KÖZLEMÉNY

Regiszteradatok felhasználási lehetőségei a kohorszkutatásban

Veroszta Zsuzsanna

ÖSSZEFOGLALÓ

A tanulmány a magyarországi gyakorlatban rendelkezésre álló adminisztratív adatok kutatási célú felhasználási lehetőségeit tekinti át, egy speciális vizsgálati területre, a születési kohorszvizsgálatokra fókuszálva. Ezen belül is a 2018 elején, a KSH Népességtudományi Kutatóintézetben indult longitudinális, survey módszertanon alapuló születési kohorszvizsgálat, a Kohorsz '18 kutatás szempontjából összegzi az adminisztratív adatfelhasználási lehetőségeket, adatkapcsolási eljárásokat és az azokat meghatározó szabályozó környezeti feltételeket. Az elemzés eredményei általánosabb szinten a jelenleg tervezett, avagy zajló adminisztratív adatokkal is dolgozó survey típusú kutatások számára nyújthatnak új szempontokat.

Tárgyszavak: születési kohorszvizsgálat, longitudinális, adminisztratív adat, adatintegráció

Veroszta Zsuzsanna, KSH Népességtudományi Kutatóintézet
E-mail: veroszta@demografia.hu

BEVEZETÉS

Tanulmányunk célja a regiszteradatok (adminisztratív adatok) kutatási becsatornázási lehetőségeinek áttekintése és mérlegelése a jelenlegi magyarországi gyakorlat alapján, elsősorban egy speciális vizsgálati területre, a születési kohorszvizsgálatokra vonatkoztatva. Jóllehet az áttekintés közvetlenül a magyarországi születési kohorszvizsgálat (Kohorsz '18 kutatás[1]) adatfelhasználási lehetőségeinek optimalizálását célozza meg, általánosabb szinten feltétlenül hasznosítható a jelenlegi szabályozási-intézményi környezetben zajló adminisztratív adatokkal is dolgozó survey típusú kutatások tágabb körében is.

A vizsgálódás alapját adó Kohorsz'18 Magyar Születési Kohorszvizsgálat adatfelvétele – a KSH Népességtudományi Kutatóintézet nagyszabású kutatási vállalkozásaként – 2018 elejétől zajlik. A felmérés során ciklikus egyéni megkeresésekre kerül sor a 2018–2019-ben gyermeket vállalók relatíve nagy, 10%-os országos reprezentatív mintáján. Ennek során elsőként a várandósság alatt, majd a megszületett gyermek fél-, másfél- és hároméves korában zajlik a kutatás survey szakasza, amelyhez a kutatók már a tervezés, majd a mintavétel során az adminisztratív adatátvétel számos kiegészítő elemével számoltak. Ezen – ezidáig csak kismértékben megvalósított, többnyire tervezés alatt álló – adatkapcsolások sokszínű voltát és az azokkal járó akadályokat, dilemmákat is végigköveti az alábbi tanulmány, amelynek fő célja mégis az, hogy – adott esetben – a gyermekek magyarországi felnövekedésével kapcsolatba hozható, jelenleg rendelkezésre álló hazai adminisztratív adatokról komplexen gondolkodva minél inkább kiaknázható forgatókönyveket vázoljon fel.

A VIZSGÁLAT KERETEI

Áttekintésünkben az adminisztratív adatok kutatási felhasználásához kapcsolódó elméleti háttér és általános gyakorlat ismertetése helyett a kohorszkutatási programot támogató lehetséges adatköröket, eljárásokat és lépéseket vesszük számba. Ennek megfelelően vizsgálódásunk az alábbi szűkebb területeket érinti:

– Milyen törvényi keretek és szervezeti háttér határozzák meg a regiszteradatok kutatási felhasználását?

[1] A Magyar Születési Kohorszvizsgálatról a KSH Népességtudományi Kutatóintézet (www.demografia.hu) és a kutatás honlapja (www.kohorsz18.hu) nyújt részletes tájékoztatást (a kutatás előkészítését lásd: Veroszta 2018).

- Melyek a kohorszkutatás számára releváns, a hazai gyakorlatban már megvalósult adatkapcsolások? Ezeknek mely elemét hasznosíthatja a kutatás?
- Melyek a kohorszkutatás számára adatkapcsolásként esetlegesen szóba jöhető adatbázisok és adatkörök?
- Melyek a regiszteradatok becsatornázásának lehetséges eljárásai a kohorszkutatáson belül?
- Hogyan szakaszolható a regiszteradatok bevonása a kohorszkutatás felépítésében?
- Végezetül milyen problémákkal, korlátokkal szükséges számolnunk a különböző adatbevonási gyakorlatok esetében? Ezek milyen stratégia vagy eljárás mentén oldhatóak fel?

A REGISZTERADATOK KUTATÁSI FELHASZNÁLÁSÁNAK FOGALMI KÖRNYEZETE

Az adminisztratív adatokhoz való kutatási célú hozzáférés szabályozási környezetének áttekintése előtt a definíciós alapok tisztázása szükséges. A tanulmány során alkalmazott definíciós készletünket az alábbiakban határozzuk meg:

Az *adat* „az információ formalizált módon való megjelenítése, amely alkalmas feldolgozásra, továbbításra, közlésre, értelmezésre" (KSH 2014, Gárdos 2015).

Az *információ* olyan objektumokra (tényekre, eseményekre, dolgokra, folyamatokra vagy a gondolati világ elemeire) vonatkozó ismeret, amely definiált, tehát értelmezésének kerete meghatározott (Gárdos 2015).

A *statisztikai adat* a „valós világ egyedeinek tulajdonságaira vonatkozó statisztikai megfigyelések, illetve további statisztikai műveletek eredménye" (KSH 2014). Az adatok egyedi azonosíthatósága ez esetben nem cél, csak a sokaság jellemzését célzó adatgyűjtés és feldolgozás eszköze (Gárdos 2015).

A *közérdekű adat* a közfeladatot ellátó szerv vagy személyek kezelésében lévő, tevékenységükre vonatkozó nem személyes adat, rögzített információ, ismeret. A közérdekű adatok kezelésének főszabálya a nyilvánosság és a hozzáférés. A közérdekű adatok az Avtv. alapján nyilvánosak, kiadásukat bárki igényelheti (Avtv. alapján Eötvös Károly Intézet 2006, Székely 2015, Cseres-Gergely – Scharle 2008).

Egyedi adat az egyénnel vagy szervezettel közvetlen kapcsolatba hozható adat (A statisztikáról szóló 1993. évi XLVI. törvény alapján Cseres-Gergely – Scharle 2008).

A *személyes adat* a természetes személlyel kapcsolatba hozható adat és az adatból levonható, egyénre vonatkozó következtetés. A kritérium a kapcsolatba hozhatóság, mivel az adat személyes jellege az adatkezelés során mindaddig fennmarad, amíg a kapcsolat helyreállítható. Személyes adatok például a természetes (név, lakcím, anyja neve, szülési dátum) és mesterséges azonosítók (adószám, TAJ, személyi szám). A személyes adatok kezelésének fő szabálya az önrendelkezés (Avtv alapján Eötvös Károly Intézet 2006, Székely 2015, Cseres-Gergely – Scharle 2008).

Az anonimizált *mikroadat vagy elemi adat* szintén egyéni szintű, tehát egy adott alanyra vonatkozik, de a közvetlen és a közvetett azonosíthatóság lehetősége nélkül, azaz személyes jellegétől megfosztva áll rendelkezésre. Az azonosíthatóság lehetőségének minimalizálása érdekében módosítási eljárások alkalmazása is szükséges lehet (a KSH adatvédelmi szabályzata (KSH VII/2005. (SK)2.) alapján Cseres-Gergely – Scharle 2008).

A *metaadat* más adatokat ír le, illetve határoz meg, például a statisztikai adatrendszerben használt fogalmakat, nómenklatúrákat, a felhasznált adatforrások leírását, az adatelőállítás módszertanát.

Elsődleges adatforrásnak az adott szervezet saját statisztikai adatgyűjtéseit tekintjük. *Másodlagos adatforrás* minden olyan adatállomány, amelynek esetében az adatgyűjtő személy vagy szervezet nem azonos az adatokat használó szervvel, e két funkció (adatgyűjtés és -felhasználás) tehát különválik.

