

Szintetikus adatok a biztonságos egészségügyi kutatásokban

Budapest, 2025. december 17. – A BME kutatói is részt vettek a SECURED projektben, melyben azt vizsgálták 9 európai ország 18 intézményének tudósai, miként lehet az egészségügyi adatokat szintetikus adatokká alakítani úgy, hogy a kutatási alkalmazásuk biztonságos legyen.

A végéhez közeledik az európai együttműködésben megvalósuló SECURED projekt, amely az elmúlt két évben az egészségügyi adatok védelmével és – a szintetikus adatok alkalmazásán keresztül – etikai aggályoktól mentes felhasználásával foglalkozott. A 9 európai ország 18 partnerét összefogó, az Európai Unió által finanszírozott projekt kutatói azt vizsgálták, miként lehet az egészségügyi adatokat úgy átalakítani, hogy a kutatási alkalmazásuk biztonságos legyen. Magyarországon a BME és a Semmelweis Egyetem Egészségügyi Menedzserképző Központ szakértői vettek részt benne.

A SECURED az angol Scaling Up Secure Processing, Anonymization and Generation of Health Data for EU Cross-Border Collaborative Research and Innovation kifejezés (Az egészségügyi adatok biztonságos feldolgozásának, anonimizálásának és generálásának kiterjesztése a határokon átnyúló uniós együttműködésen alapuló kutatás és innováció érdekében) rövidítése. Fő célja az volt, hogy az adatanonimizáció és a szintetikus adatok létrehozása révén segítse az egészségügyi szolgáltatások fejlesztését, a kutatási együttműködések biztonságos megvalósítását, és ezáltal hozzájáruljon a betegek jobb ellátásához az Európai Unióban. Nyugat- és dél-európai kutatóintézetek, kórházak mellett részt vett bennük többek között az Amszterdami Egyetem, a Corki Egyetem, a Sassari Egyetem és a Leuveni Katolikus Egyetem.

A szintetikus adatok olyan, számítógépes rendszerek által mesterségesen előállított adatok, amelyek hasonlítanak a valós adatokhoz, de nem tartalmaznak személyes információkat, például a páciens nevét vagy egyéb adatait. Ezek az akár képi, szöveges vagy videóalapú adatok lehetővé teszik az egészségügyi modellek fejlesztését és tesztelését anélkül, hogy a páciensek érzékeny adatai veszélybe kerülnének. Miközben ugyanis a digitális eszközök – például az elektronikus kórlapok vagy a távgyógyászati rendszerek – átalakítják a betegellátást, új kockázatokat is teremtenek az adatvédelmi incidensek, a magánélethez fűződő jogok sérülése révén.

„Az adatvezérelt egészségügy és a mesterséges intelligencia korában a szintetikus adatok kulcsszerepet játszanak abban, hogy a fejlesztések adatvédelmi kockázatok nélkül, mégis valóságghű környezetben történhessenek meg, így hidat képeznek a klinikai alkalmazás és a kutatási innováció között. A szintetikus adatok lehetővé teszik, hogy az egészségügyi mesterséges intelligencia fejlesztése során valóságghű, de adatvédelmi szempontból biztonságos környezetben végezhesünk kutatásokat, ezáltal felgyorsítva az algoritmusok tesztelését és a klinikai alkalmazhatóság felé vezető utat” – mondta Pollner Péter, Semmelweis Egyetem Egészségügyi Menedzserképző Központ munkatársa.

„A biztonságos adatmegosztás nemcsak technológiai kérdés, hanem bizalomépítés is az egészségügyi szereplők között. A SECURED projektben azt kerestük, hogyan lehet ezt a bizalmat valós, működő megoldásokon keresztül megerősíteni. A bizalom megerősítése éppen ott kezdődik, ahol a leggyakoribb a félreértés: a szintetikus adatok önmagukban nem garantálnak teljes adatvédelmet” – figyelmeztetett Ács Gergely, a BME Kriptográfia és Rendszerbiztonság Laboratóriumának kutatója. „Amikor a szintetikus adatok szóba kerülnek, sokan azt gondolják: »Nem valósak, tehát már nem is érzékeny adatok.« Ez a feltételezés viszont nagyon veszélyes. A kihívás lényege az, hogy a fejlesztések során két cél között kell egyensúlyoznunk, ami általában kompromisszumokkal jár: egyrészt az adat legyen pontos és hasznos, megtartva az eredeti statisztikai tulajdonságait, másrészt legyen anonim és GDPR-megfelelő. tehát ne legyen köthető fizikai személyhez. Sajnos, a kettő kielégítése kompromisszumok nélkül nem lehetséges” – magyarázta a kutató.

A probléma lényege, hogy a szintetikus adatokat előállító generatív modellek megőrzik az eredeti adatok bizonyos tulajdonságait, ami viszont érzékeny információkat tárhat fel. „A szintetikus adatok valójában aggregált információk összességei. Ha ezeket az aggregált információkat kombinálni tudjuk, felfedhetünk egyedi titkokat” – tette hozzá Pejő Balázs, aki posztdoktori kutatóként vesz részt a projektben.

A veszély forrása maga a gépi tanulás működési elve: „Általában a modell másként viselkedik azon az adaton, amit már »látott«, hiszen azon pontosabb a jóslása. Ezt a memóriát kihasználva, egyszerű statisztikai próbákkal el lehet dönteni, hogy melyik adaton javasol a modell szignifikánsan pontosabban, és így valószínűsíthető, hogy azt az egyedi adatot már látta. Ez a teszt jelzi az egyénekről történő adatszivárgást, rámutatva arra, hogy a szintetikus adatok önmagukban nem garantálják a titoktartást” – mondta Ács Gergely.

A SECURED projekt során kidolgozott módszertan lehetővé teszi, hogy több szervezet – például különböző egészségügyi intézmények – biztonságosan megosszák egymással az elemzések eredményeit a zárt csatornáikon keresztül, anélkül, hogy a nyers, érzékeny adatok elhagynák a rendszereiket. Ez az eljárás új lehetőségeket teremt az etikus egészségügyi kutatások és az orvostanhallgatók gyakorlati képzése számára egyaránt. Ehhez a projekt a megoldást nem a szintetikus adatban, hanem az azt körülvevő technológiában kereste. A biztonságos alkalmazáshoz tehát olyan további technológiai védőhálókra van szükség, mint például a homomorf titkosítás.

A SECURED projekt eredményei és módszertana nyíltan elérhető lesz, hogy azokat az egészségügyi intézmények és kutatók szélesebb körben is alkalmazhassák az egészségügyi adatok biztonságos és etikus kezelése érdekében.

További információ: BME Kommunikációs Igazgatóság kommunikacio@bme.hu