Hozzászólás a „Nagyszámú laboratóriumi vérvizsgálati eredmény exploratív jellegű vizsgálata rangkorrelációval" című közleményhez

Egy nagymintás exploratív vizsgálat érvényessége a limitációi tükrében

Brys és munkatársai e lap hasábjain a 2019. 12. számban a 609–614. oldalon megjelent „Nagyszámú laboratóriumi vérvizsgálati eredmény exploratív jellegű vizsgálata rangkorrelációval" című közleményükben egy igen nagy méretű, laborvizsgálati eredményeket tartalmazó adatbázist elemeztek a korrelációszámítás eszközével *(1)*. Vizsgálati módszerük számos félreértést és kisebb-nagyobb hibát tartalmaz, amely igen lényegesen korlátozza eredményeik érvényességét. A jelen hozzászólásban szeretnénk felhívni a figyelmet a legfontosabb problémákra: arra, hogy a vizsgálat teljesen különböző betegcsoportokat (sőt, potenciálisan még egészségeseket is) mos egybe, hogy a „big data" címke használata indokolatlan, hogy a kutatás az ismételt méréseket nem kezeli, hogy az erőelemzés teljesen zavaros és fölösleges is abban a formában, ahogy a cikkben megjelenik, hogy a szokásos nullhipotézis-szignifikanciateszt típusú eljárások alkalmazása tökéletesen értelmetlen ekkora mintanagyság mellett, hogy az eloszlásvizsgálatok használata indokolatlan, és végezetül, hogy a cikkben összekeveredik a változók normalitása kapcsolatuk linearitásával. Hozzászólásunk nem kérdőjelezi meg az „exploratív vizsgálatok" létjogosultságát, ám fel kívánjuk hívni a figyelmet arra, hogy a feltáró jelleg mit sem változtat az elméleti szigoron (sőt, csak még fontosabbá teszi azt). Fontos, hogy módszertanilag helyesen, a kutatás limitációit világossá téve mutassák be és értékeljék a kutatás tudományos eredményeit.

Bevezetés

Brys és munkatársai cikkükben *(1)* egy nagy méretű, laborvizsgálati eredményeket tartalmazó adatbázist elemeztek páronkénti korrelációk felhasználásával. Feltétlenül üdvözölnünk kell a tényt, hogy az ilyen jellegű adatbázisok számítógépes feldolgozás számára elérhetővé válnak, és orvosbiológiai hasznosításuk is megjelenik a hazai irodalomban.

Azonban sajnálatos módon a hivatkozott cikk több szempontból is rossz példát mutat, mert olyan módszertani félreértések, meg nem értések és kifejezett hibák találhatóak benne, melyek az eredmények érvényességét nagyon erősen limitálják – ez azonban nem kellő hangsúllyal, illetve a legtöbb esetben egyáltalán nem derül ki a cikkből.

Mivel a jövőben várhatóan csak növekedni fog a hasonló adatbázissal rendelkező, azt feldolgozni tervező kutatók száma, úgy véljük, hogy a teljes magyar orvoskutatói közösség számára hasznos, ha egy hosszabb levélben foglaljuk össze a cikk kapcsán kiemelendő legfontosabb módszertani problémákat.

A felfedező kutatás célja

A szerzők ezt írják: „[a]z [...] exploratív orvosi adatelemzési kutatás célja az volt, hogy megvizsgáljuk; valós, értelmezhető eredményekhez vezet-e a változók közötti korrelációk vizsgálata e nagy nem véletlen mintán", majd kicsit később ezt: „[a] PubMed-en a különböző, vonatkozó kulcsszavakkal (például: »blood test«, »statistical«, »exploratory data analysis«) kapott találatok között feltáró jellegű kutatást bemutató közleményt nem találtunk". Talán nem véletlenül – gondolhatja az olvasó. Önmagában egy puszta korrelációbecslés nem bír a felfedezés erejével akkor sem, ha statisztikai szempontból minden helyesen történik. Nem véletlen, hogy rákeresve ugyanezekre a kulcsszavakra, általában olyan cikkekre bukkanunk, amelyek biomarkereket kutatnak vagy mintázatokat keresnek a különböző betegségek diagnosztizálására. A korreláció pontbecslése önmagában már csak azért sem túl érdekes, mert több – vélhetően élesen eltérő – alcsoportot mindenféle megkülönböztetés nélkül összeöntve számolják a szerzők. Ilyenek a betegek és az egészségesek, sőt, a betegeken belül az összes létező betegség is egy kalap alatt szerepel ebben a cikkben. Nagyon erősen kérdéses, hogy mire használható *bármilyen* eredmény, mely a kisujjtöröttektől a tüdőrákosokon és a pszichotikusokon át a vesekövesekig terjedő, *és* még néhány egészségest is tartalmazó halmaz *egészére* vonatkozik... Gondoljunk bele, a korreláció elvileg akár még ellentétes előjelű is lehet az egyes alcsoportokban *(2)*, de ezt így soha nem fogjuk megtudni!

