מדריך מקצועי להתממת מידע רפואי

סופיה דראושה
רועי כהן
ד"ר מיטל אבגיל צדוק
רון חרמון
ד"ר אלכס בלכמן
עוז שנהב

תמנ"ע, יחידת ה-Big Data של משרד הבריאות, אגף טכנולוגיות דיגיטליות ודאטה

מחלקת רגולציה, אגף בריאות דיגיטלית

מחלקת חדשנות וטכנולוגיה, הרשות להגנת הפרטיות

תוכן העניינים

רקע למסמך 3
התממה וזיהוי חוזר 3
תהליך ההתממה 4
1. מיפוי המידע 4
2. הסרת מזהים ישירים 4
3. יישום שיטות התממה 4
4. הערכת הסיכון לזיהוי חוזר 5
5. ניהול מתמשך של סיכוני זיהוי וחשיפה 5
התממה ושימושיות 6
אמצעים משלימים להתממה: עיצוב לפרטיות, אבטחת מידע ותהליכים ארגוניים 7
נספח א: שיטות ודוגמאות להתממת מידע 8
השמטת ערך – Attribute Suppression 8
פסאודונימיזציה/קידוד – Pseudonymization/Coding 8
הכללה – Generalization 9
הזחת נתונים – Indentation 10
התממה פונקציונלית – Functional Anonymization 10
נספח ב: הדגמה של שיטות התממה לסוגי מידע נבחרים 12
התממת כתובות 12
התממת טקסט חופשי 12
התממת מידע קליני 13
התממת קבצי תמונה (כגון תצלומים, קבצי דימות או סריקות MRI) 14
מקורות מידע מרכזיים 14

רקע למסמך

התממה היא תהליך שמטרתו להביא להסרת מאפיינים או שינוי ערכים הנכללים במאגר המידע על מנת לצמצם או למנוע זיהוי של נושא המידע.[1] נתונים רפואיים מכילים מידע אישי בעל רגישות מיוחדת,[2] לעתים בהיקפים גדולים במיוחד, והתממה של מידע רפואי היא כלי משמעותי לצורך הגנה על פרטיותם של האנשים שנתוניהם מופיעים במאגרי המידע.

מסמך זה מיועד לאנשי המקצוע המנחים, מבצעים ומאשרים התממה של מידע רפואי. מטרת המסמך היא לסקור את השיטות הקיימות, לספק דוגמאות רלוונטיות למידע רפואי ולהציג שיקולים ואתגרים נפוצים בתהליכי ההתממה. המסמך מבוסס על פרסומים של רשויות הגנת פרטיות בישראל ובעולם המפורטים בסוף מסמך זה ועל הניסיון שנצבר בשירותי התממה שניתנו למערכת הבריאות הישראלית בכלל ובפרט, במסגרת פרויקט תמנ"ע (תשתית מחקר לנתוני עתק) של משרד הבריאות. המסמך מתמקד בשיטות יישומיות בתחומי המידע ואינו מפרט היבטים משפטיים הכרוכים בעיבוד מידע אישי.

בהיבט המשפטי יובהר, כי עצם עיבוד המידע האישי המזוהה בתוך הארגון, לרבות תוך שימוש בשיטות התממה, כפוף להוראות חוק הגנת הפרטיות, התשמ"א-1981 ולהוראות תקנות הגנת הפרטיות (אבטחת מידע), התשע"ז-2017. על ארגונים לוודא כי עיבוד המידע עומד בהוראות החוק והתקנות וכן להטמיע תהליכים ארגוניים נוספים להגנה על המידע.

התממה וזיהוי חוזר

התממה היא תהליך שנועד לצמצם את הסיכון לפגיעה בפרטיות, תוך שמירה על האפשרות להפיק תובנות סטטיסטיות או מחקריות מן הנתונים. עם זאת, התממה איננה פעולה פשוטה של מחיקת פרטים, אלא תהליך שיטתי הדורש ניתוח של מאפייני המידע ושל הסיכונים האפשריים לזיהוי חוזר.

בתהליך ההתממה חשוב להבחין בין מזהים ישירים ומזהים עקיפים:

מזהים ישירים (Direct identifiers) הם נתונים העשויים לאפשר לזהות אדם ללא צורך בשילוב עם נתונים נוספים, למשל: שם, מספר זהות, כתובת, מספר טלפון או אחד הפרטים המזהים הנכללים בהגדרת "מידע אישי" בחוק הגנת הפרטיות.

מזהים עקיפים (Indirect identifiers), לעומת זאת, אינם מצביעים לבדם על אדם מסוים, אך בשילוב עם נתונים נוספים הם עשויים להוביל לזיהויו. דוגמאות לכך הן מקום עבודה, מקצוע, תאריך לידה, או אזור מגורים. כאשר מצרפים מספר מזהים עקיפים יחד, נוצרת לעיתים "טביעת אצבע" ייחודית המאפשרת זיהוי של אדם יחיד, גם אם כל שדה כשלעצמו אינו מאפשר את זיהויו.

זיהוי חוזר (Re-identification) הוא תהליך שבו נחשפת זהותו של אדם, קבוצה מסוימת או כלל בני האדם שמידע על אודותיהם נכלל במאגר המידע, הגם שהמידע המצוי במאגר אינו מאפשר את זיהוים כשלעצמו, על סמך מזהים ישירים או עקיפים. כך, גם אם המידע אינו כולל מזהים ישירים, ניתן לקשור בין הנתונים לבין אדם מסוים באמצעות הצלבה עם מקורות חיצוניים, ידע מוקדם, או ניתוח סטטיסטי מתקדם. היכולת לבצע זיהוי חוזר תלויה במאפייני הנתונים עצמם ובתוך כך כמותם, איכותם, רמת הפירוט שלהם, ומידת הייחודיות של הרשומות. ככל שהמידע מפורט יותר, כך עולה הסיכון שניתן יהיה לזהות את נושאי המידע בדרך של זיהוי חוזר.

[1] מדריך לטכנולוגיות מגבירות-פרטיות, הרשות להגנת הפרטיות, 2025.
[2] ראו הגדרת "מידע אישי" ו-"מידע בעל רגישות מיוחדת" בסעיף 3 לחוק הגנת הפרטיות, התשמ"א-1981.

