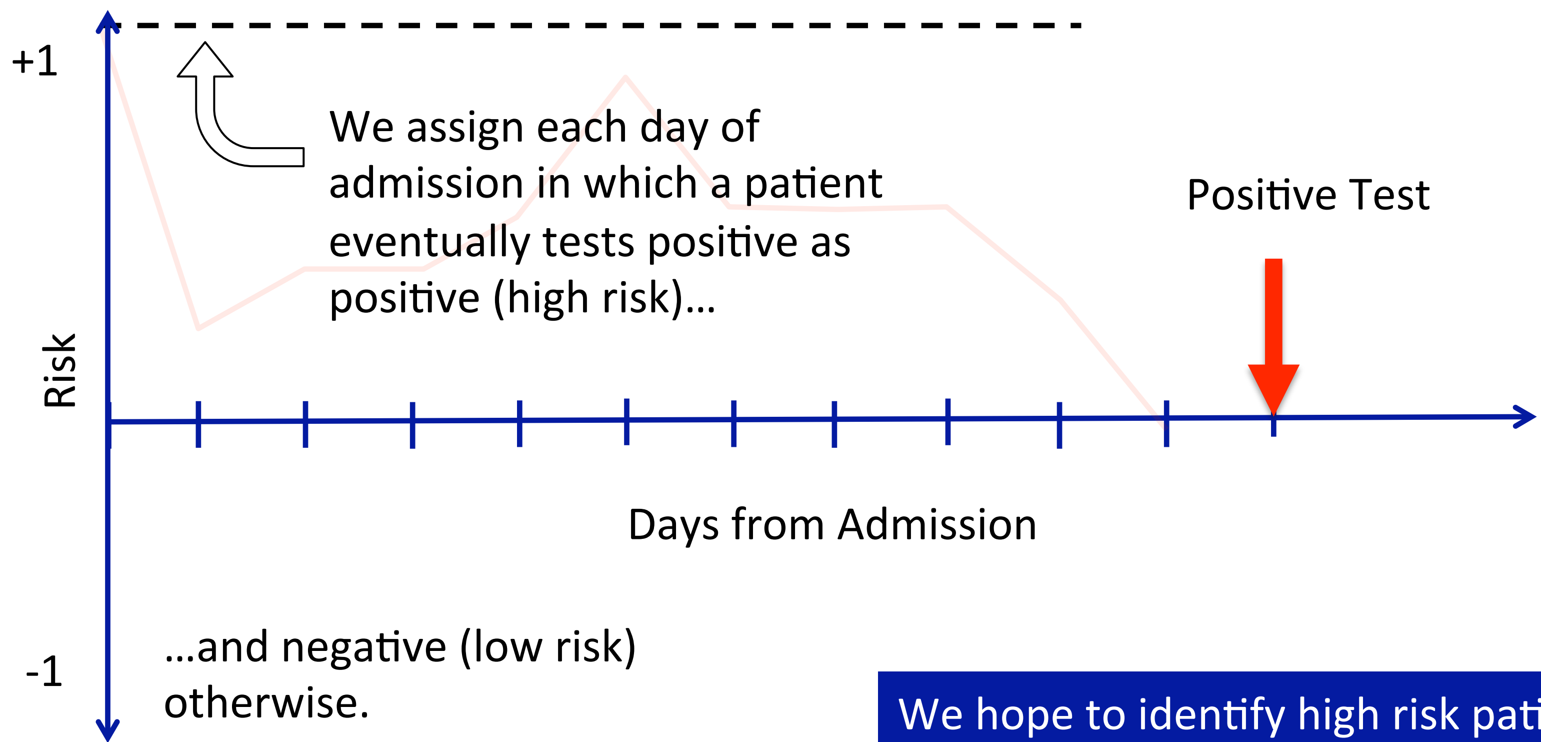
Labeling the Data



Labeling the Data

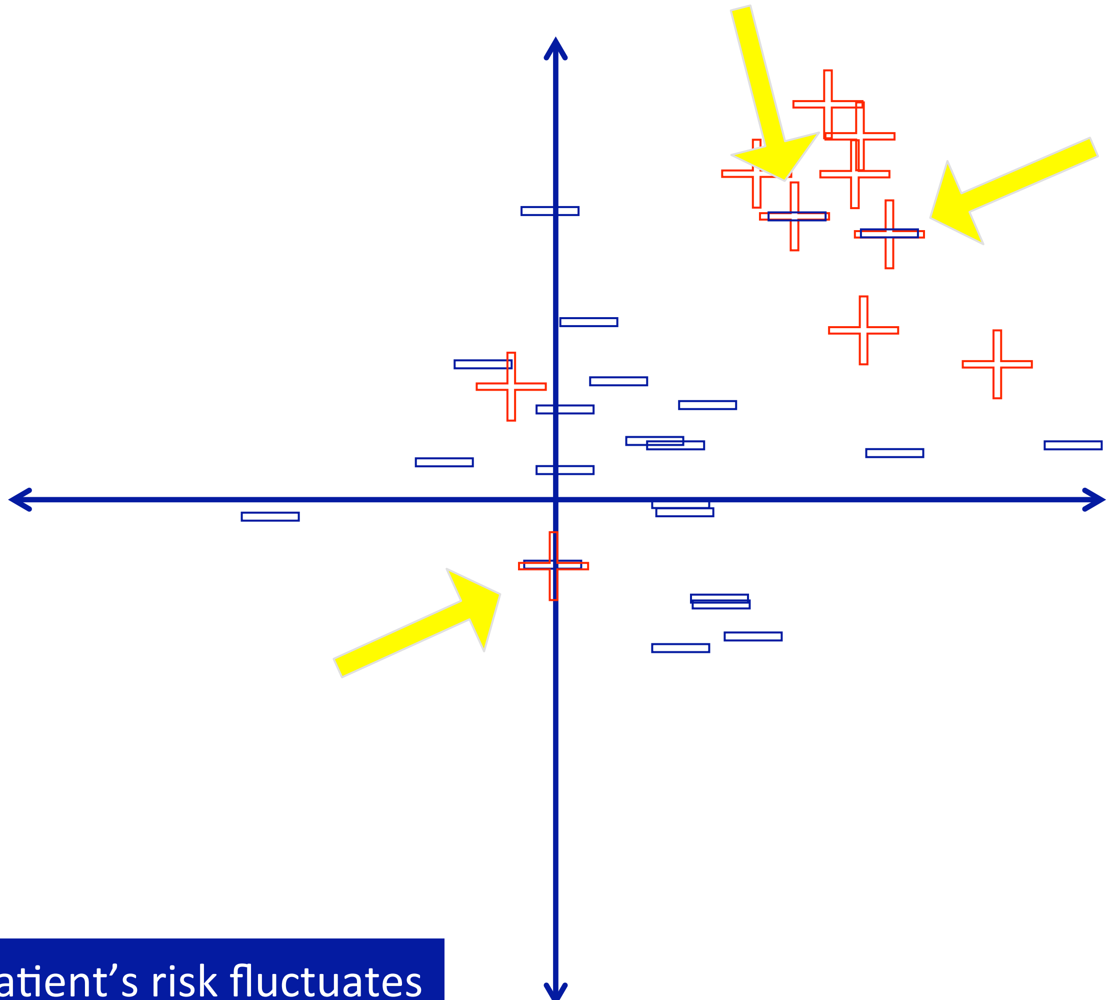


