אוניברסיטת בר-אילן

הערכת סיכון לפקקת ורידים תסחיפית ודימום חמור תוך חקירת שקלול התמורות ביניהם בעזרת אלגוריתמי שערוך הסתברות

ואדים ליטבינוב

עבודה זו מוגשת כחלק מהדרישות לשם קבלת תואר מוסמך

בפקולטה למדעי החיים של אוניברסיטת בר-אילן

רמת-גן תשפ"ב

אוניברסיטת בר-אילן

הערכת סיכון לפקקת ורידים תסחיפית ודימום חמור תוך חקירת שקלול התמורות ביניהם בעזרת אלגוריתמי שערוך הסתברות

ואדים ליטבינוב

עבודה זו מוגשת כחלק מהדרישות לשם קבלת תואר מוסמך

בפקולטה למדעי החיים של אוניברסיטת בר-אילן

רמת-גן תשפ"ב

עבודה זו נעשתה בהדרכתו של

פרופ' רון אונגר

מן הפקולטה למדעי החיים של אוניברסיטת בר-אילן.

תודות

ראשית, ארצה להודות ולהקדיש עבודה זו למשפחתי. אמי לריסה, סבי בנימין, אחותי אלה וסבתי המנוחה מרה, על התמיכה הגדולה בהשכלה אקדמית ובמחקר.

בנוסף, אני מודה מאוד לפרופ' רון אונגר וד"ר שחף שיבר על התמיכה, ההכוונה והעזרה במהלך תקופה מאתגרת. אודה גם לחברי למעבדה על העצות והדיונים המפרים: ירון גפן, אורית אדטו, עדן מימון, צילה רייך, מיקה עולמי, ד"ר תרצה דוניגר, וגם לד"ר אפי כהן על העזרה בדיוקים המתמטיים.

ולבסוף לשירן בן-מאיר, אשר מלווה אותי מתחילת המסלול ועד סופו.

תוכן עניינים

תקציר

פקקת ורידים תסחיפית (Venous thromboembolism – VTE) הינה גורם מוות משמעותי ברחבי העולם, כאשר רוב המקרים קשורים לשהייה בבית החולים עצמו. ה-VTE מתרחש כאשר ישנו קריש דם בווריד‌ים העמוקים (Deep-vein thrombosis – DVT), לרוב ברגליים או באגן, שמתנתק ממיקומו ההתחלתי ונסחף בזרם הדם.

הסיכון ל-VTE תלוי במשתנים שונים ידועים וביניהם גיל, סרטן פעיל ומצבים רפואיים מסוימים. במהלך השנים נבנה מדד בשם Padua Prediction Score אשר מבוסס על סכימה ממושקלת של משתנים ואמור לעזור לרופאים להעריך סיכון זה, ובעזרת הערכת סיכון זו הרופאים יחליטו האם לטפל במאושפז במדללי דם או לא.

מאגרי הנתונים אשר עבדנו עליהם הורכבו ממאושפזים במרכז הרפואי רבין והתקבלו על ידי ד"ר שחף שיבר מהמחלקה הפנימית.

אחת ממטרותינו הייתה לכייל את משקלם של משתני ה-Padua ולהוסיף משתנים חדשים על מנת לשערך את הסיכון בצורה יותר מדויקת. לאחר יצירת המודל החדש, נבחנו ביצועיו לעומת ביצועי ה-Padua, ונערך מבחן סטטיסטי על מנת לוודא שההבדל בין הביצועים מובהק.

כיוון שיש חשש משימוש מוגזם במדללי דם אשר יביאו לדימומים מסכני חיים, רצינו ליצור כלי אשר ישערך גם את הסיכון לכך. לאחר יצירת המודל הנוסף, הוגדר מדד ה-ACU (Anti-Coagulant Utility) אשר מביא לידי ביטוי את שני צידי המטבע בעת ההחלטה האם לטפל במדללי דם או לא. בכדי לתת מוטיבציה לסוג מסוים של מודלים מתחום למידת המכונה נעשה ניסיון להוכיח תאורטית שמודלים בעלי קליברציה טובה יתרמו ל-ACU במשימת חלוקת המאושפזים לכאלה שתהיה להם תועלת ממדללי דם, וכאלה שיכולים להינזק מהם בסבירות גבוהה. על מנת להוכיח שה-ACU הצליח במשימתו, נעשה שערוך תוחלות לאוכלוסיות המאושפזים המדממים, המאושפזים אשר עברו אירוע VTE ומאושפזים בריאים בעזרת ממוצעי המדגמים ממאגר הנתונים ובעזרת יצירת רווח סמך מסביב לממוצעים אלו.

מבוא

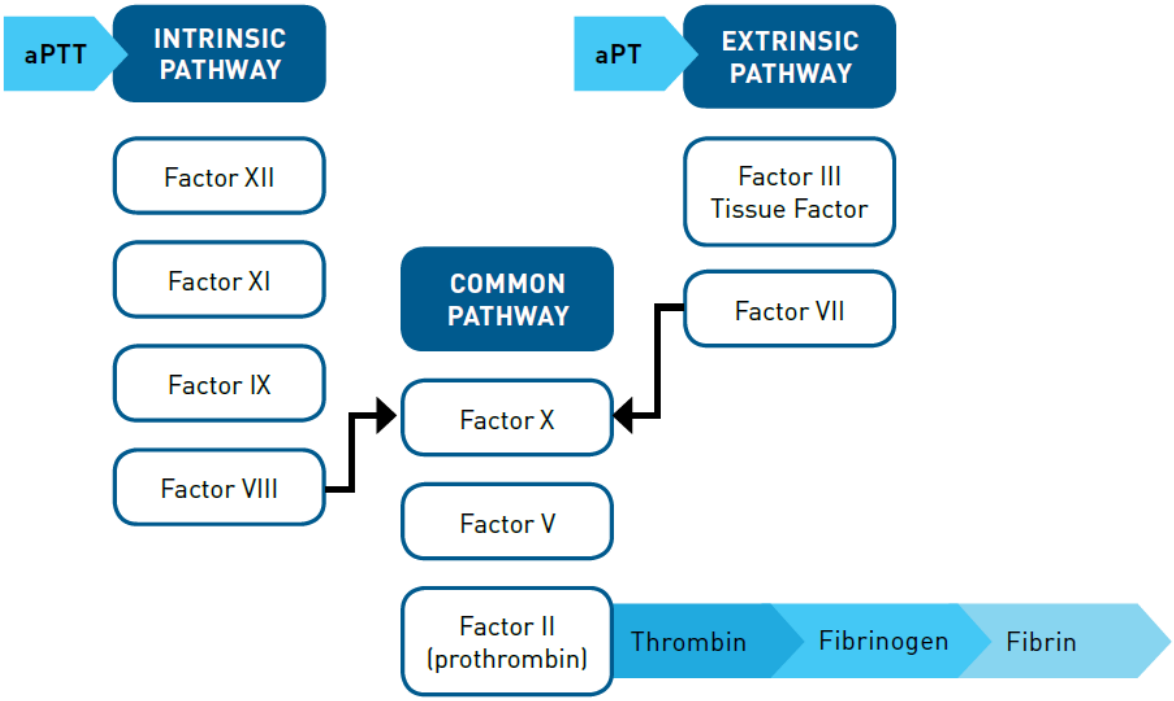
רקע ביולוגי

קרישי דם יכולים להשפיע על מחזור הדם בורידים. כאשר מתייחסים ל-VTE לעתים רבות נכלל בתוכו גם DVT (פקקת ורידים עמוקים - Deep Vein Thrombosis) וגם PE (תסחיף ריאתי - Pulmonary Embolism) שהינו מקרה פרטי של סחף קריש דם לאיברים אחרים לאחר התנתקות שלו מהווריד המקורי. במקרה של PE, קריש הדם מגיע למערכת הדם הריאתית ועלול לגרום להפרעות באספקת החמצן לרקמות. קרישי דם יכולים בנוסף להשפיע ולהפריע למחזור הדם בזרועות, מוח, כליות, כבד, מעי ועוד. אירוע VTE מושפע ממשתנים רבים הכוללים גם מצבים רפואיים שמהווים קרקע פוריה לקרישיות, ביניהם מצבים המועברים בתורשה או נרכשים במהלך החיים[1-5].

ישנם רכיבים רבים בגוף אשר מוגדרים כמעודדי קרישה (רכיב פרותרומבוטי) ונוגדי קרישה (אנטיתרומבוטי), אשר במצב רגיל האיזון ביניהם מגביל את הקרישיות. במצב של הפרת האיזון אירועי קרישיות יכולים להתרחש. בנוסף, קריש יכול להיווצר כתוצאה מאי סדירות של דופן כלי הדם[6].

הומאוסטזיס של הדם מבוקר על ידי ממשקים של אנזימים וחלבוני קרישה. רוב גורמי (פקטורי) הקרישה הם קודמנים (פרקורסורים) של אנזימים שאינם פעילים במצב נורמלי. ישנם 12 פקטורי קרישיות ידועים - ביניהם, ניתן למנות את הפקטורים: fibrinogen (factor I), prothrombin (factor II), tissue factor (TF; factor III) ו-calcium (factor IV)[6].

שתי התפיסות העיקריות של יצירת קריש הם coagulation cascade והמודל התאי. המודל המסורתי הוא הראשון, אך ראיות עכשוויות מהוות תימוכין גם למודל התאי.



איור 1. תרשים של coagulation cascade. aPT – activated prothrombin time, aPTT – activated partial thromboplastin time[6].

המסלול מחולק לפנימי וחיצוני, אשר מתכנסים למסלול משותף בהפעלת פקטור X. פקטור X משופעל ופקטור V מייצרים את הקומפלקס אשר מייצר פרותרומבין. פרותרומבין מתחלק לשני חלבונים קטנים יותר והופך לתרומבין, שבתורו מוביל ליצירת פיברינוגן שהוא הפרקורסור של פיברין, אשר מקדם הצטברות טסיות דם ומייצב את הקריש[6].

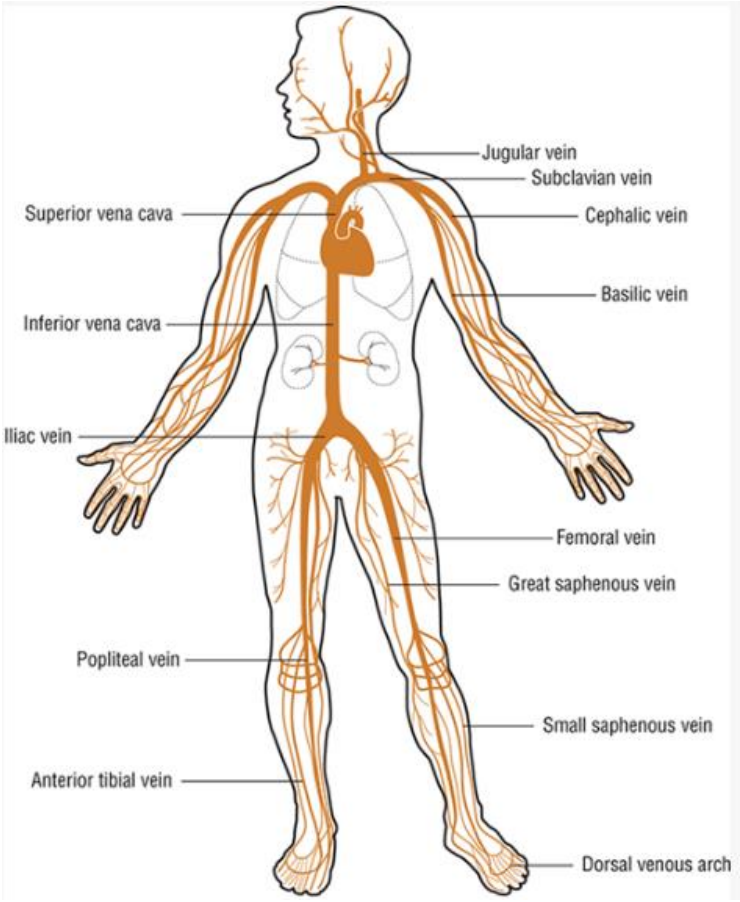
המודל התאי מערב שלבים של התחלה, הגברה, הרחבה וייצוב. ההתחלה מתרחשת עם ביטוי של TF בכלי דם שניזוקו. ביטוי זה מקדם ממשק של פקטורים VII ו-IX עם פקטור X. שפעול של פקטור X מוביל לייצור פרותרומבין שמייצר כמות קטנה של תרומבין. כמות קטנה זו מגבירה את האות מעודד הקרישה ע"י שפעול טסיות דם וקו-פקטורים של קרישיות. השלב הבא הוא ההרחבה של יצירת תרומבין על ידי הצטברות של פקטורים Va ו-VIIIa על דופן טסיות הדם. לבסוף, פקטור X משופעל תומך בפרץ יצירת התרומבין, שעוזר לייצוב הקריש[6, 7].

ניתן לחלק אירועי קרישיות לשניים: לאירועים בגלל גירוי (provoked) ולאירועים אשר אינם בשל גירוי (unprovoked). אירוע ללא גירוי יכול להיחשב לאחד שהתרחש בשל תרומבופיליה מורשת, גיל מבוגר או

מין זכר[8]. אירועים כתוצאה מגירוי כוללים VTE שהתרחש בשל גורם סיכון. בנוסף, אירועים כתוצאה מגירוי יכולים להתרחש כתוצאה מסיבות חולפות או מתמשכות. VTE כתוצאה מסיבות חולפות אמור להתרפא לאחר טיפול הולם[8, 9].

סרטן פעיל, אי ספיקת לב (CHF), השמנת יתר חולנית ודליות הן דוגמאות לגורמי סיכון עיקשים וממושכים[9]. לעומת זאת ריתוק למיטה למשך יותר מ-3 ימים, חוסר תנועתיות, טיפול באסטרוגן, חבלה או ניתוח, היריון, ופציעת רגל או אגן שמקושרים עם חוסר תנועתיות הינם גורמי סיכון זמניים[10].