Fontos rögzíteni, hogy a „felfedező kutatással" önmagában semmiféle probléma nincsen, de másrészről azt is világosan látni kell, hogy ez nem fügefalevél, mellyel eltakarhatóak a problémák, a „felfedező kutatás" nem valamiféle varázscímke, amit ráragasztva egy módszertanilag bármilyen hibákkal bíró kutatásra, az hirtelen informatívvá válik. Félreértés ne essék: az nem gond, ha egy cikknek módszertani limitációi vannak (a világon

minden cikknek vannak!), a probléma az, ha ez nem kerül bevallásra, és emiatt nem világos, hogy a cikk eredményeinek mekkora a bizonyító ereje.

A minta véletlenségéről

A szerzők megállapítják, hogy „nem random adatok[ból]" dolgoztak, és láthatóan úgy érzik, hogy emiatt mentegetőzniük kell (pedig nem – de erről picit később). Csakhogy ebbe a mentegetőzésbe aztán mindenféle fogalmat belekevernek, a szociológiai minta reprezentativitásától kezdve a rétegezésen át a placebokontrollos vizsgálatokba való betegbesorolásig. Egyiknek sincs túl sok köze a jelen kérdéshez. Egy kirívó példa: „A minta nem volt véletlen jellegű (random), hiszen a mintába az egészséges és a beteg emberek bekerülési valószínűsége nem volt egyenlő" – eszerint, ha egészséges és beteg emberek egyforma valószínűséggel kerültek volna be, akkor – ebből a szempontból – random lett volna a minta? (Akkor azokkal a vizsgálatokkal, amelyekben *kizárólag* betegek vesznek részt, szükségképp valami gond van...?)

Végül is ez leginkább egy retrospektív kohorszvizsgálathoz hasonlít, és nem baj, hogy nem random, a baj a fogalomzavar. Az orvosbiológiai vizsgálatok mintája nem kell, hogy random legyen; ez még a klinikai kísérletekre is igaz: az ilyenek mintája általában *nagyon* nem random, de ez nem feltétlenül okoz problémát, ha relatív – és nem abszolút – mutatókat használunk.

Nem mind big, ami nagy

A szerzők mintája a szó szokásos statisztikai értelmében nagy, sőt rendkívül nagy – de másrészről *egyáltalán nem* „big data" (noha még a dolgozat kulcsszavai között is szerepeltették ezt). Kérdés persze rögtön, hogy e fogalomnak mi a definíciója, a talán legáltalánosabban használt megközelítés azt mondja, hogy az, ami „hagyományos" számítási eszközökkel már nem kezelhető. Rögtön látszik persze, hogy ez nem túl egzakt meghatározás (mi definiálja a „hagyományos" számítási kapacitást?), de egy dolgot rögzíthetünk: 2,3 millió megfigyelés 49 változóra, úgy, hogy egyszerű egyváltozós vizsgálatokat és kétváltozós korrelációkat kell számolnunk, *nem* big data. Ez hagyományos eszközökkel is feldolgozható, amit elég jól igazol, hogy a szerzők – feltesszük – teljesen hagyományos eszközökkel, szuperszámítógép és speciális programok, algoritmusok nélkül dolgozták fel...

Csak egy, a nagyságrendeket szemléltető példa gyanánt: hagyományos eszközökkel feldolgozható olyan adatbázis, melyben 10 millió megfigyelési egységre illesztünk egy olyan statisztikai modellt, amelynek 8 ezer (!) becsülendő paramétere van... *(3)*.

