Az érzékeny kutatási adatok megosztása a személyre szabott orvoslás gyakorlatában

Molnár Viktor dr. ▪ Cs. Sági Judit dr. ▪ Molnár Mária Judit dr.

**Semmelweis Egyetem, Általános Orvostudományi Kar,
Genomikai Medicina és Ritka Betegségek Intézete, Budapest**

Az egészségügyi és az orvosbiológiai kutatások adatainak széttagoltsága az adatvezérelt döntéseken alapuló, személyre szabott orvoslás egyik akadálya. A fejlődéshez a méretben és komplexitásban is rendkívüli, ám töredezett egészségügyi adatkincs hatékony kiaknázását, illetve az intézményeken vagy akár határokon is átívelő adatmegosztást biztosító technológiák szükségesek. A biobankok nemcsak a minták archívumai, hanem adatintegrációs központok is egyúttal. A biobankok adatainak együttműködésben történő elemzése értékesebb következtetéseket ígér. Az adatok megosztásának előfeltétele a harmonizáció, azaz a minták egyedi klinikai és molekuláris jellemzőinek egységes adatmodellben és standard kódokkal történő leképezése. Az egészségügyben keletkezett információk ezekben a közös sémára illesztett adatbázisokban válnak elérhetővé a gépi tanulás számára, így a módszer az együttműködés során a személyes adatokat tiszteletben tartó felhasználásra is lehetőséget ad. Az érzékeny egészségügyi adatok újraértékelése elképzelhetetlen a személyes adatok védelme nélkül, amelynek jogi és koncepcionális kereteit a GDPR- (General Data Protection Regulation) és a FAIR- (findable, accessible, interoperable, reusable) elvek jelölik ki. Az Európában működő biobankok számára a BBMRI-ERIC (Biobanking and Biomolecular Research Infrastructure – European Research Infrastructure Consortium) kutatási infrastruktúra fejleszt közös irányelveket, amelyhez hazánk 2021-ben mint Magyar BBMRI Csomópont csatlakozott. Első lépésben a biobankok szövetségében kapcsolódhatnak össze a széttagolt adathalmok, ahol sokrétű kutatási cél által motivált, igényesen összerendezett adatkészletek válnak hozzáférhetővé. Ezt követően, a betegellátás valós környezetében keletkezett adatok magasabb szinten történő értékelése is lehetővé válik, így a klinikai vizsgálatok szigorú keretek között generált bizonyítékai új szintre kerülhetnek. Közleményünkben a „federált” adatmegosztásban rejlő lehetőségeket mutatjuk be a Semmelweis Egyetem biobankjainak közös projektje kapcsán.
Orv Hetil. 2023; 164(21): 811–819.

Kulcsszavak: precíziós medicina, biobank, adatmegosztás, elosztott, személyes adatvédelmet biztosító eljárások

Sharing sensitive research data in the practice of personalised medicine

Fragmentation of health data and biomedical research data is a major obstacle for precision medicine based on data-driven decisions. The development of personalized medicine requires the efficient exploitation of health data resources that are extraordinary in size and complexity, but highly fragmented, as well as technologies that enable data sharing across institutions and even borders. Biobanks are both sample archives and data integration centers. The analysis of large biobank data warehouses in federated datasets promises to yield conclusions with higher statistical power. A prerequisite for data sharing is harmonization, i.e., the mapping of the unique clinical and molecular characteristics of samples into a unified data model and standard codes. These databases, which are aligned to a common schema, then make healthcare information available for privacy-preserving federated data sharing and learning. The re-evaluation of sensitive health data is inconceivable without the protection of privacy, the legal and conceptual framework for which is set out in the GDPR (General Data Protection Regulation) and the FAIR (findable, accessible, interoperable, reusable) principles. For biobanks in Europe, the BBMRI-ERIC (Biobanking and Biomolecular Research Infrastructure – European Research Infrastructure Consortium) research infrastructure develops common guidelines, which the Hungarian BBMRI Node joined in 2021. As the first step, a federation of biobanks can connect fragmented datasets, providing high-quality data sets motivated by multiple research goals. Extending the approach to real-word data could also allow for higher level evaluation of data generated in the real world of patient care, and thus take the evidence generated in clinical trials within a rigorous framework to a new level. In this publication, we present the potential of federated data sharing in the context of the Semmelweis University Biobanks joint project.

Keywords: precision medicine, biobank, data sharing, privacy-preserving federated learning

Molnár V, Cs. Sági J, Molnár MJ. [Sharing sensitive research data in the practice of personalised medicine]. Orv Hetil. 2023; 164(21): 811–819.

(Beérkezett: 2023. február 6.; elfogadva: 2023. március 21.)

Rövidítések

BBMRI-ERIC = (Biobanking and Biomolecular Research Infrastructure – European Research Infrastructure Consortium) Biobank és Biomolekuláris Kutatási Infrastruktúra – Európai Kutatási Infrastruktúra Konzorcium; BNO = Betegségek Nemzetközi Osztályozása (az ICD magyar változata); CT = (computed tomography) komputertomográfia; DP = (differential privacy) differenciált adatvédelem; EHR = (electronic health records) elektronikus egészségügyi nyilvántartások; ERN = (European Reference Networks) Európai Referenciahálózatok; ETL = (extract, transform, load) kinyerési, átalakítási és betöltési folyamat; FAIR = (findable, accessible, interoperable, reusable) megtalálható, hozzáférhető, átjárható, újrahasznosítható; FLamby = (Federated Learning AMple Benchmark of Your cross-silo strategies) valós egészségügyi adatokból létrehozott adattár; GDPR = (General Data Protection Regulation) általános adatvédelmi rendelet; HE = (homomorphic encryption) homomorf titkosítás; HPO = (human phenotype ontology) emberifenotípus-ontológia; ICD = (International Classification of Diseases) Betegségek Nemzetközi Osztályozása; MRI = (magnetic resonance imaging) mágnesesrezonancia-képalkotás; OMIM = (Online Mendelin Inheritance in Man) a 'Mendeli öröklődés emberben' projekt online adatbázisa; PROM = (patient-reported outcome measure) beteg által közölt eredménymutató; RWD = (real world data) valós környezetben keletkezett adatok; RWE = (real world evidence) valós környezetben keletkezett evidencia; SEFA = Semmelweis Federált Adattárház; SMC = (secure multiparty computation) biztonságos több résztvevős számítások

Az adatmegosztás helye a medicina fejlődésében

A klinikai környezetben végzett kutatások vagy akár a mindennapi egészségügyi ellátás során keletkező tapasztalatok személyes jellege, érzékenysége kihívás elé állítja az együttműködő szakembereket, akik adott páciensre vonatkozó, egyéni szintű adatokat szeretnének megosztani és azokon közös elemzéseket végrehajtani. Az orvosbiológia területén az adatok megosztására, kollaboratív újrahasznosítására egyre nagyobb az igény.