תהליך ההתממה

תהליך ההתממה מבוסס על חמישה שלבים מרכזיים:[3] (1) מיפוי המידע, (2) הסרת מזהים ישירים, (3) יישום שיטות התממה, (4) הערכת הסיכון לזיהוי חוזר, ו-(5) ניהול מתמשך של סיכוני הזיהוי והחשיפה. גישה זו מדגישה את האיזון הדרוש בין שמירה על פרטיות נושאי המידע לבין שמירה על השימושיות והערך של הנתונים לצורכי מחקר, תכנון או קבלת החלטות. נבקש לסקור בקצרה שלבים אלה:

1. מיפוי המידע

לפני כל פעולה של התממה יש למפות את סוגי הנתונים שבמאגר המידע, להבין את מטרות השימוש בהם, ולזהות אילו נתונים עשויים להוביל לזיהוי ישיר או עקיף של נושאי המידע. בתוך כך, נדרשת סקירה של מבנה הנתונים, רמות הפירוט, סוגי המשתנים (דמוגרפיים, רפואיים, התנהגותיים וכדומה), וכן בחינת מקורות המידע החיצוניים שיכולים לשמש לצורך הצלבה.

בשלב זה יש לשאול מספר שאלות מרכזיות:

1. האם הגישה למידע תהיה בסביבה פתוחה (לדוגמא, פרסום באינטרנט) או סגורה ומאובטחת?
2. מהי מטרת השימוש במידע לאחר ההתממה? מי אמור להשתמש במידע ובאיזה אופן?
3. אילו פרטים נדרשים לשם מימוש אותה מטרה, ואילו פרטים מיותרים וניתן להשמיט אותם?
4. האם קיימים שדות ייחודיים (כמו מספרי רישוי, תאריכים מדויקים, או מיקומים גיאוגרפיים) שעשויים להוביל לבידול של אדם (כלומר, זיהוי של אדם מסוים שנתוניו קיימים במאגר המידע)?
5. מהו גודל הקבוצה שלה נוגע המידע (אוכלוסייה רחבה או קבוצה קטנה)?

מיפוי זה מאפשר להבין את פוטנציאל הזיהוי החוזר הטמון במידע ולבחון מראש את סוגי הסיכונים. לדוגמה, מאגר מידע רפואי הכולל נתונים על מטופלים מעיר קטנה עשוי להיות רגיש בהרבה ממאגר מידע כלל-ארצי שבו יש ריבוי של מקרים דומים.

2. הסרת מזהים ישירים

מזהים ישירים הם נתונים העשויים לאפשר לזהות את האדם באופן מידי, ללא צורך בהצלבה עם מידע נוסף. בין אלה נכללים שם מלא, מספר זהות, כתובת מגורים, מספר טלפון, כתובת דוא"ל, תצלום פנים, טביעת אצבע, או כל מזהה ביומטרי אחר. בארה"ב, חוק ניידות ואחריות ביטוח בריאות (HIPAA) מציין 18 קבוצות של מזהים הנדרשים להסרה כתנאי הכרחי (אך לא מספיק) לכך שהמידע ייחשב כלא ניתן לזיהוי חוזר.[4]

הסרת המזהים הישירים או קידודם הם שלבים הכרחיים בכל תהליך התממה, אך אינם מספיקים בפני עצמם.

חשוב לציין כי הסרה של מזהים ישירים צריכה להיעשות תוך שמירה על עקיבות (consistency) בנתונים. אם אותו אדם מופיע בכמה רשומות, יש לוודא שהמזהה הפנימי או הקוד המחליף שלו (פסאודונים) עקבי בכל הרשומות, כדי לאפשר ניתוחים סטטיסטיים מבלי לחשוף את זהותו.

3. יישום שיטות התממה

לאחר הסרת המזהים הישירים, יש ליישם שיטות התממה ביחס למידע שנותר במאגר המידע. בין השיטות הנפוצות בהתממת מידע רפואי:

1. **השמטת ערך** (Attribute Suppression): מחיקת ערך (עמודה מטבלה) שאינו נדרש למטרת עיבוד המידע.
2. **פסאודונימיזציה/קידוד** (Pseudonymization/Coding): החלפת פרטים מזהים בערכים פיקטיביים או בקודים.
3. **הכללה** (Generalization): החלפת ערכים מדויקים בערכים כלליים יותר, לדוגמה המרת גיל מדויק לטווח גילאים (30-34).

---

[3] לפי מדריך להתממה של רשות הגנת המידע בסינגפור (PDPC): https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/guide-to-basic-anonymisation-(updated-24-july-2024).pdf.

[4] לפירוט קבוצות של מזהים הנדרשים להסרה לפי חוק HIPAA: קישור (סעיף 164.514 (b)).

4. **הזחת נתונים** (Indentation): הוספה או החסרה של מרווח מוגדר לכל נתון מספרי.
5. **התממה פונקציונלית** (Functional Anonymization): יצירת ערך חישובי שמשמר את הנדרש עבור צורך מחקרי/אנליטי תוך כדי הסרת מידע שיש בו סיכון לזיהוי.

פירוט שיטות אלו מופיע **בנספח א'**. השיטות נבדלות זו מזו ברמת ההשפעה על איכות הנתונים וברמת ההגנה שהן מספקות ולכן יש לשקול מאפיינים אלו מול מאפייני המידע ומידת דיוק ופירוט הנתונים הנדרשת לשם השגת המטרה המבוקשת.

בנוסף לשיטות התממה המפורטות במסמך זה קיימות גישות ושיטות נוספות, המתוארות במדריכים של הרשות להגנת הפרטיות (ראו ה"ש 1) ורשויות הגנת פרטיות נוספות, ביניהן רשות הגנת המידע בסינגפור שמפרטת בהרחבה שיטות והיבטים שונים של התממה (ראו ה"ש 3).

