Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder

Todd Lingren, Pei Chen, Joseph Bochenek, Finale Doshi-Velez, Patty Manning-Courtney, Julie Bickel, Leah Wildenger Welchons, Judy Reinhold, Nicole Bing, Yizhao Ni, William Barbaresi, Frank Mentch, Melissa Basford, Joshua Denny, Lyam Vazquez, Cassandra Perry, Bahram Namjou, Haijun Qiu, John Connolly, Debra Abrams, Ingrid A. Holm, Beth A. Cobb, Nataline Lingren, Imre Solti, Hakon Hakonarson, Isaac S. Kohane, John Harley, Guergana Savova

Published on July 6th, 2016

Chen Arviv, 3.1.17

מוטיבציה:

- השכיחות של ASD היא 1 ל-68 בקרב ילדים מתחת לגיל 8*

- למחקרים גדולים שמבוססים על EHRs באופן רחב היקף יש פוטנציאל לגלות יחסים יותר מורכבים כמו קלסטרים בקרב לוקים ב ASD ולתרום לחקר ה-ASD בצורה משמעותית: כיווני טיפול יותר מבטיחים, הבנת סיכונים עתידיים.

- מחקרים כאלה נתקלים בבעיה בחילוץ ASD cohort אמין: גוזל הרבה משאבים לאשר באופן ידני שכל אחד ב cohort הוא אכן לוקה ב-ASD.

- על מנת לנצל את הכוח של EHR data רחב היקף, אנחנו חייבים להיות מסוגלים לבצע חילוץ cohort בעל אמינות גבוהה בצורה אוטומטית.

*"Prevalence of autism spectrum disorder among children aged 8 years–autism and developmental disabilities monitoring network"
http://onlinelibrary.wiley.com/doi/10.1002/aur.1506/abstract;jsessionid=1D3031860ECE02F62C2DCBBCBE43BCFB.f02t01

מטרות המחקר:

- שימוש ב EHRs של מספר מוסדות כדי לקדם את המחקר של ASD על ידי:

- פיתוח אלגוריתם אוטומטי לחילוץ ASD cohort בעל תוקף

- בדיקה של comorbidities (תחלואות נלוות) בקרב לוקים ב ASD.

- <u>האלגוריתמים שנראה:</u>

1. Rule-based algorithm.
2. Machine learning algorithm: Predictive classifier.

<u>מטרתם:</u> חילוץ ASD cohort אמין.

- נתחיל מלהציג את השלבים המקדימים המשותפים ל-2 האלגוריתמים.

שלב 1: initial cohort

- ICD-9 Codes (International Classification of Diseases 9th edition=) רשימה סטנדרטית של קודים שמציינים אבחון של מחלות.
- משפרת את הקונסיסטנטיות בקרב רופאים ברשומות הרפואיות לצורך תשלום, מעקב של רופאים שונים וגם לצורך מחקרים רפואיים.

2015 ICD-9-CM Diagnosis Codes >

- 001-139 Infectious And Parasitic Diseases
- 140-239 Neoplasms
- 240-279 Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
- 280-289 Diseases Of The Blood And Blood-Forming Organs
- 290-319 Mental Disorders
- 320-389 Diseases Of The Nervous System And Sense Organs
- 390-459 Diseases Of The Circulatory System
- 460-519 Diseases Of The Respiratory System
- 520-579 Diseases Of The Digestive System
- 580-629 Diseases Of The Genitourinary System
- 630-679 Complications Of Pregnancy, Childbirth, And The Puerperium
- 680-709 Diseases Of The Skin And Subcutaneous Tissue
- 710-739 Diseases Of The Musculoskeletal System And Connective Tissue
- 740-759 Congenital Anomalies
- 760-779 Certain Conditions Originating In The Perinatal Period
- 780-799 Symptoms, Signs, And Ill-Defined Conditions
- 800-999 Injury And Poisoning
- V01-V91 Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services
- E000-E999 Supplementary Classification Of External Causes Of Injury And Poisoning

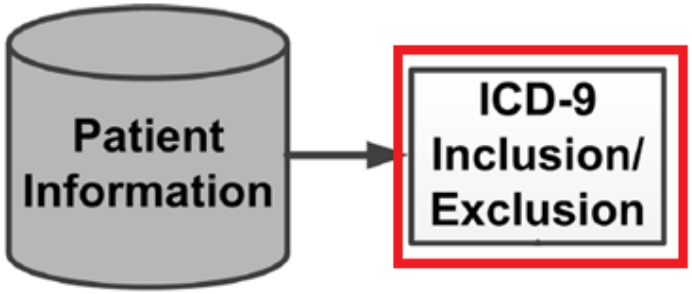
שלב 1: initial cohort

- **<u>סינון על פי ICD-9 Codes</u>**: מתוך ה EHR Databases של בתי החולים BCH (Boston Children's Hospital) ו-CCHMC (Cincinnati Children's Hospital Medical Center) יצרו cohort התחלתי שהיה מורכב מכל הפציינטים שהיה להם ברשומה הרפואית <u>לפחות</u> <u>אחד</u> מהקודים המאבחנים ICD-9 הבאים: (inclusion codes)
  - 299.0 = Autism
  - 299.80 = Asperger's
  - 299.9 = Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS)

<u>ולא</u> היה להם קוד ICD-9 שמתייחס לאחד מהבאים – (exclusion codes)

Rett Syndrome, Childhood Disintegrative Disorder, Schizophrenia, Tuberous Sclerosis, and Fragile X Syndrome, severe Intellectual Disability (cognitively delayed or IQ < 40), psychiatric illnesses

- סה"כ: BCH: 14,758 , CCHMC: 4,229
- אנחנו נמצאים כאן:



שלב 2: gold standard set

- ה-cohort ההתחלתי למרות הסינון עדיין מכיל מקרים של false-positives:
פציינטים שברשומה הרפואית שלהם מופיע קוד ICD-9 כאמור,
אבל הם לא לוקים ב-ASD.

- מתוך ה-cohort ההתחלתי (BCH: 14,758 , CCHMC: 4,229) נבחרו 302 פציינטים
(BCH: 150 , CCHMC: 152) בצורה רנדומלית.
המידע הרפואי של כל פציינט נבחן בצורה ידנית על ידי הרופאים כדי ליצור gold standard set =
קבוצה וולידית מולה נבחן את תוצאות האלגוריתמים.

- **יש 4 לייבלים ב- gold standard set:**

Yes– לוקה ב-ASD.

No– לא לוקה ב- ASD.

Maybe– מקרה שאינו וודאי, יש איזושהיא ראייה לכך.

Unknown– מקרה שבו אין מספיק מידע כדי לבצע החלטה בהתייחס למקרה.

שלב 2: gold standard set

- ב-BCH היו 2 רופאים שעברו על תיקי הפציינטים בצורה ידנית, ואילו ב-CCHMC 3 רופאים.
- <u>הם פסקו שמתוך ה gold-standard set:</u>
  - Yes : 44% (60% מתוך CCHMC ו- 26% מתוך BCH)
  - No : 10% (16% מתוך CCHMC ו- 5% מתוך BCH)
  - Maybe : 36% (18% מתוך CCHMC ו- 55% מתוך BCH)
  - Unknown : 10% (6% מתוך CCHMC ו- 13% מתוך BCH)

- **<u>Inter Annotator Agreement (IAA):</u>**

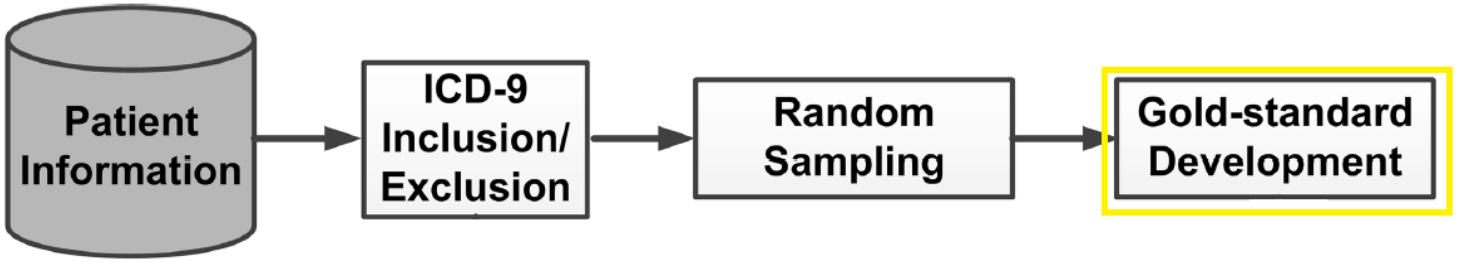
כדי לבדוק את מידת ההתאמה בין קלסיפיקציות של רופאים שונים באותו מרכז (=IAA)
נלקחה קבוצת בדיקה של 20 פציינטים מ-BCH ו-40 פציינטים מ-CCHMC.
הדרך הכמותית לחשב את מידת ההתאמה בין קלסיפיקציות של רופאים שונים היא F-measure.

שלב 2: gold standard set

- מידת ההתאמה שנמצאה בקרב 3 הרופאים ב-CCHMC:
  0.969 עבור Yes, 0.896 עבור No, 0.883 עבור Maybe, 0.75 עבור Unknown.

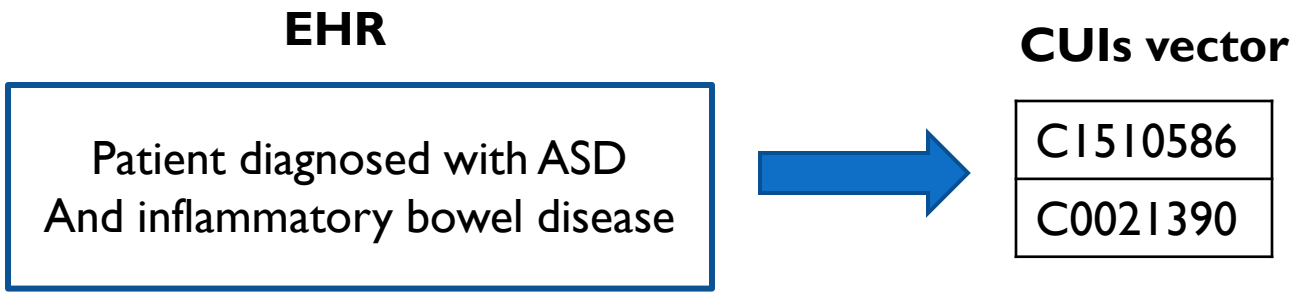
- מידת ההתאמה שנמצאה בקרב 2 הרופאים ב-BCH:
  0.927 עבור Yes, 0.625 עבור No, 0.86 עבור Maybe, 0.75 עבור Unknown.

- אנחנו נמצאים כאן:



Inputs to EHRs algorithms

- השתמשו במערכת לעיבוד שפה טבעית Apache cTakes כדי למדל את הרשומה הרפואית של כל פציינט מה initial cohort לווקטור של CUIs (= Concept Unique Identifiers) מתוך UMLS (Unified Medical Language System).