A kutatások új eredményeit (1) a szakirodalom, azaz minőségükben ugyan kontrollált (peer-review), de sokszor túlságosan fókuszált, összesített adatokat publikáló, nem szabványosan tagolt (nem strukturált) közlemények, illetve (2) az adatbázisok, vagyis az egységes nevezéktanra és merev sémára felépített, de egyedi rekordjaikat tekintve változó minőségű adattáblák reprezentálják. Mind a közlemények, mind az adatbázisok adatai esetében érvényes, hogy az eltérően definiált szempontok, leíró változók miatt nehezen összesíthetők, illetve ilyenkor is többnyire csak információveszteséggel lehetséges az összevonás. A gyűjtemények egyéni szintű, kutatási célú kiaknázásához a személyes adatok védelme mellett az átfogó adatharmonizáció nehézségeit is szükséges megoldani.

Az adatmegosztás gyakorlatában az egészségügyi ellátás kapcsán keletkezett és a kutatások során gyűjtött, változatos fókuszú információ harmadik fél számára válik hozzáférhetővé. A nagyobb elemszámú megfigyelésen nyugvó következtetések jelentősebb statisztikai erejét, megbízhatóságának javulását könnyű belátni. Ezek az adatösszevonások, habár egyes elemeik a megfigyelési szempontokra, centrumspecifikus torzításokra nézve igen változatosak, mégis összességében átfogóbb képet képesek nyújtani, ami a jobb megértés lehetőségén túl akár az adott vizsgálati célokon túlmutató felfedezésekhez is vezethet (transzferhatás). A megfigyeléseket magasabb szinten integráló kutatások javuló minősége, robusztussága közvetlenül kiaknázható többek közt a kezelési lehetőségek kockázat-haszon elemzésében, a kutatási ciklusok gyorsításában vagy skálázásában. Az interoperábilis (azaz különböző informatikai rendszerek közötti együttműködésre alkalmas formátumú) és kontrollált minőségű kutatási adatok megosztása a munkacsoportok közötti, akár nemzetközi együttműködéseknek is az előfeltétele lehet.

A napi gyakorlatba bevezethető adatvezérelt orvosi döntésekhez vagy akár a népegészségügyi programok tervezéséhez egyaránt a megosztással egyesülő adatkészletek nagyobb számhalmazaira van szükség. A megosztott adatokon történő újraelemzés vagy több adathalmaz metaanalízise a másodlagos hipotézisek tesztelésének, továbbá az adatok oktatási célú felhasználásának lehetőségét kínálja, és így válhatnak az orvostudomány személyre szabottságának motorjaivá. Az adatmegosztás kulturális szemléletváltása során a folyamatban részt vevők a hogyanra koncentrálják figyelmüket, és már nem arra, hogy miért kell megosztani az adatokat, amivel közelebb kerülünk a „nyílt tudomány világának" (open science world) víziójához.

Jelenleg az adatok megosztásának akadályaként (1) az adatvédelmi, adatbiztonsági aggályok mellett (magánélet, személyazonosság, illetve a stigmatizáció elleni védelem) (2) a hosszú távú ösztönzők hiányát (az adatszolgáltatás a kutatási projekt lezárásával jellemzően megszakad), illetve olyan változatos (3) kutatófüggő tényezőket említenek, mint a helyi szabályozás, egységes protokollok, közzétételi irányelvek, a jógyakorlatok hiányosságai vagy a széttagolt koordináció és finanszírozási források [1–5].

A biobankok a minőségi kutatás pillérei

Az alapkutatási, preklinikai eredmények gyakorlati kiaknázásának jelentős korlátja a kutatási eredmények bizonytalan megismételhetősége. Ennek hátterében gyakran a felhasznált biológiai minták egyenetlen minősége áll [6]. Erre ad választ a biobankok működését vezérlő elv: (1) a biológiai minta protokoll szerinti feldolgozása és konzerválása, valamint (2) a kapcsolódó adatok egységes adatmodellnek megfelelő gyűjtése legalább a biobank adott gyűjteményére vonatkozóan. Ezen szempontok jelentőségének felismerése újabb paradigmát hívott életre, amely FAIR (findable, accessible, interoperable, reusable) hozzáférési elveken alapszik, ahol az adat „megtalálható, hozzáférhető, átjárható és újrahasznosítható" marad.

A biobankokban a minták és a hozzájuk tartozó leíró adatok egyedi példányai adott definíciónak megfelelő gyűjteményekbe rendezve találhatók. A minta ez esetben azt a biológiai anyagot jelenti, amely az élőben megfigyelhető, keresztmetszeti állapotot igyekszik rögzíteni és későbbi mérési célpontok – akár még ki sem fejlesztett módszerek – számára információveszteség nélkül megőrizni. A biobankok sokszínűségét demonstrálja az, hogy milyen sokféle módon csoportosíthatók. A felhasználás célja szerint beszélhetünk populációs (a népességet reprezentáló gyűjtemény egészséges és főként gyakori, komplex betegségekkel élő donorokkal) vagy éppen betegségorientált biobankról. Mindezek lehetnek keresztmetszeti, eset-kontroll vagy időben történő követést (hosszmetszeti, longitudinális) megvalósító gyűjtemények. A tárolt minta típusa valamilyen szövet, testfolyadék vagy származékaik, amelynek segítségével genetikai-genomikai, akár multiomikai vagy jól hozzáférhető klinikai adatok mentén biomarker-felfedező típusú kutatásokhoz illő kérdéseket tehetünk fel. Retrospektív, prospektív, adaptív, sőt virtuális kohorszokat is górcső alá vehetünk a minőséget biztosító biobanki működésnek köszönhetően [7].