4. הערכת הסיכון לזיהוי חוזר

לאחר יישום שיטות ההתממה השונות על מאגר מידע, יש להעריך את מידת הסיכון לזיהוי חוזר, שהוא באמור, הסיכון לחשיפת לזהותו של אדם, קבוצת בני אדם או כלל בני האדם שמידע על אודותיהם נכלל במאגר המידע. הערכת הסיכון לזיהוי חוזר הוא תהליך המשלב מספר רב של תחומים ומושפע ממאפייני המידע, אופן ההנגשה של המידע ופרמטרים רבים נוספים. ניתן לקרוא על שיטות להערכת הסיכון בין היתר במסמך ההנחיות של מחלקת הבריאות ושירותי האנוש של ארצות הברית.[5]

לדוגמה, נניח שמאגר מידע רפואי עבר התממה והוסרו ממנו מזהים ישירים כמו שם ומספר זהות, אך נותרו בו נתונים כגון גיל, מין, יישוב ותאריך אשפוז. כשלעצמו, המאגר עשוי להיתפס כמותמם. אולם אם המאגר נגיש לגורם חיצוני, ניתן להצליב אותו עם מאגרי מידע פומביים – למשל פרסומים ברשתות חברתיות, אתרי חדשות מקומיים או מאגרי מידע של הרשות המקומית וכך להגיע לזיהוי חוזר של נושא המידע.

כך, פרסום פומבי על "תושב/ת יישוב קטן בן/בת 47 שאושפז/ה ביום מסוים" עשוי לאפשר לצמצם את קבוצת האנשים הרלוונטית עד לאדם יחיד, ובפועל לחשוף את זהותו ואת מצבו הרפואי. ככל שהמאגר כולל אוכלוסייה קטנה יותר, נתונים ייחודיים יותר או תאריכים מדויקים יותר, כך גדל הסיכון לזיהוי חוזר.

דוגמה מפורסמת לזיהוי חוזר מתוך מאגר מידע שעבר התממה ידועה בשם 'מקרה מושל מסצ'וסטס'. במקרה זה פורסמו נתונים רפואיים שעברו התממה, אך באמצעות הצלבה עם מקורות מידע נוספים התאפשר להגיע לזיהוי חוזר של רשומות רפואיות, ביניהן רשומה של מושל מסצ'וסטס עצמו.[6]

5. ניהול מתמשך של סיכוני זיהוי וחשיפה

השלב האחרון אינו חד-פעמי אלא תהליך מתמשך של ניהול סיכונים, ולפיכך יש להתייחס אל התממה **כתהליך** (להבדיל מפעולה או סדרת פעולות שמבוצעות באופן חד-פעמי עובר לתחילת השימוש במידע). יש להכיר בכך שסיכוני הזיהוי החוזר משתנים עם הזמן. יכולות אנליטיות חדשות, זמינות של מקורות נתונים נוספים, או שינוי בהקשרים חברתיים עשויים להפוך מאגר מידע בטוח לכאורה לפגיע. לכן, על הארגון לפתח מדיניות ניהול סיכונים לזיהוי חוזר, הכוללת בחינה תקופתית של רמת ההתממה ובתוך כך בדיקה מחדש של הנתונים לאור התפתחויות טכנולוגיות או זמינות של מידע חדש, בחינת השיטות שהופעלו ותיקוף ההנמקות לבחירתן. כמו כן, יש להבטיח שימוש באמצעים ארגוניים נאותים ואבטחת מידע ברמה הולמת, כולל בקרת גישה והרשאות, הגבלת הגישה למידע המותמם רק למורשים ובתנאים מוגדרים מראש, הדרכת עובדים וחוקרים ביחס לסיכוני שימוש וזיהוי חוזר.

---

[5] הנחיות של מחלקת הבריאות ושירותי האנוש של ארצות הברית (באנגלית: Department of Health and Human Services, בראשי תיבות: HHS) בנוגע לשיטות להסרת זיהוי של מידע בריאותי מוגן בהתאם לחוק הפרטיות של חוק ניידות ואחריות ביטוח בריאות (HIPAA) – קישור.

[6] הרחבה על המקרה מופיעה במדריך טכנולוגיות מגבירות-פרטיות (דוגמה: הצלבת מידע לזיהוי חוזר של מידע שעבר התממה – מושל מסצ'וסטס).

התממה ושימושיות

תהליך ההתממה כרוך לעיתים בהשמטת רשומות, שינוי ערכים, עיגול נתונים או הכללתם בקטגוריות רחבות יותר. מטבע הדברים, כל פעולה שמטרתה לצמצם את פוטנציאל הזיהוי פוגעת גם בשימושיות של הנתונים. לכן, בבואנו לבצע התממה יש לאזן בין הגנה על הפרטיות לבין שמירה על ערכו המדעי, המחקרי או התפעולי של המידע.

לפי עקרון מזעור המידע (Data Minimization), שהוא אחד מעקרונות היסוד של תחום הגנת הפרטיות, יש לצמצם את כמות הנתונים לאלה ההכרחיים בלבד לשם הגשמת המטרה שלשמה נאספו. עם זאת, צמצום יתר, למשל כתוצאה ממחיקת משתנים חשובים או הכללה גסה מדי, עלול להפוך את הנתונים לחסרי ערך למדע, למחקר או למדיניות הציבורית. על כן, כדאי להתאים את אופן יישום ההתממה לרמת השימושיות הנדרשת, כמובן תוך תכנון מושכל והקפדה על רמת הגנה גבוהה ככל הניתן על פרטיותם של מי שהמידע על אודותיהם מצוי במאגר המידע.

לדוגמה, בבסיס נתונים רפואי שנועד לחקור את הקשר בין גיל המטופלים לבין הצלחת טיפול אונקולוגי, יש לבחון מהם טווחי הגילאים שביחס אליהם מתבצע המחקר, ולבצע עיגול של המטופלים שנתוניהם נכללים במאגר בהתאם. כך למשל, ייתכן שניתן להסתפק בחלוקת הגילאים לעשורים שונים (השפעת הטיפול בגילאי 30-40, 50-60 וכו'), מבלי לפגוע בערכם של הנתונים הנדרשים לצורך המחקר, תוך צמצום המידע אשר נחשף ביחס למטופלים והקטנת הסיכון לזיהויים החוזר. אם יימצא כי לצורך קבלת מענה לשאלת המחקר לא ניתן לקבץ את גילאי המטופלים, יהיה צורך לבחון האם קיים נתון אחר במאגר המידע שניתן להכליל, על מנת להקטין את הסיכון לזיהויים החוזר של המטופלים.