התסמינים הטיפוסיים של DVT הם כאב חד צדדי ברגל, אדמומיות ונפיחות. PE מאופיין בכאב בחזה, קוצר נשימה, טכיפנאה וטכיקרדיה. כיוון שתסמינים אלה אינן ספציפיים ל-DVT ו-PE, נדרשת בדיקה אובייקטיבית על מנת לקשר אותם לאבחנות אלו. D-dimer הינו תוצר לוואי של דגרדציה של פיברין, וערכיו גבוהים בחולים עם VTE. בנוסף, חשוב לציין כי משתמשים בונוגרפיה ובאנגיוגרפיה של הריאות שהינן הבדיקות המדויקות ביותר לאבחון VTE[11].



איור 2. מערכת הורידים בגוף בה קריש דם יכול להסתם[11].

נוגדי קרישה (הנקראים גם נוגדי פקקת או מדללי דם) הינם דרך יעילה למנוע ולטפל במקרים של VTE ובהורדת הסיכון לשבץ בחולים עם פרפור עליות (arterial fibrillation). אולם, חסימת פקטורים במסלול הקרישה טומנת בחובה סיכון לדימומים חמורים. כיוון שלעתים חולים אשר נוטלים תרופות אלה נדרשים לניתוח או פעולה פולשנית אחרת, נושא זה מקבל משנה תוקף. באופן מסורתי, אנטגוניסטים של ויטמין K (VKA) אשר באופן עקיף משביתים את פקטורי הקרישה II, VII, IX ו-X היו עיקר מדללי הדם במשך תקופה ארוכה. בין השאר בגלל בעיות הדימום, לאחרונה ה-FDA אישר תרופה אשר הופכת את הפעילות של VKA, ושמה 4-Factor Prothrombin Complex Concentrate (4F-PCC). בעשור האחרון חלה עליה במדללי דם שאינם מבוססים על ויטמין K (NOAC), שיעדם הינו פקטורים מסוימים (פקטור IIa ו-Xa) בצורה ישירה. במהירות פותחו תרופות כדי להפוך גם את פעולתם, מה שמעיד על כך שמדללי דם יכולים להפוך לחרב פיפיות ואין להשתמש בהם באופן מופרז[12]. VTE נחשב כסיבת מוות ברת מניעה

במאושפזים במחלקות, וישנן הערכות כי כ-75% ממקרי המוות הקשורים ב-VTE נבעו מ-VTE המקושר לשהייה בבית החולים. ישנה המלצה להשתמש במדללי דם ככלי מניעה ל-VTE, בהתבסס על גורמי סיכון שונים[13, 14].

בעוד שבמאושפזים מנותחים (כירורגיים) הראיות מצביעות כל כך שמדללי דם מובילים לירידה בתמותה[15], המצב אינו בהכרח זהה במאושפזים אחרים. במחקר שבוצע נדמה כי אמנם ישנה ירידה מובהקת סטטיסטית ב-DVT, אך הנ"ל אינו מלווה בירידה בתמותה הכללית. יתרה מכך, מחקרים הראו כי מתוך 1000 חולים אשר טופלו במדלל דם מסוג הפרין בעל משקל מולקולרי נמוך, אמנם 3 מקרים של PE נמנעו, אך עם זאת 9 מקרים של דימום התרחשו, כאשר 4 ממקרים אלו נחשבו לדימום חמור[16, 17].

על מנת לפשר ביחס העלות-תועלת של מדללי דם יש צורך לבחור בקפידה את מושא הטיפול. לשם כך, מספר מדדים הוגדרו - Padua score[18], IMPEOVE VTE risk score[19], ו-GENEVA score[20]. מדד ה-Padua הוא שילוב של מודל קיים[21], יחד עם מסקנות והנחיות שפורסמו לגבי גורמי סיכון נוספים[22, 23].

מדד ה-Padua מחושב על ידי סכימה ממושקלת של משתנים בינאריים שמהווים גורמי סיכון ומבוצע על ידי הרופא בעת קבלה למחלקה. במדד הנ"ל ניקוד של 3 ומטה מהווה סיכון נמוך ל-VTE, ו-4 ומעלה מהווה סיכון גבוה. המשתנים ומשקלותיהם:

1. סרטן פעיל – 3 נקודות
2. VTE קודם – 3 נקודות
3. תנועתיות נמוכה – 3 נקודות
4. מחלת קרישיות ידועה – 3 נקודות
5. חבלה או ניתוח שהתרחש לאחרונה (חודש או פחות מכך) – 2 נקודות
6. גיל מבוגר (70 ומעלה) – נקודה
7. כשל לבבי או נשימתי – נקודה
8. אוטם שריר הלב או שבץ מוחי – נקודה

9. דלקת חמורה או מחלת פרקים – נקודה
10. השמנת יתר חולנית (BMI מעל 30) – נקודה
11. בעת טיפול הורמונלי – נקודה

מדד ה-Padua הינו המדד אשר משרד הבריאות הישראלי הנחה להשתמש בו בכל בתי החולים, לגבי כל מאושפז במחלקות הפנימיות בהקשר של הערכת סיכון ל-VTE. עם זאת, למרות הנחיה זו, המדד אינו חף מבעיות. גודל המדגם אשר המדד נוצר בהתבסס עליו היה קטן באופן יחסי וכלל 1,180 מאושפזים. כמו כן, רוב אירועי הקרישיות שדווחו במחקר (30/37) התרחשו במאושפזים עם סרטן פעיל ו/או אירוע VTE קודם, דבר המעיד על כך שהמדד אינו מאוזן[24].

רקע חישובי

אלגוריתמי למידת מכונה מתבססים על שיטות סטטיסטיות ומתמטיות ומנסים לקבוע כלל החלטה (במקרה של סיווג) על פי דוגמאות מקבוצת האימון שניתנו להם. זאת, לעומת אלגוריתמי בינה מלאכותית קלאסיים, בהם דרך הפעולה צריכה להיות מוגדרת היטב לכל קלט שהוא. תכליתם של אלגוריתמי למידת מכונה הינה בעיקר לזהות דפוסים וללמוד את ההתפלגויות של הדוגמאות השונות, גם אם דפוסים אלו אינם ידועים מראש לתוכניתן. מדובר בלמידה מפוקחת (Supervised learning) - המערכת משתמשת בדוגמאות מתויגות בקבוצת הלמידה (Training set) ומוצאת את הדפוסים שעובדים טוב בקבוצה זו, ולאחר מכן בוחנים אותה בקבוצת המבחן (Test set) כאשר שם יש דוגמאות שלגביהן המערכת אינה חשופה לתיוג.

בעיות אחרות אשר ניתנות לפתרון בעזרת למידת מכונה:

- רגרסיה – חיזוי משתנה כמותי.
- חלוקה לצברים – גיבוש קבוצות פריטים כאשר אין את התיוגים בקבוצת האימון.
- מערכות זיהוי דיבור – זיהוי הברות ומילים.
- מערכות המלצה – מערכות מוניטין במנסות לחזות טעם של משתמשים.

נהוג לייצג את וקטור המאפיינים (Features) כ-x ואת התיוג של הדוגמה כ-y, כאשר המטרה הינה למצוא פונקציה (הידועה גם בתור היפותזה, מסווג ומודל) שתקיים $f_w(x) = \hat{y} = y$. על מנת למצוא את f מודדים את העלות (Loss) בין החיזוי של f לתיוג האמיתי. המטרה היא למזער את התוחלת של העלות בשימוש ב-f באוכלוסייה. תוחלת העלות נקראת Risk[25]:

$$\mathfrak{R}(f_w) = \int \ell(y, f_w(x)) d\mathbb{P}(x, y)$$

עם זאת, כיוון שבסטטיסטיקה אין לנו גישה להתפלגות האוכלוסיה כולה $\mathbb{P}(x, y)$ אינו ידוע ועלינו להסיק מידע מהמדגם שבידינו בלבד. כעת, ה-Risk שהוא תוחלת מוחלף בפונקציה אחרת שנרצה למזער אותה והיא הממוצע של העלות (הביצועים של המודל) על המדגם. עקרון זה נקרא Empirical Risk Minimization (ERM)[25]:

$$\bar{\ell}(S_n, f_w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_w(x_i))$$

אחד הסימנים ללמידה מוצלחת הינו פער הכללה קטן, כאשר המודל לא שינן את הדוגמאות אלא למד דפוסים שיעזרו לו בקבוצת המבחן שמייצגת את העולם האמיתי. פורמלית, בהשתמש באי שוויון הופדינג[26]:

$$\mathbb{P}(\left|\mathfrak{R}(f_w) - \bar{\ell}(S_n, f_w)\right| \geq \epsilon) \leq \delta = 2e^{-2n\epsilon^2}$$

כאשר מטרתנו היא שההסתברות שהפער בין הביצועים של המודל f עם משקלות w בזמן האימון על קבוצת n הזוגות הסדורים של דוגמאות ותיוגן S_n, לביצועים שלנו על האוכלוסייה, יהיה גדול או שווה ל-ϵ (שנרצה שייצג מספר קטן וחיובי), תהיה חסומה מלעיל ב-δ (גם כן בשאיפה לקטן וחיובי). ניתן לראות כי ככל שגודל המדגם שמיוצג על ידי n גדל, כך החסם נהיה הדוק יותר ורמת הביטחון בביצועי המודל - גדלה.

בשנת 1987 פורסם מאמר שטען שאמנם ניתן יהיה להעזר בבינה מלאכותית ברפואה לקבלת החלטות לוגיות ומוגדרות היטב כגון האם לתת תרופה נגד לחץ דם כפונקציה של מדדים מסוימים, אך רוב הבעיות ברפואה הן מורכבות מדי מכדי ללכוד את כל ההיבטים ולאגד אותם כקלט מוגדר היטב לאלגוריתם[27].

בשנים הרבות שעברו מאז התברר שלמרות התחזית הפסימית הזו, למידת מכונה יכולה למלא תפקיד מאד חשוב ברפואה המודרנית.

היתרון העיקרי של אלגוריתמים מסוג למידת מכונה על פני רופאים בשר ודם הינו היכולת להבחין ביחסים רבים ומורכבים בין כל המאפיינים השונים של חולה לבין עצמם, או בין המאפיינים לתוצאה או האבחנה הרפואית. לעומת זאת, חיסרון בולט של האלגוריתם הינו הצורך שלו בכמויות גדולות של דוגמאות על מנת ללמוד ולהכליל את שמוצג בפניו בתמונה (לדוגמה), כאשר מנגד - לתינוק אנושי מספיקות פעמים בודדות להיחשף ליצור או חפץ בכדי ללמוד מה הוא, וכיוצא מכך - לזהות פריט נוסף מהסוג שלו בפעם הבאה.

במסגרת הדוגמאות המעשיות, ניתן להתבונן על רשתות נוירונים עמוקות שמסווגות נגעי עור סרטניים ברמת הצלחה של רופא ואף למעלה מכך[28], הערכת סיכון לדלקת הנגרמת מחיידק הקלוסטרידיואידס דיפיצילה, הידוע בתור מחולל מחלה המקושר לבתי חולים[29], חיזוי תמותה במהלך שנה לאחר תסמונת כלילית חריפה (Acute Coronary Syndrome)[30], ניבוי לידה נרתיקית לאחר ניתוח קיסרי[31] ועוד.

מאגרי הנתונים

ישנם שני מאגרים אשר עליהם בוצע המחקר. שני המאגרים התקבלו באדיבותה של ד"ר שחף שיבר מהמרכז הרפואי רבין.

קריטריוני הכללה במאגר:

1. מטופלים אשר התקבלו למחלקה הפנימית או הגריאטרית.
2. האשפוז נמשך יותר מ-48 שעות.

קריטריוני אי-הכללה במאגר:

1. שימוש כרוני במדללי דם לשם מטרות אחרות (כגון פרפור פרוזדורים, מסתם לב מלאכותי וכו').
2. אינדיקציה חדשה לתחילת טיפול במדללי דם בעת קבלה.
3. מטופלים מנותחים במחלקה.
4. מטופלים אשר חוו לאחרונה שבר או ניתוח.

5. מטופלים בעלי פילטר וריד נבוב תחתון (ICV Filter).
6. התוויות נגד מוחלטות לשימוש במדללי דם (דימום פעיל, ספירת טסיות דם פחות מ-50,000 למיקרוליטר, חבלה חמורה, היסטוריה של דימום תוך-גולגולתי).

חשוב לציין כי במידה ומטופל התאשפז יותר מפעם אחת, אחד האשפוזים נבחר באקראי והוכלל במאגר.

המאגר הראשון כלל 18,890 אשפוזים בין השנים 2012-2017. ב-142 מקרים שונים הייתה אבחנה של VTE וב-3,102 אבחנה של דימום חמור. המאגר השני כלל 61,670 אשפוזים בין השנים 2018-2021. ב-663 מקרים שונים הייתה אבחנה של VTE וב-1,130 אבחנה של דימום חמור. ניתן לשים לב שתדירות הדימומים קטנה פי 3 לערך בין המאגרים, למרות שהמאגר עצמו גדל בסדר גודל דומה. הסבר אפשרי הוא שמתן רב יותר של תרופות נוגדות חומצה למאושפזים גרמו לתופעה זו כיוון שהן עוזרות נגד דימומים במערכת העיכול.

מטרות המחקר

כפי שהוזכר במבוא, טיפול מופרז במדללי דם במאושפזים עלול להוביל לדימומים מסכני חיים - ולשם כך ישנו שימוש במודל Padua על מנת לברור מי זקוק יותר לטיפול ומי פחות. כיוון שמודל זה אינו מושלם, אנו נתמקד במטרות הבאות:

1. להשתמש בכל 11 המשתנים של Padua, אך לבדוק האם מישקולים שונים של המשתנים יכולים להביא לביצועים טובים יותר מהמודל במקורי.
2. בפרט, האם המשתנים של סרטן פעיל ו-VTE קודם מספיקים לבדם להביא לביצועים המניחים את הדעת בניבוי VTE.
3. הוספת משתנים חדשים אשר לא נכללו במודל ה-Padua המקורי, וניסיון לשפר את הניבוי באופן מובהק סטטיסטית.