A biobankokban tárolt gyűjtemények példányainak ritkasága, egyedi módon fókuszált vagy esetleg átfogó jellege igen értékessé teszi őket, mivel a leképezett állapot kutatása gyakran csak számos centrum összefogásával képzelhető el. A biobankok feladata az emberi biológiai mintapéldányok nyílt hozzáférésének biztosítása lenne, ez az elv azonban jogilag korlátozottan alkalmazható. A hozzáférés valójában csak nagyon specifikus adatkészletekre lenne biztosítható, amelyekben az egyedi minták és rekordok anonimizáltak, és az újraazonosítás kockázata igen csekély, illetve ahol a dinamikus tájékozott beleegyezést egy adatvédelmi kockázatokat kezelni képes rendszer támogatja.

A biológiai mintát leíró adatok gyűjtése

Nemcsak a biológiai minta, de a kapcsolódó leíró adatok esetében is hasonló problémát jelent az információk konzerválása, illetve az ehhez szükséges standardizáció igénye. A technológiai fejlődés, a digitalizáció térnyerésével az orvostudományi kutatások adatgazdái kénytelenek voltak felismerni, hogy képtelenek lépést tartani az adatgyűjtés és -tárolás egyre bővülő lehetőségeivel, nem tudják megfigyeléseiket az elemzési igényeknek megfelelő strukturált adattárakba integrálni. Az ebből származó törekvés vezetett el az *egészségügyi adatvagyon* koncepciójához, és eközben a változatos adattípusok befogadásához rugalmasan alkalmazkodni képes megoldások is születtek. Ezek között az egyik irányzat egy nem kizárólag strukturált adat befogadására képes tároló kialakítására helyezi a hangsúlyt, amellyel jól skálázható adatgyűjtés valósítható meg, ahol a feldolgozás, rendszerezés csak halasztva történik, és az egységesítés, az adatharmonizáció szempontjait majd az utólagos elemzés igényei fogják diktálni. Így például egy egészségügyi adaton futó retrospektív elemzésben a kísérleti csoportok kialakítása és az azokba történő mintabesorolás nem a gyűjtés, hanem csak az adatok feldolgozása kapcsán történhet. A fejlődés másik irányát az *adattárházak* képviselik, ahol az adatok olyan adatmodell és terminológia szerint kapnak formázást, melyben már összekapcsolhatóvá válnak a különböző adatforrásokból származó adatokkal. Ennek megfelelően az adattárházban tárolandó adatokat eleve strukturálják, tisztítják, harmonizálják és integrálják, még mielőtt azok az egységes nyilvántartásba kerülnének. Miután az új információ az adattárházban meghatározott struktúrába és formátumokba kerül, a bővülés meghatározott időközönként történő tárházba való adatbetöltéssel valósul meg (ETL = extract, transform, load [kinyerési, átalakítási és betöltési] folyamat). Végeredményben az új, rendezetlen rekordokat halmozó tárolók és a merev adattárházak hiányosságai hibrid megoldást hívtak életre, az előnyöket ötvöző „adatkincstárház" (data lakehouse) modellt, amely egy köztes tárolási réteget alkalmaz, így fenntartva az eredeti adatok integritását [8–10].

A biobankok hálózatba kapcsolása

A biobankok magasabb szinten történő integrációjával, hálózatokba kapcsolásával nagyobb számú, adott feltételeknek jobban megfelelő biológiai minta vagy adat rövidebb idő alatt szerezhető be, illetve a tanulmányok növekvő számának és összetettségének kiszolgálása válik lehetővé. Az európai biobankhálózat, a BBMRI-ERIC (Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium; https://www.bbmri-eric.eu/) ezen a területen vezető szerepet játszó integrátor szervezet, amely infrastrukturális támogatást biztosít a csatlakozó kutatások számára. Feladatát képezi a biobanki működés irányelveinek fejlesztése, minőségirányítási standardok és információtechnológiai megoldások kidolgozása, emellett a szervezet a kapcsolódó etikai, jogi és társadalmi kérdésekben is állást foglal.

A biobankok hálózatában egy-egy különleges minta vagy gyűjtemény elérhetőségét erre a célra kifejlesztett alkalmazások támogatják. Például a *Locator* biológiai minták és kapcsolódó adatok keresését teszi lehetővé a hálózatba csatolt biobankokban, így logikai kapcsolókkal a donor életkora, a minta típusa, tárolási hőmérséklete stb. feltételek kombinálhatók, a releváns jelöltekre szűkítve a keresést. A *Negotiator* szolgáltatás már a következő lépést segíti, hatékony kommunikációs platformot biztosítva a biobankok és a mintákat, vagy adatokat igénylő kutatók számára. Szabványosítja és ezzel leegyszerűsíti a megfelelő minták és adatok kutatási célokra történő beszerzésének folyamatát, ami különösen akkor hasznos, ha a kutatónak több lehetséges biobankjelölttel kellene felvennie a kapcsolatot.

A BBMRI-ERIC *Directory* (címtár) a világ legnagyobb biobank-katalógusa, amelyben a gyűjteményeket kezelő, együttműködésre nyitott biobankok aggregált, összesített információkat osztanak meg magukról. A BBMRI-ERIC hálózatában az egyes regionális központokat a nemzeti csomópontok (national nodes) alkotják, amelyek segítenek az egyedi gyűjteményeket gondozó biobankok regisztrációjában, és ezzel elősegítik láthatóvá válásukat a hasonló érdeklődésű kutatások számára [11, 12]. A Semmelweis Egyetem 2021-ben csatlakozott a BBMRI-ERIC-hez a magyar csomóponton keresztül, amelynek alapító tagjai a négy orvosegyetem: a Semmelweis Egyetem, a Debreceni Egyetem, a Pécsi Tudományegyetem és a Szegedi Tudományegyetem, valamint a Dél-pesti Centrumkórház és a Richter Gedeon Nyrt. (www.bbmri.hu).