Az *adminisztratív adatforrások* olyan, közfeladatot ellátó szervezet által fenntartott adatgyűjtemények, amelyek célja a szervezet (nyilvántartási, engedélyezési, jogosultsági) közfeladatainak ellátása. Az *adminisztratív adatok* jellemzője előbbi célból a célcsoport teljes lefedettsége és egyedi azonosíthatósága.

A REGISZTERADATOK KUTATÁSI FELHASZNÁLÁSÁNAK SZABÁLYOZÁSI KÖRNYEZETE

A kohorszvizsgálat esetében tehát a nyilvántartási, engedélyezési, jogosultsági céllal rögzített adminisztratív adatok kutatási becsatornázása a cél. Ennek érdekében szükséges áttekintenünk a lehetőségeket meghatározó hazai szabályozási környezetet. A törvényi háttér összegzése során kiemelt alábbi jogszabályok bemutatása esetében a formális hivatkozás/idézés helyett a felhasználhatóság szempontjait vesszük alapul.

Adatvédelmi törvény

A személyes adatok védelméről és a közérdekű adatok nyilvánosságáról szóló 1992. évi LXIII. törvény (röviden adatvédelmi törvény [Avtv]) szabályozza a közérdekű adatokat, amelyek kezelésének főszabálya a nyilvánosság és bárki számára való hozzáférhetőség. A törvény személyes adatként definiál ugyanakkor minden, az érintett személlyel kapcsolatba hozható adatot, valamint az adatból levonható, az érintettre vonatkozó következtetést. A közérdekű adatok kezelésének főszabálya a nyilvánosság, a személyes adatoké pedig az önrendelkezés (Székely 2015). A törvény a személyes adatok, illetve a közérdek védelmének konfliktusában jellemzően a személyes adatok védelmének ad prioritást. Eszerint személyes adatokat akkor lehet kezelni (továbbadni, összekapcsolni), ha ehhez az érintett hozzájárul, vagy törvény megengedi. (Cseres-Gergely – Scharle 2008). A kohorszkutatás számára a közérdekű adatok hozzáférhetőségének biztosítása lehet releváns törvényi hivatkozási alap.

Elektronikus információszabadság törvény

A 2005. évi XC. törvény az elektronikus információszabadságról, a közinformációk megismerését elősegítő új eszközök és szolgáltatások létrehozásáról rendelkezik. Idetartozik a közérdekű adatok kötelező internetes közzététele, a jogszabályok és a jogszabályalkotás, valamint a bírósági határozatok nyilvánossága. A törvény fő célja az *egyedi* adatok, illetve dokumentumok internetes hozzáférhetőségének elősegítése az adatbázisok felállítása és működtetése révén (Székely 2015). A szabályozás fontos eleme a közzététel és a kereshetőség mellett az, hogy bármely nyilvános közérdekű adat igényelhető elektronikus úton. Az elektronikus információszabadság megvalósítása során az elektronikus közzététel és az adatok fellelhetőségét biztosító, metaadatokkal dolgozó közadatkereső rendszer megkönnyíti az adatok megtalálását, hozzáférését (Eötvös Károly Intézet 2008). A törvény az információszabadságról szóló törvénybe olvadt be, de rendelkezései közül a hozzáférés módjának szabályozása – az elektronikus adatközlés kötelezettsége – a kohorszkutatás számára továbbra is releváns lehet.

Döntéselőkészítési adatokról szóló törvény

A 2007. évi CI. törvény a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról az Európai Parlament és a Tanács 2003/98/EK. számú irányelvéhez igazodik. Eszerint a közigazgatási szervek birtokában levő adatok és más dokumentumok elsődleges céljukon túli felhasználását célozza. A törvény szerint az államnak és az államigazgatási szerveknek kötelessége, hogy a gazdaságra és a társadalmi folyamatokra kiható döntéseiket előzetesen mérlegeljék, eredményüket utólagosan vizsgálják és ehhez adminisztratív céllal gyűjtött adatokat is felhasználjanak. A szabályozás az anonim összekapcsolás céljára kormányrendeletben meghatározott szervezetet jelöl ki (Fodor – Veroszta 2011).

A törvény alapján a másodlagos adatfelhasználás hatásvizsgálatok révén a közpénzek hatékonyabb kezelését segíti elő, valamint a közös (közpénzen létrehozott) adatvagyon felhasználását társadalmi jólétet növelő célokra (pl. kutatásra). Emellett a személyes adatok védelmét (önrendelkezés jogát) is biztosítania kell. Ennek érdekében, a szabályozás alapján kizárólag anonimizált mikroszintű adatok átadása valósulhat meg, és az anonimizálást az adatkezelő kérésre köteles elvégezni. Eszerint tehát az adatkezelők mikroszintű adatot is kötelesek kiadni (azonosításra alkalmas adatok esetén HASH algoritmusalapú anonimizálás után). A potenciális felhasználók (államigazgatási szakértők) az igényléshez törvényi felhatalmazást kapnak. Mikroszintű adat továbbra sem adható át abban az esetben, ha a személyesség nem szüntethető meg (Cseres-Gergely – Scharle 2008).

Az adatbázisok összekapcsolásának kérvényezéséhez miniszteri vagy kormányhivatal-vezetői engedély szükséges, piaci szereplők számára az adatkérés nem biztosított. Ugyanakkor a kapcsolt adatbázisok az eljárás lezárultával közérdekű adattá válnak, így a közadattárból mindenki számára igényelhetők és ingyenesen hozzáférhetők (Szabó 2011). Ebben a törvényben jelenik meg tehát az adatok összekapcsolásának lehetősége és keretrendszere, ami mind az adatkérés, mind az esetleges adatkapcsolás szempontjából alapvető jelentőségű a kohorszkutatás számára.

Nemzeti adatvagyonról szóló törvény

A nemzeti adatvagyon körébe tartozó állami nyilvántartások fokozottabb védelméről szóló 2010. évi CLVII. törvény 1. § 1. pontja a „nemzeti adatvagyon" kategóriáról rendelkezik. Ezt a közfeladatot ellátó szervek által kezelt közér-

dekű adatok, személyes adatok és közérdekből nyilvános adatok összességeként definiálja (Székely 2015). A szabályozás részletesebb áttekintése a közérdekű adatok védelmére való hivatkozás kapcsán lehet fontos a kohorszkutatás kapcsán.

Információszabadságról szóló törvény

Az információs önrendelkezésről és az információszabadságról szóló 2011. évi CXII. törvény (Infotv.) szerint közérdekű adat a „közfeladatot ellátó szerv vagy személy kezelésében lévő és tevékenységére vonatkozó, vagy közfeladatának ellátásával összefüggésben keletkezett, a személyes adat fogalma alá nem eső, bármilyen módon vagy formában rögzített információ vagy ismeret". E szabályozáson belül két főszabály jelenik meg, amelyek egyike az állam átlátszósága, másik az állampolgár átlátszatlansága, azaz elszámoltathatóság a köz és információs önrendelkezés az egyén szintjén. A kétpólusú személyes/közérdekű adatkoncepció több okból folyamatosan finomodott (pl. közérdekből nyilvános adatok definiálása) (Székely 2015). E szabályozás a kutatásba bevont, közfeladatot ellátó szervek körének indoklásához nyújthat hivatkozási alapot. Az EU Általános adatvédelmi rendeletének (GDPR) hatályba lépését (lásd alább) az infotörvény számos ponton módosítással követte.

Közadat törvény

A közadatok újrahasznosításáról szóló 2012. évi LXIII. törvény az Európai Unió tagállamai számára kötelezően implementálandó irányelv alapján egyfelől meghatározza a „közadat" definícióját („az információs önrendelkezési jogról és az információszabadságról szóló törvényben meghatározott közérdekű adat és közérdekből nyilvános adat"); másrészt kiterjeszti a hozzáférés szabályait. Eszerint a közadat újrahasznosítás céljából történő rendelkezésre bocsátása „a közadathoz az igénylő részére biztosított olyan hozzáférés, amely lehetővé teszi az igényelt közadat újrahasznosítását az igénylő számára, ideértve különösen a közadat adathordozón vagy elektronikus úton történő egyszeri vagy rendszeres átadását, a közadatot tartalmazó adatbázishoz történő közvetlen hozzáférés biztosítását…" (Székely 2015). A törvény a közadatokhoz való hozzáférés és az újrahasznosítás szabályozásában jelentős. Fontos, hogy megjelenik benne az adatátadás kötelezettsége és rendszeressége.