כמו כן, נרצה ליצור מדד חדש אשר ישקלל בתוכו גם את הסיכון ל-VTE וגם את הסיכון לדימום חמור. לשם כך נצטרך להשתמש באלגוריתם למידת מכונה שהוא הסתברותי מטבעו - בכדי ליצור את שני המודלים, וננסה לתת צידוק תאורטי לבחירת סוג זה של אלגוריתמים. בהמשך לכך, על מנת להראות שהמדד המשולב הצליח להבחין בין קבוצות שצפויה להן תועלת ממדללי דם לעומת קבוצות שצפוי נזק, נבצע ניתוחים סטטיסטיים של המדד המשולב בין קבוצות המטופלים השונות.

שיטות

כלים יישומיים

Python – שפת תכנות דינמית נפוצה. לשפה זו ישנן חבילות רבות אשר מאפשרות טיפול יעיל ונוח במטריצות, מבני נתונים, אלגוריתמים שונים ומבחנים סטטיסטיים. החבילות אשר נעשה בהן שימוש במחקר הן Pandas, NumPy, Sklearn, Parfit, SciPy, Matplotlib, Seaborn, Statsmodels.

מדד תועלת ממדללי דם (ACU)

בכדי לגבש המלצה מושכלת האם לטפל במדללי דם או לא, יש צורך לקחת בחשבון גם את הסיכון לדימומים. לשם כך, נגדיר את מדד ה-ACU (Anti-Coagulant Utility) שמורכב משני מודלים. נסמן ב-$\widehat{\mathbb{P}}$ הסתברות משוערכת, ב-v ו-b נסמן VTE ודימום חמור בהתאמה, ו-x_i יהיה וקטור של t משתנים (features) של מאושפז i.

$$\widehat{\mathbb{P}}(v|x_i) - \widehat{\mathbb{P}}(b|x_i) \stackrel{\text{def}}{=} ACU(x_i)$$

$$ACU\colon \mathbb{R}^t \to [-1,1]$$

בהינתן מאושפז i, נשערך את ההסתברות שלו לעבור אירוע VTE ונחסיר מהנ"ל את שערוך ההסתברות שלו לדימום. התוצאה המתקבלת הינה ציון הרווח שלנו מטיפול במדללי דם. בהמשך לכך, כאשר הרווח שנקבל הינו שלילי, אזי ניתן לפרש זאת בתור הנזק שיגרם למאושפז מטיפול זה.

בחירת משתנים

תהליך זה ידוע בשם Feature Selection. מטרת התהליך הינה לצמצם את כמות המשתנים במאגר שאינם מועילים לסיווג. הסיבה לכך עשויה להיות שאין למשתנה החוזה מתאם עם משתנה המטרה הנחזה, או שיש משתנים נוספים אשר במתאם עם משתנה זה (Multicollinearity[32]). מטרת התהליך הינה להוריד את מורכבות המודל – דבר אשר עשוי לסייע ליכולת ההכללה שלו, ובנוסף יוריד את סיבוכיות הזמן של האימון[33]. לדוגמה, סיבוכיות זמן האימון של רגרסיה לוגיסטית הינה $\mathcal{O}(n \cdot d)$, כאשר d הוא מספר המשתנים. במחקר זה בוצע שימוש בעיקר בשתי שיטות היוריסטיות: השיטה הראשונה הינה Sequential Forward Selection - כאשר בתחילה משתנה אחד אשר מסווג באופן הטוב ביותר את משתנה המטרה

נבחר. לאחר מכן, נבחר משתנה נוסף מהמשתנים הנותרים אשר מוסיף את מירב הביצועים לביצועי המשתנה הראשון שנבחר, וכך הלאה. השיטה השניה הינה Sequential Backward Selection, כאשר רעיונה המרכזי הינו רעיון דומה לשיטה הראשונה, כאשר ההבדל הינו שבשיטה זו מתחילים עם כל המשתנים, ובכל איטרציה מחסירים משתנה אחד שהוא הכי פחות מועיל לחיזוי.

ערכים חסרים

במאגרים שעבדנו עמם, היו בין היתר משתנים עם ערכים חסרים רבים. בהמשך לכך, משתנים אשר היו בהם מעל 33% ערכים חסרים - הושמטו מהמאגר ולא נעשה בהם שימוש לניבוי. לגבי שאר המשתנים בוצעה השלמת נתונים חסרים באופן הבא: למשתנים שמיים או סודרים בוצע Mode Imputation. למשתנים כמותיים בוצע Median Imputation כיוון שהחציון פחות רגיש לערכי קיצון לא מייצגים מאשר הממוצע. בהמשך לכך - נראה כי בהערכה גסה המשתנים הכמותיים קיימו:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

מספר משתנים הופיעו במאגר אחד ולא הופיעו באחר – ולכן, אם לאחר שילוב שני המאגרים למשתנה מסוים היו מעל 33% ערכים חסרים המשתנה הושמט מהמאגר גם כן.

הערכת ביצועים

לשם הערכת הביצועים של המודלים שלנו נפצל את הנתונים לקבוצת אימון (Training set) וקבוצת מבחן (Test set). המודל נבנה על סמך קבוצת האימון ומוערך לפי קבוצת המבחן וזאת על מנת להימנע מהתאמת יתר (Overfitting), תוך שימוש במדד נכונות (Accuracy) אשר נגזר ממטריצת הערפול (Confusion matrix).

טבלה 1. מטריצת ערפול.

	Class 1 Predicted	Class 0 Predicted
Class 1 Actual	True Positive (TP)	False Negative (FN)
Class 0 Actual	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

כאשר מאגר הנתונים הוא מאוד לא מאוזן כמו במקרה שלנו, מדד ה-accuracy יפגע, וזאת כיוון שהמונה יהיה כמעט תמיד זניח ביחס למכנה. לכן נשתמש ב-balanced accuracy:

$$Balanced\ Accuracy = \frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} = \frac{TPR + TNR}{2}$$

במדד לעיל, ניתן לראות שלשתי המחלקות יש משקל שווה.

המדד לעיל מתייחס לביצועי המודל כאשר ישנו ערך סף אחד לחיזוי, שהחל ממנו החיזוי יהיה 1 ומתחתיו החיזוי יהיה 0. כאשר נרצה להתחשב בכל ערכי הסף האפשריים, נשתמש בעקומת ROC (Receiver Operating Characteristic). בעקומה זו, ציר ה-x הינו FPR וציר ה-y הינו TPR, וכך - לכל ערך סף יש נקודה במרחב זה שמייצגת את הביצועים. השטח שמתחת לעקומה זו נקרא AUC (Area Under the Curve), וככל שהוא מתקרב ל-1, המודל בכללותו הינו יותר טוב, קרי - יותר TPR ופחות FPR[34].

נציין כי מודל ה-Padua לא נבנה בהסתמך על המאגרים שלנו, ולכן נוכל להשתמש בכל המאגר כדי למדוד את ביצועיו, וזאת ללא חשש להתאמת יתר.

מבחנים סטטיסטיים

מבחן t למדגמים תלויים

על מנת להוכיח שהמודל שלנו לניבוי VTE מוצלח יותר מ-Padua באופן מובהק סטטיסטית בוצע מבחן t למדגמים תלויים (Dependent T-Test for Paired Samples). ביצענו k-fold cross validation, כאשר על כל fold נמדדו הביצועים של המודל שלנו והביצועים של Padua וחושב ההפרש ביניהם. בהמשך מתקבל ממוצע ההפרשים $\bar{d}$, והמבחן בודק האם הוא מובהק. על מנת להתקרב להנחת ההתפלגות הנורמלית, נבחר $k = n = 30$ בכדי לקיים:

$$\bar{d} \sim \mathcal{N}(\mu, {}^{\sigma^2}/_{n})$$

לפי משפט הגבול המרכזי.

מבחן שקילות מסוג שני מבחני t חד כיווניים (TOST)

חשוב לציין כי לעתים נתקלנו בבעיה כאשר רצינו לאחד את שני המאגרים. בעיה משמעותית זו הייתה שחלק מהמשתנים הכמותיים (שהיו לרוב מדדי דם כאלה או אחרים) שהתייחסו לאותו מדד בשני המאגרים, היו מוגדרים באופן שונה. לדוגמה המשתנה אלבומין, היה מוגדר כ"אלבומין בעת קבלה לאשפוז" במאגר הישן, ולעומת זאת משתנה זה היה מוגדר במאגר החדש כ"אלבומין ממוצע במהלך האשפוז".

במקרים שבהם היה חשוב להשתמש במשתנה על מנת לנבא אחד משני המצבים הרפואיים שאנו דנים בהם, השתמשנו בשיטת TOST (Two One Sided t Tests), וזאת על מנת לאחד בין המשתנים ולתת לכך צידוק סטטיסטי. המבחן הנ"ל למעשה בודק האם שני המדגמים נלקחו מאותה התפלגות. נקודה זו שונה מרוב המבחנים הסטטיסטיים הסטנדרטיים - בהם השערת ה-0 הינה ששני המדגמים נלקחו מאותה התפלגות, ומטרת החוקר היא להראות שזה לא כך - ולדחות את השערת ה-0. ההשערות ב-TOST:

$$\mathcal{H}_0\colon\ \mu_2 - \mu_1 \leq -\theta \ \ \text{or} \ \ \mu_2 - \mu_1 \geq \theta$$

$$\mathcal{H}_1\colon -\theta < \mu_2 - \mu_1 < \theta$$

על מנת לדחות את השערת ה-0 מתקבלים שני p values: האחד שההפרש בין התוחלות גדול מ-θ לכיוון מסוים, והשני גדול מ-θ לכיוון אחר. כלומר כאשר ה-p values שמתקבלים הינם מובהקים וקטנים מספיק, אנו יכולים להסיק ששני המדגמים נלקחו מאותה התפלגות עד כדי קבוע, כלומר - שתי התפלגויות שונות שהפרש התוחלות שלהן הוא עד כדי קבוע מסוים. ניתן לעשות מבחן זה תוך כדי הנחה שהשונויות של שתי ההתפלגויות שוות או שונות.

רווח סמך

על מנת לבדוק שה-ACU הממוצע שונה בין אוכלוסיית הקרישיות, אוכלוסיית המדממים ואוכלוסיית הבריאים (המאושפזים שלא עברו VTE ולא דיממו), נחשב רווח סמך עם $\alpha = 0.05$ לערך ה-ACU הממוצע בכל fold ב-k-fold cross validation, כאשר התהליך בוצע לכל $k \in \{2,3,4,5\}$. אם רווח הסמך

לא יחפוף בין האוכלוסיות נוכל להסיק כי באופן מובהק קיים שוני, תוחלת האוכלוסיות שונה מבחינת התועלת שלהם ממדללי דם, והמדד שלנו הצליח.

קליברציה

קליברציה (כיול) של מסווג הינה היכולת של המסווג להיות אמין לא רק בניבוי שלו, אלא גם ברמת הביטחון שלו בסיווג. בהנתן וקטור משתנים x ומסווג f עם פלט הסתברותי בין 0 ל-1, ומשתנה המטרה y שיכול לקבל ערכים 0 או 1, אם f הוא בעל קליברציה מושלמת[35]:

$$\mathbb{P}(y = 1 \mid f(x) = \pi) = \pi$$

$$\forall \pi \in [0,1]$$

לדוגמה, אם המסווג שלנו הפיק 20 פעמים הסתברות של 0.9, ולפי הגדרת הקליברציה הוא צודק ב-90% מהמקרים, ב-18 מהמקרים הללו התיוג יהיה $y = 1$, דהיינו המודל צודק. בהמשך לכך, בשני מקרים המודל ישגה והדוגמאות יהיו עם $y = 0$.

נציין שמודל בעל קליברציה טובה לא מעיד בהכרח שיהיה לו accuracy גבוה. לדוגמה למודל שמטיל מטבע וכך יוצר תחזיות יש קליברציה מושלמת, כיוון שהוא יפיק רק הסתברויות של 50%, והוא אכן יהיה צודק ב-50% מהמקרים כיוון שהוא מנחש באקראי מבין שתי תוצאות אפשריות. ה-accuracy של מודל זה גם יהיה 50% וזה נמוך מאוד.

ישנם אלגוריתמים רבים שניתן להשתמש בהם על מנת לחזות VTE ודימום ממאגר נתונים. ישנם אלגוריתמים שמפיקים רמות ביטחון אמינות ובעלות קליברציה טובה, ויש גם כאלו שלא.

כעת, ננסה לתת מוטיבציה לבחור למדד ה-ACU אלגוריתמים שמפיקים מודל בעל קליברציה טובה, וזאת על ידי הוכחה תאורטית שאם המודלים הם בעלי קליברציה טובה, אזי המודל המשולב בהכרח יעניק תועלת גבוהה יותר ממדללי דם לאלה שעברו VTE מאשר לאלה שדיממו.

מוטיבציה תאורטית לבחירת אלגוריתם

טענה לאוכלוסיות

אנו טוענים שאם שני המודלים, אחד לחיזוי VTE והשני לחיזוי דימום, הם בעלי קליברציה מושלמת, אז התוחלת של התועלת ממדללי דם (ACU) באוכלוסיית הקרישיות היא גדולה מתוחלת התועלת של אלה שיעברו דימום חמור. כלומר מדובר כאן בשני מקרים של תוחלת מותנית, תוחלת ה-ACU בהנתן והמאושפז עבר VTE (לכן $v = 1$), גדולה מתוחלת ה-ACU בהנתן והמאושפז דימם ($b = 1$). פורמלית:

$$\mathbb{E}(ACU(x) \mid v = 1) > \mathbb{E}(ACU(x) \mid b = 1)$$

הוכחה לטענה זו נמצאת בנספח א'.

אך כמובן שהמדגמים בעולם האמיתי הם בגודל סופי, לכן נוכיח גם עבורם.

טענה למדגמים בגודל סופי

כעת נטען שאם שני המודלים הם בעלי קליברציה מושלמת, אז גם במדגם בגודל סופי ממוצע ה-ACU של המאושפזים בקבוצת הקרישיות $\mathcal{V}$ יהיה גדול מממוצע הקרישיות בקבוצת המדממים $\mathcal{B}$. פורמלית:

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} ACU(x_i) > \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} ACU(x_j)$$

ההוכחה נמצאת בנספח ב'.