Labeling the Data



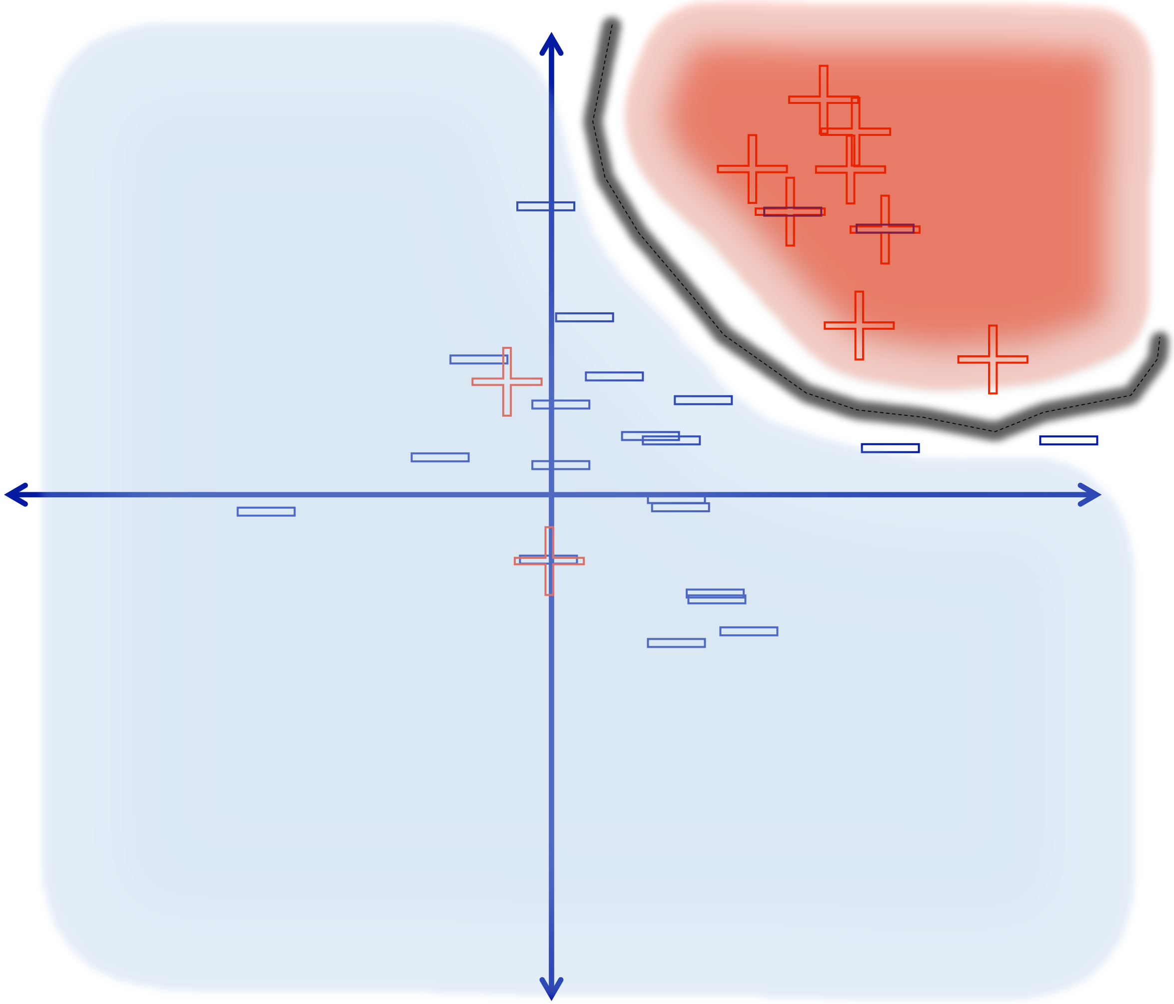
We hope to identify high risk patients as early as possible.