2013. évi ccxx. törvény az állami és önkormányzati nyilvántartások együttműködésének általános szabályairól

Az „interoperabilitás törvény" célja az adminisztratív nyilvántartások regiszterének kialakítása, a nyilvántartások elektronikus információs rendszerének létrehozása. A regiszternek tartalmaznia kell a nyilvántartás és a nyilvántartó azonosító adatait, a nyilvántartás vezetéséről és adattartalmáról rendelkező jogszabályra hivatkozást, a nyilvántartott adatok megnevezését és a nyilvántartás együttműködésének módját a többi nyilvántartással (Gárdos 2015). A szabályozás a tekintetben lehet fontos a kutatás számára, amennyiben a nyilvántartások kezelhető, egységesség – ezáltal az összekapcsolhatóság – felé mutató felépítését támogatja.

Az EU általános adatvédelmi rendelete (General Data Protection Regulation, GDPR)

Az Európai Parlament új adatvédelmi szabályait 2016/679/EU számon az EU Általános adatvédelmi rendelete (General Data Protection Regulation, GDPR) összegzi. A 2018 májusától életbe lépett rendelet gondoskodik a természetes személyek személyes adatok kezelése tekintetében történő védelméről, illetve az e típusú adatok szabad áramlásának szabályozásáról. Az online adatforgalom által indokolt aktualitását jól példázza, hogy a rendelet már nemcsak az azonosított, hanem az azonosítható (pl. IP-cím, cookie alapján) személyes adatokat is hatálya alá veszi. Kutatási – tehát nem piaci – célú adatgyűjtés esetén is fontos fejlemény a hozzájárulási kötelezettségek kibővülése, amelynek következtében minden adatközlő direkt és konkrét beleegyezését adja adatainak kezeléséhez. Kutatási célú adatátvétel esetén különösen fontos szempont, hogy a személyes adatok kezelésére vonatkozó hozzájárulások megszerzését adatkezelési célonként külön is biztosítani kell. A rendelet emellett az álnéven történő adatkezelés lehetőségét nyitja meg mint jogszerű és célszerű adatkezelési technikát a személyes adatok védelme érdekében.

HAZAI REFERENCIAKUTATÁSOK

Az alábbiakban néhány, a 2007. évi CI. törvény által biztosított adatkapcsolási lehetőségen alapuló hazai kutatást veszünk számba. A kutatások teljes

körű és részletes bemutatására e helyütt nem törekszünk. A cél néhány – a kohorszkutatás szempontjából esetleg irányadó – minta kiemelése a későbbi részletesebb feldolgozás előtt.

Az MTA–KRTK munkaerőpiaci adatkapcsolása

Az államigazgatási adatok eddigi legjelentősebb kutatási célú összekapcsolására 2012-ben került sor. A bevont öt államigazgatási szerv adatainak összekapcsolása a népesség munkaerőpiaci életútjának követését célozta meg.

Az adatgazdák az Országos Egészségbiztosítási Pénztár, az Országos Nyugdíjbiztosítási Főigazgatóság, a Nemzeti Adó- és Vámhivatal, a Nemzeti Munkaügyi Hivatal és az Oktatási Hivatal voltak. Az alapsokaság a 2003-ban 5 és 74 éves kor közötti népesség, a lefedett csoport e sokaság fele, a lefedett időszak pedig 2003–2011 volt. Az adatok gyakorisága havi bontású, a minta nagysága 4,6 millió fő. A kutatás célja életútvizsgálat, ennek megfelelően a bevont adatcsoportok a demográfiai jellemzők mellett az igénybe vett ellátásokat és az érvényes jogviszonyokat, valamint a foglalkoztatási, kereseti adatokat fedték le (forrás: http://adatbank.krtk.mta.hu).

Az adatkapcsolás a munkaerőpiaci adatbázisok integrációjának általános tapasztalatai mellett az életút megragadása szempontjából jelenthet fontos igazodási pontot a kohorszkutatás számára.

Az Educatio Nonprofit Kft. adatkapcsolása

A másik jelentős adatkapcsolás-sorozat a Diplomás Pályakövetési Rendszer (DPR) keretében zajlott 2011-ben, 2013-ban és 2014-ben.[2] A vizsgálatok célja a Felsőoktatási Információs Rendszerben (FIR) diplomát szerzettként regisztráltak munkaerőpiaci és képzési életútjának rögzítése. Ez esetben tehát demográfiai-oktatási-munkaerőpiaci adatok összekapcsolásáról van szó. A mintaképző adatbázis (FIR) meghatározott évfolyamainak teljes lefedettségű adatkapcsolása jött létre a kapcsolati kódként alkalmazott TAJ, illetve az adóazonosító jel alapján (az első, pilot szakaszban az összekapcsolás személyes adatok kombinációja révén valósult meg).

[2] Az Educatio Nonprofit Kft. feladatait – köztük a Diplomás Pályakövetési Rendszer működtetését – 2016-tól az Oktatási Hivatal (OH) vette át. A program továbbra is folytatódik, az OH keretében jelenleg is zajlanak FIR alapú, pályakövetési célú, adminisztratív adatokon alapuló adatkapcsolások.

A pályakövetési célú adatintegráció első szakasza 2011-ben egy végzett évfolyamra (2009) vonatkozóan négy adatközlő szervezet adatait egyesítette, amelynek során a végzést követő első év státuszát vizsgálta (Fodor – Veroszta 2013). A második adatintegrációs eljárás 2013-ban három adatközlő szervezet bevonásával – a 2010-es kilépő évfolyam adatait kapcsolta a végzést követő 2. év státuszainak rögzítésével (Nyüsti – Veroszta 2014). A harmadik fázis 2014-ben hét adatközlő összehangolásával, két kilépő évfolyamon zajlott (2010-ben és 2011-ben abszolváltak), akiket egyrészt a 2013. májusi státuszuk alapján vizsgált, másrészt képzési és munkaerőpiaci jellemzőiket 2009-ig visszamenőleg vezette vissza, dinamikus elemzési nézőpont alkalmazását téve lehetővé (Nyüsti – Veroszta 2015). A pályakövetési célú adatkapcsolások mintanagysága 50 ezer fő (2011), 57 ezer fő (2013), 137 ezer fő (2014) volt. A bevont adatszolgáltatók: FIR, APEH, OEP, FSZH (2011), FIR, NAV, OEP (2013), FIR, Diákhitel Központ Zrt., MÁK, NAV, NMH, OEP, ONYF (2014).

Mindezek mellett – szintén a DPR keretében – 2014-ben a képzési és a munkaerőpiaci átmenet, illetve életút részletes vizsgálatát lehetővé tévő adatkapcsolásra került sor, amelynek mintaképző adatbázisa a Közoktatás információs rendszer (KIR), alapsokasága pedig az 1988–1994 között született teljes kohorsz (840 ezer fő). A KIR–FIR–ONYF adatait összesítő adatkapcsolás vizsgálati időszaka 2012. január–2013. november között zajlott le.[3]

Az adatkapcsolás referenciát jelenthet a különböző típusú – tehát nem pusztán munkaerőpiaci regisztrációs céllal létrejött – adminisztratív adatkörök összekapcsolásához speciális célcsoporton.

Az MTA–KRTK, NAV–NMH összekapcsolása

A szintén az MTA–KRTK által végzett 2012-es adatkapcsolás során a Nemzeti Munkaügyi Hivatal bértarifa-felvételben szereplő munkavállalók adataihoz kapcsolták az adott személyt alkalmazó vállalatnak a NAV társasági adóbevallás nyilvántartásában szereplő mérleg- és eredménykimutatásából származó adatait. Az adatok a 2002–2011-es évekre vonatkoztak, gyakoriságuk éves bontású. A vizsgálat alapsokaságát a Magyarország területén 4 fő felett működő jogi személyiségű gazdasági szervezetek, továbbá

[3] Az adatközlő szervek elnevezése során a megvalósult adatkapcsolások esetében az eljárás évében aktuális intézményneveket alkalmazzuk. Az adatforrások specifikálásakor a frissített, az azóta zajlott intézményváltásokat figyelembe vevő elnevezésekkel élünk.

2002., 2005., 2007. és 2008. években a nonprofit szervezetek teljes, illetve 2002-től részmunkaidős dolgozói alkották. Az adatkapcsolás évenkénti 120–180 ezres munkavállalói és 9–11 ezres vállalati létszámot fedett le. A kutatás a munkavállalók demográfiai adatai mellett azok foglalkoztatási és kereseti viszonyait tartalmazta egyéni szinten, valamint ehhez kapcsolódóan a foglalkoztató vállalatok létszám-, tulajdon- és ágazati viszonyait, illetve mérleg- és eredménykimutatását (forrás: http://adatbank.krtk.mta.hu).