A személyes adatok a kutatási célú adatmegosztásban

Az egyént egyértelműen azonosítja személyes adatainak, például név, lakóhely, születési hely és idő stb. kombinációja. Az egészségügyi rekordok érzékeny adatokat jelentenek tele testi, értelmi vagy lelki állapotra vonatkozó információkkal, a genetikai adatok pedig a betegségek örökletes kockázatáról árulkodhatnak, ami szintén kiemelt védelmet indokol. A kutatások és a rutindiagnosztika átfogó genomikai mérései során azonosított ultraritka variánsok, haplotípusok kezelése valóban olyan kihívást jelent, amelyet már nem lehet egyszerűen egy számsorozattal való megjelöléssel, anonimizálással kezelni. Az adott kutatásban részt vevő személy azonosítására alkalmas adatok köréhez tartozik tehát az átfogó genetikai profil is (például egy exomszekvenálás ritka variánsai), amelyben egy ritka, etnikumra, családra, egyénre jellegzetes genetikai variáns segítségével az adott személyhez tartozó rekordok hasonló pontossággal azonosíthatók, mint például a születési dátuma alapján.

A páciens a saját egészségügyi adatainak tulajdonosa [13]. Az adatok integritásának megőrzése mellett a magánélet védelmének fenntartása is kiemelt cél. A jelenkori elektronikus adatkezelés biztonsági kihívásai életre hívták a GDPR-t (General Data Protection Regulation), az Európai Unió új általános adatvédelmi rendeletét. Célja, hogy megfelelő egyensúlyt teremtsen az egyének személyes adatainak védelme és ezen adatok szabad áramlása között.

Kódolással, pseudonym (álnevesített) azonosítók segítségével klinikai adatok is megoszthatóvá válnak, ilyenkor a személyt azonosító nevek, számok helyett kódokat közvetítünk. Biztonságosan és a kutatási adatoktól elkülönítve, csak egy „kódkulcs” képes az adott személyt az adataival összekapcsolni [14]. Ugyanígy a biológiai minta is egy kód alatt jelenik meg a vizsgálatok, elemzések teljes vertikumában. Ezek a kódok megfejthetők maradnak, így az egyénnel indokolt esetben újra fel lehet venni a kapcsolatot (reconnection), szemben az egyirányú, anonimizációval történő kódolással, amelynél erre utólag már nincs lehetőség. Amint az adatok valóban anonimizáltak, és az egyének így már nem azonosíthatók, az adatok nem esnek a GDPR hatálya alá, és szabadabban felhasználhatók. Kihívást jelent a genetikai tulajdonságok olyan hatékony maszkírozása, amely már lehetetlenné teszi az egyén ritka variánsok és kombinációik alapján történő azonosítását, ez azonban összetettebb megoldást sürget.

Tanulás az adatokból, de az adatok nélkül?

A mesterséges intelligencia, ezen belül a gépi tanulás korszerű megoldásai nemcsak a genetikai adatok értelmezését, hanem az érzékeny kutatási, ezen belül a személyt azonosító adatokat, így a genetikai profilt is biztonságosan kezelő megosztását is segíthetik. A megosztott adatokon (vagyis federáltan) tanuló algoritmusokkal operáló módszerek egészségügyi célú alkalmazása igen ígéretes, különösen, ha a személyes adatokat tiszteletben tartó felhasználás is megvalósul az együttműködés során (privacy-preserving federated learning), hiszen így válik lehetővé a klinikai adat átfogó kiaknázása. Segítségével egy jelentős akadály, az adatvédelmi jogszabályok és a szakemberek aggályai miatti korlátozott hozzáférhetőség problémája áthidalhatóvá válik. A gépi tanuló algoritmusok kiképzéséhez ráadásul megfelelő mennyiségű (és persze minőségű) tanító adatra van szükség, ami összefügg a megbízhatósággal, és aminek a kontrollja szintén a klinikai gyakorlatba lépés feltétele [15].

A széttagoltan elérhető adatok egyesítése persze egyszerűen megoldható lenne a harmonizált adatok egyetlen központi adatbázisba történő gyűjtésével, ekkor azonban sérülhetne az érzékeny és kritikus adatokhoz méretezhető biztonság. A gépi tanulás viszont képes kezelni a személyes adatok védelmének problémáját, mivel a matematikai modell tanulása során akár az egyedi adatok biztonságosan helyben maradhatnak. Eközben csupán a modell „utazik”, a nyers adatok azonosításra alkalmas formában egyáltalán nem is kerülnek megosztásra. Ebben az a nagyszerű, hogy a gépi tanulás algoritmusa a fragmentált adatkészleteken anélkül képes tanulni, hogy

azok elmozdításra kerülnének a lokális adattárolóból [16].

Az orvosi leírásokban található megállapításokhoz standard címkék rendelhetők, ezzel a strukturálatlan szöveg a kódolással egy gépi tanulási folyamat számára megfelelő bemenetet képez, olvashatóvá válik. Az alkalmazott kódokat betegségszinten a diagnózisok vagy éppen a tünetek szintjén használatos szótárak, ontológiák adják. Az egészségügyi adatokhoz kapcsolható ontológiák hierarchikus felépítése a bizonytalanságokat, az általánosan vagy éppen nagyon is részletesen, nagy információtartalommal megadható megállapításokat is segít kifejezni. Ebből pedig számszerűen kifejezhetők a fenotípusok, egyének vagy csoportok közötti szemantikus hasonlóságok.

A kódolás folyamata, azaz a fogalmak, jellemzők azonosítása és standard címkékhez rendelése, egy közös nyelvre alakítása jelenti az egyik szűk keresztmetszetet. A következtető algoritmus kiképzéséhez, azaz hogy például képessé váljon egy CT-felvétel kiértékeléséhez, a radiológusoknak először fárasztó címkézési munkát kell végezniük a képeken található jellemzőkön. A képalkotókból származó adatkészletek szakértői tudással történő gazdagítása grandiózus, szakismeretet igénylő, nagyon nehezen delegálható feladat. A tapasztalt szakértői erőforrások bevonása a napi orvosi gyakorlat mellett ugyanis nem könnyen valósítható meg [17]. Automatizálható megoldásokkal persze elérhető, hogy a szövegállományok elemzésével szótárakat generáljunk a címkézéshez, például a betegségek nevének felismeréséhez. Még a hasonló, gépesített asszisztenciával támogatott harmonizálási folyamat végén is egy adatkurátor vagy adatmenedzser, majd egy szakterületi adattudós áll [18]. Feladatuk két lépésre (és szintre) is bontható: elsőként az adatok strukturálása, minőségének ellenőrzése, kódolása történik, majd ezeket a félkész adatokat lehet ellenőrizni, magasabb szinten áttekinteni, továbbértelmezni [13].