אמצעים משלימים להתממה: עיצוב לפרטיות, אבטחת מידע ותהליכים ארגוניים

המסמך אינו עוסק באמצעי אבטחת מידע ותהליכים ארגוניים ויש לפנות למקורות נוספים להגדרות מקיפות בתחומים אלו. עם זאת, חשוב לזכור שהגנה על פרטיות הנבדקים והמטופלים איננה מתחילה ונגמרת בתהליך ההתממה בלבד. תכנון הפרויקט בהתאם לגישת העיצוב לפרטיות עוד בשלב המקדמי, לפני תחילת העיבוד והעבודה על הנתונים.[7]

על בעל השליטה במאגר המידע לוודא כי עיבוד המידע במאגר עומד בדרישות הדין, ובכלל זה בהוראות חוק הגנת הפרטיות, והוראות תקנות הגנת הפרטיות (אבטחת מידע). בתוך כך, נדרשת הקמה של סביבה ארגונית וטכנולוגית מאובטחת, אשר מצמצמת את הסיכון לגישה בלתי מורשית, לדליפת מידע או לשימוש לרעה במידע רפואי רגיש, ולהגנה מפני סיכוני אבטחת מידע נוספים. יש לוודא הפרדה, בהיקף ובמידע הסבירים האפשריים, בין המערכות המחוברות לסביבה הארגונית והטכנולוגית ולמאגר המידע, אשר ניתן מהן לגשת למידע, לבין מערכות מחשוב אחרות המשמשות את בעל השליטה.

חשוב להבטיח אחסון בסביבות עבודה מאובטחות, תוך שימוש בערוצי תקשורת מוצפנים כגון רשתות וירטואליות פרטיות (VPN) או פרוטוקולי אבטחת תקשורת (TLS). בסביבות אלה חשוב, בין היתר, להגדיר הגבלות טכניות ואמצעי הגנה למניעת העתקת נתונים, צילום מסך, חיבור התקנים ניידים למערכות או הורדה של קבצים רגישים למחשבים אישיים. כך תישמר שליטה מלאה במידע גם בעת עבודה שוטפת של צוותי מחקר וניתוח נתונים.

מרכיב נוסף הוא בקרה על הוצאת קבצים מן הסביבה המאובטחת. כל בקשה לייצוא נתונים חייבת לעבור תהליך בדיקה שבו נבחן אם הקובץ המיוצא עבר התממה כנדרש ואינו כולל מידע אישי או מזהים ישירים. לעיתים נהוג להטמיע מערכת "שער יציאה" (Data Release Gateway) המוודאת את עמידת הקבצים בדרישות פרטיות לפני שחרורם.

כמו-כן, במקרים הנדרשים, יש להבטיח קיומם של אישורים לעריכת המחקרים, כגון אישורי ועדת הלסינקי למחקר רפואי או חוזי התקשרות המגדירים זכויות, אחריות וחובות ביחס לשימוש בנתונים. מסמכים אלה מגדירים את גבולות השימוש, את מי שמוסמך לגשת למידע, ואת התנאים שבהם ניתן לעבדו או לשתפו.

דוגמה ליישום אמצעים משלימים, המהווים רובד נוסף שתכליתו לצמצם את הסיכון לפגיעה בפרטיותם של נושאי המידע, היא בפרויקט תמנ"ע (תשתית מחקר לנתוני עתק) של משרד הבריאות. מודל הגישה למידע של תמנ"ע מבוסס על תפיסת Five Safes הנמצאת בשימוש במספר מדינות ומתוארת, בין היתר, בפרסום של נציב המידע האוסטרלי.[8] המסגרת בוחנת חמישה נושאים דרך מענה לחמש שאלות מרכזיות בחלק מהיערכות כוללת להגנה על המידע האישי:

1. **פרויקטים בטוחים**: האם השימוש המיועד במידע הוא הולם?
2. **אנשים בטוחים**: האם ניתן לסמוך על החוקרים שישתמשו במידע בצורה הולמת? מהם המסמכים (חוזים, או אחרים) שיש לנו עם המשתמשים?
3. **נתונים בטוחים**: האם קיים סיכון לחשיפה בנתונים עצמם?
4. **הגדרות בטוחות**: האם אופן מתן הגישה מגן מפני שימוש בלתי מורשה?
5. **תוצאות בטוחות**: האם תוצאות הפלט הסטטיסטיות אינן יוצרות סיכון לזיהוי חוזר וחשיפה?

אמצעים אלה יוצרים יחד מערכת הגנה רב-שכבתית שתכליתה למזער ככל הניתן את הפגיעה בפרטיות לכל אורך מחזור החיים של המידע, מהשלב הראשוני של איסוף המידע ועד השלב שבו נדרשת הפצה של תוצאות המחקר או הניתוח.[9]

---

[7] ראו באתר הרשות להגנת הפרטיות – עיצוב לפרטיות ומדריך עזר לביצוע תסקיר השפעה על פרטיות.

[8] O'Keefe, Christine & Otarepec, Stephanie & Elliot, Mark & Mackey, Elaine & O'hara, Kieron. (2017). The de-identification decision-making framework.

[9] ראו מדריך עזר לממונה על הגנת הפרטיות במחזור חיי המידע באתר הרשות להגנת הפרטיות.

נספח א: שיטות ודוגמאות להתממת מידע

הבחירה בשיטת ההתממה המתאימה תלויה במכלול רחב של גורמים: מאפייני הנתונים עצמם (כגון רמת הרגישות, כמות המשתנים הייחודיים ומידת ההטרוגניות של המדגם), מנגנוני ההגנה והבקרה הקיימים בארגון (כגון סביבה מאובטחת, בקרת גישה או נהלי שיתוף מידע), וכן רמת הסיכון המוגדרת על פי הערכת סיכוני הפרטיות. ההתממה אינה תהליך טכני בלבד אלא החלטה ניהולית, המחייבת איזון בין הגנה על הפרטיות לבין שמירה על ערך המחקר, הדיוק הסטטיסטי והשימושיות של הנתונים. נסקור ונדגים שיטות נפוצות להתממה.