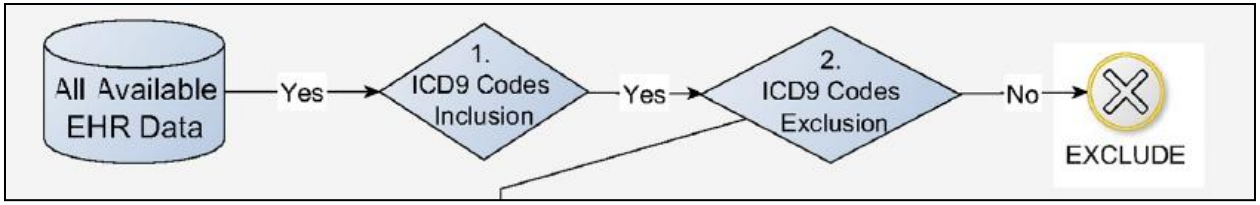


C1510586 (עבור ASD) ו C0021390 (עבור inflammatory bowel disease).

- אורך הוקטור הממוצע של פציינט היה: ב-CCHMC: 344 CUIs , ב-BCH: 627 CUIs.
- הוקטורים האלו הם האינפוט לשני אלגוריתמים שונים שנציג של חילוץ cohort אוטומטי.

Rule-based algorithm

שלב 1: יצירת ה-initial cohort כפי שהצגנו קודם באמצעות סינון על פי ICD-9 Codes.



כעת בידינו ה-initial cohort שדיברנו עליו. (BCH: 14,758 , CCHMC: 4,229)

שלב 2: מיפוי של CUIs לסימפטומים ממילון* לסימפטומי ASD.

ביצוע מיפוי של CUIs לסימפטומים ממילון לסימפטומי ASD – מילון שמכיל רשימה מקיפה של סימפטומים של ASD ע"פ הקריטריונים המאבחנים ASD של DSM-IV (= Diagnostic and Statistical Manual of Mental Diseases 4th Edition)

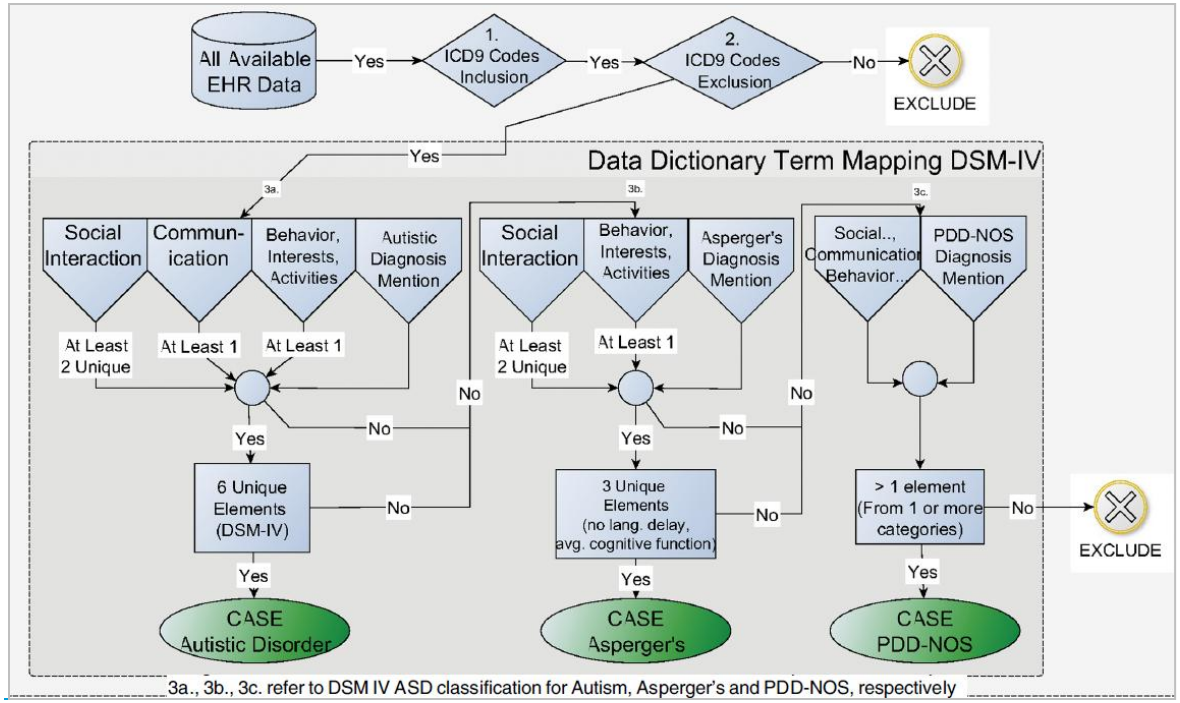
שלב 3: Rule-based cohort selection

עבור כל פציינט: האלגוריתם ספר כמה סימפטומים של ASD (שהיו מיוצגים על ידי CUIs) מופיעים בווקטור של הפציינט.

* http://jamanetwork.com/journals/jamapediatrics/fullarticle/485907

Rule-based algorithm

- אם לפציינט היו לפחות 6 סימפטומים (כך שלפחות 2 מ-Social Interaction, לפחות 1 מ- Communication, ולפחות 1 מ- Behavior, Interests and Activities) אז הפציינט נקבע כ Autistic Disorder Case.
- אם פציינט לא התאים לקריטריון של autistic disorder (3a) התבצעה הערכה של הסימפטומים עבור Asperger's Syndrome (3b) – אם הוא התאים, הוא נקבע כ Asperger's Case.
- אם פציינט לא התאים לקריטריון של תסמונת אספרגר, התבצעה הערכה של הסימפטומים עבור PDD-NOS (3c) - אם הוא התאים, הוא נקבע כ PDD-NOS Case.
- אם פציינט לא התאים לאף אחד מהקריטריונים 3a-c, הוא לא אף אחד מהמקרים.



הקביעה שניתנת בהינתן הסימפטומים מבוססת על הקריטריון של DSM-IV לאיבחון של ASD, תסמונת אספרגר ו -PDD NOS.

Machine learning algorithm

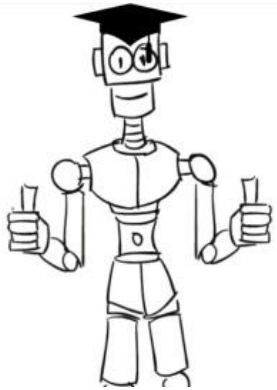
- כדי לפתח את האלגוריתם הזה, חילקו את ה-gold standard set (302) ל-3 קבוצות (table 1):
  Training Set -60%, Development Set -20% , Testing/Evaluation Set -20%.

Table 1. Distribution of patients

	Training	Development	Test	Total
BCH	95	27	28	150
CCHMC	87	34	31	152
Combined	182	61	59	302

BCH–Boston Children's Hospital; CCHMC–Cincinnati Children's Hospital and Medical Center.

- הסיווגים ללייבלים השונים התבצעו ע"י שימוש במערכת SVM (Support Vector Machines) דו שלבית:



Stage 1: בשלב הראשון הפציינטים סווגו לאחת מ-2 קטגוריות:

Unknown או No **או** Maybe או Yes

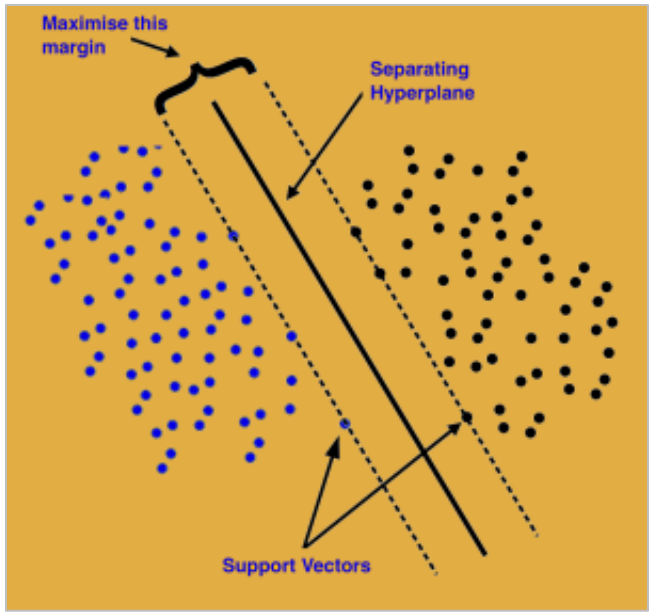
Stage 2: הפציינטים שבקטגוריה:

Maybe או Yes

סווגו לקטגוריה Yes **או** לקטגוריה Maybe

Machine learning algorithm

- **SVM:** מודל של למידה חישובית שמשתמש בקבוצה של דגימות, features ולייבלים, כך שכל דגימה משתייכת ללייבל 1 או לייבל 2, על מנת ליצור מסווג בינארי שבהינתן דגימה חדשה, היא תסווג ללייבל הנכון.
- מבוסס על הרעיון של מציאת ה-hyperplane שמפריד dataset בצורה הטובה ביותר ל-2 מחלקות.
- דגימות חדשות מסווגות לפי המחלקה שהן נופלות אליה.



Machine learning algorithm

- קבוצת ה features ההתחלתית הם ה- CUIs שמופיעים באיחוד של כל הווקטורים ב- gold standard set.
- על הקבוצה הזו ביצעו feature selection: כדי להקטין את המימד של ווקטורי המטופלים ובכך לזרז את זמן ביצוע האלגוריתם.
- ביצעו זאת על ידי שימוש במבחן חי בריבוע: χ^2
- בסטטיסטיקה מבחן חי בריבוע משמש כדי לבדוק האם יש תלות בין 2 משתנים. במקרה שלנו אנו רוצים לבדוק האם יש תלות בין feature לללייבל אליו משויך הפציינט.
- לכל feature חישבו את ה- chi square value, ודירגו את ה-features על פי הערכים מהגבוה לנמוך:
- ככל של feature יש chi square value יותר גבוה זה מעיד שיש תלות יותר מובהקת בין השניים.

features: לאחר התהליך נקבע כי ה- features set יהיה מורכב מ 250 features שה-chi square value שלהם הוא הכי גבוה. (upper-bound)

labels: Yes, Maybe, No, Unknown

Machine learning algorithm

שלבים עיקריים:

1. **Training**

על ה-Training Set המודל התאמן ויצר מסווגים עם אפשרויות שונות של features sets ו- cost parameters.

2. **Development**

על ה-Development Set נבדקו המסווגים (ששונים זה מזה ב features sets ו-cost parameters) על מנת לבחור את המסווג שנותן את הביצועים הטובים ביותר.

3. **Test**

הערכת המסווג על ה- Test Set.