אלגוריתם

בחירת אלגוריתם למידת מכונה ליצירת המודלים שלנו תלויה בכך שהמודל הנוצר יתאפיין ב-Explainability (יכולת הסברה) של הניבויים, ושהמודל יהיה בעל קליברציה טובה עד כמה שניתן. המאפיין הראשון חשוב מכיוון שבניבוי מצבים רפואיים צריך לדעת מדוע כל סיווג ניתן. בעולם הרפואי ישנה העדפה להמנע ממודלים שהם "קופסה שחורה", ופעמים רבות רוצים לדעת איזה גורם סיכון (משתנה חוזה) מקושר חזק יותר לאבחנה הרפואית ואיזה מקושר חלש יותר, האם הקשר הוא חיובי או שלילי ועוד. המאפיין השני חשוב בגלל הטענות לעיל שהוכחו ונתנו מוטיבציה תאורטית למודלים בעלי קליברציה טובה.

אלגוריתמים מסוג רשתות נוירונים עמוקות אמנם מפיקים לרוב מודלים בעלי קליברציה טובה, אך הן לא מקיימות את המאפיין הראשון מכיוון שאינן נותנות משקל לכל משתנה ויש בין שכבת הקלט לשכבת הפלט משקלות רבות נוספות, מה שפוגע ביכולת ההסברה של המודל. אלגוריתם SVM אמנם יכול לדרג את חשיבות המשתנים לסיווג, אך מפיק מודל בעל קליברציה לא מוצלחת ולכן לא עומד במאפיין השני. בנוסף אלגוריתמי Naïve Bayes ו-Boosted Trees מפיקים מודלים בעלי קליברציה לא מוצלחת גם כן. אנו בחרנו להשתמש באלגוריתם רגרסיה לוגיסטית שמקיים את שני המאפיינים[36].

הרגרסיה הלוגיסטית מסווגת דוגמה בעזרת שערוך ההסתברות שלה[37]:

$$\mathbb{P}(y=1|X;W)=\sigma(w_0+w_1\cdot x_1+\cdots+w_t\cdot x_t)$$

$$\mathbb{P}(y=0|X;W)=1-\mathbb{P}(y=1|X;W)$$

כיוון שלמשתנים מוצמדות משקלות w, ניתן לדרג אותם לפי סדר חשיבותם לסיווג. σ היא פונקציית ה-Sigmoid:

$$\sigma(z)=\frac{1}{1+e^{-z}}$$

$$\sigma\colon \mathbb{R}\to(0,1)$$

לפונקציה זו תכונות רבות אשר עוזרות במהלך שלב הלמידה של האלגוריתם, ביניהן: היא חסומה בין 0 ל-1 (מה שעוזר לתת פרשנות הסתברותית לסיווגים), מונוטונית עולה וגזירה בכל נקודה עם נגזרת חיובית. כלל ההחלטה לרוב הוא 0.5 באופן אינטואיטיבי, אך ניתן לשנות אותו בהתאם לצורך:

$$\sigma(w_0+\cdots+w_t\cdot x_t)>0.5 \qquad \Leftrightarrow$$

$$w_0+\cdots+w_t\cdot x_t>0 \qquad \Leftrightarrow$$

$$\hat{y}=1$$

כלל ההחלטה הינו לינארי, דבר המגביל את כוחה של הרגרסיה הלוגיסטית למצוא יחסים מורכבים ולא לינאריים בין המשתנים.

הדרך לשערך את המשקלות הינה בעזרת מיקסום פונקציית תועלת הסתברותית, כאשר הקליברציה הטובה של המודל שמתקבל קשורה לעובדה זו. אנו נרצה למקסם את ה-likelihood (נראות, סבירות) של התיוגים בהנתן הדוגמאות במדגם. הדוגמאות נדגמו באופן בלתי תלוי ולכן ישנה מכפלה בין ה-likelihood של הדוגמאות. עם זאת, כיוון שמכפלה היא מסובכת יותר מסכום מבחינה חישובית ופונקציית log היא מונוטונית עולה, אנו נעבוד עם log likelihood[38]:

$$\ell og\mathcal{L}(W) = \ell og \prod_{i=1}^{n} \mathbb{P}(y_i|X_i;W) = \sum_{i=1}^{n} \ell og\mathbb{P}(y_i|X_i;W)$$

כאמור, אנו נרצה למצוא את W שימקסם את ההסתברות לצפות בתיוגים שנתונים לנו במדגם. כיוון שמשתנה הסיווג הוא בינארי נביא ביטוי מפורש ל-likelihood של דוגמה מסוימת:

$$\mathbb{P}(y|X;W) = \ \sigma(w_0 + \cdots + w_t \cdot x_t)^y (1 - \sigma(w_0 + \cdots + w_t \cdot x_t))^{1-y}$$

על מנת למצוא את W אשר ממקסם את ה-likelihood גוזרים את הפונקציה לפי W ומשווים ל-0. כיוון שאין פתרון סגור למשוואה נשתמש ב-gradient ascent כדי לעדכן את W בקצב η (קבוע הלמידה) בכיוון שמשפר את הביצועים.

$$\frac{\partial \ell og\mathcal{L}(W)}{\partial W} = \sum_{i=1}^{n} (y_i - \sigma(w_0 + \cdots + w_t \cdot x_t^i))X_i$$

$$W_{s+1} \leftarrow W_s + \eta \cdot \frac{\partial \ell og\mathcal{L}(W_s)}{\partial W}$$

תוצאות

משקלות חדשים למשתני Padua

גורמי הסיכון שנמצאים במדד Padua הנם בעלי משקלים שונים הנעים מ-1 עד 3, וכל משקל הינו מספר שלם. עובדה זו גרמה לנו להניח כי ייתכן שישנה דרך למשקל אותם מחדש ולהניב ביצועים טובים יותר. כיוון שרק במאגר הישן (n=18,890) היו כל המשתנים של Padua, השתמשנו רק בו על מנת לבצע משימה זו. כאמור בחלק "שיטות", השתמשנו ב-k-fold cross validation עם k=30. הפונקציה שמתקבלת עם המשקלות החדשים של משתני Padua לאחר למידה על כל המאגר עם רגרסיה לוגיסטית:

$$g_v(x) = -1.09 +$$

$$-1.228 \cdot Thrombophilia +$$

$$1.536 \cdot Malignant +$$

$$3.528 \cdot VtePrior +$$

$$0.233 \cdot ReducedMobility +$$

$$0.652 \cdot Over70 +$$

$$0.016 \cdot HeartRespiratoryFailure +$$

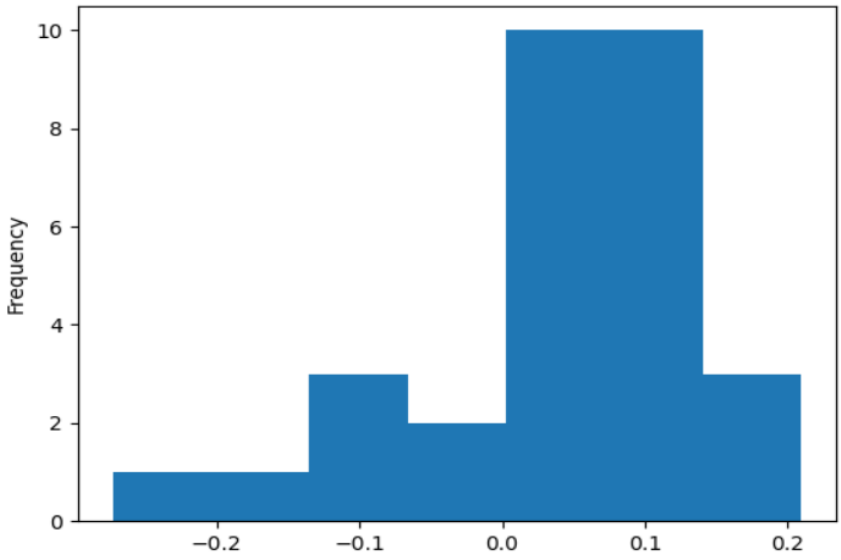
$$0.254 \cdot InfectionRheumatologicalDisorder +$$

$$0.246 \cdot Obesity +$$

$$-0.856 \cdot OngoingHormonalTreatment$$

$$f_v(x) = \frac{1}{1 + e^{-g_v(x)}}$$

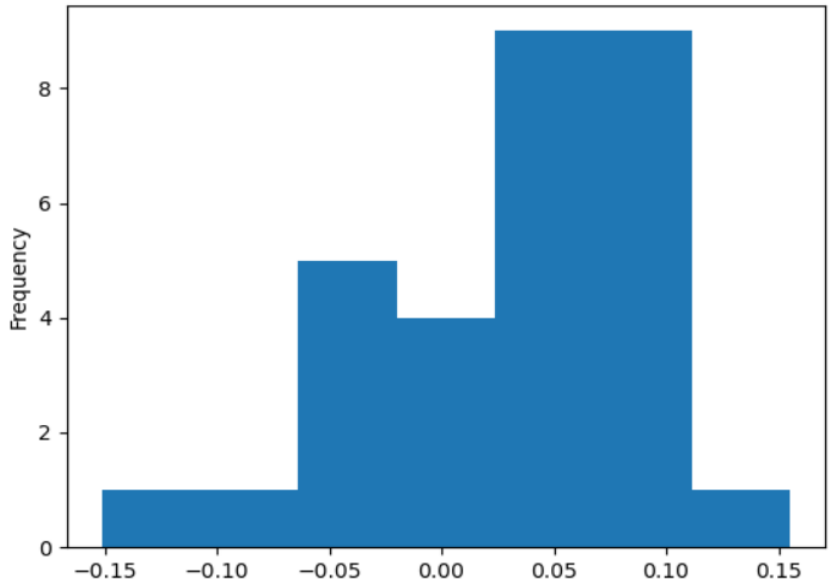
ה-balanced accuracy הממוצע של המודל שלנו בכל ה-folds היה 0.683 ושל Padua היה 0.647.



איור 3. מציג את התפלגות ההפרשים ב-folds השונים בין ה-balanced accuracy של המודל עם המשקולות החדשים ל-Padua.

ניתן לראות שברוב המקרים יש יתרון של כ-0.05 עד 0.12 של המודל עם המשקלות החדשות לעומת המשקלות הישנות. כאשר בודקים מובהקות של ההפרש הממוצע בין המודלים עם מבחן t לתלויים, מתקבל p value של 0.0419.

ה-AUC הממוצע של המודל שלנו בכל ה-folds היה 0.755 ושל Padua היה 0.724.



איור 4. מציג את התפלגות ההפרשים ב-folds השונים בין ה-AUC של המודל עם המשקלות החדשים ל-Padua.

גם במקרה של AUC ניתן לראות שברוב המקרים למודל שלנו יש יתרון על Padua. כאשר בודקים מובהקות של ההפרש הממוצע בין המודלים עם מבחן t לתלויים, מתקבל p value של 0.008.

כפי שניתן לראות בפונקציה לעיל המשקל הגבוה ביותר ניתן לגורמי סיכון של "סרטן פעיל" ו-"VTE קודם". בשביל לחקור נקודה זו יותר לעומק, בדקנו מה יקרה אם נעלה את המשקל של שני משתנים אלו ל-4 נקודות במסגרת ה-Padua ונשתמש בהם בלבד. כלומר מספיק אחד משני גורמי הסיכון האלה בכדי להחשב בסיכון גבוה ל-VTE.

ה-balanced accuracy של המודל המכווץ על כל המאגר היה 0.645 לעומת 0.648 של Padua.

ה-AUC של המודל המכווץ היה 0.645 לעומת 0.717 של Padua.

טבלה 2. סיכום שינוי התיוגים במעבר מ-Padua למודל המכווץ בעל שני המשתנים.

	$\hat{v}=0 \rightarrow \hat{v}=1$	$\hat{v}=1 \rightarrow \hat{v}=0$
n	1302	3416
$v=1$	26	35
$b=1$	326	440

ניבוי יחד עם משתנים חדשים

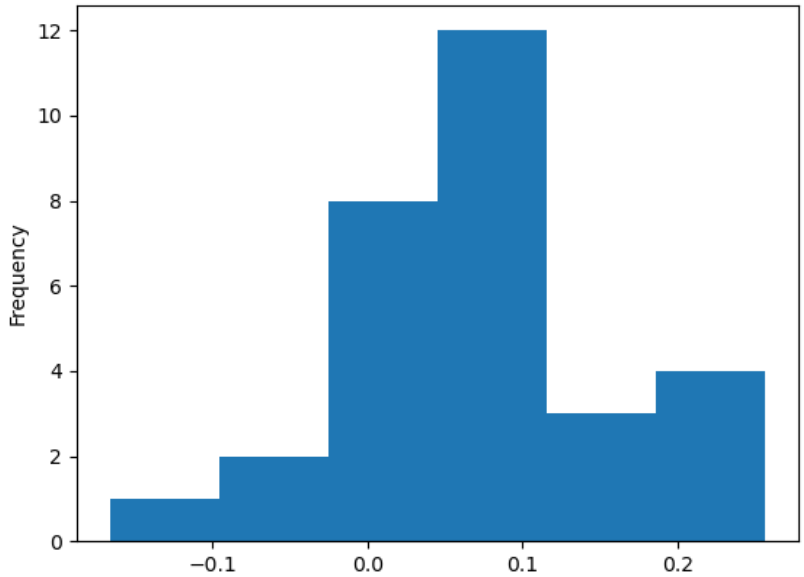
חמישת המשתנים שנבחרו על ידי אלגוריתם SFS הינם: Sex, Malignant, VtePrior, Albumin, Aids.

על מנת להשתמש במשתנה אלבומין, היה צורך בלהשתמש במבחן TOST שעליו הובא פירוט בחלק "שיטות". במאגר הישן רמת האלבומין הוגדרה בתור "אלבומין בעת קבלה" ובמאגר החדש הוגדרה בתור "אלבומין ממוצע במהלך האשפוז". הממוצע של משתנה "אלבומין בעת קבלה" היה 3.9245 וסטיית התקן הייתה 0.5969. לעומתו הממוצע של משתנה "אלבומין ממוצע במהלך האשפוז" היה 3.729 וסטיית התקן הייתה 0.632.