השמטת ערך – Attribute Suppression

מחיקת ערך (עמודה מטבלה) שאינו נדרש למטרת עיבוד המידע. לדוגמה, אם המידע כולל את שם המטופל, מחלה מאובחנת ותוצאות בדיקות מעבדה ומטרת המחקר היא לבחון את השפעת סוג המחלה על רמות הסוכר בדם, ניתן להסיר את השדה המזהה (שם המטופל) ולשמור רק על משתנים רלוונטיים למחקר. כך מושגת הפחתת סיכון משמעותית מבלי לפגוע בתועלת המחקרית.

	A	B	C	D	E
1	Patient_id	age2010	machoz	gender	disease
2	1	26	מרכז	זכר	0
3	2	29	דרום	זכר	0
4	3	33	תל אביב-יפו	זכר	0
5	4	82	ירושלים	נקבה	0
6	5	28	דרום	זכר	0
7	6	81	דרום	נקבה	0
8	7	61	חיפה	זכר	0
9	8	27	דרום	נקבה	0
10	9	39	מרכז	זכר	0

<

	A	B	C	D
1	age2010	machoz	gender	disease
2	26	מרכז	זכר	0
3	29	דרום	זכר	0
4	33	תל אביב-יפו	זכר	0
5	82	ירושלים	נקבה	0
6	28	דרום	זכר	0
7	81	דרום	נקבה	0
8	61	חיפה	זכר	0
9	27	דרום	נקבה	0
10	39	מרכז	זכר	0

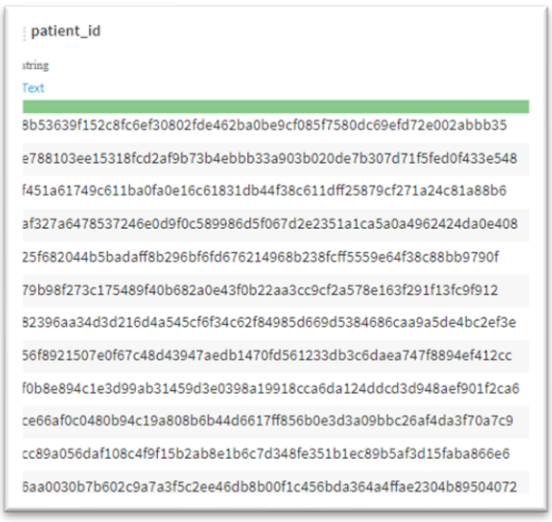
דוגמא להשמטה של משתנה תעודת זהות (Patient_id) באקסל .
מימין – סט הנתונים המקורי, **משמאל** – סט הנתונים החדש ובו המשתנה Patient_id הוסר).

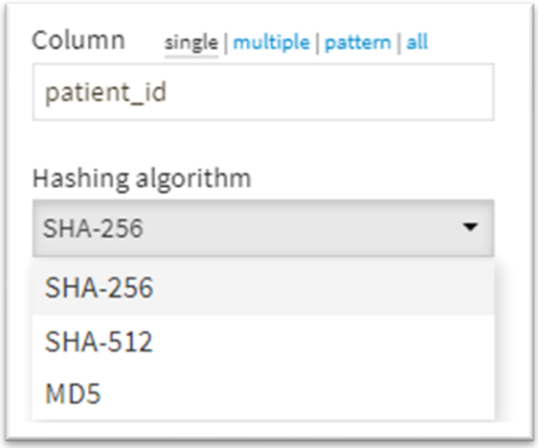
יצוין כי השמטה של שם המטופל ומספר תעודת הזהות בלבד, לא יובלו, ככלל, להיחשב כאמצעי התממה מספקים, לשם הגנה על פרטיותם של המטופלים שמידע על אודותיהם מצוי במאגר המידע.

פסאודונימיזציה/קידוד – Pseudonymization/Coding

החלפת נתונים מזהים (ישירים או עקיפים) בערכים פיקטיביים או בקודים. לדוגמה, ניתן להחליף מספרי זהות, שמות או כתובות במחרוזות אקראיות או ברצפים אלפאנומריים כדי להגן על המידע האישי אך לאפשר קישור בין מספר רשומות השייכות לאותו אדם. ישנם אופני יישום רבים של מנגנוני פסאודונימיזציה, וביניהם פרוטוקולי ערבול (hashing), חלקם זמינים בקוד פתוח.[10]

[10] בשימוש בקוד פתוח ישנם יתרונות רבים, אך בצידם גם סיכונים בהיבטי הגנת הפרטיות ואבטחת המידע שחשוב להכיר. לעקרונות שעל פיהם יש לפעול בעת שימוש בקוד פתוח ראו בקישור.



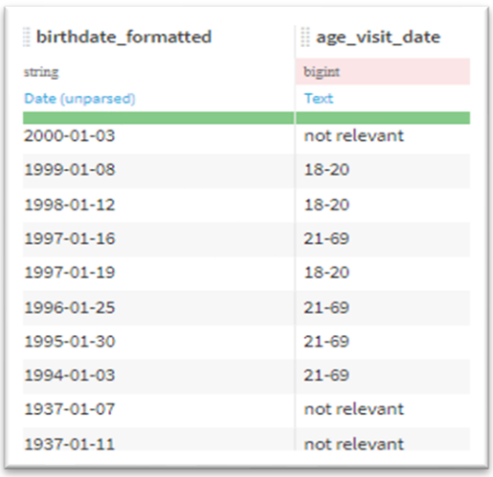


דוגמא ליצירת מזהה (פסאודונים) באמצעות קידוד משתנה תעודת זהות (patient_id).
מימין - בחירה בין מספר פרוטוקולי ערבול אפשריים.
משמאל – תוצאת הערבול עבור המשתנה.