Az adatkapcsolás fenti eljárása a kohorszkutatás számára az eltérő elemzési szintű – mikro-makro, vagyis egyéni és intézményi – adatok összekapcsolási gyakorlatának lehet egyik referenciapontja.

A POTENCIÁLIS ADATBÁZISOK ÉS ADATKÖRÖK

Az alábbiakban a Magyar Születési Kohorszvizsgálathoz kötődő külső adatkapcsolás szempontjából esetlegesen szóba jöhető adminisztratív adatbázisokat tekintjük át. Az adatbázisok ismertetésében ismét nem törekszünk teljeskörűségre, a cél minden esetben a hozzáférés lehetőségének, a szervezeti háttérnek, a kapcsolódás technikai lehetőségének és a vizsgálat számára lényeges adatkörök átgondolása. Emellett több esetben megfogalmazzuk a további tisztázásra váró, kérdéses pontokat is.

A Kohorsz ’18 esetében az adatfelvétel mintaképző adatbázisát a védőnői rendszerben működő, makroadatokat tartalmazó védőnői jelentésösszesítők korábbi évekből származó körzetszintű statisztikái alkotják, összekapcsolva körzetszintű, gazdasági-fejlettségi KSH-makroadatokkal (Kapitány 2018). Tekintettel arra, hogy a továbbiakban adatkapcsolások számára szóba jöhető, születési adatokat tartalmazó adatbázisok becsatornázása még nem valósult meg, az előbbi adatbázis mellett a védőnői és szülésértesítő rendszerhez kapcsolódó egyéb adatforrások összegzése is indokolt (Rohr 2016 gyűjtése alapján). Figyelembe véve, hogy ezen adatbázisok mindegyike feltehetően tartalmazza majd az alapvető demográfiai háttérváltozókat, ezen adatok rendelkezésre állását más adatbázisokból külön nem vizsgáljuk, azt adottnak vesszük.

Születésértesítő rendszer

A 2015. február 15-én indult elektronikus alapú adminisztratív rendszer a születést követő kórházi nyilvántartást és az adatok védőnői/háziorvosi

rendszerbe történő becsatornázását szolgálja. Az adatgyűjtés és -rögzítés a kórházi védőnőkhöz kapcsolódik, akik minden születés után egységes adatrögzítő felületen gyűjtik, majd továbbítják az adatokat a területi védőnőknek, értesítési kötelezettséggel a háziorvosok számára. Az adatbázis két fő eleme a születésértesítő és hazabocsátási adatlap.

Szervezet: Állami Népegészségügyi és Tisztiorvosi Szolgálatot (ÁNTSZ).

Adatok formája: elektronikus mikroadat.

Célcsoport: szülő nők és megszületett gyermekek.

Lefedettség: kötelezően teljes körű az adott célcsoporton.

Adatközlő: kórházi védőnő.

Kapcsolati kód: anyai TAJ (a gyermek TAJ-t ez alapján ideiglenesen generálják).

Főbb változók:

- születés intézménye (makroadatok kapcsolhatók),
- területi védőnő,
- anya adatai (TAJ, demográfiai adatok),
- terhesgondozás (védőnői gondozás volt-e a várandósság alatt),
- szülés (a várandósság hete, szülési fájdalomcsillapítás, a szülés módja, beavatkozások, BNO-listáról szülés alatti szövődmény és gyermekágyi betegség),
- szülészetről távozás ideje és helye,
- újszülött adatai (demográfiai adatok, élve- vagy halvaszületés, generált TAJ, születési testtömeg és testhossz, távozási testtömeg, fejkörfogat, mellkörfogat, Apgar-teszt, szülési sérülések, beavatkozások, fejlődési rendellenességek, újszülött betegségek, táplálás),
- gyermekvédelmi jelzés és intézkedés.

A születésértesítő rendszer a születési adatok egyéni szintű becsatornázását teszi lehetővé a kutatás számára. Ennek megvalósulási feltétele az adatfelvételt megelőzően a válaszadó várandós beleegyezező nyilatkozaton adott hozzájárulása.[4]

[4] A születéseket rögzítő adatbázist produkál még a KSH-hoz tartozó élveszületési lap is, valamint a Tauffer-statisztika (Állami Egészségügyi Ellátó Központ), ugyanakkor – tekintve, hogy ezek rögzítése nem TAJ-alapú – e keretek között alapvetően makroadatként alkalmazhatók.

Várandósgondozási rendszer – védőnői jelentésösszesítő

A hazai várandósgondozási rendszerben keletkező, a védőnők által rögzített adatok a kohorszkutatás szempontjából két fő típust alkotnak. A ciklikus (havi gyakoriságú) védőnői összesítő jelentések aggregált adatai teljes lefedettséggel, kötelező adatszolgáltatás révén elektronikusan elérhetőek, statisztikai célokat szolgálnak a munkáltató és szakfelügyelet felé. Az adatközlés excel alapú formanyomtatványon zajlik. Az összesítő elkészítése a védőnők által végzett egyéni szintű dokumentáción alapul, amely azonban egységes elektronikus rendszerben nem rögzített, sőt elektronikus rögzítettsége is esetleges.

Szervezet: Állami Népegészségügyi és Tisztiorvosi Szolgálatot (ÁNTSZ).
Adatok formája: elektronikus makroadat.
Célcsoport: terhesgondozás alatt álló nők, kisgyermekes nők és megszületett gyermekek.
Lefedettség: kötelezően teljes körű az adott célcsoporton.
Adatközlő: területi védőnő.
Kapcsolati kód: makroszintű.
Főbb változók:

- 12. hétig felvett várandósok száma,
- gondozásban részesültek száma,
- környezeti mutatók (dohányzás, gyermekbántalmazás előfordulási gyakorisága),
- újszülöttek száma,
- koraszülések,
- rendellenességek,
- súlyeloszlás,
- csecsemőhalálozás.

A védőnői jelentésösszesítők alapján készülő körszetszintű területi statisztikák képezik az alapját a Kohorsz ’18 kutatás kiinduló mintájának.

Várandósgondozási rendszer – védőnői törzslapok

A hazai várandósgondozási rendszerben keletkező, a védőnők által rögzített egyéni szintű adatok egy része elektronikusan rögzített (törzslapok), más része papíralapú dokumentáció. Az adatok nem alkotnak egységes adatbázist, hozzáférésük tehát védőnői szinten oldható meg.

Szervezet: Állami Népegészségügyi és Tisztiorvosi Szolgálatot (ÁNTSZ).
Adatok formája: elektronikus vagy papíralapú mikroadat.
Célcsoport: terhesgondozás alatt álló nők, kisgyermekes nők és megszületett gyermekek.
Lefedettség: kötelezően teljes körű az adott célcsoporton.
Adatközlő: területi védőnő.
Kapcsolati kód: anyai TAJ, gyermek TAJ.
Főbb változók/adatforrások:

- törzslapok adatai,
- munkanaplóadatok,
- csecsemőnyilvántartó-adatok,
- oltásnyilvántartó-adatok,
- várandósgondozási könyv adatai (demográfiai adatok, terhesgondozást végző szakorvos, háziorvos, védőnő, szülés várható időpontja, vércsoport- és antitestvizsgálat eredményei, előző terhességek adatai, anamnézis, leletek, terhesség jellemzői, terhesgondozási naptár, diagnosztikai vizsgálatok eredményei stb.).

Az adatok – ez esetben egységes adatbázisról nem beszélhetünk – a teljes vizsgálati populációra vonatkozóan adnak részletes leírást a terhesség lefolyásáról, az anya, majd a megszületett gyermek fizikai és környezeti jellemzőiről. A várandósgondozási könyv adatainak egyéni szintű rögzítésére – szintén a válaszadó anya írásos hozzájárulása mellett – az első adatfelvételi hullámban védőnői együttműködéssel kerül sor.

Egységes Védőnői Informatikai Rendszer – eVIR

Az eVIR fejlesztésére a Koragyermekkori program (TÁMOP 6.1.4 kiemelt projekt) keretében került sor, az informatikai fejlesztés tervezett indulása 2016 nyara volt, ám számos technikai nehézség miatt az adatgyűjtés teljes körű lefedettséggel nem valósult meg – egyénsoros, becsatornázható adatokat nem tud a kutatás számára produkálni. A programban a védőnői törzslapok becsatornázása a komplex koragyermekkori adatbázis egyik elemét képezte. A teljes adatbázis a 0–7 éves gyermekek adatainak folyamatos regisztrációját és követéses vizsgálatát célozta meg a szociális ellátórendszer teljes körén belül. A rendszer modulokra tagolódik, más-más adatközlőt (gyermekorvost, szakorvost, szociális gondozót, gondozót, szülőt) érintve. Az eVIR esetében ez a területileg illetékes védőnőket jelentené.