Learning the Decision Boundary



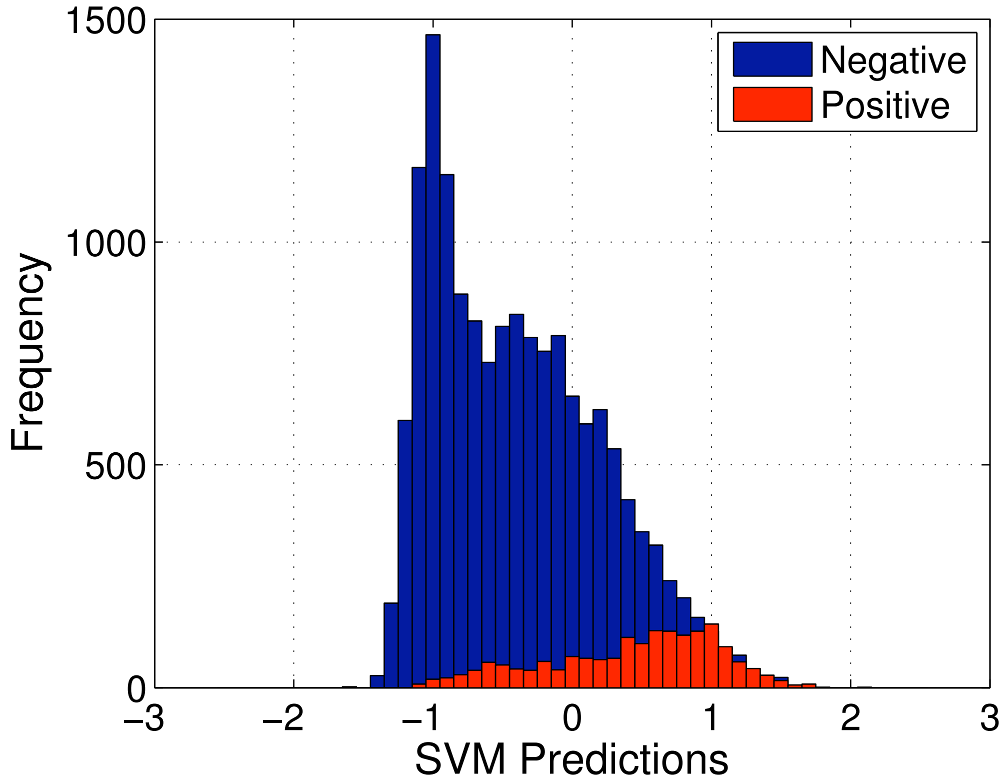
We expect a patient's risk fluctuates
→ noise in the training labels

Learning the Decision Boundary



Note: Simplified illustration. We learn a linear hyperplane in the high dimensional feature space.