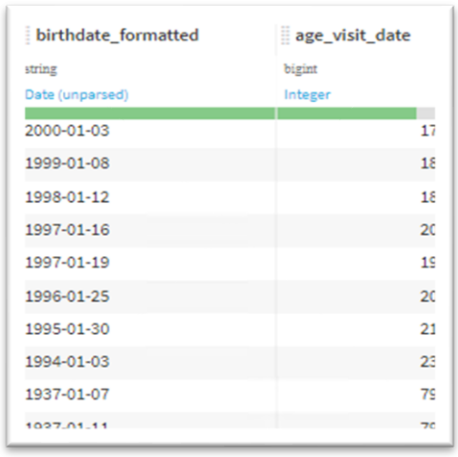
הכללה – Generalization

טשטוש מידע אישי על ידי הפחתת רמת הפירוט של הנתונים. במקום לשמור ערכים מדויקים, הנתונים מקובצים לקטגוריות, טווחים או קבוצות בעלות משמעות כללית יותר. תהליך זה מפחית את הסיכון לזיהוי חוזר של אדם מסוים, משום שהמאפיינים האישיים שלו "נבלעים" בתוך קבוצה רחבה של פרטים דומים.

<u>הכללה - דוגמה 1</u>: במקום לציין במסד הנתונים את הגיל המדויק של המטופל, ניתן להציג טווחי גיל, למשל: מתחת ל-18, 19-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, מעל 90 שנה.



birthdate_formatted	age_visit_date
string	bigint
Date (unparsed)	Text
2000-01-03	not relevant
1999-01-08	18-20
1998-01-12	18-20
1997-01-16	21-69
1997-01-19	18-20
1996-01-25	21-69
1995-01-30	21-69
1994-01-03	21-69
1937-01-07	not relevant
1937-01-11	not relevant



birthdate_formatted	age_visit_date
string	bigint
Date (unparsed)	Integer
2000-01-03	17
1999-01-08	18
1998-01-12	18
1997-01-16	20
1997-01-19	19
1996-01-25	20
1995-01-30	21
1994-01-03	23
1937-01-07	79

<u>הכללה - דוגמא 2</u>: הכללת משתנה גיל בעת הביקור (age_visit_date). במקום מספרים מדויקים של גיל, משתנה הגיל בקובץ לקבוצות בעלות משמעות למחקר:
מימין – מאגר המידע המקורי, **משמאל** – הכללה של המשתנה לקטגוריות (18-20, 21-69 ולא

בביצוע הכללה יש לשים לב לנקודות הבאות:

1. שמירה על רלוונטיות קלינית מול הצורך המחקרי. למשל, במחקר שעוסק בילדים, לא נכון לעשות הכללה של הגיל לקטגוריות של מעל גיל 65 ומתחת לגיל 65.
2. התאמה של רמות הפירוט בהכללה. למשל, הכללה של נתוני גיל לטווחים של 5 שנים עשויה להתאים במחקרים רבים במבוגרים, אך במחקר המתמקד בילדים ייתכן שיידרש טווח קצר יותר.
3. קטגוריות צרות או ייחודיות במיוחד עשויות להגביר את הסיכון לזיהוי חוזר. כדי להימנע מכך, יש לבחון כמה פרטים מקובצים לאותה הקטגוריה. באם ישנה קטגוריה בה ישנו מספר מצומצם של פרטים – ייתכן שהקטגוריה צרה מדי ויש לשקול למזג אותה לקטגוריה רחבה יותר.
4. יש לוודא שכלל הערכים מוכלים בתוך הקטגוריות השונות, ולא מפספסים אף נתון. כלומר, שכל ערך בקובץ המקורי עובר המרה לאחת הקטגוריות.

הזחת נתונים – Indentation

הוספה או החסרה של מרווח מוגדר לכל נתון מספרי. המטרה היא לשמר את היחסים בין הנתונים השונים בתוך קובץ הנתונים תוך ביצוע שינוי של הנתונים המקוריים על פי קבוע מוגדר לשם הגנה על הפרטיות.

ההזחה יכולה להיעשות באופן אקראי שאינו חשוף למשתמשים, ייתכן בטווח מסוים. למשל, לכל התאריכים בקובץ הנתונים ניתן להוסיף או להפחית בין 1-4 שבועות בצורה אקראית. מניסיוננו, בהזחת התאריכים חשוב לשקול שההזחה תהיה בכפולות של 7 כדי לא לשנות את היום בשבוע (לדוגמה, אשפוזים בסופי שבוע עשויים להיות בעלי מאפיינים שונים לעומת אשפוזים בימי חול).

עקרונות שיש לקחת בחשבון כאשר מבצעים הזחה:

1. **עקביות**: יש לוודא שהשינוי שהוחל יהיה עקבי לאורך כל הקובץ. למשל אם מזיחים תאריך אחד בטווח של שבועיים, גם תאריכים אחרים בקובץ צריכים להיות מוזחים באותו הערך. העקביות דואגת לכך שהסדר היחסי במרווחי הזמן בין התאריכים אינם משתנים לאחר ההתממה. למשל, במחקר העוסק בדיכאון לאחר לידה, המרווח העקבי בין תאריך הלידה להופעת התאריכים הקליניים צריך להישמר גם אם התאריכים עצמם מוזחים.
2. **רלוונטיות מחקרית:** יש לשקול את רמת הפירוט הנדרשת בנתונים. ההזחה יכולה להיות לפי קבוע המתאים להקשר מחקרי, למשל – בימים, שבועות או שנים.
3. **שיקוף ההזחה:** יש ליידע את העוסקים בקובץ המותמם, מהו המשתנה לפיו נעשתה ההזחה וטווח ההזחה – למשל לציין שההזחה של התאריכים היא עד ארבעה שבועות.
4. **מניעת זיהוי ההזחה:** במקרה וישנם נתונים ייחודיים שהם ערכי ברירת מחדל, כמו למשל, התאריך 01/01/1900 באשר אין תאריכים בקובץ הנתונים – יש להסירם ו/או להימנע מהזחת התאריך 01/01/1900, על מנת שחוקר הנתונים לא יוכל להסיק את ערך ההזחה.

E
visit_start_date
5/27/2011
3/23/2019
1/5/2022
4/23/2018
6/4/2014
1/16/2011
5/20/2022
6/30/2020
9/22/2012

<

E
visit_start_date
4/29/2011
2/23/2019
12/8/2021
3/26/2018
5/7/2014
12/19/2010
4/22/2022
6/2/2020
8/25/2012

דוגמא להזחת נתונים.
מימין – הערכים המקוריים של עמודת תחילת הביקור (visit_start_date).
משמאל - הערכים שמקבלים החוקרים לאחר הזחה של 28 ימים עבור כל התאריכים.