A valós környezetben keletkezett egészségügyi adatok

A valós környezetben keletkezett adatok (real world data – RWD) és az ezek feldolgozásából származó ismeret, evidencia (real world evidence – RWE) egyre nagyobb szerepet játszanak az egészségügyi döntésekben, így irányelvek megfogalmazásakor, egy készítmény vagy eljárás hatásosságának, biztonságosságának felmérésekor. Ezek „a betegek egészségi állapotára és/vagy az egészségügyi ellátás nyújtására vonatkozó adatok, amelyek különböző forrásokból rutinszerűen gyűjthetők” (https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence). A valós környezetben keletkezett adatok jól kiegészíthetik a klinikai vizsgálatok adta szigorú keretek között történő megfigyeléseket. A randomizált vizsgálatokhoz képest különbséget jelent, hogy az RWD-k (1) megfigyeléses jellegűek, (2) gyakran strukturálatlanok és szempontjaik az intézmények között eltérőek, a források eredeti célja miatt (3) nem törekszenek a teljességre, sőt (4) végpontokat rendszerint nem is tartalmaznak, ráadásul sajnos (5) különböző típusú mérési hibáknak és torzításoknak is ki vannak téve.

Az elektronikus egészségügyi nyilvántartási rendszerek fejlődése (electronic health records – EHR) bőséges valós adatot szolgáltat. Ezek tekintélyes mérete akár ritka eseményekre épülő kockázatbecslést is támogathat, vagy akár az expozíciók, beavatkozások és kimenetek közötti összefüggéseket segíthetnek retrospektíven feltárni. A valóságot jobban leképező, szélesebb látképbe regiszterek adatai vagy éppen a betegtől származó, kérdőívekkel felvehető életminőségi vagy funkcionális leírók (patient-reported outcome measures – PROM), továbbá képalkotók, elektrofiziológiai műszerek vagy akár hordható elektronikai eszközök nyers kimenetei is beletartozhatnak [19–21].

Számos kihívással kell szembenézni az elmélet gyakorlatba való ültetéséhez: az adat előkészítése, a modell frissítése, korrekciója jelentős és folyamatos ráfordítást igényel. A közvetlenül mérhető előnyök a pénzügyi területen, a mobilkommunikációs és hordható eszközök kapcsán sokkal gyorsabb fejlődést biztosítanak a gépi tanulási megoldásoknak. Az egészségügy valós adatai esetében jelenleg a valós kísérleti adatbázisok kialakításának és kezelésének optimalizálásánál tart. A FLamby (Federated Learning AMple Benchmark of Your cross-silo strategies) különböző méretű, struktúrájú és természetű valós egészségügyi adatok (szövettan, CT- és MRI-képek, táblázatba foglalt leíró adatok, túlélési elemzés adatai, dermoszkópia) által alkotott, természetes módon felosztott adattárat annak érdekében hozták létre, hogy az alkalmazott kutatási problémák számára teszterületet kínáljon, és valós környezetet szimulálva tegye próbára a megismételhetőséget (https://owkin.com/publications-and-news/blogs/bridging-the-gap-between-federated-learning-theory-and-practice-with-real-world-healthcare-datasets).

Tanulás elosztva és mégis szövetségben

A gépi tanulás szerepe az egészségügyben tehát a nagy és komplex adathalmazok alapján a csoportok közötti különbségek megfelelő általánosítóképességgel történő azonosítása. Az orvosi döntéseket támogató modellek kialakításához jól általánosító, túlillesztés nélküli betanítás szükséges, ami sokféle adathoz való hozzáférést igényel, ezek pedig általában elszigeteltek, és szétszórva vannak a különböző egészségügyi intézmények között.

A hálózatokat itt decentralizáltan működő, ugyanakkor egymással összekapcsolt csomópontok alkotják. A szövetségben történő gépi tanulás alkalmazása lehetővé teszi, hogy a lokális adatokat a hálózat különböző csomópontjain anélkül kérdezzék le, hogy maguk a nyers adatok elhagynák eredeti biztonságos helyüket. Eközben a felek valójában „csak” közbenső modellparamétereket

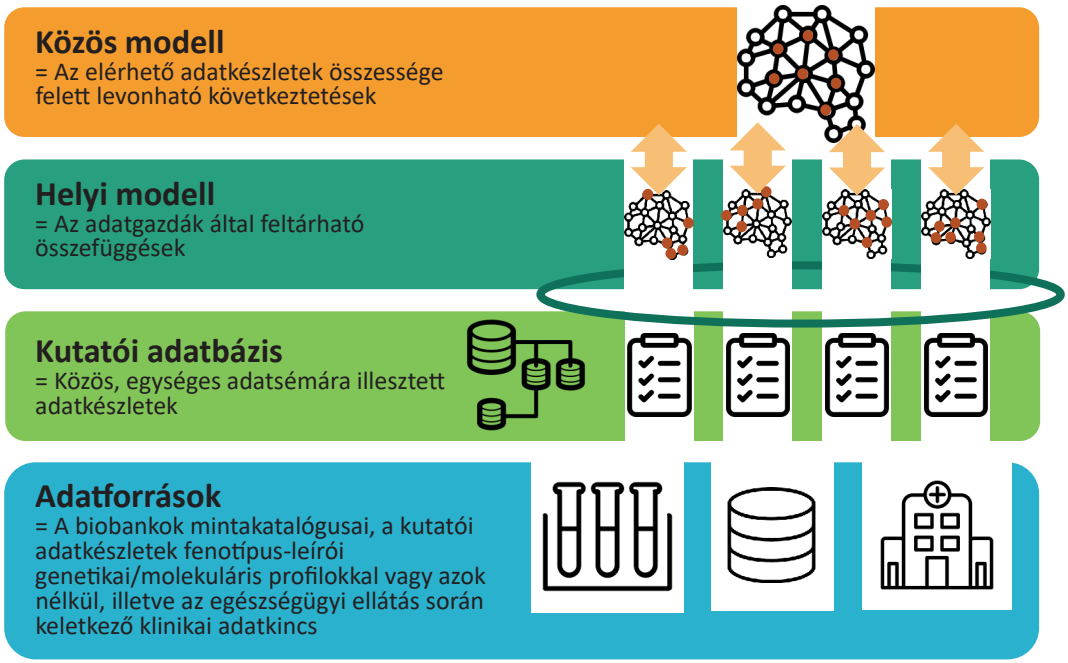
cserélnek egymás között. A modell és a felek között kicserélt köztes értékek persze az érzékeny adatok helyben tartása és a kicserélt adatok mennyiségének csökkentése mellett is ki vannak téve adatvédelmi kockázatoknak, amit kriptográfiai megoldásokkal lehetséges kontrollálni. Ezek a módszerek különféle stratégiákon alapulnak, mint például a differenciált adatvédelem (DP – differential privacy), a biztonságos több résztvevős számítások (SMC – secure multiparty computation) vagy a homomorf titkosítás (HE – homomorphic encryption) [22–24].