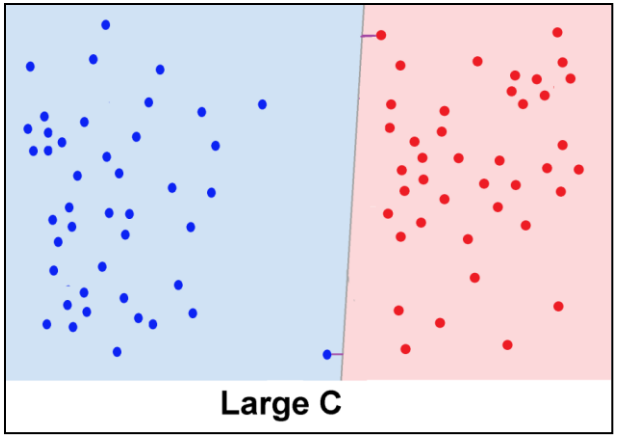
Machine learning algorithm

Development – Cost parameter

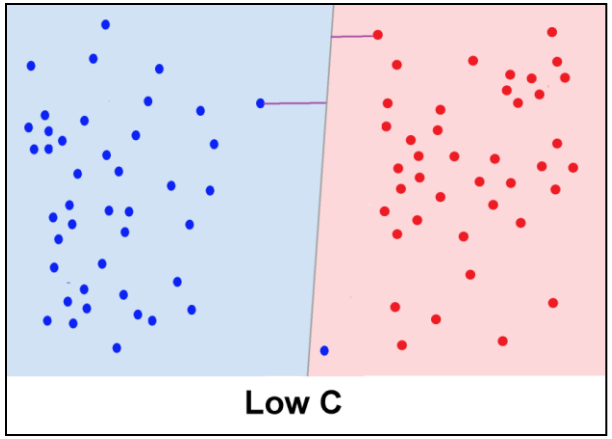
המטרה של מודל SVM היא ליצור מסווג בינארי שיצליח לסווג נכונה דגימות עתידיות.
לשם כך רוצים:

1. Hyperplane כך שה-margin המינימלי על פני הנקודות הקיימות יהיה הגדול ביותר.
2. Hyperplane שמפריד בצורה נכונה כמה שיותר נקודות קיימות.

ה-פרמטר C קובע כמה חשוב לנו להקפיד על מס' 2 גם אם זה יגבה מחיר של מס' 1. למשל:



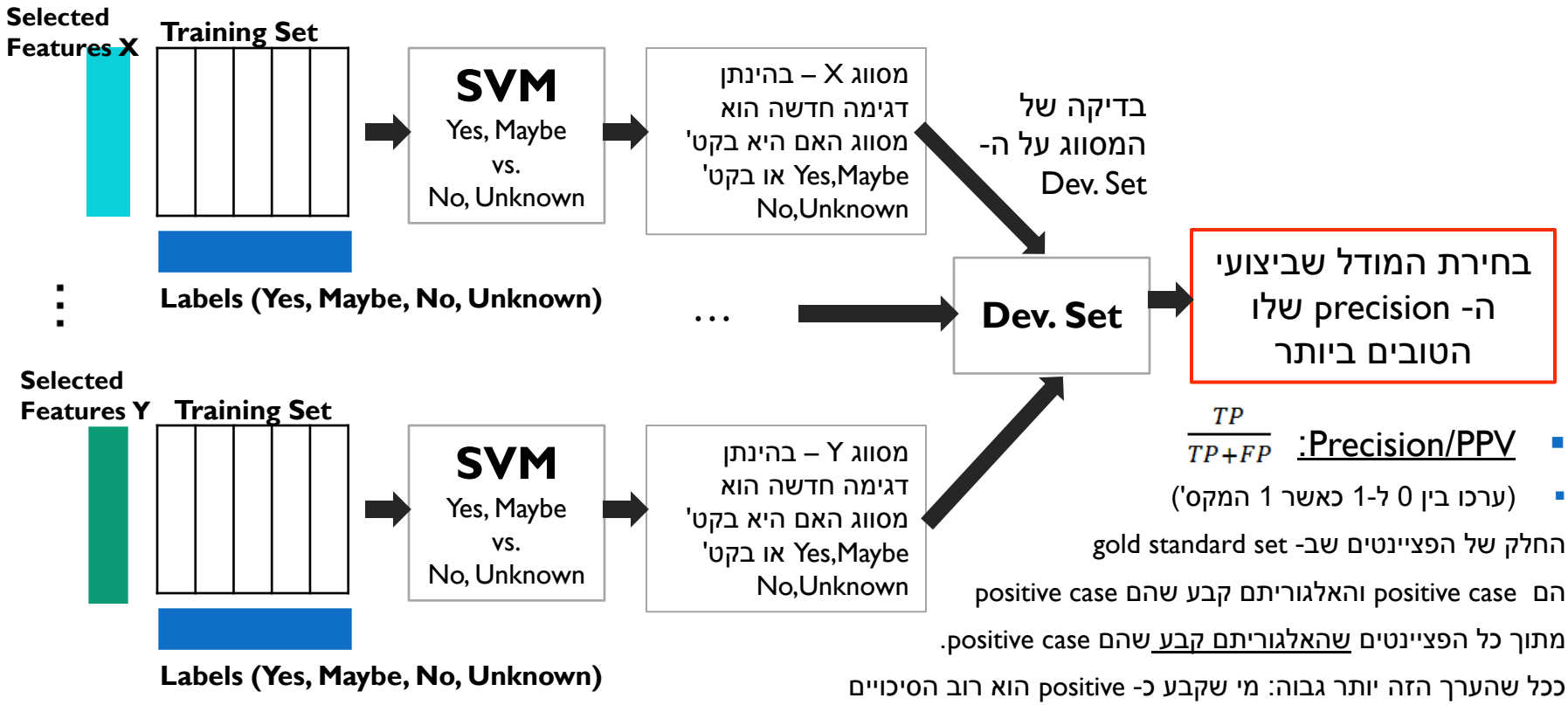
ערך c גדול נותן סיווג יותר טוב, אבל ה-margin המינימלי יותר קטן



ערך c נמוך נותן margin מינימלי די גדול (בסגול)
אבל במחיר של סיווג שגוי של דגימה כחולה

Machine learning algorithm

Training and Development (Stage I):



- Precision/PPV: $\frac{TP}{TP+FP}$
- (ערכו בין 0 ל-1 כאשר 1 המקס')

החלק של הפציינטים שב- gold standard set הם positive case והאלגוריתם קבע שהם positive case מתוך כל הפציינטים שהאלגוריתם קבע שהם positive case.

ככל שהערך הזה יותר גבוה: מי שקבע כ- positive הוא רוב הסיכויים באמת positive, יש פחות טעויות של false positivies.

Stage 1

Test of Stage I

הערכת המסווג על ה- Test Set.

ווקטור של פציינט מה- Test Set

המסווג הנבחר (שביצועי ה- precision שלו היו הטובים ביותר על ה- Dev. Set)

המסווג מחזיר תשובה-

Maybe או Yes

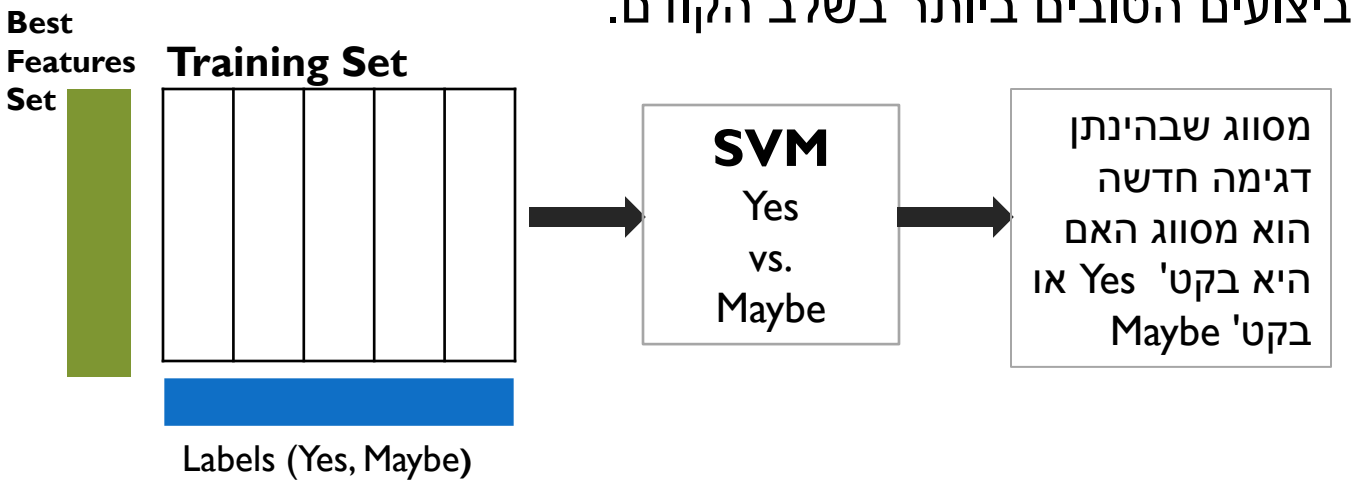
או

Unknown או No

נראה בהמשך כיצד העריכו את ביצועיו.

Stage 2

עבור יצירת המסווג של Stage 2 השתמשו ב- features set וב- cost parameter שנתנו את הביצועים הטובים ביותר בשלב הקודם.



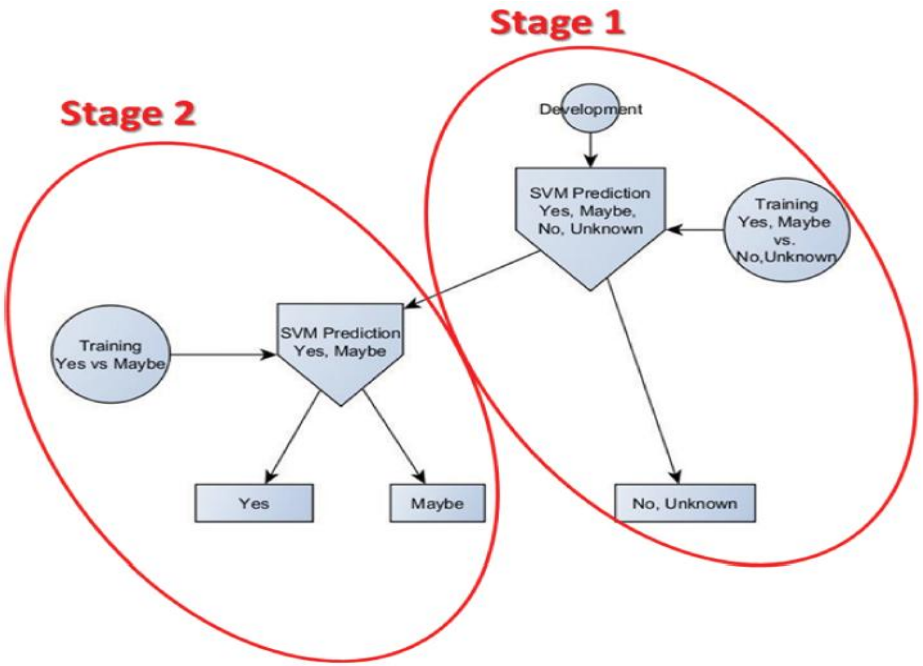
Test of Stage 2

הערכת המסווג על ה- Test Set.