החוקרים אינם יודעים את ערך ההזחה ונאמר להם שההזחה היא קבועה בטווח של עד חודש מתאריך המקור.

פונקציונלית – Functional Anonymization

יצירת ערך חישובי שמשמר את הצורך המחקרי/אנליטי תוך כדי הסרת מידע שיש בו סיכון לזיהוי.

לדוגמה, במחקרים בהם הצורך של החוקרים הוא במשך השהות של מטופלים במוסד טיפולי, ניתן לחשב מראש עבור החוקרים את מספר ימי האשפוז ולא להעביר את התאריכים עצמם. כך, החוקרים יוכלו לנתח קשרים בין משך האשפוז לבין משתנים אחרים (בגון גיל, סוג טיפול או תוצאות רפואיות), מבלי להיחשף לתאריכי אשפוז מדויקים שעלולים לאפשר זיהוי חוזר, במיוחד באוכלוסיות קטנות או באירועים חריגים.

באותו אופן, ניתן לייצג תאריך של בדיקה רפואית כמספר הימים שעברו מאז האשפוז. במקרים מסוימים ניתן אף לייצג תאריך לידה של נושא מידע על ידי ציון הגיל (עם זאת, במקרים אחרים ייצוג התאריך כגיל לא ייחשב להתממה מספקת).

נספח ב: הדגמה של שיטות התממה לסוגי מידע נבחרים

התממת כתובות

נתונים גיאוגרפיים הם רגישים בהיותם מאפשרים זיהוי של הפרט, ובחלקם אף מהווים מזהה ישיר המאפשר זיהוי חוזר ואיתור של בני אדם במסדי נתונים. קיימות מספר גישות לביצוע פעולות התממה ביחס לנתונים גיאוגרפיים; יש לבחור בשיטה המתאימה ביותר לחוקרים ולצורך המחקרי:

1. **הסרה**: לעיתים למשתמשים במידע אין צורך בנתונים גיאוגרפיים ולכן אפשר להסיר אותם מהקובץ.
2. **הכללה**: ניתן לספק אזורי מגורים כלליים שאין להסיק מהם את הכתובות המקורית – למשל נפה, מחוז, שם הישוב, השכונה, האזור הסטטיסטי או המיקוד. יש לשים לב שאם מדובר באזור עם רזולוציה מאוד גבוהה או ייחודית (כמו מיקוד, למשל), לעיתים בעזרת משתנים אחרים הנמצאים בקובץ או תוך הצלבה עם מידע נוסף ניתן לזהות את הפרט. לכן, יש לנקוט משנה זהירות במידה ונדרשת עבודה עם רזולוציה גבוהה ויש לבחון את הסיכון לזיהוי חוזר יחד עם שאר המשתנים בהקשר זה.
3. **מידע סיכומי**: במחקרים בהם הצורך של החוקרים הוא במרחק בין כתובות (למשל, המרחק בין כתובת הבית לכתובת המוסד הטיפולי), ניתן לחשב עבור החוקרים מראש את המרחק ולייצר משתנה שערכיו הם קילומטרים, במקום חשיפת הכתובת.

באופן כללי, ביצוע פעולות התממה ביחס לנתונים גיאוגרפיים כרוך בהמרת הנתונים הגיאוגרפיים לנתונים אשר מאפשרים, מחד גיסא, להסיק מסקנות על אזור המגורים – במקרים בהם אזור המגורים הוא משתנה רלוונטי לשאלת המחקר, ומאידך גיסא, מביאים לצמצום הסיכון לפגיעה בפרטיות נושאי המידע, באופן אשר לא יאפשר זיהוי חוזר של המידע תוך שימוש באמצעים סבירים.

התממת טקסט חופשי

ברשומות רפואיות ישנם שדות רבים עם טקסט חופשי המכיל מידע קריטי לצרכי טיפול, ולעיתים עולה הצורך לעשות בו שימוש לצרכי מחקר. יחד עם זאת, טקסט חופשי יכול להכיל מידע רגיש ביותר לגבי הפרט, כולל פרטים מזהים או אבחנות שיש להתמים.

ההתממה של מידע מובנה (structured data) מתמקדת בהסרת שדות המאפשרים זיהוי ישיר או עקיף של נושאי המידע. ואולם, בהתממת טקסט חופשי (unstructured data), האתגר גדול יותר משום שאין שדות מזהים מוגדרים מראש, ולכן יש להבין את המשמעות וההקשר של המילים על מנת להחליט אילו מהן יש להסיר. חשוב לזכור שמחיקה מיותרת עשויה לאיין את המשמעות הקלינית של הטקסט, למנוע שימוש עתידי או אף ליצור הטיה במודלים המאומנים עליו. בנוסף, פעולת התממה של טקסט חופשי דורשת זמן ניכר אם היא מבוצעת ידנית, ולכן לא ריאלי לבצע אותה על מאגרי מידע גדולים. לעומת זאת, אימון מודלים מבוססי למידת מכונה דורש כמות גדולה מאוד של טקסטים חופשיים. על כן, קיימת חשיבות רבה בהתממה על ידי אלגוריתמים, ביכולת להעריך את איכותה, וכן ביצירת מנגנונים מאוזנים שמפחיתים את החשש לפגיעה בפרטיות בדרכים נוספות. למידע נוסף ניתן לעיין במדריך של הרשות להגנת הפרטיות ליישום טכנולוגיות מגבירות-פרטיות במערכות בינה מלאכותית (PETs AI).[11]

כך, על מנת לנסות לשמור על פרטיות של מטופלים מצד אחד ומתן אפשרות לחוקרים להסיק מסקנות מצד שני, ניתן להשתמש באלגוריתמים המבצעים עיבוד שפה (NLP) ומסירים שדות רגישים. אחד הכלים הקיימים הוא אלגוריתם

[11] קישור למדריך: https://www.gov.il/he/pages/pets_ai

SAFE HARBOR,[12] שפותח על ידי מיקרוסופט וצוות תמנ"ע, ונמצא בשימוש במספר ארגוני בריאות ומשרדי ממשלה כיום.[13]

<u>דוגמאות לפונקציונאליות ההתממה:</u>

שמות: שמות פרטיים, אמצעיים ושמות משפחה יותממו ויוחלפו ב <שם_>.