Szervezet: Országos Tisztiorvosi Hivatal (OTH).
Adatok formája: elektronikus mikroadat.
Célcsoport: 0–7 éves gyermekek.
Lefedettség: kötelezően teljes körű az adott célcsoporton.
Adatközlő: területi védőnő.
Kapcsolati kód: gyermek TAJ.
Főbb változók/adatforrások:

- törzslapok,
- nyilvántartások,
- rizikókérdőívek.

Tekintettel arra, hogy a rendszer még nem működik, a kohorszkutatás számára vélhetőleg időben nem lesz lehetséges az adatok alapadatként történő felhasználása. A későbbi becsatornázás lehetősége miatt érdemes a fejlesztést követni.

Nemzeti Adó- és Vámhivatal (NAV)

A NAV adatai munkáltatókra és munkavállalókra egyaránt vonatkoznak. A munkavállalói adatállomány munkáltatói bejelentésen alapulva fedi le a foglalkoztatottak teljes körét. Az adatok munkáltatókra és vállalkozókra is kiterjednek a társaságiadó-bevallás adatállománya révén. A munkáltatói járulékbevallási adatok havi, a társaságiadó-bevallás éves gyakoriságú.

Szervezet: Nemzeti Adó- és Vámhivatal (NAV).
Adatok formája: elektronikus mikroadat.
Célcsoport: foglalkoztatottak, vállalkozók és munkáltatók.
Lefedettség: teljes körű az adott célcsoporton.
Adatközlő: munkáltató, vállalkozó.
Kapcsolati kód: TAJ, munkáltató adószáma.
Főbb változók:

- munkáltatói adatok (társasági adó) (ágazat, tulajdonviszony, vállalatméret, telephely/székhely, árbevétel, jegyzett tőke, exportbevétel, bérköltség),
- munkavállalói adatok (08-as bevallás) (jövedelem, FEOR, biztosítási jogviszony kezdete és vége, alkalmazás minősége, heti munkaóra),
- vállalkozói adatok (58-as bevallás) (telephely, jövedelem, vállalkozás formája).

Az adatbázis becsatornázásával egyfelől egyénhez kötött munkáltatói adatokat nyerhetünk, amelyek további makrováltozókkal bővíthetők. Másfelől hozzáférhetővé válik a vállalkozói tevékenység. Harmadrészt pedig elérhetővé válnak a survey módszerrel nehezebben megszerezhető (regisztrált) jövedelmi adatok a foglalkoztatottak körében. Ezen adatok az anya szülés előtti és munkaerőpiaci visszatérése utáni időszakának vizsgálatában lehetnek fontosak.

Nemzeti Munkaügyi Hivatal (NMH) – NGM

Az intézmény országos kiterjedtségű, állami állásközvetítő szervezetrendszert működtet. Az NMH álláskeresői regiszteréből származó adatok a munkakereséshez, a munkavállaláshoz nyújtott támogatásokat – ezáltal tehát lényegében a munkanélküliség előfordulását is – mutatják. Az adatok rögzítik az egyes személyek bekerülését az álláskeresői regiszterbe, illetve az álláskeresési-járadékregiszterbe. A rendszer az igénybe vett munkaerőpiaci képzések adatait is rögzíti. 2016 óta a Nemzeti Munkaügyi Hivatal (NMH) foglalkozás-egészségügyi, illetve a munkahigiénés szakterületeit az Országos Tisztifőorvosi Hivatal, míg a munkavédelmi és foglalkoztatási szakterületeit a Nemzetgazdasági Minisztérium (NGM) vette át. A képzési terület a Nemzeti Szakképzési és Felnőttképzési Hivatalhoz (NSZFH) került.

Szervezet: Nemzeti Munkaügyi Hivatal (NMH), Nemzetgazdasági Minisztérium (NGM).

Adatok formája: elektronikus mikroadat.

Célcsoport: álláskeresők.

Lefedettség: célcsoport regisztrált tagjai.

Adatközlő: álláskereső.

Kapcsolati kód: TAJ.

Főbb változók:

- álláskeresői regiszterből származó adatok (munkaügyi kirendeltség, első regisztráció dátuma, ellátás típusa, ellátás lezárulása, iskolai végzettség, keresett állás, igényelt kereset, utolsó munkaviszony megszűnésének dátuma),
- álláskeresésijáradék-regiszter adatai (pénzbeli ellátás adatai),
- munkaerőpiaci képzések adatai (képzés helye, típusa, neve, lezárása).

Az adatállomány erőssége a munkanélküli periódusok és azok körülményeinek adminisztratív rögzítése. Ez egy ciklikus adatkérésen alapuló jelzőrendszer felépítése kapcsán lehet releváns.

Magyar Államkincstár (MÁK)

A MÁK adatain belül a gyes-, gyet- és csp-adatbázisokból gyermekvállalási adatok állnak rendelkezésre az ellátást igénybe vevők köréről.

Szervezet: Magyar Államkincstár (MÁK).
Adatok formája: elektronikus mikroadat.
Célcsoport: gyermekgondozási ellátásokat igénybe vevők.
Lefedettség: célcsoporton teljes.
Adatközlő: ellátásra jogosult.
Kapcsolati kód: TAJ.
Főbb változók:

- folyósításban érintett gyermekek születési éve,
- az ellátás típusa,
- az ellátás időtartamának kezdete és vége,
- az ellátás havi összege,
- egyedülállóság ténye,
- gyermekek száma.

Az adatállomány erőssége a gyermekvállalásra és -nevelésre vonatkozó adatok adminisztratív rögzítése. Ez egy ciklikus adatkérésen alapuló jelzőrendszer felépítése kapcsán lehet releváns.

Nemzeti Egészségbiztosítási Alapkezelő (NEAK)

A NEAK (korábban Országos Egészségbiztosítási pénztár – OEP) adatsorai munkaviszonyra vonatkozó adatokat biztosítanak, elsősorban a foglalkoztatásra és a foglalkoztatottra vonatkozóan.

Szervezet: Nemzeti Egészségbiztosítási Alapkezelő (NEAK).
Adatok formája: elektronikus mikroadat.
Célcsoport: foglalkoztatottak.
Lefedettség: teljes körű az adott célcsoporton.
Adatközlő: munkáltató.
Kapcsolati kód: TAJ.
Főbb változók:

- személyhez kötődő adatok (lakcím, állampolgárság, családi állapot, elhalálozás),
- munkaviszonyhoz kötődő adatok (foglalkozás 4 jegyű FEOR-kódja, alkalmazás minősége, foglalkoztatás kezdete és vége, külföldi munkavállalás),

– pénzbeli ellátási adatok (táppénz, tgys-, gyed-ellátás típusa, kezdete, vége).

Az adatbázis a kérdőíves kutatási elem lehetséges kiegészítéseként szolgáló foglalkoztatási adatok mellett (foglalkozás neve, típusa) a munkaerőpiaci események regisztrálására is szolgálhat. Ilyen értelemben megint csak a jelzőrendszer részeként (munkába állás, táppénz) érdemes lehet számba venni a kohorszkutatás számára. Bár a rendszer a külföldi munkavállalást is rögzíti, ennek bejelentettsége jelenleg töredezett.

Bértarifa-felvételek

A Pénzügyminisztérium (korábban NFSZ majd NGM) bértarifa-felvételei telephely szintű munkáltatói makroadatokat tartalmaznak, amelyek egyéni szinten nem integrálhatók, de a munkáltató kódszintű azonosítása esetén kapcsolhatók. A bértarifa-felvételek a verseny- és a költségvetési szférában éves rendszerességgel gyűjtik az egyéni szintű bér-, illetmény- és kereseti adatokat. A költségvetési intézmények dolgozóira vonatkozó adatok egy részét a MÁK biztosítja, így ennek lefedettsége csaknem teljes körű, a többi dolgozóról mintavételes eljárás során szolgáltatnak adatot. A versenyszféra esetében kb. 10%-os mintát vesznek minden nemzetgazdasági ágban a legalább 5 főt foglalkoztató vállalkozások/nonprofit szervezetek alapsokaságán.