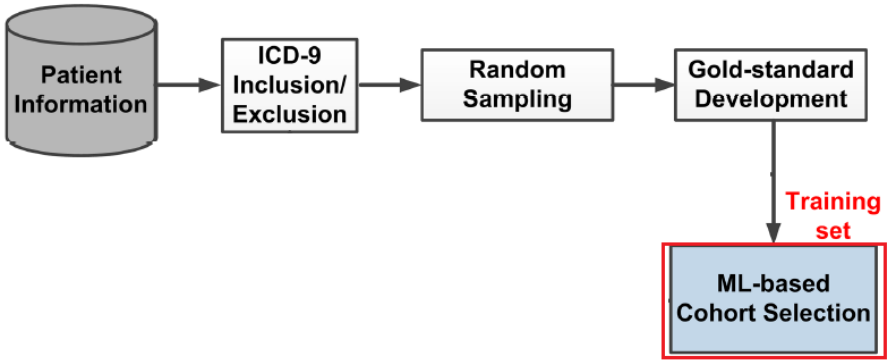
באופן דומה ל- Stage 1, בודקים את המסווג על וקטורי הפציינטים שב- Test Set.

עבור כל וקטור, המסווג מחזיר תשובה: Yes או Maybe, ונראה כיצד העריכו את ביצועיו.

Summary



- סיכום השלבים:



- אנחנו נמצאים כאן:

Results: Rule-based algorithm

- בדקו את ביצועי האלגוריתם על ה- gold standard set. (היא שימשה כ- **Evaluation Set**)

- **<u>הבדיקה נעשתה בצורה בינארית:</u>** (positive case/negative case)
  - ASD, Asperger's, PDD-NOS - positive case
  - הלייבלים Yes ו-Maybe של ה gold standard set positive case (והשאר negative)

- **<u>אז קיימים 4 מקרים אפשריים:</u>**

1. האלגוריתם קבע שהפציינט הוא positive case ואכן לפציינט יש לייבל של positive case (כלומר Yes או Maybe) = True Positive
2. האלגוריתם קבע שהפציינט הוא positive case אך לפציינט יש לייבל של negative case (כלומר No או Unknown) = False Positive
3. האלגוריתם קבע שהפציינט הוא negative case ואכן לפציינט יש לייבל של negative case (כלומר No או Unknown) = True Negative
4. האלגוריתם קבע שהפציינט הוא negative case אך לפציינט יש לייבל של positive case (כלומר Yes או Maybe) = False Negative

Results: Rule-based algorithm

מושגים (בדיקת טיב האלגוריתם):

- **Precision/PPV**: $\frac{TP}{TP+FP}$ (ערכו בין 0 ל-1 כאשר 1 המקס')

- החלק של הפציינטים שהאלגוריתם קבע שהם positive case והלייבל שלהם אכן שייך ל-positive case (במקרה שלנו Yes, Maybe) מתוך כל הפציינטים שהאלגוריתם קבע שהם positive case.
- ככל שהערך הזה יותר גבוה: רוב המקרים שקבע כ- positive הם באמת positive, יש פחות טעויות של false positives.

- **Recall/Sensitivity**: $\frac{TP}{TP+FN}$ (ערכו בין 0 ל-1 כאשר 1 המקס')

- החלק של הפציינטים שהאלגוריתם קבע שהם positive case והלייבל שלהם אכן שייך ל-positive case (במקרה שלנו Yes, Maybe) מתוך כל הפציינטים שהלייבל שלהם שייך ל-positive case.
- המדד מעיד על כמה מהמקרים שהם positive cases הוא הצליח להחזיר שהם positive cases. כמה מקרים כאלה הוא "איתר" ביחס לכל ה positive cases הקיימים.

Results: Rule-based algorithm

- **F1-Measure:** $$F1 - Measure = \frac{2 * precision * recall}{precision + recall}$$

- מדד לדיוק של התוצאות.

הערך שלו הוא בין 0 ל-1, כאשר הערך הגבוה ביותר הוא 1 (כאשר precision ו-recall שניהם 1).

תוצאות:

Table 3. Rule Based Results

Evaluation Set	Precision/PPV	Recall/Sensitivity	F1-Measure	Area under ROC Curve
BCH	0.885	0.891 (14)	0.888	0.642
CCHMC	0.840	0.622 (48)	0.715	0.599
Combined	0.866	0.758 (62)	0.808	0.579
CHOP (independent validation on 50 patients)	0.849	0.737 (10)	0.788	0.659

*מוסד נוסף (CHOP= Children's Hospital of Philadelphia) בדק את ביצועי האלגוריתם באופן עצמאי עם initial cohort של 50 אנשים.

Results: ML algorithm

- בדקו את ביצועי האלגוריתם על ה- gold standard set. (היא שימשה כ- **Evaluation Set**)

Stage 1: בשלב זה המסווג סיווג את הפציינטים לאחת מ-2 קט': Maybe או Yes או Unknown או No

הבדיקה נעשתה בצורה בינארית: (positive case/negative case)

- הקטגוריה Maybe או Yes ⬅ positive case (ו- Unknown או No negative case)

- הלייבלים Yes ו-Maybe של ה-gold standard set ⬅ positive case (והשאר negative case)

[כמו ה-rule-based algorithm]

גם כאן קיימים 4 מקרים אפשריים:

1. True Positive
2. False Positive
3. True Negative
4. False Negative

Results: ML algorithm

Stage 2: בשלב זה המסווג סיווג את הפציינטים שב- Maybe או Yes לאחת מ-2 קט': Yes או Maybe

הבדיקה נעשתה בצורה בינארית: (positive case/negative case)

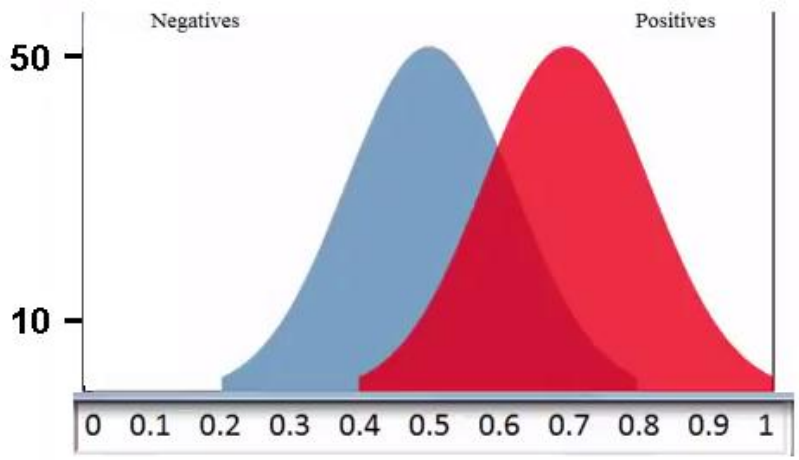
- הקטגוריה Yes ⟸ positive case (ו- Maybe negative case).

- הלייבל Yes של ה-gold standard set ⟸ positive case (והלייבלים Unknown,No,Maybe הם negative case).

- **גם כאן קיימים 4 מקרים אפשריים:**
  1. True Positive
  2. False Positive
  3. True Negative
  4. False Negative

Results: ML algorithm

- **Area under ROC Curve:** דרך נפוצה כדי להמחיש בצורה ויזואלית את הביצועים של מסווג בינארי.
- הסבר על ידי דוגמא היפותטית: תהי מחלה כלשהיא, ונניח שיצרנו מסווג בינארי שלכל פציינט מחזיר באיזו הסתברות הוא חולה במחלה .



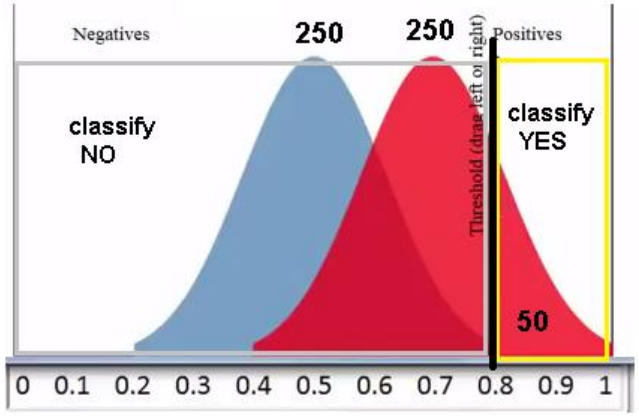
כל נקודה
היא פציינט

- הקבוצה שנלקחה לצורך בדיקת המסווג מורכבת מ-250 positives ו- 250 negatives:
- 250 נקודות אדומות = פציינטים שהלייבל האמיתי שלהם הוא positive
- 250 נקודות כחולות = פציינטים שהלייבל האמיתי שלהם הוא negative
- ציר x: ההסתברויות שהמסווג ניבא, ציר y: מספר הפציינטים (כמו בהיסטוגרמה)
- לדוגמא: יש 10 פציינטים שהמסווג החזיר שהם חולים בהסת' 0.3 והלייבל האמיתי שלהם הוא negative.

Results: ML algorithm

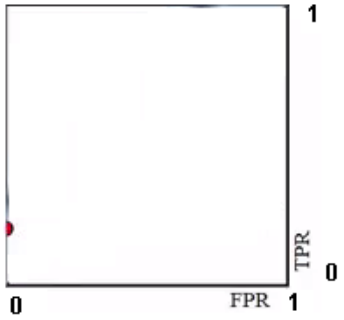
- **ROC Curve:** מתאר את הTPR (ציר y) מול ה-FPR (ציר x) עבור כל classification threshold אפשרי.

- עבור threshold ב-0.8: הסת' 0.8 ומעלה יסווג כ-positive, והסת' קטנה מ-0.8 יסווג כ-negative.



True Positive Rate (TPR) = $\frac{TP}{TP+FN}$ = 50/250 = 0.2

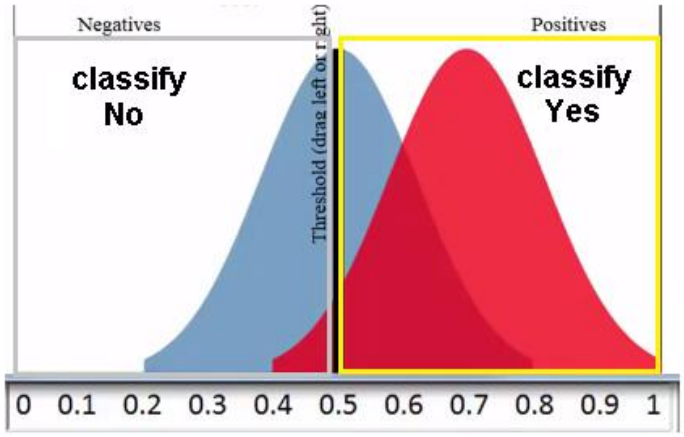
False Positive Rate (FPR) = $\frac{FP}{FP+TN}$ = 0/250 = 0



- נציב את הנקודה (0, 0.2):

Results: ML algorithm

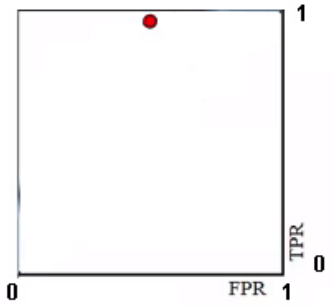
- עבור threshold ב-0.5: הסת' 0.5 ומעלה יסווג כ-positive, והסת' קטנה מ-0.5 יסווג כ-negative.