Daily Risk -> SVM Continuous Predictions

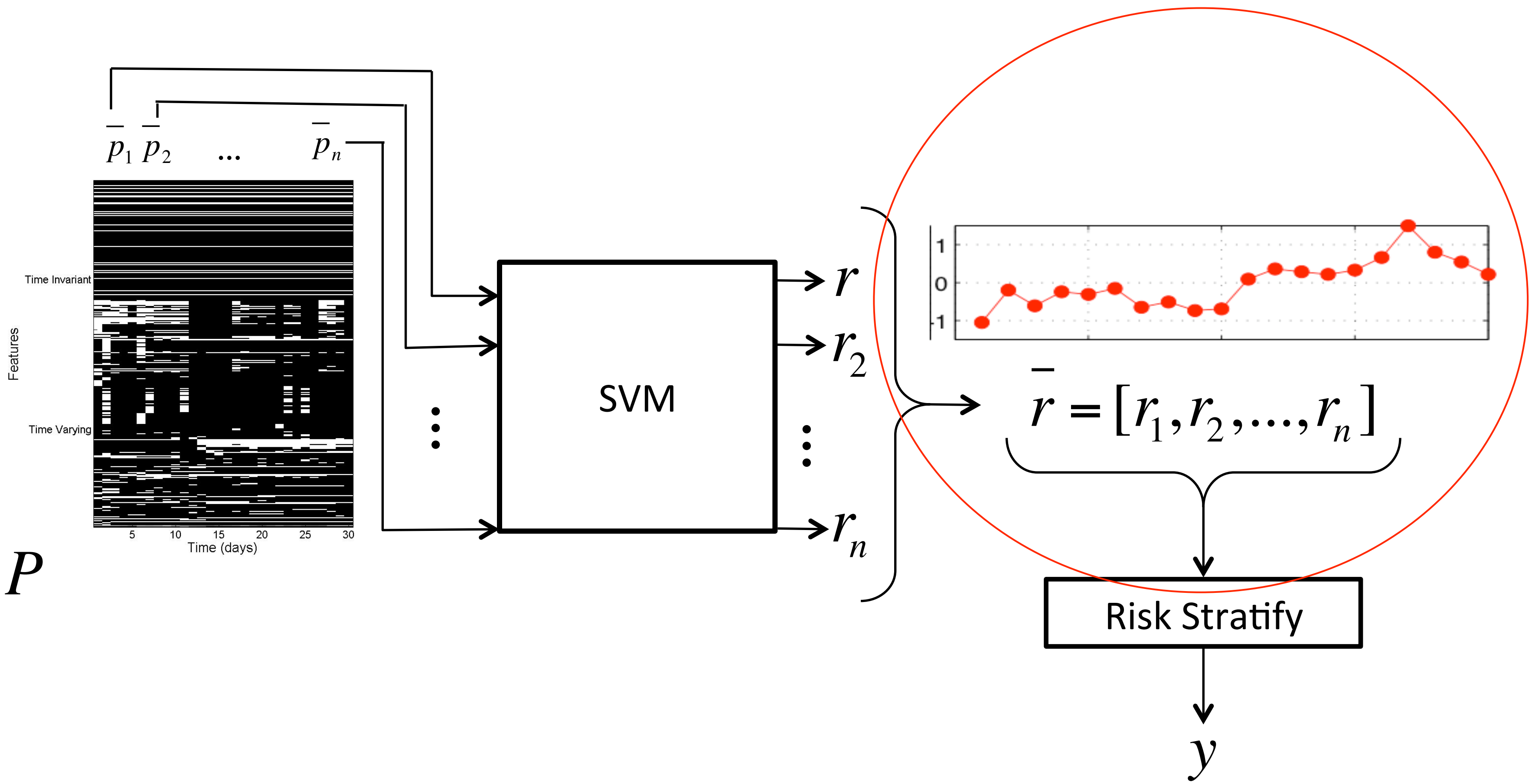


(Training data)

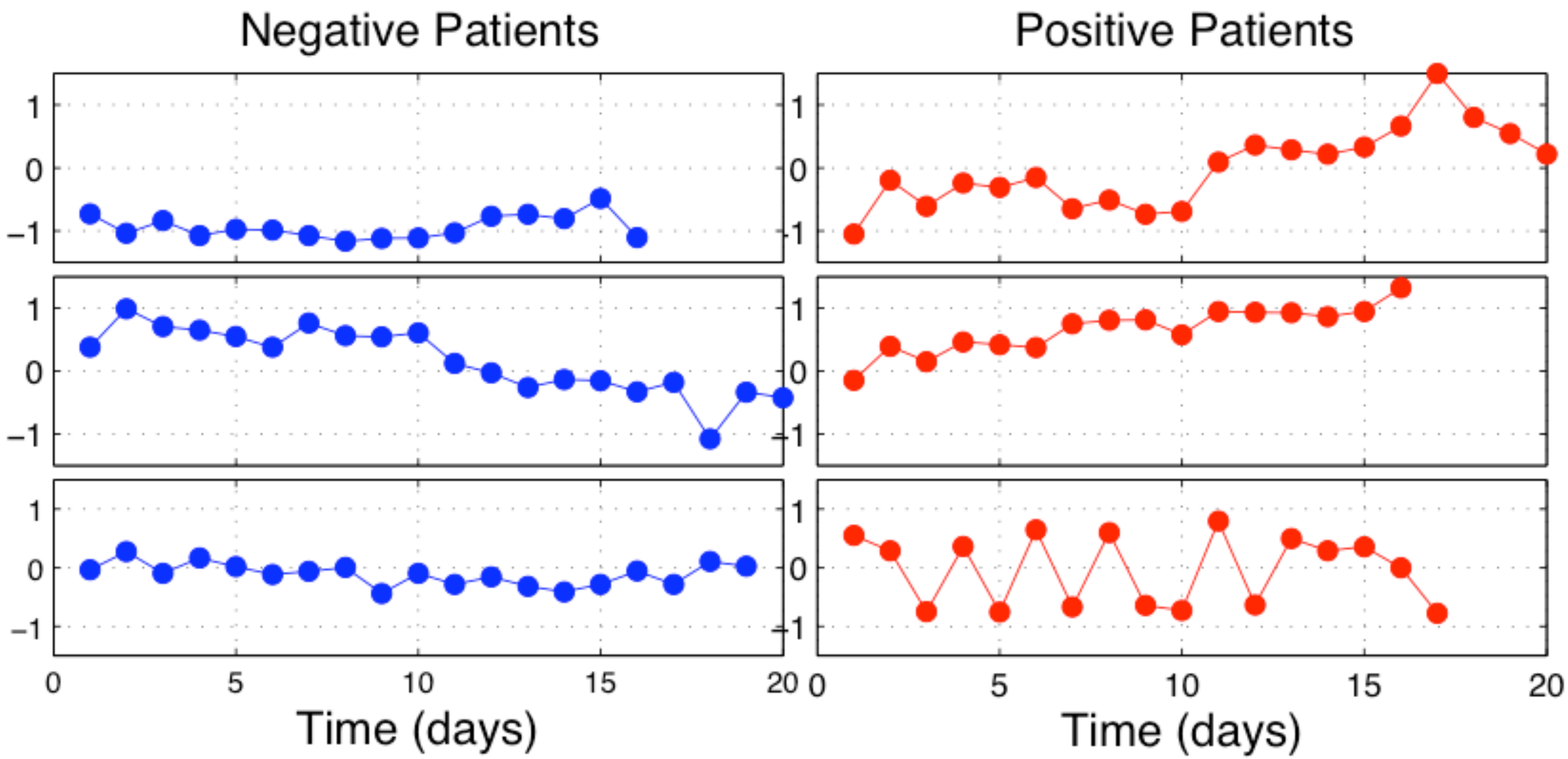
We consider the distance each feature vector lies from the SVM decision boundary this results in a **continuous** prediction for each day.