A biobankok közötti adatmegosztás az elosztott statisztikai lekérdezések számára ideális kísérleti terep. A biobankok adatainak szövetségi hálózatba integrálásával egy adott feltételeknek megfelelő biológiai anyag (például egy ritka progresszív idegrendszeri kórkép korai fázisában vett liquorminta) elérhetősége válik lekérdezhetővé, illetve a kapcsolódó kutatási célú betegadatok újraelemzése, újrahasznosíthatósága lesz megoldható, ráadásul mindeközben a részt vevő munkacsoportok a saját szabályaik szerint működhetnek, és az adatok védelme is biztosított.

A kutatások során keletkező heterogén adatok általános modellekké fejlesztése több rétegen keresztül valósul meg, amelyek ráadásul folyamatosan fejlődnek az új információk beérkezésével és szerves beépülésével. Az együttműködés alapja a visszajelzés a végfelhasználó adatgazdákig *(1. ábra)*.

A biológiai mintákat konzerváló nemzeti adattárházak és a nemzetközi konzorciumok szeparáltan működtek a közelmúltig. Ezek indulásukkor konszenzuselvek alapján kezdték meg a minták és adatok gyűjtését, előre lefektetve az abban részt vevők szerepét, feladatait és jogkörét. Olyan betegségek esetén, amelyeknél kis esetszámokkal kell dolgozni, az egyesülés egyre kevésbé kerülhető meg. A már működő adattárházak kapcsolódása esetén számos különbséget kell konszenzusra hozniuk egy együttműködés során, például a kialakított új közös rendszerben a jogosultságok kiosztása is bonyolítja az együttműködést [14, 25, 26].

Az eddigi legnagyobb globális, gépi tanulással operáló erőfeszítésről számol be egy tanulmány a közelmúltból, amelyben többparaméteres MRI-képeken tanulva a glioblastoma szöveti képalkotó által leképezett alterületeinek, sebészi vagy radioterápiás tervezéshez fontos határainak kijelölésére egy jól általánosító modell kifejlesztését célozták meg. A gépi tanuláshoz sikerrel kapcsolták be az egyébként ritka daganat 6314 esetének harmonizált adatait, hat kontinens 71 különböző centrumát. Az egyesítő elemzéshez a szakértő együttműködő klinikusoknak egységes annotációs protokollt kellett követniük, amely már az adatok előfeldolgozására is kiterjedt, hogy figyelembe vehessék a különböző MRI-berendezések képalkotási különbségeit. Ez a tanulmány demonstrálta, hogy a módszer lehetővé teszi ritka esetek tapasztalatainak globális megosztását, amellyel az irodalomban szereplő összes eddigi esetet meghaladó számú megfigyelés, sőt metaadat vonható össze, ráadásul úgy, hogy azok mindvégig a kórházi tűzfalak mögött maradnak. Az átfogó adatok ereje a sokszínűségben állhat: az értékelésből kitűnik, hogy még az eleve nagy esetgyűjteménnyel rendelkező centrumok lokális modelljei számára is előnyös javulást jelentett a közös tanulás [27]. Egy további friss tanulmány a tripla negatív ritka emlődaganatok neoadjuváns kemoterápia változatos válaszának



1. ábra A széttagolt adatkészletek bekapcsolása együttműködési hálózatokba. A személyes adatvédelmi garanciákkal biztonságosan kiaknázható egészségügyi adatkincs, amely valósághű modelleket és pontosabb következtetéseket eredményez

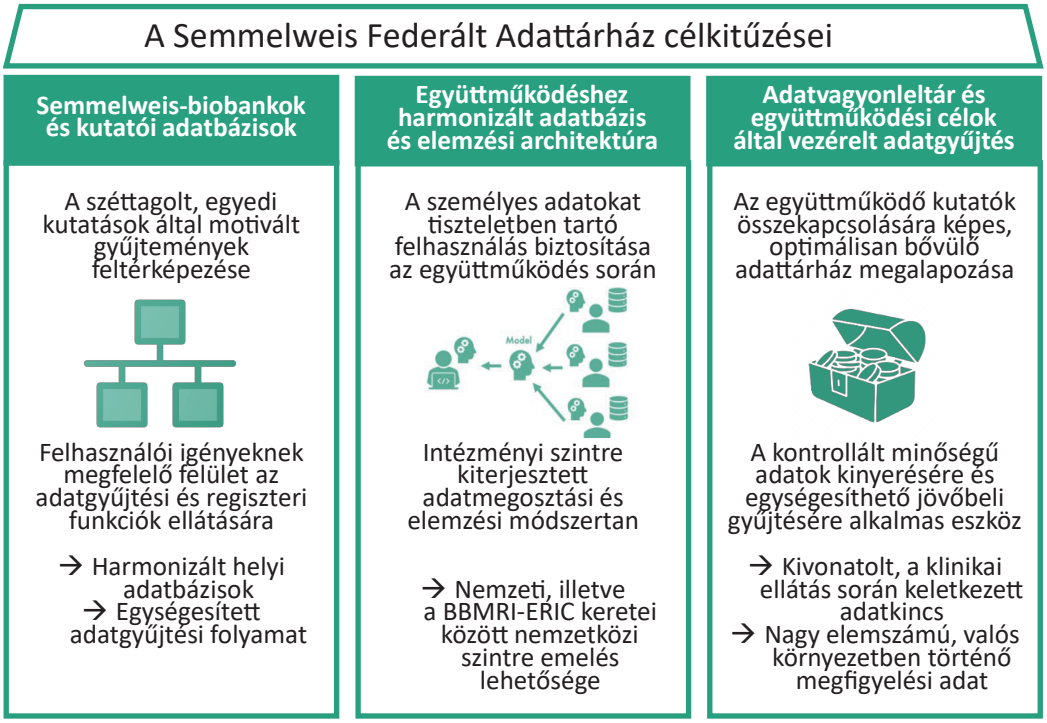
megértését célzó, elosztott modelltanulási módszerről számol be. Itt szövettani teljes digitális képek mély tanulásával (deep learning), azaz a képek több rétegben kivonatolható tulajdonságain alapuló, mesterséges neurális hálózatok rétegeiben leképezett, általánosító következtetésekkel dolgoztak, megkerülve az időigényes szakértői annotációt. Ezzel együtt az értelmezhetőség sem veszett el teljesen, ismert és potenciális biomarkerek – mint az apokrin daganatsejtek, az infiltráló lymphocyták, illetve a fibrosis és a daganatsejtek elrendeződése – jelentőségét sikerült számszerűsíteni [28].