True Positive Rate (TPR) = $\frac{TP}{TP+FN}$ = 235/250 = 0.94

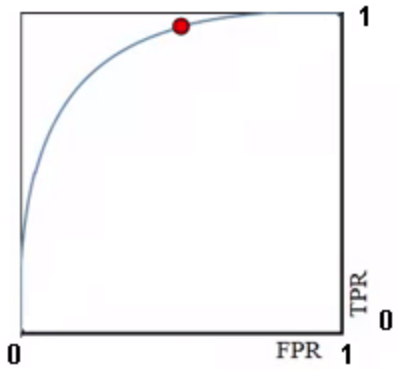
False Positive Rate = $\frac{FP}{FP+TN}$ =125/250 = 0.5

- נציב את הנקודה (0.5, 0.94):



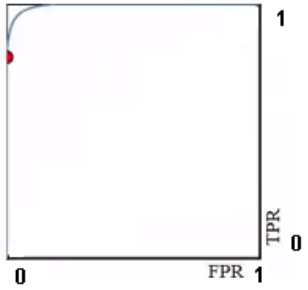
Results: ML algorithm

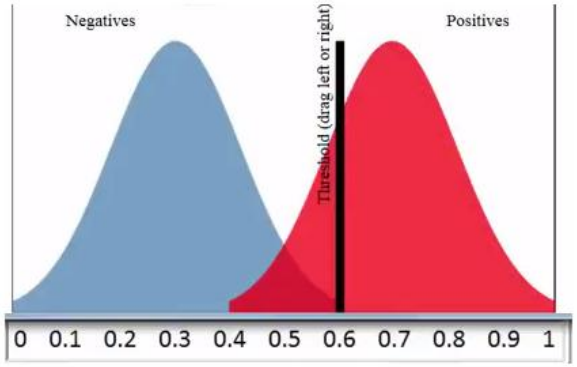
- בסופו של דבר כדי ליצור את כל ה- ROC Curve צריך לחשב FPR מול TPR עבור כל ה- classification threshold האפשריים מ-0 ל-1.
- <u>בדוג' שלנו נקבל:</u>



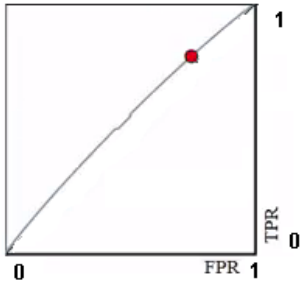
- **<u>Area under ROC Curve:</u>** השטח מתחת ל-ROC Curve. מהווה מדד לביצועי המסווג.
- ככל שהשטח קרוב יותר ל-1: המסווג יותר טוב.
- אם השטח קרוב ל-0.5: המסווג חסר תועלת משמעותית. נראה דוגמאות-

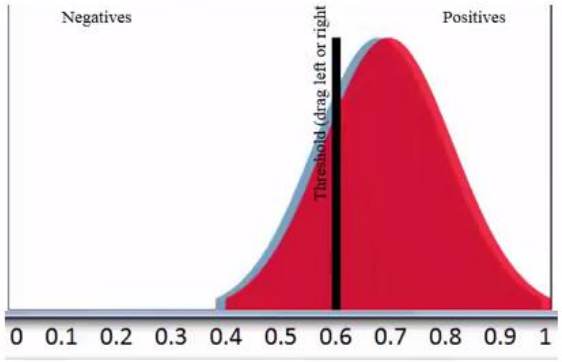
Results: ML algorithm





- מסווג מוצלח: AUC קרוב ל- 1





- מסווג לא מוצלח: AUC קרוב ל-0.5

Results: ML algorithm

תזכורת מושגים (בדיקת טיב האלגוריתם):

- **Precision/PPV**: $\frac{TP}{TP+FP}$ (ערכו בין 0 ל-1 כאשר 1 המקס')

- החלק של הפציינטים שהאלגוריתם קבע שהם positive case והלייבל שלהם אכן שייך ל-positive case (במקרה שלנו Yes, Maybe) מתוך כל הפציינטים שהאלגוריתם קבע שהם positive case.
- ככל שהערך הזה יותר גבוה: רוב המקרים שקבע כ- positive הם באמת positive, יש פחות טעויות של false positives.

- **Recall/Sensitivity**: $\frac{TP}{TP+FN}$ (ערכו בין 0 ל-1 כאשר 1 המקס')

- החלק של הפציינטים שהאלגוריתם קבע שהם positive case והלייבל שלהם אכן שייך ל-positive case (במקרה שלנו Yes, Maybe) מתוך כל הפציינטים שהלייבל שלהם שייך ל-positive case.
- המדד מעיד על כמה מהמקרים שהם positive cases הוא הצליח להחזיר שהם positive cases. כמה מקרים כאלה הוא "איתר" ביחס לכל ה positive cases הקיימים.

Results: ML algorithm

- חלק A של הטבלה מציג את תוצאות ביצועי המודל עבור stage 1.
- ב- stage 2 כפי שהוסבר, התבצע סיווג ל-Yes ול- Maybe.
- חלק B של הטבלה מציג את תוצאות הביצוע המשולב של שלבים 1 ו-2, כאשר הלייבל Yes של ה- gold standard הוא positive case וכל שאר הלייבלים negative case.

Table 2. Best Machine Learning Results on Test Set.

A: One Stage–Yes, Maybe vs. No, Unknown					
Training Set	**Test Set**	**Precision/PPV**	**Recall/Sensitivity**	**F1-Measure**	**Area under ROC Curve**
1st stage: Combined	BCH	0.726	0.852	0.784	0.533
1st stage: CCHMC	CCHMC	0.66	0.813	0.728	0.545
1st stage: Combined	Combined	0.864	0.836	0.785	0.583
B: Two Stage, Yes vs. No, Maybe, Unknown					
Training Set	**Test Set**	**Precision/PPV**	**Recall/Sensitivity**	**F1-Measure**	**Area under ROC Curve**
2nd stage: Combined	BCH	0.780	0.783	0.780	0.762
2nd stage: Combined	CCHMC	0.799	0.769	0.780	0.733
2nd stage: Combined	Combined	0.786	0.769	0.761	0.770

BCH–Boston Children's Hospital; CCHMC–Cincinnati Children's Hospital and Medical Center; PPV–positive predictive value; ROC–Receivers Operator Characteristic.

Results: ML algorithm

grid search שביצעו על מנת לבדוק אפשרויות שונות של features ו- cost parameters על ה- development set:

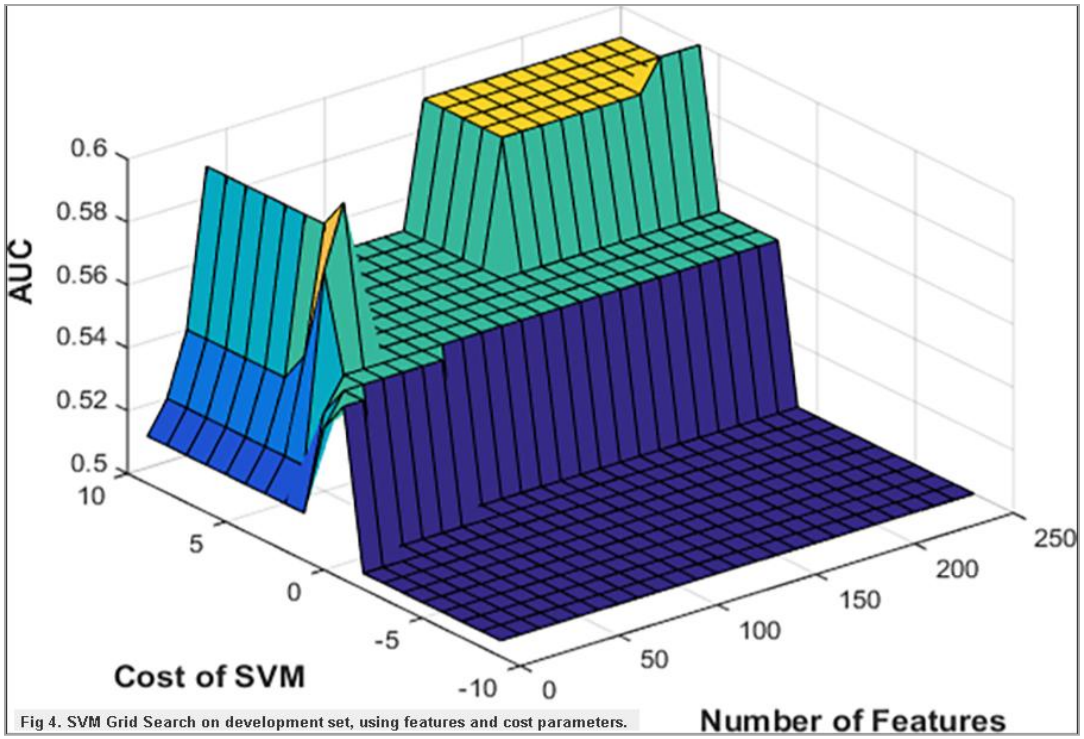
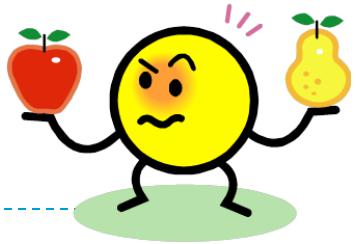


Fig 4. SVM Grid Search on development set, using features and cost parameters.

- גם מעל 250 features הערך נשאר אותו דבר ולא השתנה ולכן הם החליטו לקבוע upper-bound של 250 features.

Results: Comparison



1. תוצאות של rule-based algorithm:

Table 3. Rule Based Results

Evaluation Set	Precision/PPV	Recall/Sensitivity	F1-Measure	Area under ROC Curve
BCH	0.885	0.891 (14)	0.888	0.642
CCHMC	0.840	0.622 (48)	0.715	0.599
Combined	0.866	0.758 (62)	0.808	0.579
CHOP (independent validation on 50 patients)	0.849	0.737 (10)	0.788	0.659

2. תוצאות של machine learning algorithm:

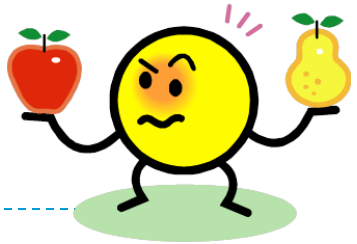
Table 2. Best Machine Learning Results on Test Set.