הפונקציה והמשקלות המתקבלות לאחר אימון על כל המאגר עם רגרסיה לוגיסטית:

$$g_v(x) = 2.64 +$$
$$0.424 \cdot Sex +$$
$$1.177 \cdot Malignant +$$
$$3.317 \cdot VtePrior +$$
$$-1.027 \cdot Albumin +$$
$$1.694 \cdot Aids$$
$$f_v(x) = \frac{1}{1 + e^{-g_v(x)}}$$

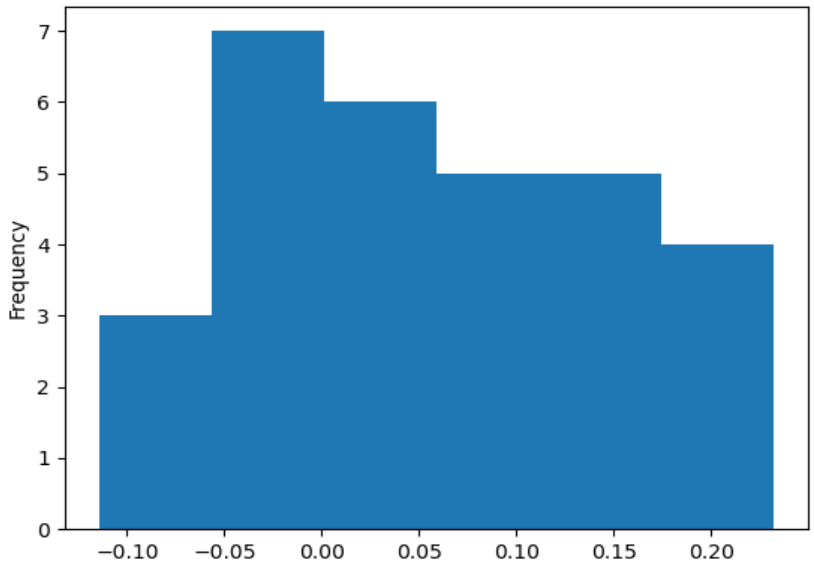
ה-balanced accuracy הממוצע ב-folds השונים של המודל עם המשתנים החדשים היה 0.722 לעומת 0.649 של Padua.



איור 5. מציג את התפלגות ההפרשים ב-folds השונים בין ה-balanced accuracy של המודל עם המשתנים החדשים ל-Padua.

כאשר בודקים מובהקות של ההפרש הממוצע בין המודלים עם מבחן t לתלויים, מתקבל p value של 0.0002.

ה-AUC הממוצע ב-folds השונים של המודל עם המשתנים החדשים היה 0.776 לעומת 0.718 של Padua.



איור 6. מציג את התפלגות ההפרשים ב-folds השונים בין ה-AUC של המודל עם המשתנים החדשים ל-Padua.

כאשר בודקים מובהקות של ההפרש הממוצע בין המודלים עם מבחן t לתלויים, מתקבל p value של 0.0014.

חיזוי דימום

שבעת המשתנים אשר אלגוריתם SFS בחר בתור הקשורים ביותר לדימום הינם: CongestiveHeartFailure, SevereKidneyDisease, Malignant, Aids, Creatinine, Albumin, Function1Ind2Unind

במאגר הישן הגדרת משתנה הקריאטינין הייתה "קריאטינין מקסימלי במהלך האשפוז" ובמאגר החדש הגדרתו הייתה "קריאטינין ממוצע במהלך האשפוז". בוצע תהליך דומה של מבחן TOST כמו שנעשה על אלבומין. הממוצע של משתנה "קריאטינין מקסימלי במהלך האשפוז" היה 1.35 וסטיית התקן הייתה 1.364. לעומתו הממוצע של משתנה "קריאטינין ממוצע במהלך האשפוז" היה 1.37 וסטיית התקן הייתה 2.53.

הפונקציה והמשקלות המתקבלות לאחר אימון של רגרסיה לוגיסטית על כל המאגר:

$$g_b(x) = 0.942 +$$

$$-0.699 \cdot CongestiveHeartFailure +$$

$$-1.62 \cdot SevereKidneyDisease +$$

$$1.53 \cdot Malignant +$$

$$-1.03 \cdot Aids +$$

$$0.21 \cdot Creatinine +$$

$$-0.44 \cdot Albumin +$$

$$0.22 \cdot Function1Ind2Unind$$

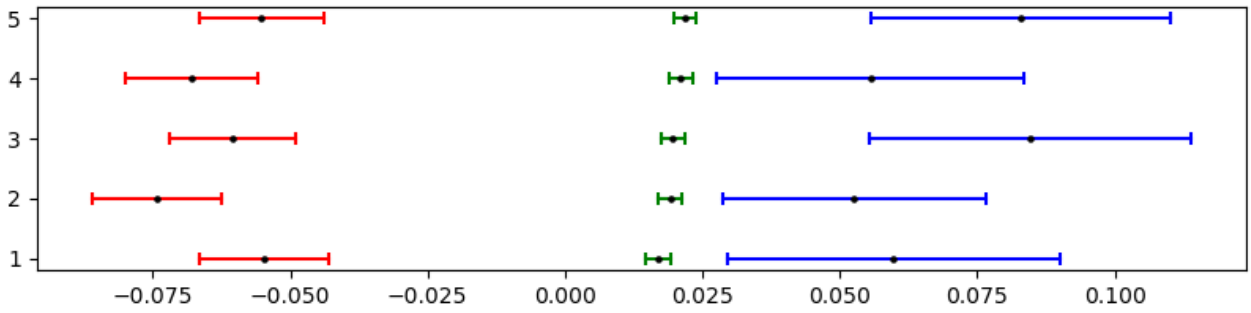
$$f_b(x) = \frac{1}{1 + e^{-g_b(x)}}$$

ה-balanced accuracy הממוצע עם k=10 fold cross validation היה 0.657, וה-AUC הממוצע היה 0.706.

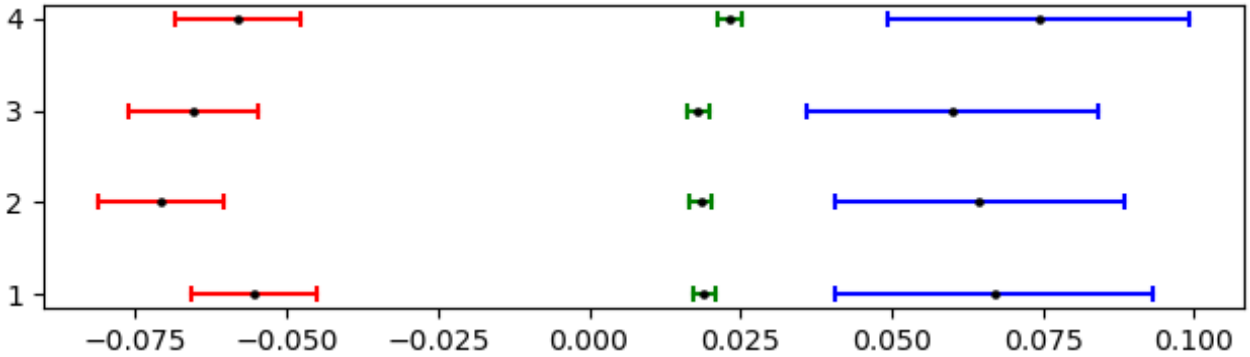
מדד תועלת ממדללי דם (ACU)

על מנת לבחון את מובהקות ממוצעי התועלת ממדללי דם (ACU) השתמשנו ברווח סמך לתוחלות שלושת האוכלוסיות אשר מעניינות אותנו: אוכלוסיית הקרישיות שאלה המאושפזים אשר עברו VTE, אוכלוסיית המדממים דימום חמור ואוכלוסיית הבריאים שלא עברו אף אחד מהמצבים הרפואיים לעיל.

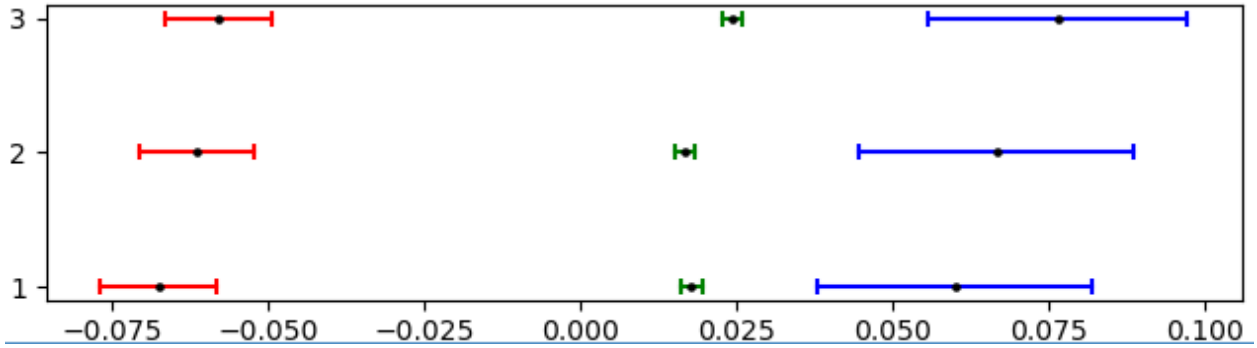
בכל האיורים הבאים ציר ה-x הוא התועלת ממדללי דם (ACU), ציר ה-y הוא מספר ה-fold ב- k-fold cross validation. הנקודה השחורה הינה הממוצע של המדגמים והקטע שהם מוכלים בו זה רווח הסמך לתוחלות עם $\alpha = 0.05$. אדום מייצג את אוכלוסיית המדממים, ירוק את אוכלוסיית הבריאים, וכחול את אוכלוסיית הקרישיות.



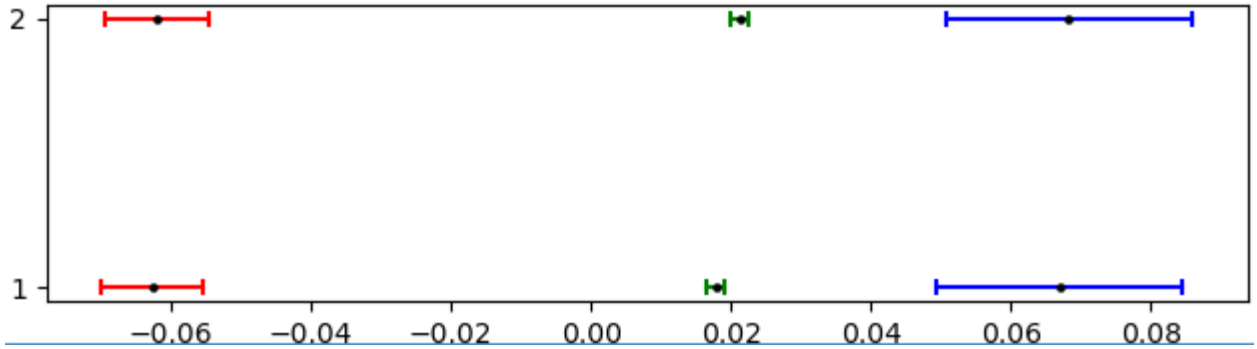
איור 7. רווחי סמך לתוחלות עם α של 0.05. ציר ה-x הוא ערך ה-ACU, ציר ה-y הוא מספר ה-fold עם k=5 fold cross validation.
הצבע הכחול מייצג את אוכלוסיית הקרישיות, הירוק את אוכלוסיית הבריאים, ואדום את אוכלוסיית המדממים.



איור 8. רווחי סמך לתוחלות עם α של 0.05. ציר ה-x הוא ערך ה-ACU, ציר ה-y הוא מספר ה-fold עם k=4 fold cross validation.
הצבע הכחול מייצג את אוכלוסיית הקרישיות, הירוק את אוכלוסיית הבריאים, ואדום את אוכלוסיית המדממים.



איור 9. רווחי סמך לתוחלות עם α של 0.05. ציר ה-x הוא ערך ה-ACU, ציר ה-y הוא מספר ה-fold עם k=3 fold cross validation.
הצבע הכחול מייצג את אוכלוסיית הקרישיות, הירוק את אוכלוסיית הבריאים, ואדום את אוכלוסיית המדממים.



איור 10. רווחי סמך לתוחלות עם α של 0.05. ציר ה-x הוא ערך ה-ACU, ציר ה-y הוא מספר ה-fold עם k=2 fold cross validation.
הצבע הכחול מייצג את אוכלוסיית הקרישיות, הירוק את אוכלוסיית הבריאים, ואדום את אוכלוסיית המדממים.

דיון ומסקנות

ההוכחות התאורטיות לגבי מדד ה-ACU והקשר שלו למודלים בעלי קליברציה טובה מקנות מוטיבציה חזקה לשימוש במודלים מסוג זה במקרה שלנו. יתרה מכך, ניתן להכליל זאת למקרה כללי כאשר מנסים לנבא שני מצבים רפואיים כאשר מתן תרופה לאחד מגדילה את הסיכוי להתרחשות השני, לדוגמה מתן תרופה מדכאת חיסון על מנת להקטין סיכוי להופעת מחלה אוטואימונית, כאשר אותה תרופה לא מומלצת למי שמועָד לזיהומים חמורים. בנוסף, אפשר לטעון שבמודלים משולבים מסוג זה רצוי יהיה להקריב מעט accuracy אם ישנה הבטחה לקליברציה טובה יותר במודל אחר.

בחלוקה חדשה של משקלים למשתני ה-Padua, התקבל כי ישנה חלוקה המניבה balanced accuracy ו-AUC גבוה יותר מביצועי ה-Padua באופן מובהק. במסגרת החלוקה מחדש, נראה כי ישנם שני משתנים אשר חשובים במיוחד והם "סרטן פעיל" ו-VTE קודם". לאור תוצאות אלו ניתן לשער כי 11 זה מספר רב מדי כאשר מנסים לנבא סיכון ל-VTE, וכי השפעת חלק מהמשתנים היא זניחה לעומת השפעת המשתנים אחרים.