$$r_d = \overline{w} \bullet \overline{p_d} - b$$

Our Approach to Risk Stratification



Example Risk Processes



Using Risk Processes for Risk Stratification

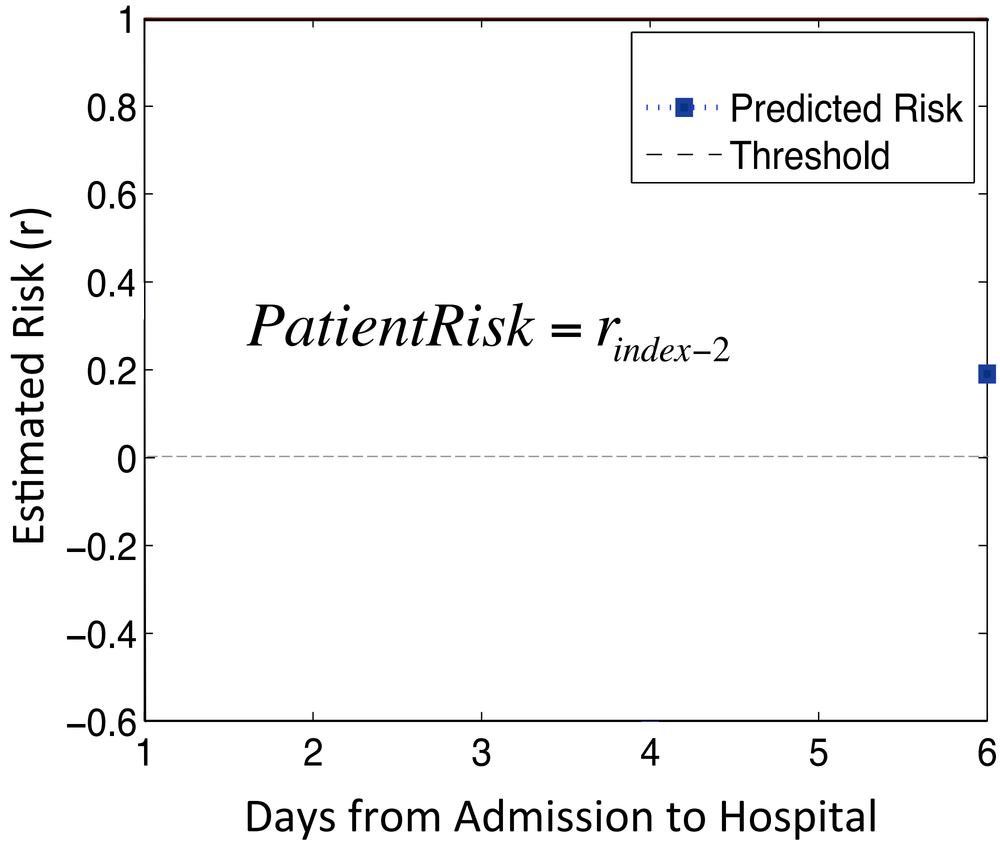
- Instantaneous approach:
  - Analogous to typical risk stratification approaches (Dubberke et al., 2011)
  - Considers value of risk process only on day of prediction
- Cumulative approach:
  - Combine estimates from all previous days