A: One Stage–Yes, Maybe vs. No, Unknown					
Training Set	**Test Set**	**Precision/PPV**	**Recall/Sensitivity**	**F1-Measure**	**Area under ROC Curve**
1st stage: Combined	BCH	0.726	0.852	0.784	0.533
1st stage: CCHMC	CCHMC	0.66	0.813	0.728	0.545
1st stage: Combined	Combined	0.864	0.836	0.785	0.583
B: Two Stage, Yes vs. No, Maybe, Unknown					
Training Set	**Test Set**	**Precision/PPV**	**Recall/Sensitivity**	**F1-Measure**	**Area under ROC Curve**
2nd stage: Combined	BCH	0.780	0.783	0.780	0.762
2nd stage: Combined	CCHMC	0.799	0.769	0.780	0.733
2nd stage: Combined	Combined	0.786	0.769	0.761	0.770

BCH–Boston Children's Hospital; CCHMC–Cincinnati Children's Hospital and Medical Center; PPV–positive predictive value; ROC–Receivers Operator Characteristic.

לאלגוריתם Rule-based ביצועים טובים יותר מבחינת Precision.

Baseline Comparison



- **<u>השוואת התוצאות שהתקבלו מול ה-baseline:</u>**
- <u>baseline</u>: אלגוריתם שמשתמש רק ב-ICD-9 Codes רלוונטים (כפי שדיברנו בהתחלה) על מנת לחלץ ASD cohort.

- <u>תוצאות האלגוריתם הזה:</u>

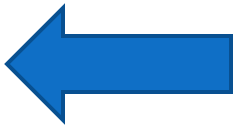
Table 4. Baseline Results (ICD-9 codes) on Test Set.

Test Set	Precision/PPV
BCH	0.273
CCHMC	0.645
Combined	0.460

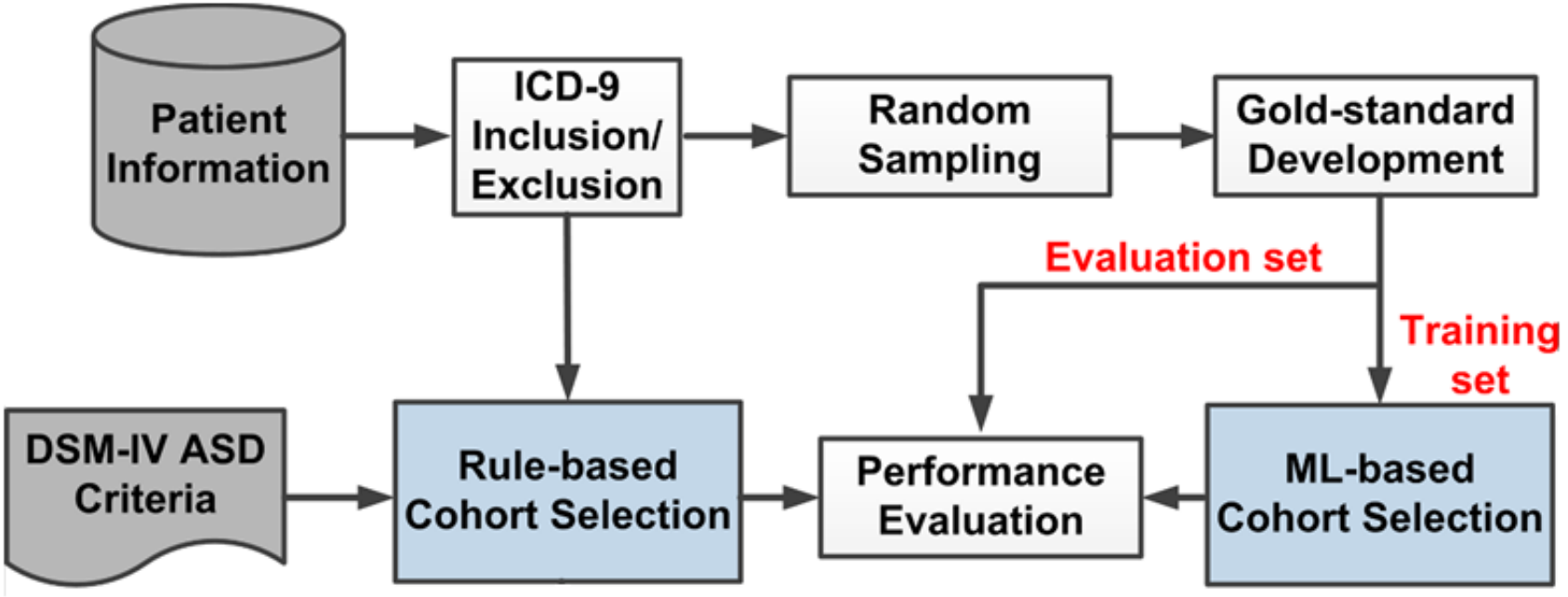
Baseline Comparison

Precision / Test Set	Baseline	Rule-Based
BCH	0.273	0.885
CCHMC	0.645	0.840
Combined	0.460	0.866

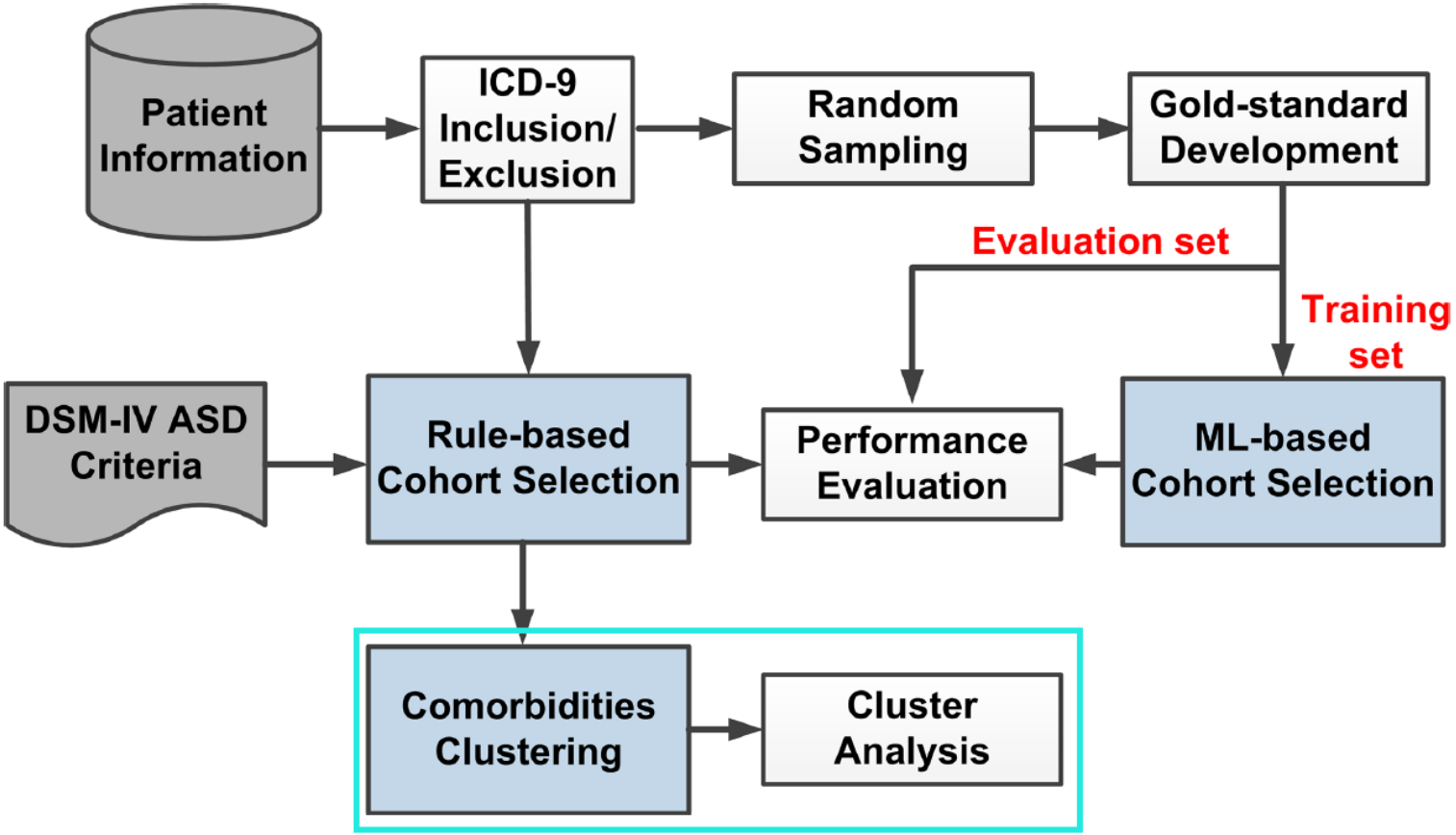
אם מסתמכים רק על ICD-9 Codes לחילוץ ASD cohort, מקבלים cohort הרבה פחות טהור, עם הרבה שגיאות של false positives.



So far…


Patient Information
ICD-9 Inclusion/ Exclusion
Random Sampling
Gold-standard Development
Evaluation set
Training set
DSM-IV ASD Criteria
Rule-based Cohort Selection
Performance Evaluation
ML-based Cohort Selection

What's next:


Patient Information
ICD-9 Inclusion/ Exclusion
Random Sampling
Gold-standard Development
Evaluation set
Training set
DSM-IV ASD Criteria
Rule-based Cohort Selection
Performance Evaluation
ML-based Cohort Selection
Comorbidities Clustering
Cluster Analysis

Comorbidities Clustering

- **Comorbidity=** תחלואה נלווית- הימצאות של מספר מחלות כרוניות בו זמנית.

- מחקר שנעשה בעבר* מצא שאנשים עם ASD, בהשוואה לאוכלוסיה הכללית, נוטים הרבה יותר להיות מאובחנים בתחלואות נילוות מסויימות כמו למשל Epilepsy, Schizophrena, Inflammatory Bowel Disease, Cranial Abnormalities.
- מחקר נוסף** השתמש ב-clustering כדי להראות את קיומם של קלסטרים בקרב אנשים שלוקים ב-ASD, כאשר כל אדם מאופיין על ידי קבוצת התחלואות הנילוות שלו.

- במחקר הזה ביצעו מספר שיפורים בהשוואה למחקר הקודם**:

1. קבוצת הלוקים בASD התקבלה על ידי שימוש באלגוריתם שהוצג במאמר, שמספק דגימה נקייה יותר של לוקים ב ASD.
2. ה- dataset הוא יותר גדול בצורה משמעותית – דאטה מכמה בתי חולים.
3. השוואה בין כמה סוגי clustering ושימוש ב- visuazliation algorithms.

* Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis – http://pediatrics.aappublications.org/content/133/1/e54

** The Co-Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033224

Comorbidities Clustering

- ה- data ההתחלתי: ה- initial cohort (BCH: 14,758 , CCHMC: 4,229) וגם dataset נוסף על מנת להעריך את הכלליות של התוצאות: VUMC: 6,482 (גם הוא לאחר סינון על פי ICD-9 Codes)

- **<u>שלבים:</u>**

1. הרצת האלגוריתם rule-based על כל הדאטה הזה.
   האלגוריתם קבע כי 87.7% מ-BCH (12,949), 70.7% מ-CCHMC (2,988),
   ו- 72.8% מ-VUMC (4,721) הם ASD Case.

2. המרת ICD-9 Codes של הפציינטים שנקבעו כ- ASD Case ל- **PheWAS categories** (=Phenotype Wide (Association Study .
   כעת כל פציינט מיוצג על ידי וקטור של PheWAS Codes.

3. **<u>k-means clustering:</u>** שימוש באלגוריתם k-means לביצוע clustering של הפציינטים (המיוצגים על ידי הpheWAS code vectors) <u>בכל בית חולים בנפרד.</u>
   [תזכורת: לחלק את הדגימות ל-k קלסטרים כך שכל דגימה משתייכת לקלסטר הכי קרוב אליה, וקלסטר מיוצג על ידי ממוצע הדגימות שבו]

Comorbidities Clustering

- בתהליך הקלסטרינג ניסו אפשרויות שונות של בין 2 עד 20 קלסטרים כדי למצוא את ה- clustering עם ה- silhouette coefficient הגבוה ביותר.
- **Silhouette value:** ערך בטווח [-1,1] שמציין עד כמה אובייקט מתאים ל-cluster אליו הוא שייך. ערך גבוה מעיד שהאובייקט מתאים בצורה טובה לקלסטר (cohesion/similarity) שהוא שייך ולא מתאים בצורה טובה לקלסטרים השכנים (separation). אם לרוב האובייקטים ב- clustering יש ערך גבוה אז ה-clustering נעשה בצורה טובה, ואם להרבה אובייקטים יש ערך נמוך או שלילי אז ב- clustering יש יותר מדי או מעט מדי קלסטרים.
- איך מחשבים את הערך הזה?

- לכל נקודה i בדאטה נחשב:
- **a(i)** = ממוצע המרחק של i לכל הנקודות בקלסטר שהוא שייך אליו.
- לכל קלסטר אחר d (i לא נמצא בו) נחשב את ממוצע המרחק של i לכל הנקודות בקלסטר, ונסמן ב- **b(i)** את המינימום על פני הממוצעים האלה שחישבנו.
- **Silhoutte value** של i:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

Comorbidities Clustering

- לאחר ביצוע clustering על ידי k-means, ביצעו clustering על ידי 2 אלגוריתמים נוספים: DBSCAN ו- Agglomerative Clustering והשוו בין התוצאות.
- שני האלגוריתמים נתנו הפרדה (sepeartion) פחות טובה בין הקלסטרים.

- נרחיב על האלגוריתם של Agglomerative Clustering.

Comorbidities Clustering

Agglomerative Clustering:

נסמן ב-n את מספר ה- data points .

1. בשלב הראשון כל דאטה פוינט x_i תהיה בקלסטר משלה $c_i = \{x_i\}$.
   נסמן ב- C את קבוצת כל הקלסטרים.
2. באופן איטרטיבי עד שכל הדאטה יהיה בקלסטר אחד:
   2.1. נמצא זוג קלסטרים c_i ו- c_j שהם הכי קרובים אחד לשני,כלומר: $D(c_i, c_j) = \min_{k,l} D(c_k, c_l)$
   על פי Distance Metric שנגדיר.
   2.2. נמזג את c_i ו- c_j לקלסטר חדש: c_{i+j} .
   2.3. נוציא מהקבוצה C את c_i ו- c_j ונכניס אליה את c_{i+j} .

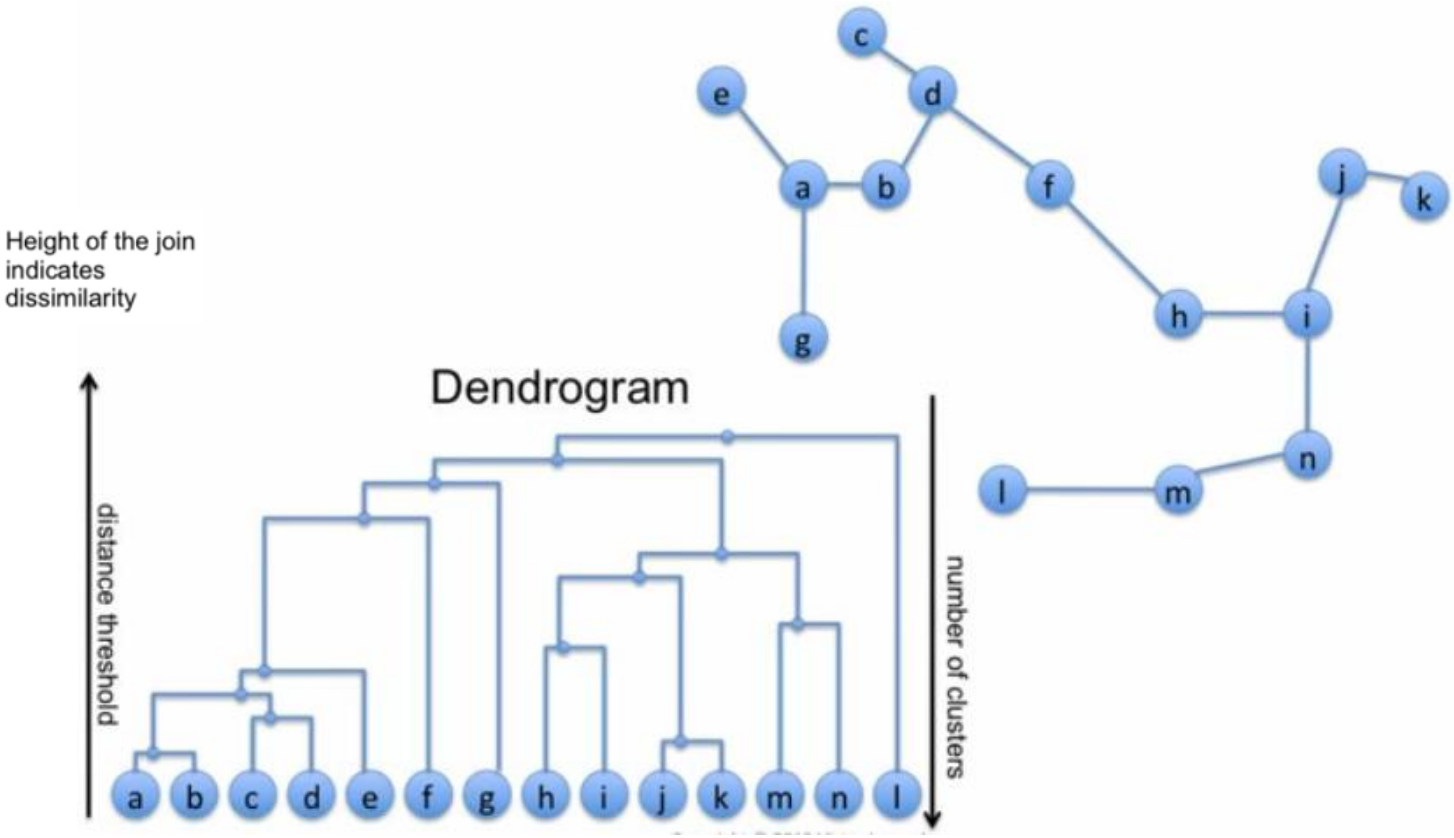
<u>הערה:</u>
יש כל מיני אפשרויות למדוד מרחק בין 2 קלסטרים. למשל:

- Average Linkage: ממוצע המרחקים של כל 2 נקודות משני הקלסטרים.
- Single Linkage: המרחק בין 2 הנקודות הקרובות ביותר זו לזו בין שני הקלסטרים.
- Complete Linkage: המרחק בין 2 הנקודות הרחוקות ביותר זו מזו בין שני הקלסטרים.

- התהליך נותן עץ היררכיה של קלסטרים.

Comorbidities Clustering

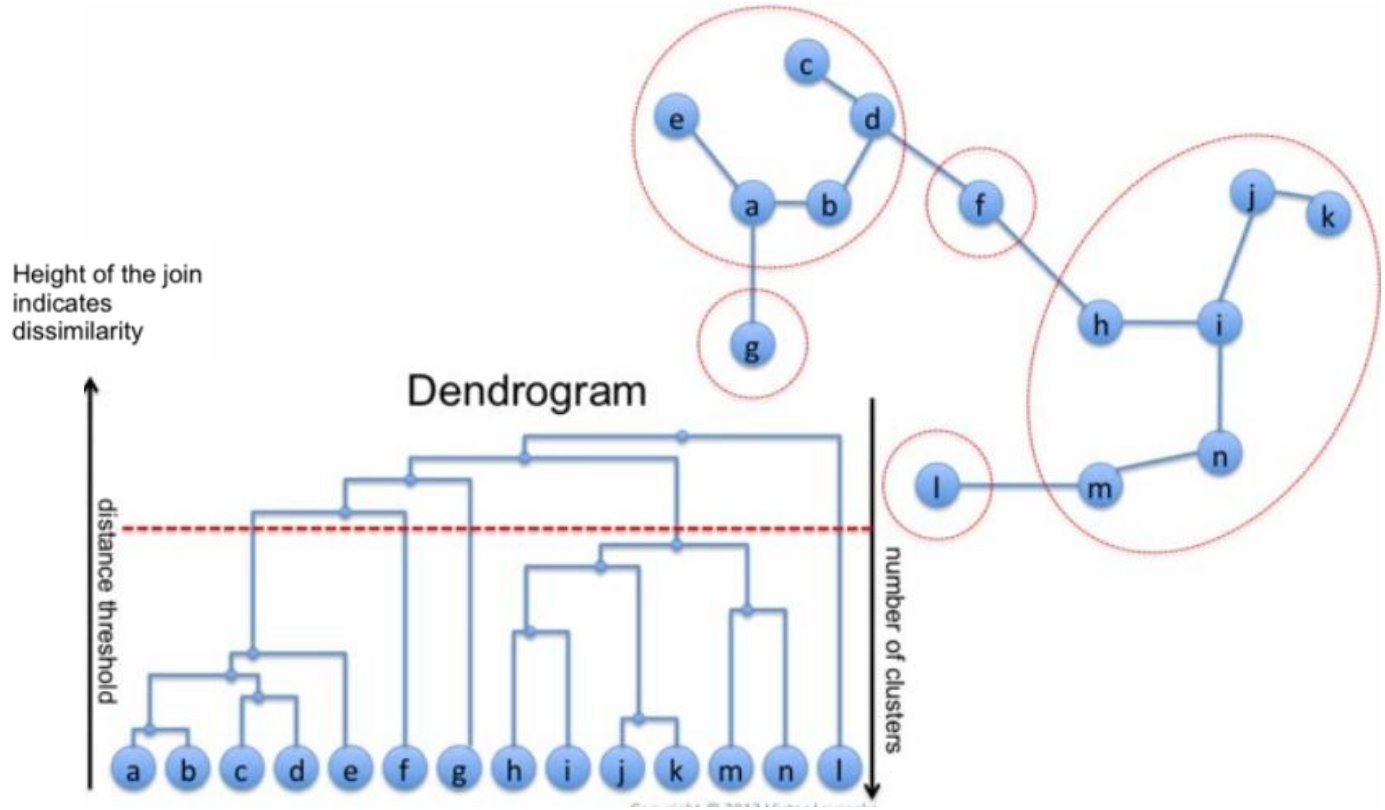
דוגמא: נשתמש ב single linkage כמרחק בין 2 קלסטרים.