כאשר השתמשנו במשתני "סרטן פעיל" ו-"VTE קודם" בלבד במודל המכווץ, ה-balanced accuracy היה שקול ואף גבוה במעט מ-Padua, תוצאה אשר מעידה על כך שהמשקל שלהם צריך לעלות מעל 3 הנקודות שיש להם במודל המקורי. לעומת זאת ה-AUC של שניהם בלבד הוא נמוך יותר. סיבה אפשרית היא שיכולת החלוקה של מרחב הקלטים הוא נמוך הרבה יותר עם שני משתנים בינאריים בלבד מאשר עם 11 משתנים בינאריים. כמו כן, במקרה שלנו יהיה ערך סף אחד בלבד להחלטת הסיווג, כיוון שהגדרנו שמספיק שאחד משני המשתנים יהיה חיובי על מנת לסווגו כ-1, ועקב סיבה זו ה-AUC שווה ל- balanced accuracy במודל המכווץ. כאשר מתבוננים על טבלת הפירוט של המאושפזים שסיווגם השתנה, רואים שבמעבר למודל החדש מפספסים נטו 9 מקרי VTE $(26 - 35 = -9)$, אך מצד שני מרוויחים נטו 114 מדממים $(440 - 326 = 114)$ ומצילים אותם מכך שהמודל הישן היה ממליץ לתת להם מדללי דם, ככל הנראה להחמיר את הדימום - ולפגוע בהם עוד יותר.

כאשר הייתה יכולת להתרחב אל משתנים חדשים לחיזוי הדימום, הביצועים היו טובים עוד יותר בהשוואה ל-Padua והמובהקות הייתה גבוהה אף יותר. נבחרו 5 משתנים, מה שמחזק את הטענה ש-11 המשתנים של Padua זה מספר רב מדי. בנוסף גם משתנה הנגזר מבדיקת דם, רמת אלבומין, הראה קשר של יחס הפוך לסיכוי ל-VTE, כאשר יחס הפוך זה נגזר מהמשקל השלילי שלו במסווג. Padua מתייחס למשתנים בינאריים בלבד מסוג אבחנות רפואיות ואינו לוקח בחשבון משתנים כמותיים כגון זה והנ"ל יכול להיות חיסרון אפשרי נוסף. נקודה מעניינת נוספת נגעה למשתנה ה-Sex, כאשר הערך 1 סימל זכר והערך 2 נקבה. בפונקציית המסווג ניתן לראות כי המשקל של משתנה זה הינו חיובי, מה שמצביע על כך שדווקא נקבות יותר נקשרו לאירועי VTE - בניגוד לנאמר בספרות המבוא. ניתן לשער כי המספר הנמוך יחסית של אירועי VTE מתוך המדגמים (0.75% במאגר הישן ו-1.08% במאגר החדש) יכול להביא לסטיה סטטיסטית מההתפלגות באוכלוסיה, ויש להתייחס לקשר זה בספקנות. העובדה כי למשתנה המדובר היה את המשקל הנמוך ביותר מחמשת המשתנים במודל מחזקת טענה זה.

יכולת חיזוי הדימום הייתה מוגבלת עם ביצועים מוגבלים, אך אנו משערים כי גם מוגבלות זו עדיפה על פני לא להתחשב בסיכון לדימום חמור כלל וכלל כאשר ניגשים להחלטת מתן מדללי דם. נקודה מעניינת הינה שמשתנה ה-Aids הופיע בשני המודלים, כאשר במודל לניבוי VTE הקשר הינו ישר, ולעומת זאת במודל לניבוי דימום הקשר הינו הפוך. ניתן לשער כי איידס גורם לשיבוש ולפעילות יתר של מנגנון הקרישיות, ודרוש לכך מחקר נוסף. עוד משתנה שהופיע בשני המודלים הוא Malignant, דבר אשר יכול להעיד על כך שסרטן פוגע במערכת הדם באופן כללי ולא לכיוון מסוים. משתנה האלבומין הופיע גם הוא בשני המודלים, כאשר בשניהם הוא היה בעל מקדם שלילי, אך במודל לניבוי דימום משקלו היה קטן יותר וייתכן שזוהי חריגה סטטיסטית כתוצאה ממגבלות המדגמים.

מדד ה-ACU הראה הבדלים מובהקים בין האוכלוסיות בכל קבוצות המבחן ובכל ה-k שנבחנו. ערך ה-k קבע לכמה קבוצות מבחן המאגר ושלושת הקבוצות שעניינו אותנו (קרישיות, דימום ובריאים) יחולקו. תוצאה מעשית זו מזכירה את מה שהתקבל בהוכחה התאורטית למדגמים סופיים, כיוון שגם במקרה זה ה-ACU הממוצע לקבוצת המדממים היה שלילי בעוד לקבוצת הקרישיות הוא היה חיובי. כאשר מעיינים

ברוחב רווח הסמך בין ה-k השונים ניתן להבחין בכך שהרווח קטֵן ככל ש-k יורד. הבחנה זו מתקיימת כיוון שככל ש-k יורד כך מחלקים את המדגם לפחות קבוצות, דבר אשר מגדיל את קבוצת המבחן. הגדלת ה-n תקטין את גבולות הרווח כי $\sqrt{n}$ נמצא במכנה של גבולות אלו. בנוסף, ניתן לראות כי הרווח של קבוצת הקרישיות הוא הגדול ביותר בכל k ובכל fold, והנ"ל נובע מכך שמאושפזים אשר עברו VTE היו נדירים במאגרים, וההשפעה של גודל הקבוצה על רוחב הרווח תוארה לעיל. יחד עם העובדה כי דימומים חמורים יכולים להיות מסכני חיים לא פחות ואף יותר מ-VTE במקרים מסוימים, תוצאות אלו תומכות בטענה שצריך להתחשב גם בסיכון לדימום בעת ההחלטה האם לטפל במדללי דם או לא.

ביבליוגרפיה

1. Souto, J. C. et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am. J. Hum. Genet. 67, 1452–1459 (2000).

2. Ariëns RA, de Lange M, Snieder H, Boothby M, Spector TD, Grant PJ. Activation markers of coagulation and fibrinolysis in twins: heritability of the prethrombotic state. Lancet. 2002 Feb 23;359(9307):667-71. doi: 10.1016/S0140-6736(02)07813-3. PMID: 11879863.

3. Larsen, T. B. et al. Major genetic susceptibility for venous thromboembolism in men: a study of Danish twins. Epidemiology 14, 328–332 (2003).

4. Heit, J. A. et al. Familial segregation of venous thromboembolism. J. Thromb. Haemost. 2, 731–736 (2004).

5. Zöller, B., Ohlsson, H., Sundquist, J. & Sundquist, K. Familial risk of venous thromboembolism in first-, second- and thirddegree relatives: a nationwide family study in Sweden. Thromb. Haemost. 109, 458–463 (2013).

6. Palta S, Saroa R, Palta A. Overview of the coagulation system. *Indian J Anaesth*. 2014;58(5):515-523. doi: 10.4103/0019-5049.144643.

7. Hoffman M, Monroe DM. Coagulation 2006: a modern view of hemostasis. *Hematol Oncol Clin North Am*. 2007;21(1):1-11. doi: 10.1016/j.hoc.2006.11.004

8. Kearon C, Akl EA, Ornelas J, et al. Antithrombotic therapy for VTE disease: CHEST guideline and expert panel report. *Chest*. 2016;149(2):315-352. doi: 10.1016/j.chest.2015.026.

9. Kearon C, Ageno W, Cannegieter C, Cosmi B, Geersing GJ, Kyrle PA; Subcommittees on Control of Anticoagulation, and Predictive and Diagnostic Variables in Thrombotic Disease. Categorization of patients as having provoked or unprovoked venous thromboembolism: guidance from the SCC of ISTH. *J Thromb Haemost*. 2016;14(7):1480-1483. doi: 10.1111/jth.13336.

10. Phillippe, Haley M. "Overview of venous thromboembolism." *The American journal of managed care* 23.20 Suppl (2017): S376-S382.

11. Witt DM, Clark NP, Vazquez SR. Venous thromboembolism. In: DiPiro JT, Talbert RL, Yee GC, Matzke GR, Wells BG, Posey L. eds. *Pharmacotherapy: A Pathophysiologic Approach. 10th ed.* New York, NY: McGraw-Hill Medical; 2017:231-260.

12. Milling TJ Jr, Ziebell CM. A review of oral anticoagulants, old and new, in major bleeding and the need for urgent surgery. Trends Cardiovasc Med. 2020

Feb;30(2):86-90. doi: 10.1016/j.tcm.2019.03.004. Epub 2019 Mar 26. PMID: 30952383; PMCID: PMC6763385.

13. Cohen AT, Agnelli G, Anderson FA, et al. VTE Impact Assessment Group in Europe (VITAE). Venous thromboembolism (VTE) in Europe. The number of VTE events and associated morbidity and mortality. Thromb Haemost. 2007 Oct;98(4):756-64. PMID: 17938798

14. Treasure T, Hill J. NICE guidance on reducing the risk of venous thromboembolism in patients admitted to hospital. J R Soc Med. 2010 Jun;103(6):210-2. PMID: 20513894

15. Collins R, Scrimgeour A, Yusuf S, et al. Reduction in fatal pulmonary embolism and venous thrombosis by perioperative administration of subcutaneous heparin. Overview of results of randomized trials in general, orthopedic, and urologic surgery. N Engl J Med. 1988 May 5;318(18):1162-73. PMID: 3283548

16. Alikhan R, Bedenis R, Cohen AT. Heparin for the prevention of venous thromboembolism in acutely ill medical patients (excluding stroke and myocardial infarction). Cochrane Database Syst Rev. 2014 May 7;(5):CD003747. PMID: 24804622

17. Lederle FA, Zylla D, MacDonald R, et al. Venous thromboembolism prophylaxis in hospitalized medical patients and those with stroke: a background review for an American College of Physicians Clinical Practice Guideline. Ann Intern Med. 2011 Nov 1;155(9):602-15. PMID: 22041949

18. Barbar S, Noventa F, Rossetto V, et al. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score. J Thromb Haemost. 2010 Nov;8(11):2450-7. PMID: 20738765

19. Spyropoulos AC, Anderson FA Jr, FitzGerald G, et al. IMPROVE Investigators. Predictive and associative models to identify hospitalized medical patients at risk for VTE. Chest. 2011 Sep;140(3):706-714. PMID: 21436241

20. Nendaz M, Spirk D, Kucher N, et al. Multicentre validation of the Geneva Risk Score for hospitalised medical patients at risk of venous thromboembolism. Explicit ASsessment of Thromboembolic RIsk and Prophylaxis for Medical PATients in SwitzErland (ESTIMATE). Thromb Haemost. 2014 Mar 3;111(3):531-8. PMID: 24226257

21. Kucher N, Koo S, Quiroz R, et al. Electronic alerts to prevent venous thromboembolism among hospitalized patients. N Engl J Med. 2005 Mar 10;352(10):969-77. PMID: 15758007

22. Geerts WH, Bergqvist D, Pineo GF, et al. Prevention of venous thromboembolism: American College of Chest Physicians Evidence-Based

Clinical Practice Guidelines (8th Edition). Chest 2008; 133(Suppl. 6): 381–453. PMID: 18574271

23. Cardiovascular Disease Educational and Research Trust. Prevention and treatment of venous thromboembolism. International Consensus Statement (guidelines according to scientific evidence). Int Angiol 2006; 25: 101–61. PMID: 16763532

24. Vardi M, Haran M. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score: a rebuttal. J Thromb Haemost. 2011 Jul;9(7):1437-8. doi: 10.1111/j.1538-7836.2011.04305.x. PMID: 21501378.

25. Vapnik V. Principles of risk minimization for learning theory. Advances in neural information processing systems. 1991;4.

26. Hoeffding W. Probability inequalities for sums of bounded random variables. InThe collected works of Wassily Hoeffding 1994 (pp. 409-426). Springer, New York, NY.

27. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine. Where do we stand? N Engl J Med. 1987 Mar 12;316(11):685-8. doi: 10.1056/NEJM198703123161109. PMID: 3821801.

28. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115-118. doi: 10.1038/nature21056. Epub 2017 Jan 25. Erratum in: Nature. 2017 Jun 28;546(7660):686. PMID: 28117445; PMCID: PMC8382232.

29. Wiens J, Campbell WN, Franklin ES, Guttag JV, Horvitz E. Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile. Open Forum Infect Dis. 2014 Jul 15;1(2):ofu045. doi: 10.1093/ofid/ofu045. PMID: 25734117; PMCID: PMC4281796.

30. Hadanny A, Shouval R, Wu J, Gale CP, Unger R, Zahger D, Gottlieb S, Matetzky S, Goldenberg I, Beigel R, Iakobishvili Z. Machine learning-based prediction of 1-year mortality for acute coronary syndrome☆. Journal of Cardiology. 2022 Mar 1;79(3):342-51.

31. Lipschuetz M, Guedalia J, Rottenstreich A, Persky MN, Cohen SM, Kabiri D, Levin G, Yagel S, Unger R, Sompolinsky Y. Prediction of vaginal birth after cesarean deliveries using machine learning. American journal of obstetrics and gynecology. 2020 Jun 1;222(6):613-e1.

32. Alin A. Multicollinearity. Wiley interdisciplinary reviews: computational statistics. 2010 May;2(3):370-4.

33. Bermingham, M., Pong-Wong, R., Spiliopoulou, A. et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci Rep 5, 10312 (2015). https://doi.org/10.1038/srep10312.

34. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006 Jun 1;27(8):861-74.

35. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. InInternational conference on machine learning 2017 Jul 17 (pp. 1321-1330). PMLR.

36. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. InProceedings of the 22nd international conference on Machine learning 2005 Aug 7 (pp. 625-632).

37. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013 Apr 1.

38. Le Cam L. Maximum likelihood: an introduction. International Statistical Review/Revue Internationale de Statistique. 1990 Aug 1:153-71.

נספחים

נספח א' - הוכחה לאוכלוסיות

נגדיר אוכלוסיית מאושפזים כאשר כל מאושפז מסומן ב-X, שמורכב מוקטור משתנים (features) x, משתנה בינארי v שמגדיר האם המאושפז שייך לאוכלוסיית הקרישיות (VTE) ומשתנה בינארי b שמגדיר האם המאושפז שייך לאוכלוסיית המדממים. פורמלית $X = (x, v, b)$.