    E.g., constant, linear, and quadratic weighted averages

Evaluating Instantaneous Approach

Patient tests positive for *C. diff* on day 8

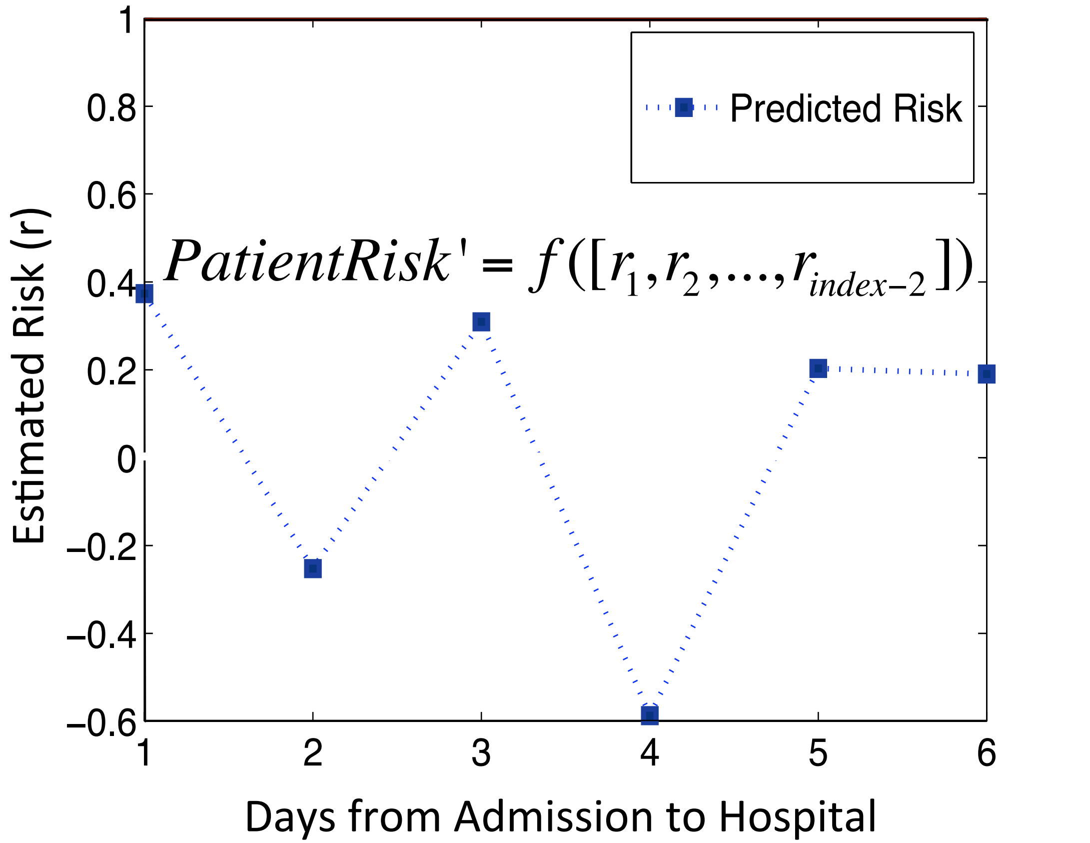
- Consider **instantaneous** estimate for patient risk at a constant distance before the index event e.g., 2 days



Compute classifier performance by sweeping the decision threshold from min to max.

Evaluating Cumulative Approach

Patient tests positive for *C. diff* on day 8



- **Combine** estimates for patient risk from the time of admission up to a constant distance from the index event e.g., 2 days

Compute classifier performance by sweeping the decision threshold from min to max.