Biobankok szövetsége a Semmelweis Egyetemen

Az elosztott adatkészletek kialakítására törekvő hazai példa a kialakítás alatt álló Semmelweis Federált Adattárház (SEFA). Fő célkitűzése egy fenntartható, orvostudományi adatgyűjtést intézményi szinten integráló együttműködési modell megalapozása a Semmelweis Egyetemen *(2. ábra)*.

A biobankokhoz csatlakozó kutatói adatbázisok közös nyelvét a fenotípus leírásában azok a standard kódok jelentik, amelyek (1) betegség (például ritka betegségekben az ORPHANET- vagy OMIM-kód, a gyakoriaknál az ICD-10- vagy az ezen alapuló BNO-kódkészlet), illetve (2) tünet mint megfigyelhető rendellenesség, akár laborértékben vagy elektrofiziológiai paraméterben való eltérés (HPO – human phenotype ontology) szintjén is képesek leírni az elsődlegesen vizsgált és társuló patológiás eltéréseket/állapotokat. A szélesebb és átfogó adatkészleteken történő tanulás előnye mellett az együttműködés eközben magának a szakértői tudásnak a megosztására is lehetőséget ad. Az adatok harmonizálása során jól kontrollálható, amikor az egészségügyi adatban található eredeti állítást (például a páciensnek „gyakran vannak légúti fertőzései") egy megfelelő ontológia szerint kódoljuk, azaz egy szótár adott kifejezésének (Recurrent respiratory infections HP: 0002205) feleltetjük meg. Itt nem lenne szükséges minden egyes lépésnél felmerülő elemi döntéshez magasan képzett kutatókat rendelni, sőt a tématerület függvényében a kódolási kérdést a különböző kompetenciaszinteknek lehetne delegálni. Márpedig a szakértői tudás optimális kihasználása kulcskérdés, ha az egészségügyi ellátórendszer rutinfeladatai során képződő adatokat kutatások számára szeretnénk kiaknázni.

A federált adatmegosztásra képes rendszer kialakításával a felhasználó adatgazda (ti. a személyes, kutatási adatok felett rendelkező kutató) képessé válik a személyes vagy egyéb minősített adatok veszélyeztetése nélkül a projektben közreműködők körében, akár intézményi szinten, például egy kísérlet megtervezéséhez szükséges áttekintésre. A standard kódolás belső hierarchiájának köszönhetően nemcsak egy konkrét diagnózisnak, hanem általánosabb vagy éppen specifikusabb kódoknak megfelelő minták és adatok száma és azok széttagoltsága is lekérdezhetővé válik. Ezzel lehetőség nyílik közre-



2. ábra A Semmelweis Federált Adattárház (SEFA) célkitűzései

BBMRI-ERIC = Biobank és Biomolekuláris Kutatási Infrastruktúra – Európai Kutatási Infrastruktúra Konzorcium

működő partnerek keresésére, amelyben a kérelmező és az adatszolgáltató szempontjai is megosztásra kerülhetnek (hasonlóan a BBMRI *Negotiator* fejlesztés alatt álló federált alkalmazásához, amelynél a biobankok közötti meghatározott feltételeknek megfelelő minták igénylése szabványosítható). Magasabb szinten többváltozós, komplex elemzések is lehetővé válnak, sőt a genetikai és a hierarchikusan kódolt fenotípusos adatok megosztásával több szakértő, kutató és adatmenedzser együttműködése, akár az adatok felügyelt, optimálisan irányított bővítése, javítása is megvalósulhat. Rizikómodellezés, molekuláris epidemiológiai elemzések, a genetikai térkép imputációja, populációspecifikus referenciaszekvencia mind elérhetővé válhat a reprezentatívabb és jobb minőségű adat birtokában. Nem kevésbé jelentős távlatot jelentenek a technológia olyan ígéretei, mint a teljes intézményi adatkincs integrálása vagy az épülő európai kutatási infrastruktúrák vérkeringésébe történő szerves bekapcsolódás.

Összefoglalás

A személyes adatokat tiszteletben tartó gépi tanulás egy új elemzési megközelítés, amely nagy léptékben teszi lehetővé több, érzékeny adatokkal rendelkező intézmény számára a bekapcsolódást, hogy az adatok összevonása nélkül közös modelleket hozzanak létre. A technológia *federált* jelzőjét, annak különböző tulajdonságait előtérbe helyezve akár szövetségi, egyesített vagy elosztott elnevezésekkel is említik. A mély tanulást alkalmazni képes módszer különösen ígéretes az egészségügy területén, mivel biztosítani képes a páciens adatainak bizalmas kezelését, hiszen az adatgazda partnerek között csak a modellek közlekednek, míg maguk az adatkészletek helyben maradnak, és továbbra is különálló helyeken tárolódnak. Az adatmegosztásnak ez a módja a kollaboratív adatgyűjtés új formája is lehet, amelynél a folyamatban a minőség és az alkalmazkodás szempontjai is érvényesülhetnek. A töredezett adatbázis által biztonságosan összekapcsolt adatkészletek felhasználásával közös, globális modell alakítható ki, amely a helyileg tárolt kutatói adatbázisok számára is visszajelzést adhat az optimális pótlás, korrekció és mélyítés irányításával. Mindezzel a valós adatokra épülő döntéstámogató klinikai felhasználások számára keletkezik jól kontrollált, de az érzékeny egészségügyi adatokat biztonságban tartani képes tanító halmaz. A modell így nemcsak pontosabb következtetések levonását teszi lehetővé, hanem annak meghatározásában is segíthet, hogy hol szükséges az adatok bővítése. Ezáltal nemcsak a kutatói adatbázisokban elhelyezett adatkészleteknek a kísérlettervezéstől a komplex elemzésekig terjedő hatékony kiaknázása, hanem a valós környezetben keletkező klinikai adatok válnak elérhetővé, és a jövő együttműködésen alapuló, optimálisan összehangolható kutatásai is előkészíthetők lesznek.