נניח שאין מקרים של X שבהם מתקיים $v = b = 1$ וגם אין $v = b = 0$. כלומר

$$\mathbb{P}(v = 1) + \mathbb{P}(b = 1) = \mathbb{P}(v = 1) + \mathbb{P}(v = 0) = 1$$

יהיו מודלים f_v לחיזוי VTE ו-f_b לחיזוי דימום כאשר לשני המודלים פלט הסתברותי. נוכיח שאם המודלים בעלי קליברציה מושלמת, כלומר

$$\mathbb{P}(v = 1 \mid f_v(x) = \pi) = \pi$$

$$\mathbb{P}(b = 1 \mid f_b(x) = \pi) = \pi$$

$$\forall \pi \in [0,1]$$

אז התוחלת של התועלת ממדללי דם (ACU) באוכלוסיית הקרישיות הינה גדולה מתוחלת התועלת של אלה שיעברו דימום חמור. כלומר מדובר בשני מקרים של תוחלת מותנית, תוחלת ה-ACU בהנתן והמאושפז עבר VTE (לכן $v = 1$), גדולה מתוחלת ה-ACU בהנתן והמאושפז דימם ($b = 1$). פורמלית:

$$\mathbb{E}(ACU(x) \mid v = 1) > \mathbb{E}(ACU(x) \mid b = 1)$$

ונניח ש-

$$f_v(x) \sim \mathcal{U}(0,1)$$

$$f_b(x) \sim \mathcal{U}(0,1)$$

כלומר ההסתברות שהמודלים פולטים מתפלגת באופן אחיד באותו קטע, כך שניתן יהיה להגיד כי

$$\mathbb{P}(f_v(x) = \pi) = \mathbb{P}(f_b(x) = \pi) = \mathcal{U}(0,1) \tag{1}$$

כלומר לכל π יש הסתברות שווה להיות מופקים מהמודלים, ונסמן את שתי ההתפלגויות כ-$\mathbb{P}(f(x) = \pi)$.

טענות עזר:

$$\mathbb{P}(v = 0 \mid f_v(x) = \pi) = \mathbb{P}(b = 1 \mid f_v(x) = \pi) = 1 - \pi \tag{2}$$

$$\mathbb{P}(b = 0 \mid f_b(x) = \pi) = \mathbb{P}(v = 1 \mid f_b(x) = \pi) = 1 - \pi \tag{3}$$

נובעות מכך כי הנחנו שכל מאושפז הוא או $v = 1$ או $b = 1$ בהכרח, ועל פי חוק ההסתברות המשלימה להנחת הקליברציה.

כעת, נתחיל לפתח את מדד ה-ACU למרכיביו הקטנים יותר באיבר התוחלת לאוכלוסיית הקרישיות (האיבר השמאלי באי השוויון)

$$\mathbb{E}(ACU(x) \mid v = 1) = \mathbb{E}(\widehat{\mathbb{P}}(v \mid x) - \widehat{\mathbb{P}}(b \mid x) \mid v = 1) =$$

$$\mathbb{E}(\widehat{\mathbb{P}}(v \mid x) \mid v = 1) - \mathbb{E}(\widehat{\mathbb{P}}(b \mid x) \mid v = 1)$$

המעבר הראשון הוא לפי הגדרת ה-ACU, והמעבר השני נובע מלינאריות התוחלת.

$$= \mathbb{E}(f_v(x) \mid v = 1) - \mathbb{E}(f_b(x) \mid v = 1)$$

ההסתברויות המשוערכות הן הפלטים של המודלים.

$$= \int_0^1 \mathbb{P}(f_v(x) = \pi \mid v = 1) \cdot \pi d\pi - \int_0^1 \mathbb{P}(f_b(x) = \pi \mid v = 1) \cdot \pi d\pi \tag{4}$$

נעזרנו בנוסחה הידועה לתוחלת מותנית של משתנה רציף $\mathbb{E}(\mathrm{X}|\mathrm{Y} = \mathrm{y}) = \int_{-\infty}^{\infty} \mathbb{P}(X = x|Y = y)xdx$.

התוחלת שווה לאינטגרל על כל π אפשרי שהמודלים יכולים להפיק, כאשר המשתנה מוגדר רק מ-0 עד 1

כיוון שהוא הסתברות. ופלט זה מכפילים בהסתברות שלו להיות מופק מהמודל בהינתן ההנחה שהמאושפזים שייכים לאוכלוסיית ה-VTE.

האיבר הראשון ב-(4):

$$= \int_0^1 \frac{\mathbb{P}(v=1 \mid f_v(x)=\pi) \cdot \mathbb{P}(f_v(x)=\pi) \cdot \pi}{\mathbb{P}(v=1)} d\pi$$

השתמשנו בחוק בייס $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$.

$$= \int_0^1 \frac{\pi \cdot \mathbb{P}(f_v(x)=\pi) \cdot \pi}{\mathbb{P}(v=1)} d\pi$$

הצבנו את הנחת הקליברציה.

$$= \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)} \int_0^1 \pi^2 d\pi$$

ההסתברות במונה קבועה לכל π ולכן ניתן להוציא אותה מהאינטגרל כתוצאה מהנחה (1), והמכנה קבוע גם כן, לא תלוי ב-π ויכול לצאת מהאינטגרל.

לבסוף לאחר שפותרים את האינטגרל מתקבל שתוחלת ההסתברות ל-VTE באוכלוסיית הקרישיות:

$$\mathbb{E}\left(\widehat{\mathbb{P}}(v|x) \;\middle|\; v=1\right) = \frac{1}{3} \cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)}$$

האיבר השני ב-(4):

$$= \int_0^1 \frac{\mathbb{P}(v=1 \mid f_b(x)=\pi) \cdot \mathbb{P}(f_b(x)=\pi) \cdot \pi}{\mathbb{P}(v=1)} d\pi$$

השתמשנו בחוק בייס, כמו מקודם.

$$= \int_0^1 \frac{(1-\pi)\cdot \mathbb{P}(f_b(x)=\pi)\cdot \pi}{\mathbb{P}(v=1)} d\pi$$

הצבנו את טענת עזר (3).

$$= \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)} \int_0^1 (1-\pi)\cdot \pi d\pi = \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)} \left(\frac{1}{2}-\frac{1}{3}\right)$$

הוצאנו את המונה והמכנה מהאינטגרל מאותם צידוקים כמו קודם לכן ופתרנו אותו.

לבסוף, מתקבל שתוחלת ההסתברות לדימום באוכלוסיית הקרישיות הינה:

$$\mathbb{E}(\widehat{\mathbb{P}}(b\,|\,x)\ |\ v=1) = \frac{1}{6}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)}$$

וסך הכל תוחלת ה-ACU לאוכלוסיה זו:

$$\mathbb{E}(ACU(x)\ |\ v=1) = \frac{1}{3}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)} - \frac{1}{6}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)} = \frac{1}{6}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)}$$

כעת באופן דומה מאוד נמצא את תוחלת ה-ACU של אוכלוסיית המדממים. כל המעברים יהיו אותו דבר כיוון שההנחה הינה ששני המודלים הם בעלי קליברציה מושלמת, רק שהפעם האינטגרלים יהיו בסדר הפוך והמכנה יהיה ההסתברות שמאושפז יהיה שייך לאוכלוסיית המדממים.

$$\mathbb{E}(ACU(x)\ |\ b=1) = \mathbb{E}(\widehat{\mathbb{P}}(v\,|\,x) - \widehat{\mathbb{P}}(b\,|\,x)\ |\ b=1) =$$

$$\mathbb{E}(\widehat{\mathbb{P}}(v|x)\ |\ b=1) - \mathbb{E}(\widehat{\mathbb{P}}(b|x)\ |\ b=1) =$$

$$\frac{1}{6}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(b=1)} - \frac{1}{3}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(b=1)} = -\frac{1}{6}\cdot \frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(b=1)}$$

לסיכום, כיוון שהסתברויות לא יכולות להיות שליליות גם המונה והמכנה משני צידי אי השוויון לא יכולים להיות שליליים, ומתקיים

$$\mathbb{E}(ACU(x)|\ v=1) = \frac{1}{6}\cdot\frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(v=1)} > -\frac{1}{6}\cdot\frac{\mathbb{P}(f(x)=\pi)}{\mathbb{P}(b=1)} = \mathbb{E}(ACU(x)|\ b=1)$$

$$\mathbb{E}(ACU(x)\ |\ v=1) > \mathbb{E}(ACU(x)\ |\ b=1) \qquad \blacksquare$$

נספח ב' - הוכחה למדגמים בגודל סופי

תהי $\mathcal{S}$ קבוצת מדגם בגודל סופי של מאושפזים שמחולקת לקבוצת מאושפזים שעברו אירוע קרישיות שנסמנה ב-$\mathcal{V}$ וקבוצת מאושפזים שדיממו דימום חמור שנסמנה $\mathcal{B}$, כך ש-

$$\mathcal{S} = \mathcal{V} \cup \mathcal{B}$$

$$\mathcal{V} \cap \mathcal{B} = \emptyset$$

כלומר, כל מאושפז ב-$\mathcal{S}$ שייך לקבוצת הקרישיות או לקבוצה הדימום, אך לא לשתיהן יחד. נסמן ב-|X| את גודל הקבוצה X, כלומר את מספר הדוגמאות בה. מהגדרת המדגם $\mathcal{S}$ מתקיים

$$|\mathcal{S}| = |\mathcal{V}| + |\mathcal{B}|$$

כל מאושפז מוגדר על ידי x שהוא וקטור המשתנים (features), v שהוא משתנה בינארי שמגדיר האם המאושפז שייך לקבוצת הקרישיות או לא, ו-b, משתנה בינארי שמגדיר האם המאושפז שייך לקבוצת המדממים או לא.

ויהיו מודלים f_v לחיזוי VTE ו-f_b לחיזוי דימום. נוכיח שאם המודלים בעלי קליברציה מושלמת, כלומר

$$\mathbb{P}\big(v = 1 \mid f_v(x) = \pi\big) = \pi$$

$$\mathbb{P}\big(b = 1 \mid f_b(x) = \pi\big) = \pi$$

$$\forall \pi \in \{\pi_1, \pi_2, \dots, \pi_k\}$$

אזי בהכרח ממוצע ה-ACU של קבוצת המאושפזים שעברו VTE גדול מממוצע ה-ACU של המאושפזים המדממים. פורמלית:

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} ACU(x_i) > \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} ACU(x_j)$$

כיוון שבעת בניית גרף קליברציה מחלקים את המדגם ל-k bins בגדלים שווים כאשר בכל bin יש את כל הדוגמאות אשר קיבלו הסתברות מסוימת מהמודל, נניח שגדלי ה-bins שווים:

$$|\mathcal{S}_{\pi 1}| = |\mathcal{S}_{\pi 2}| = \cdots = |\mathcal{S}_{\pi k}| \tag{1}$$

נסמן bin שהמודל הפיק הסתברות π_i לכל הדוגמאות שלו בתור הקבוצה $\mathcal{S}_{\pi i}$. כיוון שגדלי ה-bins שווים נוכל להשתמש בסימון כללי לגודל כל bin כזה כ-$|\mathcal{S}_{\pi}|$, ונניח שמספר ה-Bins הוא שווה גם עבור המודל שחוזה VTE וגם עבור המודל שחוזה דימום ונסמנו בתור k.

בנוסף, נניח ש-k הסתברויות π השונות שיש לנו מחלקות את כל מרחב ההסתברויות, קרי את הקטע [0,1], למקטעים שווים בגודלם כאשר k גדול מ-1. כלומר

$$\frac{1}{k-1} = \pi_k - \pi_{k-1} = \cdots = \pi_2 - \pi_1 \tag{2}$$

לדוגמה, אם $k = 3$, אז חילקנו את הקטע [0,1] לטווחים של חצי, ויש לנו:

$$\pi_1 = 0$$

$$\pi_2 = 0.5$$

$$\pi_3 = 1$$

$$\frac{1}{3-1} = 1 - 0.5 = 0.5 - 0$$

טענות עזר:

לכל $\pi \in \{\pi_1, \pi_2, \dots, \pi_k\}$ מתקיים:

$$|\mathcal{V}_{\pi i}| = |\mathcal{S}_{\pi}| \cdot \pi_i \tag{3}$$

כאשר נסמן את כל הדוגמאות ששייכות ל-$\mathcal{V}$ וקיבלו את ההסתברות π_i בתור הקבוצה $\mathcal{V}_{\pi i}$.

טענה זו נובעת מהגדרת הקליברציה. כלומר, מספר הפעמים שהמודל f_v הפיק את ההסתברות π_i וצדק (דהיינו אכן המאושפז עבר VTE ולכן הוא שייך לקבוצה $\mathcal{V}$) - שווה למספר הפעמים שהמודל חזה הסתברות זו בכללי, כפול ההסתברות. אם נחזור לדוגמה הקודמת שהובאה בהגדרת הקליברציה - אם המודל החזיר את ההסתברות 0.9 ב-20 מקרים שונים ואם המודל בעל קליברציה מושלמת - הוא יהיה צודק ב-18 מתוך 20 המקרים. וכיוון שלכל דוגמה במדגם היא $v = 1$ או שהיא $b = 1$ (כנאמר בהגדרת המדגם $\mathcal{S}$), אז מספר המדממים בדוגמה לעיל יהיה: $20 \cdot (1 - 0.9) = 20 - 18 = 2$

ובמקרה הכללי

$$|\mathcal{B}_{\pi i}| = |\mathcal{S}_{\pi}| \cdot (1 - \pi_i) \tag{4}$$

גם עבור f_b וההסתברויות שהוא מפיק האמור לעיל מתקיים כתוצאה מאותם הסברים, רק הפוך כי הפעם זו ההסתברות המשלימה:

$$|\mathcal{B}_{\pi i}| = |\mathcal{S}_{\pi}| \cdot \pi_i \tag{5}$$

$$|\mathcal{V}_{\pi i}| = |\mathcal{S}_{\pi}| \cdot (1 - \pi_i) \tag{6}$$

כאמור נרצה להוכיח ש-

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} ACU(x_i) > \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} ACU(x_j)$$

וכאשר נציב את הגדרת ACU ונפתח סוגריים

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} \widehat{\mathbb{P}}\left(v|x_i\right) - \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} \widehat{\mathbb{P}}\left(b|x_i\right) > \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} \widehat{\mathbb{P}}\left(v|x_j\right) - \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} \widehat{\mathbb{P}}\left(b|x_j\right)$$

לכן נוכל לטפל ולפתח את ארבעת הסכומים בנפרד.