Szervezet: Pénzügyminisztérium.

Adatok formája: elektronikus makroadat.

Célcsoport: foglalkoztatók.

Lefedettség: a célcsoporton vett minta.

Adatközlő: foglalkoztató.

Kapcsolati kód: munkáltatói szintű (szervezeti törzsszám, szakágazati jelzőszám).

Főbb változók:

– szervezet jellemzői (azonosító, fizikai/szellemi foglalkoztatotti létszám, teljes/részmunkaidős létszám, állami/önkormányzati/külföldi tulajdoni arány, településazonosító, TEÁOR),
– egyéni jellemzők (demográfiai adatok, belépés ideje, iskolai végzettség, szabadság, foglalkoztatási forma, szolgálati idő),
– béradatok (kereset, alapbér, bérpótlék, jutalék, órabér, munkaóra, díjak és pótlékok).

A bértarifa-felvételek makrováltozóként kapcsolt adatállományának bevonása egy esetleges munkáltatókra vonatkozó vizsgálat kapcsán merülhet fel

(például a gyermekvállalás előtti és utáni foglalkoztatói mintázatok összevetése során). Ennek feltétele, hogy a munkáltatóra vonatkozó kapcsolati kód más adatbázisból is rendelkezésre álljon.

Magyar Államkincstár (MÁK)

A MÁK (korábban ONYF) adatbázisa a biztosított és kieső munkavállalási időszakokra vagy a munkavégzésre vonatkozóan rögzít adatokat. A MÁK adattartalom a fenti munkaerőpiaci adatbázisok adatain alapul, tehát felhasználását inkább valamely adatszolgáltató kiesése esetén érdemes mérlegelni.

LEHETSÉGES REGISZTERALAPÚ ADATKAPCSOLÁSI ELJÁRÁSOK

Az alábbiakban a születési kohorszvizsgálat kapcsán esetlegesen alkalmazható adatkapcsolási eljárásokat vesszük számba. A vizsgálatban ezidáig (2018 végéig) a várandóssággal és a születéssel kapcsolatos, fentebb külön jelzett adatkapcsolások valósultak meg. A lehetőségek általános bemutatására törekedve az alábbiakban a lehetséges eljárások szakaszolása és tartalmi specifikálása helyett az eljárások elvének ismertetése a célunk példa- vagy ötletszinten, kitérve a kohorszkutatáson belüli alkalmazás esetleges lehetőségeire.

Azokat az adatkapcsolási eljárásokat tekintjük tehát át, amelyekben a bevont adatbázisok legalább egyike egyéni szintű és legalább egyike adminisztratív adatokat tartalmaz. Rendelkezésre áll továbbá az adatkapcsoláshoz egy egyedi azonosítást lehetővé tévő kapcsolati kód vagy a személyes adatok (legalább valószínűségi) azonosításra alkalmas kombinációja, amely lehetővé teszi, hogy egy adatbázis valamely megfigyelési egységének adataihoz más adatbázisból adatokat kapcsoljunk. E formai jellemzőkön túl tartalmilag az adatkapcsolást olyan eljárásnak tekintjük, amely több (vagy több időszakra vonatkozó) adatbázis információtartalmának integrálását célozza meg (Veroszta 2015).

Adminisztratív adatbázison belüli összekapcsolás

Ennek során egyazon regiszter különböző időpontokra vagy tartami elemekre vonatkozó adatait kapcsoljuk az adott megfigyelési egységek egyedi adatsora-

ihoz. Ez a longitudinális vizsgálatok alapeljárása, amely regiszteradatokon akár időben visszamenőleg is megvalósulhat. Ez esetben az adatok egyéni szintű azonosítása a rendszer adottsága. Ilyen kutatási elem lehet például a mintába kerülő nők szülés előtti munkaerőpiaci életútjának vizsgálata NEAK-adatokon.

A követéses vizsgálathoz, azaz az időben előrehaladó, rendszeres adatkapcsoláshoz az adatbázisokon belül szükséges a kapcsolati kódok biztosítása. Ez a nyilvántartás keretén belül természetesen megoldott, de a kutatás számára praktikusan azt jelenti, hogy a lekért adminisztratív adatbázist az összekapcsolás érdekében nem anonimizálják, vagy – az adatvédelmi szempontok figyelembevételével – az egyedi azonosítókat (felfejtő kódot) a későbbi adatkapcsoláshoz külön rendszerben kezelve elérhetővé teszik.

Longitudinális adatkapcsolásra kerülhet sor például abban az esetben, ha a védőnői rendszerből vett ügyfélminta adataihoz kapcsoljuk ciklikusan a gyermek ugyanabban a rendszerben rögzített fejlődési adatait. Ez esetben is felmerül az egyedi azonosítás folytonosságának biztosítása, a kutatási adatbázisra vonatkozó anonimitási kritérium mellett. Szintén ezzel az eljárással követhető a szülő nők munkaerőpiaci helyzetének későbbi alakulása a NEAK vagy NAV adatbázisán belüli ciklikus lekéréssel.

Adminisztratív adatbázisok közti összekapcsolás

A több adminisztratív – azaz teljes körű, egyéni szintű – adatbázis közti egyénsoros összekapcsolás több módon is megvalósulhat. Egyszerűbb esetben mindkét adatbázisban megtalálható az az egyedi azonosító vagy kapcsolati kód, amely az adatsorok egyéni szintű összekötését lehetővé teszi (match-merge eljárás). A kapcsolati kód hiánya esetén az adatkapcsolási eljárás a mindkét adatbázisban megtalálható személyes adatok olyan kombinációjára épül, amelyek egyéni megkülönböztetést tesznek lehetővé (deterministic linkage).

A hazai adminisztratív adatkapcsolási gyakorlatban mind az adatkapcsolásért, mind a létrejött adatbázis anonimizálásáért egy külső szervezet (NISZ Zrt.) felel. Az anonim adatkapcsolás egyirányú algoritmussal (HASH-kód alkalmazásával) történik, amely a kapcsolati kódok utólagos visszafejthetetlenségét garantálja. Az anonimizálási kötelezettség a kapcsolati kódok kiemelése mellett arra is vonatkozik, hogy a kapcsolást végző szervezet – saját szakmai mérlegelése alapján – a személyes azonosíthatóságot akár kombináció révén lehetővé tévő változókat kiemelje az adatbázisból, vagy magasabb szinten aggregálja (pl. irányítószám helyett kistérséget és településtípust rögzítve).

A kohorszkutatás kapcsán a fentiekben számba vett adatbázisok esetében – többnyire a TAJ alapján – a kapcsolatikód-alapú egyszeri kapcsolás technikailag megoldhatónak tűnik. Keresztmetszeti vizsgálat (illetve regiszterek esetében egyszeri kapcsolás) esetén azonban az anonimizálás az adatkapcsolást végző szervezet kötelezettsége, így az adminisztratív adatbázisok összekapcsolásakor nem valósítható meg a kapcsolati kódok longitudinális survey vizsgálatok során alkalmazott elkülönített tárolása. Tekintettel arra, hogy a jelenleg hatályos, egyirányú HASH-algoritmuson alapuló adatkapcsolási eljárás nem teszi lehetővé a két adatbázis közti kapcsolati kód későbbi felfejtését, az adminisztratív adatbázisok integrációján alapuló longitudinális vizsgálatokra a jelenlegi szabályozási környezetben nincs lehetőség (a problémákat összegző fejezetben kitérünk majd a szabályozásból adódó jelenlegi mozgástérre).

Adminisztratív adatok összekapcsolása makrováltozókkal

Ezekben az adatkapcsolásokban az adminisztratív adatbázisban lévő egyénsoros (illetve a megfigyelési egységre vonatkozó) mikroadatokhoz magasabb aggregáltsági szintű adatokat – külső makrováltozókat – kapcsolunk. Az adatkapcsolás alapját a megfigyelési egységek valamely más adatbázis kategorizációjának megfelelő csoportosítása képezi. Ilyen kontextuális információt nyújthatnak például a hivatalos statisztikai adatgyűjtés mutatói, indexei.

Ez az adminisztratív és a survey adatbázisokon gyakran alkalmazott eljárás minden bizonnyal a kohorszkutatásban is fontos szerepet kap majd. Kézenfekvő a településstatisztikai adatok becsatornázása, de vélhetőleg az intézményi ellátáshoz kapcsolódó makroadatok (védőnői, egészségügyi, gyermekellátási, oktatási körzetek vagy intézmények statisztikai adatai) is fontos szerepet kapnak majd a kutatás során. Ehhez a kutatási adatbázisnak a válaszadók/megfigyelési egységek intézményi/regionális hovatartozását kódszinten szükséges biztosítania.