תאריכים: תאריך מלא, חלק של היום יוחלף ב <יום_>

מידע גיאוגרפי: יישובים יוחלפו ב <מיקום_>

ארגונים ומוסדות: ארגונים ומוסדות מוחלפים ב <ארגון_>.

מספרי טלפון: מספר טלפון מוחלף ב <קשר_>.

פרטים מזהים: תעודת זהות מוחלפת ב <מזהה_>.

כתובת אימייל: כל כתובת אימייל מוחלפת ב <קשר_>.

<u>דוגמה להפעלת האלגוריתם</u>:

אני ישראלה ישראלי מתל אביב, נולדתי בתאריך 11.5.2003 בבית חולים סורוקה שנמצא בבאר שבע ולמדתי בטכניון את התואר הראשון שלי. כתובת המייל שלי היא israellah@gmail.com ומספר הטלפון שלי הוא 0512345678 ומספר ת"ז שלי הוא 123456789.

<u>אחרי התממה של טקסט חופשי:</u>

<u>אני <שם_> <מיקום_>, נולדתי בתאריך <יום_> <ארגון_> שנמצא ב<ישוב_> ולמדתי <ארגון_> שנמצא ב<ישוב_> ולמדתי <ארגון_> את התואר הראשון שלי. כתובת המייל שלי היא <קשר_> ומספר הטלפון שלי הוא <קשר_> ומספר ת"ז שלי הוא <מזהה_></u>

התממת מידע קליני

מאגרי מידע עשויים להכיל נתונים קליניים על פרטים, ובכלל זה אבחנות ופרוצדורות, ממצאים קליניים ותוצאות בדיקות, המהווים מידע רגיש ביותר על הפרט. במקרים שבהם נמצא כי על מנת לצמצם את הסיכון לפגיעה בפרטיותם של נושאי המידע, ישנו צורך בביצוע פעולות התממה ביחס למידע הקליני, ניתן להשתמש בשיטות הבאות:

1. הכללה - שימוש בקטגוריות רחבות, כמו מחלות לב, מחלת כליה, מחלות כבד, או הכללה של תוצאות כמותיות לטווח של ערכים.
2. הסרה - מחיקת נתונים קליניים ייחודיים כגון מחלות או פרוצדורות נדירות. אבחנות או פרוצדורות בשכיחות מאוד נמוכה יכולות להוביל לזיהוי הפרטים. יש להימנע מחשיפת אבחנה או פרוצדורה עם שכיחות נמוכה. ניתן להסיר אבחנות או פרוצדורות אלו או להחליף במלל "אבחנה נדירה". לחילופין, ניתן להסיר את המטופלים בעלי האבחנות הנדירות מתוך המאגר המונגש לחוקרים.
3. שימוש בהיררכיה יותר גבוהה של אבחנות או פרוצדורות – לדוגמה, בציון אבחנות רפואיות ניתן להשתמש בהיררכיה גבוהה יותר במילון ICD-9. לדוגמה, במקום הקוד 551.3 המייצג Diaphragmatic hernia with gangrene (היררכיה נמוכה) אפשר להשתמש בקוד 551 המייצג Other hernia of abdominal cavity with gangrene או 550-553 המייצג Hernia of abdominal cavity (היררכיה גבוהה).

[12] גישת נמל מבטחים (Safe Harbor) מקורה בחוק ה-HIPPA האמריקאי (HIPAA, ראו ה"ש 5) וכוללת השמטה מהמידע של 18 משפחות של מזהים ישירים ועקיפים (כגון כתובות, לוחיות רישוי, תאריכים, מזהי גלישה באינטרנט ועוד).

[13] מדובר במוצר קוד פתוח שלא מבטיח התממה מלאה. יש לתקף את הערכים המותממים לאחר הרצת הקוד, והאחריות נותרת בידי המשתמש. האלגוריתם זמין בקישור הבא: https://github.com/8400TheHealthNetwork/HebSafeHarbor.

התממת קבצי תמונה (בגון תצלומים, קבצי דימות או סריקות MRI)

קבצי DICOM (קבצי הדמיה דיגיטלית) וקבצי דימות אחרים עשויים להכיל בנוסף לתמונות גם מטא דאטה (Meta Data) עם מידע רגיש של הפרט, וכן להכיל מזהים ישירים. אלו העקרונות שיש לשים לב אליהם בהתממת קבצי תמונה הן לקבצי המטא-דאטה והן למידע החזותי:

1. הסרה או השחרה שאינה מאפשרת חשיפה של המידע ביחס לפרטים המזהים ב- Meta Data של הקובץ (תאריכים, מספר סידורי).
2. הסרת הפרטים המזהים המוטבעים על גבי הקובץ או התמונה (בגון שם, תעודת זהות וכד' של המטופל או המטפל).
3. הסרה של פרטים או אלמנטים מזהים כגון תיעוד פנים, קעקועים ומאפיינים ייחודיים של הפרט.
4. הסרה של מזהים ביומטריים כמו טביעת אצבע.

יש לזכור ולתת את הדעת על כך שקבצי תמונה עשויים להכיל פרטים ייחודיים נוספים שמזהים או עשויים לזהות את הפרט (למשל תאונה נדירה, או מצב רפואי ייחודי).

מקורות מידע מרכזיים

מסמך זה כולל מידע רב מהמקורות המנויים להלן. להרחבה על הנושאים המפורטים במסמך זה ניתן לעיין במקורות אלו:

1. מדריך לטכנולוגיות מגבירות-פרטיות, הרשות להגנת הפרטיות, 2025.
2. מדריך להתממה של הממונה על המידע בבריטניה (ICO): https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/
3. מדריך להתממה של רשות הגנת המידע בסינגפור (PDPC): https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/guide-to-basic-anonymisation-(updated-24-july-2024).pdf
4. מדריך התממה של משרד הבריאות האמריקאי (HHS): https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html