נתחיל עם ההסתברות הממוצעת ל-VTE של קבוצה $\mathcal{V}$:

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\widehat{\mathbb{P}}\left(v|x_i\right)=\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}f_v(x_i)$$

המעבר נובע מכך שההסתברות המשוערכת ל-VTE זה הפלט של המודל f_v.

$$=\frac{1}{|\mathcal{V}|}\left[\sum_{i=1}^{|\mathcal{V}_{\pi 1}|}f_v(x_i)+\sum_{i=1}^{|\mathcal{V}_{\pi 2}|}f_v(x_i)+\cdots+\sum_{i=1}^{|\mathcal{V}_{\pi k}|}f_v(x_i)\right]$$

כאן אנו חילקנו את הסכום לתתי סכומים, כאשר בכל תת סכום כל המקרים ששייכים ל-$\mathcal{V}$ והם קיבלו הסתברות מסוימת. המעבר נכון מכיוון ש- $|\mathcal{V}_{\pi 1}|+|\mathcal{V}_{\pi 2}|+\cdots+|\mathcal{V}_{\pi k}|=|\mathcal{V}|$.

$$=\frac{1}{|\mathcal{V}|}\left[\sum_{i=1}^{|\mathcal{V}_{\pi 1}|}\pi_1+\cdots+\sum_{i=1}^{|\mathcal{V}_{\pi k}|}\pi_k\right]$$

המעבר הנ"ל הינו נכון כיוון שהמודל פולט הסתברות, ובכל סכום ישנה אותה הסתברות, כי כך הגדרנו את החלוקה לתתי הסכומים במעבר הקודם.

$$=\frac{1}{|\mathcal{V}|}\left[|\mathcal{V}_{\pi 1}|\cdot\pi_1+\cdots+|\mathcal{V}_{\pi k}|\cdot\pi_k\right]$$

כל π_i הוא קבוע בכל סכום, ואנו סוכמים אותו $|\mathcal{V}_{\pi i}|$ פעמים בכל סכום, לכן ניתן להכפיל אותו במספר הפעמים שהמודל הפיק אותו בקבוצה $\mathcal{V}$.

$$=\frac{1}{|\mathcal{V}|}\left[|\mathcal{S}_{\pi 1}|\cdot\pi_1\cdot\pi_1+\cdots+|\mathcal{S}_{\pi k}|\cdot\pi_k\cdot\pi_k\right]=\frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}\left[\sum_{i=1}^{k}\pi_i^2\right]$$

הצבנו את טענת העזר (3), ובמעבר הבא יכולנו להוציא את $|\mathcal{S}_\pi|$ מחוץ לסכום, בשל כך שהוא קבוע והוא שווה לכל $|\mathcal{S}_{\pi i}|$ כנאמר בהנחה (1).

סך הכל קיבלנו

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\widehat{\mathbb{P}}\left(v|x_i\right)=\frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}\left[\sum_{i=1}^{k}\pi_i^2\right]$$

כיוון שגם מודל f_b הוא בעל קליברציה מושלמת, כל המעברים הינם למעשה זהים ומקבילים למעברים שעשינו לעיל כאשר מחשבים את ממוצע ההסתברויות לדימום בקבוצה $\mathcal{B}$, רק שבמקום טענת עזר (3) נציב את טענת עזר (5). מתקבל

$$\frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\widehat{\mathbb{P}}\left(b|x_j\right)=\frac{|\mathcal{S}_\pi|}{|\mathcal{B}|}\left[\sum_{i=1}^{k}\pi_i^2\right]$$

כעת, נחשב את ההסתברות הממוצעת לדימום עבור קבוצת קבוצה $\mathcal{V}$ שאף דוגמה שם לא דיממה. נבצע מעברים דומים חוץ משינוי אחד:

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\widehat{\mathbb{P}}\left(b|x_i\right)=\frac{1}{|\mathcal{V}|}\left[\sum_{i=1}^{|\mathcal{V}_{\pi 1}|}\pi_1+\cdots+\sum_{i=1}^{|\mathcal{V}_{\pi k}|}\pi_k\right]=\frac{1}{|\mathcal{V}|}\left[|\mathcal{V}_{\pi 1}|\pi_1+\cdots+|\mathcal{V}_{\pi k}|\pi_k\right]$$

ביצענו את אותו פירוק הסכום לפי הסתברויות שעשינו קודם, רק שבמקרה זה f_b פולט אותן והוא עושה זאת לדוגמאות שהן b=0, כלומר - אף אחד מהם לא דימם באמת.

$$=\frac{1}{|\mathcal{V}|}\left[|\mathcal{S}_{\pi 1}|\cdot(1-\pi_1)\cdot\pi_1+\cdots+|\mathcal{S}_{\pi k}|\cdot(1-\pi_k)\cdot\pi_k\right]$$

השתמשנו כאן בטענת עזר (6), שנבעה מההסתברות המשלימה, כדי להציב במקום $|\mathcal{V}_{\pi i}|$.

$$=\frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}\left[\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i\right]$$

הוצאנו את $|\mathcal{S}_\pi|$ כיוון שהוא קבוע לכל π_i כנאמר בהנחה (1).

בהמשך לכך, בסך הכל קיבלנו

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\widehat{\mathbb{P}}\left(b|x_i\right)=\frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}\left[\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i\right]$$

כיוון שגם מודל f_v הוא בעל קליברציה מושלמת, כל המעברים הינם זהים ומקבילים למעברים שעשינו לעיל כאשר מחשבים את ממוצע ההסתברויות ל-VTE בקבוצת המדממים $\mathcal{B}$. מתקבל:

$$\frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\widehat{\mathbb{P}}\left(v|x_j\right)=\frac{|\mathcal{S}_\pi|}{|\mathcal{B}|}\left[\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i\right]$$

כעת יש בידינו את ארבעת הסכומים שהרכיבו את הטענה המקורית ואת ה-ACU הממוצע של שתי הקבוצות. נציב לסיכום:

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}ACU(x_i)=\frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}\left[\sum_{i=1}^{k}\pi_i^2-\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i\right]$$

$$\frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}ACU(x_j)=\frac{|\mathcal{S}_\pi|}{|\mathcal{B}|}\left[\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i-\sum_{i=1}^{k}\pi_i^2\right]$$

אנו רואים שאם נוכיח ש-

$$\sum_{i=1}^{k}\pi_i^2>\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i \tag{7}$$

אזי ה-ACU הממוצע של קבוצת הקרישיות $\mathcal{V}$ יהיה חיובי, זאת כיוון שהחיסור שבתוך הסוגריים יהיה חיובי, וה-ACU הממוצע של קבוצת המדממים $\mathcal{B}$ יהיה שלילי כיוון שהחיסור יהיה שלילי. הנ"ל מתקיים בגלל ש- $\frac{|\mathcal{S}_\pi|}{|\mathcal{B}|}$ ו-$\frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}$ הינם חיוביים, כי גדלי קבוצות לא ריקות הינם תמיד חיוביים.

כעת נוכיח את טענה (7). הגדרנו את איברי הטורים בהנחה (2) כך שלכל k גדול מ-1, מחלקים את הטווח [0,1] ל-$(k-1)$ מקטעים כך ש:

$$\frac{1}{k-1}=\pi_k-\pi_{k-1}=\cdots=\pi_2-\pi_1$$

צריך להוכיח

$$\sum_{i=1}^{k}\pi_i^2>\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i$$

נתחיל בלחשב את החיבור של שני הטורים

$$\sum_{i=1}^{k}\pi_i^2+\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i=\sum_{i=1}^{k}\pi_i^2+\sum_{i=1}^{k}\pi_i-\sum_{i=1}^{k}\pi_i^2=\sum_{i=1}^{k}\pi_i$$

פתחנו סוגריים, פירקנו את הטור השני לשני טורים נפרדים, וחיסרנו כך שהטורים הצטמצמו ונשארנו עם אחד.

נשתמש בנוסחה לסכום טור חשבוני $\sum_{i=1}^{n} i=\frac{n(a_1+a_n)}{2}$ ונקבל

$$\sum_{i=1}^{k}\pi_i=\frac{k(\frac{0}{k-1}+\frac{k-1}{k-1})}{2}=\frac{k}{2}$$

לכן

$$\sum_{i=1}^{k}\pi_i^2+\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i=\frac{k}{2}=\frac{2k}{4}$$

אם נוכיח ש-

$$\sum_{i=1}^{k}\pi_i^2>\frac{k}{4}$$

אזי בהכרח

$$\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i<\frac{k}{4}$$

ונגיע לנדרש בהוכחה.

נפתח את הטור

$$\sum_{i=1}^{k}\pi_i^2=\frac{0^2}{[k-1]^2}+\frac{1^2}{[k-1]^2}+\cdots+\frac{[k-1]^2}{[k-1]^2}$$

הפיתוח הוא לפי ההגדרה של האיברים של הטור בהנחה (2). נמשיך עם הוצאת גורם משותף

$$= \frac{1}{[k-1]^2}(1^2 + 2^2 + \cdots + [k-1]^2)$$

ונקבץ את הסכימה בסוגריים לטור בפני עצמו

$$= \frac{1}{[k-1]^2} \cdot \sum_{i=1}^{k-1} i^2$$

נעזר בנוסחה לסכום טור ריבועי $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$ וכאשר נציב $n = k-1$ נקבל

$$= \frac{1}{[k-1]^2} \cdot \frac{(k-1)(k-1+1)[2(k-1)+1]}{6}$$

נפתח סוגריים ונפשט את הביטוי

$$= \frac{1}{[k-1]^{\cancel{2}}} \cdot \frac{\cancel{(k-1)}k(2k-1)}{6}$$

כעת נקטין את הביטוי על ידי כך שנחסיר במספר גדול יותר את המונה, ונוציא גורם משותף

$$= \frac{1}{(k-1)} \cdot \frac{k(2k-1)}{6} > \frac{k(2k-2)}{(k-1)6} = \frac{2k\cancel{(k-1)}}{\cancel{(k-1)}6} = \frac{k}{3}$$

לבסוף, קיבלנו שהטור גדול מ-$\frac{k}{3}$ שאם מגדילים את המכנה מקיים

$$\sum_{i=1}^{k} \pi_i^2 > \frac{k}{3} > \frac{k}{4}$$

כנדרש. לבסוף, הוכחנו ש-

$$\sum_{i=1}^{k} \pi_i^2 > \frac{k}{4} > \sum_{i=1}^{k} (1-\pi_i) \cdot \pi_i$$

ולכן,

$$\sum_{i=1}^{k} \pi_i^2 > \sum_{i=1}^{k} (1-\pi_i)\cdot\pi_i \qquad \blacksquare$$

לבסוף, קיבלנו

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} ACU(x_i) = \frac{|\mathcal{S}_\pi|}{|\mathcal{V}|}\left[\sum_{i=1}^{k} \pi_i^2 - \sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i\right] > 0$$

$$\frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} ACU(x_j) = \frac{|\mathcal{S}_\pi|}{|\mathcal{B}|}\left[\sum_{i=1}^{k}(1-\pi_i)\cdot\pi_i - \sum_{i=1}^{k} \pi_i^2\right] < 0$$

$$\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|} ACU(x_i) > \frac{1}{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} ACU(x_j) \qquad \blacksquare$$

Abstract

Venous thromboembolism (VTE) is a significant cause of death worldwide, with most cases being related to the hospital stay itself. VTE occurs when there is a blood clot in the deep veins (Deep-vein thrombosis - DVT), usually in the legs or pelvis, which detaches from its initial position and drifts in the blood stream.

The risk of VTE depends on various known variables including age, active cancer, and certain medical conditions. Over the years, an index called the Padua Prediction Score was built, which is based on a weighted sum of variables and is supposed to help doctors assess the risk for VTE, and to decide whether or not to treat the hospitalized patient with anti-coagulants.

The data bases we worked on were including patients from the Rabin Medical Center and were received by Dr. Shahaf Shiver from the Internal Medicine Department.

One of our goals was to recalibrate the weight of the Padua variables and add new variables in order to estimate the risk more accurately. After creating the new model, its performance was compared to the performance of the Padua, and a statistical test was conducted to verify that the difference between the performances is significant.

Since there is a risk of excessive use of anti-coagulants that will lead to life-threatening major bleeding, we wanted to create a tool that would also take into consideration this risk. After creating the additional model for predicting major bleeding, the ACU (Anti-Coagulant Utility) index was defined, which refers to both sides of the coin when deciding whether to treat using anti-coagulants or not. In order to motivate a certain type of models from the field of machine learning, an attempt was made to theoretically prove

that models with a good calibration will contribute to the ACU in the task of dividing the hospitalized into those who will benefit from blood thinners, and those who can be harmed by them with a high probability. In order to prove that the ACU was successful in its mission, we estimated the expected value of ACU for the populations of the hospitalized bleeding, the hospitalized who had a VTE event and the healthy hospitalized using the averages of the samples from the data base and by creating a confidence interval.

This work was carried out under the supervision of

Prof. Ron Unger

from the Faculty of Life Sciences,

Bar-Ilan University.

BAR-ILAN UNIVERSITY

Assessing the Risk of Embolic Venous Thrombosis and Major Bleeding while Inspecting Their Tradeoff Using Probability Estimation Algorithms

Vadim Litvinov

Submitted in partial fulfillment of the requirements for the Master's Degree
In the Faculty of Life Sciences
Bar-Ilan University

Ramat-Gan, Israel 2022

BAR-ILAN UNIVERSITY

Assessing the Risk of Embolic Venous Thrombosis and Major Bleeding while Inspecting Their Tradeoff Using Probability Estimation Algorithms

Vadim Litvinov

Submitted in partial fulfillment of the requirements for the Master's Degree
In the Faculty of Life Sciences
Bar-Ilan University

Ramat-Gan, Israel 2022