Regiszteralapú survey kutatás

Ezekben a keresztmetszeti vagy longitudinális adatkapcsolásokban a teljes körű adminisztratív adatbázis megfigyelési egységre vonatkozó egyedi adatsorai egészülnek ki survey kutatásból származó, a mintára vonatkozó szintén egyedi adatsorokkal. A regiszteralapú survey kutatás esetén a mintaképző adminisztra-

tív adatbázis már rendelkezésre álló információtartalmának összekapcsolása a csak személyes megkérdezés révén elérhető adatokkal nemcsak a két adattípus erényeit kombinálja, hanem jelentősen csökkenti a válaszadói terheket (ezáltal a költségeket) is. Ez esetben kapcsolati kódként az egyén eléréséhez szükséges adat szolgál (pl. lakcím, e-mail cím, telefonszám).

Regiszter- és survey adatok utólagos összekapcsolása

Ebben az esetben már meglévő különböző forrású, ám azonos alappopulációt lefedő adminisztratív, illetve survey adatbázisok egyedi szintű összekötéséről van szó. Ennek során a surveyben szereplő egyes válaszadók adatkombinációiból állítunk össze egyedi azonosításra alkalmas adatcsomagokat, és keressük megfelelését az adminisztratív adatbázisban szereplő elvileg teljes alappopuláció egy tagjával. Az ehhez alkalmazott módszer a valószínűségi adatkapcsolás (probabilistic record linkage), amely statisztikai eljárással azonosítja a két, azonos alappopulációt lefedő adatbázis tagjai közti kapcsolatot. A módszer (akárcsak a deterministic linkage) a mindkét adatbázisban rendelkezésre álló egyéni szintű, személyes adatok kombinációjára épül, de nem követeli meg a megfigyelési egységszintű teljes egyezést. Az eljárás megkülönböztet egyező, nem egyező és bizonytalanul egyező kapcsolódást a különböző forrású adatsorok között. Utóbbi esetben az egyezés statisztikai valószínűségével számol és ezt rendeli hozzá az egyedi adatsorhoz. Az eljárás révén a kutatási adatok információtartalma utólagosan egészül ki a regiszterekből származó információkkal, de az adatkapcsolás a szisztematikus hiba azonosításában is fontos módszertani eszköz lehet.

A valószínűségi adatkapcsolás kohorszkutatáson belüli alkalmazását abban az esetben lehet érdemes megfontolni, ha a regiszter- és survey adatok közti közvetlen, kapcsolati kódon (TAJ) avagy kontaktadaton (regiszteralapú mintán) alapuló összekötés nem tud megvalósulni.

KUTATÁSI SZAKASZOK ADATKAPCSOLÁSI LEHETŐSÉGEI

Az alábbiakban a fenti regiszteralapú kutatási eljárások és adatbázisok elhelyezésére törekszünk a hazai kohorszkutatás kutatási programján belül. Ennek során a főbb kutatási szakaszokhoz, avagy vizsgált életszakaszokhoz társítjuk a fentiekben részletezett elemeket. Nyilván – bár erre külön nem térünk ki – adott-

nak vesszük az egyes kutatási szakaszok survey adatbázisainak egymáshoz kapcsolhatóságát.

VÁRANDÓSSÁG IDŐSZAKA		
Eljárás	**Tartalom**	**Adatbázis**
Körzetszintű mintavétel (megvalósult)	Mintaképzés a háttéradatok beemelésével	Várandósgondozási rendszer, védőnői jelentésösszesítők
Survey és regiszteradatok összekapcsolása (folyamatban)	Védőnők által rögzített várandósgondozási könyv adatainak bekapcsolása	Várandósgondozási könyv – védőnői rögzítés
Makrováltozók bevonása	Települési háttéradatok kapcsolása (lakóhelyre)	KSH
SZÜLÉS		
Eljárás	**Tartalom**	**Adatbázis**
Survey és regiszteradatok összekapcsolása (folyamatban van)	Szülésre vonatkozó adatok beemelése anyai TAJ alapján	Születésértesítő rendszer
Survey és regiszteradatok összekapcsolása	Anya munkaerőpiaci részvételének retrospektív adatai	Kincstár
Makrováltozók bevonása	Települési háttéradatok kapcsolása (kórházra és/vagy lakóhelyre)	KSH
GYERMEK RENDSZERES VIZSGÁLATA		
Eljárás	**Tartalom**	**Adatbázis**
Survey és regiszteradatok összekapcsolása	Anya munkaerőpiaci életútjának vizsgálata	Kincstár

Aktuális életesemények

Egy „adminisztratív jelzőrendszerként" definiálható, egyelőre ötletszinten tervezett kutatási elem a kohorszkutatás speciális kiegészítését jelentené. A kutatási elem lényege, hogy az időben rögzített kutatási szakaszok mellett a megfigyelteket (gyermek, illetve anya) azonosítja a szociális és oktatási rendszer, valamint a munkaerőpiaci adminisztráció különböző belépési pontjain, majd erre reagálva eseti adatfelvétel végezhető. Az eljárás előnye, hogy a fejlődési szakaszokhoz igazított kutatási szakaszok mellett az egyedi életeseményekhez is igazítja az adatfelvételt. Az eljárás módszere lényegében regiszteralapú survey kutatás. A gyakorlatban ez a „jelzőrendszer" azt jelentené, hogy az adminisztratív rend-

szerek a kutatás számára visszacsatolást küldenének, amennyiben a megfigyelt (anya vagy gyermek) az adatbázisban regisztrálásra kerül. Így válna lehetővé például a gyermekek vizsgálata egységesen a bölcsődei ellátás kezdete utáni hónapokban, vagy az óvoda megkezdésekor. Szintén ez az eljárás tenné lehetővé az anyák felkeresését munkába állásuk után néhány hónappal, vagy újabb gyermekvállalásuk, sőt esetleges munkanélkülivé válásuk esetén.

E kutatási modul módszertani háttere az előbbi séma szerint az alábbiként összegezhető:

	AKTUÁLIS ÉLETESEMÉNYEK	
Eljárás	Tartalom	Adatbázis
Regiszteralapú survey	Bölcsődei gondozásba vétel jelzése	Önkormányzat
	Óvoda megkezdésének jelzése	Közoktatás információs rendszere (KIR)
	Testvérszületés jelzése	Kincstár
	Anya munkába állásának jelzése	Kincstár
	Anya munkanélkülivé válásának jelzése	NMH

Az adatgyűjtő szervektől érkező – technikailag működőképes, de reálisan nem megvalósítható – jelzés kiváltásaként a rendszeres, egymástól független (pl. negyedéves gyakoriságú) adminisztratív adatbázis-lekérésekkel az egyes életesemények bekövetkeztének gyakorisága vizsgálható az adott mintán. Ez azonban az anonimizálás következtében nem teszi lehetővé a kohorsztagok célzott megkeresését az esemény utáni egységes időkeretben. Minderre tekintettel ez esetben csak visszamenőleges vizsgálatban gondolkodhatunk: amennyiben az egyes életesemények a rendszeres vizsgálati periódusok valamelyikében felmerülnek (pl. az interjúból kiderül az anya munkába állása, munkanélkülisége), az adott almintára kérhetünk regiszteralapú kapcsolást a visszamenőleges adatokra. Ez azonban nem szolgálja az adatfelvétel életeseményekhez kötött rugalmasságát, pusztán regiszteradatok problémacentrikus gyűjtésének tekinthető.

ÖSSZEGZÉS: NYITOTT KÉRDÉSEK ÉS PROBLÉMÁK

Az alábbiakban a fentiek összegzéseként áttekintjük azon kérdések sorát, amelyek átgondolása az adminisztratív adatok kohorszkutatáson belüli felhasználásához bizonyosan szükséges. A problémák jelzésekor a jelenlegi szabályozási

környezet adta lehetőségekből/korlátokból indulunk ki, tehát a kutatás kapcsán esetlegesen születő egyedi megegyezéseket, adatátadási gyakorlatokat nem tudjuk átgondolni – ezek a speciális feltételek nyilván a későbbiekben lazíthatnak a most körvonalazott keretrendszeren. A cél egyfelől a már e szakaszban azonosítható problémák jelzése és lehetséges megoldásuk átgondolása, másfelől a tervezésre épülő szemléletmód megnyitása a későbbiekben felmerülő, további problémák felé is – ilyen értelemben cél a lista folyamatos bővítése, módosítása a kutatási terv véglegesedése során.