Credit: https://www.youtube.com/watch?v=XJ3194AmH40

Comorbidities Clustering

קובעים distance threshold, חותכים את העץ בערך הזה ומקבלים את ה-clustering:



Credit: https://www.youtube.com/watch?v=XJ3194AmH40

Comorbidities Clustering

- **כעת נציג את ה-clustering הנבחר, שהתקבל על ידי k-means:**
- ביצעו clustering ל- dataset של כל בי"ח בנפרד. (BCH, CCHMC,VUMC)
- בכל בי"ח מצאו תבנית clustering דומה:

- בכל בי"ח יש 3 או 4 קלסטרים קטנים (5-20% מתוך הset כולו) וקלסטר אחד גדול.
- בכל אחד מהקלסטרים הקטנים יש תחלואה נלווית אחת ששולטת (הכי שכיחה) והיא מבין הקטגוריות הבאות של תחלואות נלוות:

1) psychiatric problems including anxiety disorder, hyperkinetic syndrome, obsessive compulsive disorder, and depression;

2) developmental disorders including dyslexia, lack of coordination, and various disorders of the ear, skin and other bodily systems;

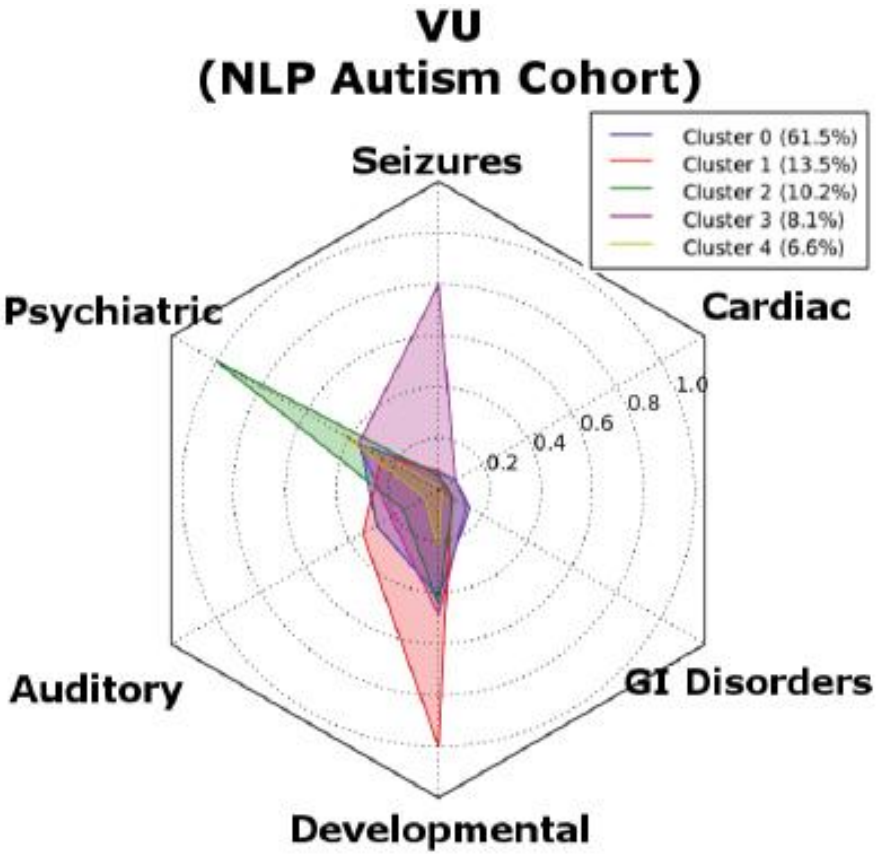
3) epilepsy and recurrent seizure.

- הקלסטר הגדול יותר שמהווה כ 60-80% מהset כולו, לא היה מאופיין בשכיחות גבוהה של איזושהיא תחלואה נלווית או קטגוריה של תחלואה נלווית.

Comorbidities Clustering

- כדי להשוות בין הקלסטרים שהתקבלו לתוצאות של מחקרים קודמים, קיבצו את PheWAS Codes על פי קבוצות comorbidities הבאות:

Seizures, Psychiatric, Auditory, Developmental, GI Disorders, Cardiac Disorders



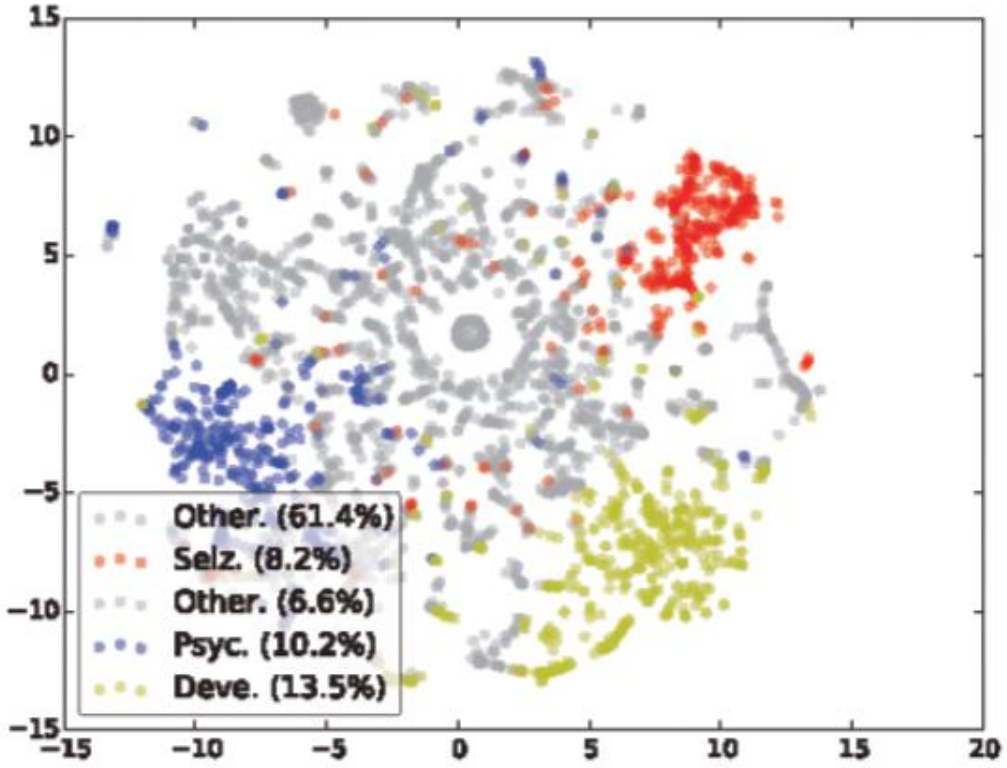
Comorbidities Clustering

- הוקטורים של הפציינטים היו ממימד גבוה (100-200 pheWAS codes).
- השתמשו באלגוריתם להקטנת מימד שנקרא t-distributed Stochastic Neighbor Embedding כדי להציג בצורה ויזואלית את וקטורי PheWAS Codes ב-2 מימדים, וכך לראות בצורה ויזואלית את ההפרדה (seperation) בין הקלסטרים.

- **<u>האלגוריתם מורכב מ-3 שלבים עיקריים:</u>**

1. האלגוריתם משתמש במרחקים שבין כל 2 אובייקטים x_i , x_j ממימד גבוה, ומגדיר הסתברות שמייצגת similarity עבור כל זוג אובייקטים x_i , x_j – אם x_i קרוב ל-x_j (similar) אז יש הסתברות גבוהה לכך שהוא יבחר אותו בתור neighbor ואם הוא רחוק (dissimilar) אז הסתברות מאוד נמוכה.
2. האלגוריתם מגדיר הסתברות דומה בין כל 2 אובייקטים y_i , y_j ממימד נמוך. מתקיים שאם y_i ו- y_j ממדלות נכון את ה-similarity שבין האובייקטים x_i ו- x_j ממימד גבוה אז ההסתברויות של 1 ו-2 יהיו שוות.
1. לכן מיקום הנקודות y_i במפה נקבע על ידי מינימיזציה של Kullback–Leibler divergence (מדד לשוני בין 2 התפלגויות) בין שתי ההתפלגויות.

Comorbidities Clustering



c) VUMC

הקטנת המימד על ידי אלגוריתם t-SNE על PheWAS codes.

-השם של כל קלסטר הוא על פי ה-comorbidity category שבשכיחות הכי גבוהה בו.

- שמות מופיעים יותר מפעם אחת כאשר יש יותר מקלסטר אחד ש'נשלט' על ידי אותה קטגוריה.

(Deve.–Developmental Disorders, Seiz.–Seizure Disorders, Psych.—Psychological Disorders).

Discussion

- אלגוריתם rule-based:
  - על ידי התייחסות ללייבל Maybe כ- positive case הם שמרו על נאמנות לקריטריון DSM-IV מכיוון שעבור מספר גבוה של מטופלים שהלייבל שלהם הוא Maybe האלגוריתם קבע שהם PDD-NOS Case.
  - מערכת לעיבוד שפה טבעית Apache cTAKES היא לא חפה מטעויות, והיא סובלת מאי דיוקים מסיבות שונות, למשל עושר לשוני – יש דרכים רבות להביע התנהגות של לוקה ב-ASD, סיבה נוספת היא ידע שלא מקודד על ידי המערכת, למשל אם ברשומה רפואית רשום שילד הולך לבי"ס מיוחד.
- באלגוריתם של ML, התוצאות עם הביצועים הטובים ביותר התקבלו מ-Combined training set, וזה מעיד על פוטנציאל שקיבוץ של data ממספר מוסדות יכול לשפר את כוח החיזוי.

Final Words

- ראינו שחילוץ קוהורט של ASD ע"פ ICD-9 Codes הוא לא אמין.
- המחקר מספק אלגוריתם אוטומטי לחילוץ קוהורט של ASD באמינות גבוהה, כלי עבור מחקרים מבוססי EHR גדולים עתידיים.
- המחקר מאשר את מה שמחקרים קודמים עם קבוצות דגימה יותר קטנות מצאו - קיימים מספר קלסטרים שונים בקרב לוקי ASD, שכל אחד מאופיין ע"י comorbidity אחר.
- האלגוריתם מקדם את היכולת לחקור אפיקי טיפול שונים שמבוססים על comorbidity clusters וכן גם סיכונים ומאפיינים מולקולרים וגנטיים.

The End