Defining the Index Event

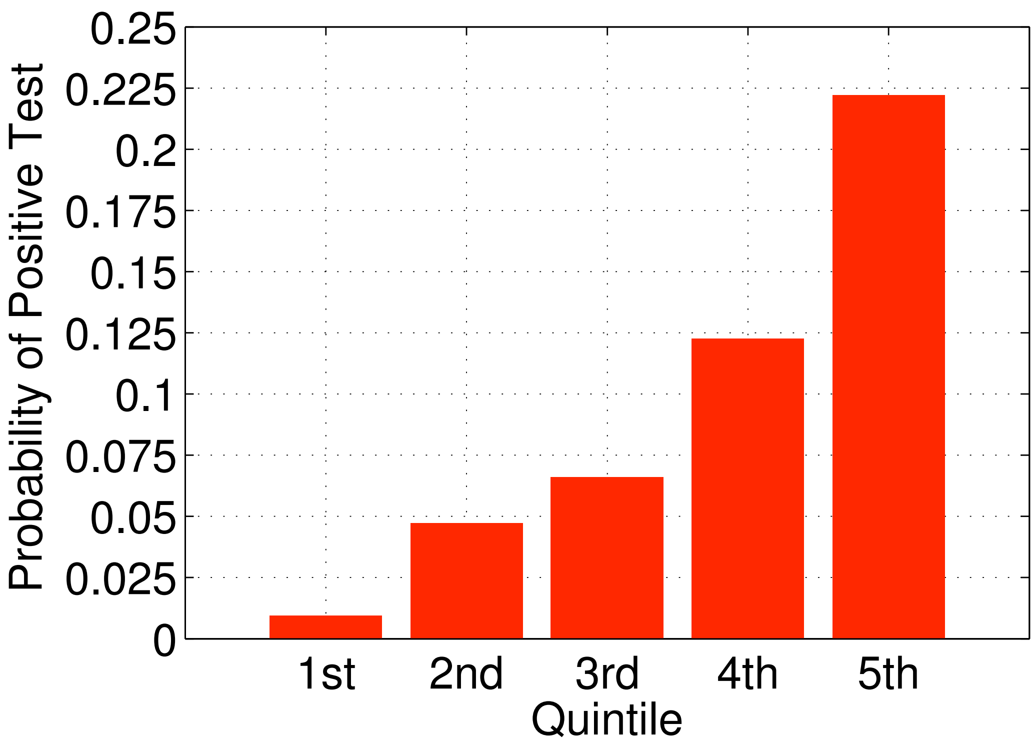
- Positive Examples → day of positive test result
  - We consider only data collected up to two days before a positive test result
- Negative Examples → midpoint of admission
  - Considering discharge as the index event can lead to deceptively good results

Results

Cumulative

Approach	Testing AUROC (95% CI)
Constant weighted avg.	0.7518 (0.69-0.81)
Linear weighted avg.	0.7444 (0.67-0.80)
Quadratic weighted avg.	0.7360 (0.67-0.80)
Instantaneous	0.6870 (0.61-0.77)

Results



Patients in the 5th quintile are at >20-fold greater risk than those in the 1st quintile!

Conclusion

- First step in analyzing how patient risk for acquiring *C. diff* may evolve during a hospitalization
  - Improvement over existing methods
- Next steps:
  - Find patterns of risk that lead to worse/better outcomes
  - Investigate application in other contexts (e.g., other HAIs, in-hospital mortality, LOS)

Acknowledgements

- Natural Sciences and Engineering Research Council of Canada (NSERC)
- National Science Foundation (NSF)
- Quanta Computer Inc.

Works Cited

- McDonald LC, et al. *Clostridium difficile* infection in patients discharged from US short-stay hospitals, 1996-2003. *Emerg Infect Dis* 2006;12:409-15.
- Pepin J, et al. Increasing risk of relapse after treatment of Clostridium difficile colities in Quebec, Canada. Clin Infect Dis. 2005 Jun 1;40(11):1591-7.
- American Cancer Society. Breast Cancer Facts & Figures 2009-2010. Atlanta: American Cancer Society, Inc.
- Tanner, J. et al. Waterlow score to predict patients at risk of developing clostridium difficile-associated disease. Journal of hospital infection, 71(3):239-244, 2009.
- Dubberke, E. et al., Development and validation of a clostridium difficile infection risk prediction model. Infect Control Hosp Epidemiol 32(4):360-366,2011.