A longitudinális adminisztratív adatkapcsolás esetében problémát okozhat, hogy a jelenlegi szabályozási környezet akadályozza az időben eltérő adatkapcsolások megvalósítását. Az eseti adminisztratív adatkapcsolásokhoz a NISZ Zrt. által alkalmazott anonimizálási eljárás nemcsak a kapcsolati kód visszafejthetetlenségét, hanem a személyes adatok azonosításra alkalmas kombinációinak megszüntetését is garantálja az átadott adatbázison. Emiatt az adatok későbbi kiegészítése akár azonos adatbázisok későbbi adataival, akár más adatbázisok adataival nem megvalósítható. Az időbeliség egyetlen feloldásának az látszik, ha minden vizsgált időpontra új adatösszekapcsolási kérést indítunk el, amely így értelemszerűen magába foglalja majd a korábbi adatkérések által lefedett időszakokat is. Ez azonban jelentősen megnöveli az adattisztítási, adatbáziskezelési költségeket.

Regiszter-survey adatkapcsolás technikailag minden olyan adminisztratív adatbázison elvégezhető, amely a vizsgálati populációt lefedi, és hozzájuk a mintaképző változók mentén egyéni azonosításra alkalmas adatokat (kontaktadatokat) ad. A problémát az anonimitás és az összekapcsolhatóság biztosítása közti feszültség jelentheti. Regiszteralapú survey esetén a regiszterből a mintába kerültek kontaktadatai (és egyéb háttéradatai) mellett egy olyan azonosításra alkalmas kód is szükséges (célszerűen TAJ-alapú), amely a későbbi kutatási szakaszok regiszter- vagy survey alapú adatai is tudnak egyénsorosan kapcsolódni. A regiszter-survey adatkapcsolás másik kérdéses pontja, hogy milyen szabályozás vonatkozik azokra az adminisztratív adatkapcsolási kérésekre, amelyek kapcsolati kódja survey kutatásból származik. Kérdés tehát, hogy pl. a kutatás későbbi szakaszában kérhető-e egyáltalán a különböző regiszterek adatainak kutatási adatbázishoz kapcsolása. Ezen személyes adatok kiadhatóságát és összekapcsolását a jelenlegi szabályozás (GDPR) alapján a válaszadó eseti hozzájárulása teszi lehetővé. A kutatás az eddig megvalósított adatfelvételek és adatkapcsolások esetében eszerint járt el.

Longitudinális kutatás esetében különös jelentőséggel bír az adatbázisok „nyitva tartása", azaz annak lehetősége, hogy a kutatás előrehaladtával az egyre

frissülő adatok egyéni szinten becsatornázhatóak legyenek. Az adminisztratív adatok körében ez jelenleg nehezen megoldható. Jóllehet a regiszteradatok kutatási felhasználása mára szabályozott keretek között megvalósulhat, a jelenlegi szervezeti és jogi közeg nem tudja kezelni az időbeli ismétlődés kérdését. Adatkapcsolások tehát adott időpontra vonatkozóan és időben visszafelé létrehozhatók ugyan, de az összekapcsolt adatállományok ez után sem időben, sem tartalmukban nem bővíthetőek. Nincs tehát lehetőség új adatkörök vagy új adatbázisok (akár survey adatok) későbbi bevonására, vagy ugyanazon adatkörök frissebb adatainak beemelésére az eredeti adatbázisba. A longitudinális kutatási design ugyanakkor épp az egyénsoros adatok időbeli bővítésén alapul, ilyen értelemben megköveteli az adatbázisok összekapcsolhatóságát, azaz nyitva tartását a későbbi kutatási szakaszok felé. A szabályozás várhatóan középtávon nem fog változni, így arra kell számítanunk, hogy az elindított adatbázisintegrációk mind külön egységet alkotnak, a későbbi kapcsolódás lehetősége nélkül. Nagy jelentőségű regiszteradatok esetében ezért csak az ismételt adatkéréseket lehet elképzelni, egyre növelt időtávokon, amelyben a mintaképző adatbázis mindig azonos, hiszen ezek későbbi becsatornázása sem megoldható az integrált adatbázison, a kapcsolati kódok eltüntetése miatt.

IRODALOM

Cseres-Gergely Zsombor – Scharle Ágota 2008: Az államigazgatásban keletkező adatok nyilvánosságáról. In Köllő *János (szerk.)*: *Áttekintés az adminisztratív adatbázisokkal és teljeskörű összeírásokkal kapcsolatos kutatási tapasztalatokról*. 15 p. Forrás: http://adatbank.krtk.mta.hu/data/egy/2.pdf Letöltés: 2018.01.07. Eötvös Károly Intézet 2006: *Hozzáférés a közszféra adataihoz*. 278 p.

Fodor Szabolcs – Veroszta Zsuzsanna 2013: Államigazgatási adatok pályakövetési célú integrációja a hazai gyakorlatban In Garai Orsolya – Veroszta Zsuzsanna (szerk.) *Államigazgatási adatbázisok a diplomás pályakövetésben*. Educatio Társadalmi Szolgáltató Nonprofit Kft., Budapest, 83–128.

Gárdos Éva 2015: Adatok és kezelésük a hivatalos statisztikában. *Educatio*, (3), 27–39.

Kapitány Balázs 2018: Az alapsokaság meghatározásának, a minta kialakításának gyakorlati lépései. In Veroszta Zsuzsanna (szerk.): *Kohorsz '18 Magyar Születési Kohorszvizsgálat módszertani leírás. A várandós kutatási szakasz előkészítése.* Kutatási Jelentések 99. KSH Népességtudományi Kutatóintézet, Budapest. DOI: 10.21543/Kut.2018.99

KSH 2014: *Módszertani dokumentáció/Fogalmak, definíciók*. Forrás: http://www.ksh.hu/apps/meta.menu?p_lang=HU&p_menu_id=220&p_param=S&p_session_id=82019363 Letöltés: 2018.01.07.

Nyüsti Szilvia – Veroszta Zsuzsanna 2014: *Diplomás pályakövetési adatok 2013 – Adminisztratív adatbázisok integrációja*. Educatio Társadalmi Szolgáltató Nonprofit Kft., Budapest, 114 p.

Nyüsti Szilvia – Veroszta Zsuzsanna 2015: *Diplomás pályakövetés 2014 – Adminisztratív adatbázisok integrációja – Gyorsjelentés*. Educatio Társadalmi Szolgáltató Nonprofit Kft., Budapest, 20 p.

Rohr Adél 2016: *A webes felületen keresztüli adatszolgáltatás lehetőségei - A védőnői adatgyűjtési rendszerek lehetőségei.* NKI Születési Kohorszvizsgálat. Prezentáció 2016. február 29. KSH NKI, Budapest.

Szabó Ákos 2011: Államigazgatási adatbázisok integrációjának hazai közege. In Garai Orsolya – Veroszta Zsuzsanna (szerk.): *Államigazgatási adatbázisok a diplomás pályakövetésben*. Educatio Társadalmi Szolgáltató Nonprofit Kft., Budapest, 47–83.

Székely Iván 2015: Közadatok és nyilvános adatbázisok: a hozzáférés kérdései. *Educatio*, (3), 40-49.

Veroszta Zsuzsanna 2015: Adminisztratív adatok társadalomkutatási kezelése. *Educatio*, (3), 3–14.

Veroszta Zsuzsanna (szerk.) 2018: *Kohorsz '18 Magyar Születési Kohorszvizsgálat módszertani leírás. A várandós kutatási szakasz előkészítése.* Kutatási Jelentések 99. KSH Népességtudományi Kutatóintézet, Budapest. DOI: 10.21543/Kut.2018.99

USING ADMINISTRATIVE DATA IN COHORT STUDIES

ABSTRACT

The study reviews the possibilities of integrating administrative and survey data for research purposes under current Hungarian conditions, focusing on a special area of birth cohort studies. Within this context it summarizes the available procedures, tools, regulations and datasets from the perspective of Growing Up in Hungary - Cohort '18, a longitudinal, survey based birth cohort study of a HCSO Hungarian Demographic Research Institute. Besides this, on a more general level, the results of the analysis can provide new perspectives of the use of administrative data in social research for current survey based studies.