Anyagi támogatás: A szerzőket a TKP2021-NVA-15, TKP2021-EGA-25, OTKA 139010 pályázatok támogatták.

Szerzői munkamegosztás: M. V.: Irodalomkutatás, kéziratírás, ábrakészítés. Cs. S. J.: Irodalomkutatás, kéziratírás. M. M. J.: A koncepció kidolgozása, kéziratírás, szakmai véleményezés. A szerzők a cikk végleges változatát elolvasták és közlésre jóváhagyták.

Érdekeltségek: A szerzőknek nincsenek érdekeltségeik.

Köszönetnyilvánítás

A Semmelweis Egyetem Genomikai Medicina és Ritka Betegségek Intézete az Európai Referenciahálózatok, az ERN Rare Neurological Disorders és az ERN Neuromuscular Disorders tagja, valamint a BBMRI-ERIC Magyar Csomópontja. A szerzők köszönik az együttműködésből fakadó folyamatos továbbképzés lehetőségét.

Irodalom

[1] Dolley S. Big data's role in precision public health. Front Public Health 2018; 6: 68.
[2] Enticott JC, Melder A, Johnson A, et al. A learning health system framework to operationalize health data to improve quality care: an Australian perspective. Front Med. 2021; 8: 730021.
[3] Institute of Medicene. Sharing clinical research data: workshop summary. National Academies Press, Washington, DC, 2013.
[4] Kriegova E, Kudelka M, Radvansky M, et al. A theoretical model of health management using data-driven decision-making: the future of precision medicine and health. J Transl Med. 2021; 19: 68.
[5] Ross JS. Clinical research data sharing. What an open science world means for researchers involved in evidence synthesis. Syst Rev. 2016; 5: 159.
[6] Esteva-Socias M, Artiga MJ, Bahamonde O, et al. In search of an evidence-based strategy for quality assessment of human tissue samples. Report of the tissue Biospecimen Research Working Group of the Spanish Biobank Network. J Transl Med. 2019; 17: 370.
[7] Coppola L, Cianflone A, Grimaldi AM, et al. Biobanking in health care: evolution and future directions. J Transl Med. 2019; 17: 172.
[8] Mate S, Kampf M, Rödle W, et al. Pan-European data harmonization for biobanks in ADOPT BBMRI-ERIC. Appl Clin Inform. 2019; 10: 679–692.
[9] Ong TC, Kahn MG, Kwan BM, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak. 2017; 17: 134.
[10] Wieder P, Nolte H. Toward data lakes as central building blocks for data management and analysis. Front Big Data 2022; 5: 945720.
[11] Krekora-Zając D, Marciniak B, Pawlikowski J. Recommendations for creating codes of conduct for processing personal data in biobanking based on the GDPR art. 40. Front Genet. 2021; 12: 711614.
[12] Litton JE. Launch of an infrastructure for health research: BBMRI-ERIC. Biopreserv Biobank 2018; 16: 233–241.
[13] Hulsen T. Sharing is caring. Data sharing initiatives in healthcare. Int J Environ Res Public Health 2020; 17: 3046.

[14] Bentzen HB, Castro R, Fears R, et al. Remove obstacles to sharing health data with researchers outside of the European Union. Nat Med. 2021; 27: 1329–1333.
[15] Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med. 2020; 3: 119.
[16] Xu J, Glicksberg BS, Su C, et al. Federated learning for healthcare informatics. J Healthc Inform Res. 2021; 5: 1–19.
[17] Ng D, Lan X, Yao MM, et al. Federated learning. A collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quant Imaging Med Surg. 2021; 11: 852–857.
[18] Pletscher-Frankild S, Pallejà A, Tsafou K, et al. DISEASES: text mining and data integration of disease-gene associations. Methods 2015; 74: 83–89.
[19] Bastarache L, Brown JS, Cimino JJ, et al. Developing real-world evidence from real-world data: transforming raw data into analytical datasets. Learn Health Syst. 2022; 6: e10293.
[20] Mahendraratnam N, Mercon K, Gill M, et al. Understanding use of real-world data and real-world evidence to support regulatory decisions on medical product effectiveness. Clin Pharmacol Ther. 2022; 111: 150–154.
[21] Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest. 2020; 130: 565–574.
[22] Dyda A, Purcell M, Curtis S, et al. Differential privacy for public health data: an innovative tool to optimize information sharing while protecting data confidentiality. Patterns (N. Y.) 2021; 2: 100366.
[23] Kumar AV, Sujith MS, Sai KT, et al. Secure multiparty computation enabled e-healthcare system with homomorphic encryption. IOP Conf Ser: Mater Sci Eng. 2020; 981: 022079.
[24] Scheibner J, Ienca M, Vayena E. Health data privacy through homomorphic encryption and distributed ledger computing: an ethical-legal qualitative expert assessment study. BMC Med Ethics 2022; 23: 121.
[25] Barnes C, Bajracharya B, Cannalte M, et al. The Biomedical Research Hub: a federated platform for patient research data. J Am Med Inform Assoc. 2022; 29: 619–625.
[26] Hallock H, Marshall SE, 't Hoen PA, et al. Federated networks for distributed analysis of health data. Front Public Health 2021; 9: 712569.
[27] Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun. 2022; 13: 7346. Erratum: Nat Commun. 2023; 14: 436.
[28] Ogier du Terrail J, Leopold A, Joly C, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nat Med. 2023; 29: 135–146.

(Molnár Viktor dr.,
Budapest, Üllői út 26., 1085
e-mail: molnar.viktor@med.semmelweis-univ.hu)

„Auxilia humilia firma consensus facit."
(Szerény eszközöket is erőssé tesz az egyetértés.)