Összefoglalva, vigyázni kell azzal (és ez messze nem csak a szóban forgó cikkre vonatkozó megjegyzés), hogy a big data ne váljon olyan kifejezéssé, amit teljesen hétköznapi eszközökkel feldolgozható méretű adatbázisokra is ráillesztünk, csak azért, mert ez egy jól hangzó hívószó mostanság és „trendivé" teszi a kutatásunkat az olvasó előtt.

Az ismételt mérések problémaköre

A szerzők ezt írják: „egy személy többször is a mintába kerülhetett (egy betegnél jellemzően több alkalommal végeznek vérvizsgálatot)". Ha jól értjük, akkor a szerzők ezt nem tudták egyértelműen beazonosítani (más helyen ugyanis azt írják, hogy csak a következő három adat volt az adatbázisban: laborvizsgálat-azonosító, vizsgálatazonosító, a vizsgálati érték eredménye – úgy tűnik egyént azonosító változó, ami alapján ezt ki lehetett volna szűrni, nem volt). Ha viszont egyes betegek többször is bekerülhettek, ráadásul nem tudjuk, hogy hányszor, milyen gyakran, akkor a becsült korreláció torzított, mert keveredik benne az egyeden belüli korreláció az egyedek közöttivel! A kettő eltérése tetszőleges lehet, még csak annyi sem biztos, hogy egyáltalán az előjelük megegyezik *(4, 5)*.

Az erőelemzésnek nincs értelme abban a formában, ahogy a cikkben szerepel

Röviden megfogalmazva: egy statisztikai próba ereje annak a valószínűsége, hogy ha a nullhipotézise a valóságban nem áll fenn („van hatás"), akkor ezt a nullhipotézist a minta alapján ténylegesen el is utasítjuk („észre is vesszük a hatást"). Ez természetesen – több egyéb paraméter mellett – függ attól, hogy a hatásnak mekkora a nagysága: mindent változatlanul tartva nagyobb hatásra nézve nagyobb lesz az erő, hiszen nagyobb hatást értelemszerűen könnyebb észrevenni. A klinikai kutatásokban gyakran látott erőelemzés lényege, hogy a fenti összefüggésre fordítva nézünk rá: *feltételezünk* egy hatást, majd kiszámoljuk, hogy *ha* tényleg akkora a hatás, akkor annak

adott erővel, azaz adott valószínűséggel (például 80%) történő kimutatásához hány beteg bevonására lenne szükség. (Mert, ha kevesebbet vonunk be, kisebb lesz az erő, ha többet, nagyobb.) És ez alapján indul a betegbevonás.

Pusztán ennyit elég tudni, hogy látszódjon, miért teljesen értelmetlen a szerzők eljárása: ebben a vizsgálatban szó nem volt arról, hogy előzetesen meg kell tervezni a bevonandó betegek számát: az jelen esetben *eleve adottság* volt! (2013-ban a BIK-laborban vizsgált betegek.) A kiszámolt 760-as szám akkor lenne érdekes, ha az lenne a helyzet, hogy a számításunk *alapján* hívunk be valamekkora számú beteget laborvizsgálatra; itt erről szó nincsen. Ha a betegeket nem a vizsgálat céljából kell, erőforrás ráfordításával, begyűjteni, akkor az egész erőelemzésnek nincsen semmi értelme: ez esetben egész egyszerűen használni kell mindenkit, aki elérhető, mindenféle számolás nélkül.

A 760-as szám tehát, bár nem hibás, teljesen felesleges jelen esetben (és tegyük hozzá, egyedül a 0,3-nek feltételezett elméleti korreláció esetén igaz). Ennek meghatározása után azonban végképp követhetetlenné válik a szerzők gondolatmenete. Először is, a 760 pontos kiszámítása után közlik, hogy „[t]ekintettel a vizsgálat exploratív jellegére" (ez nem befolyásolja az erőt), „az adatok nagy mennyiségére" (valószínűleg arra gondolnak, hogy a megfigyeléses vizsgálatoknál alkalmazott egyes eljárásoknak van minimális adatigénye, de később nem alkalmaznak ilyen eljárást) és a „standard hiba (a mintavétel eloszlásának szórása) megfelelő csökkentése" (akkor az előbbi, 760-hoz vezető számítás paraméterei mégsem csökkentették „megfelelően" a standard hibát…? de ha így van, akkor miért olyan paraméterekkel számoltak?) „érdekében 10 000-es minimális mintaelemszámot határoztunk meg". De miért? Hogyan jött ez a szám ki? Miért pont 10 000 és nem 10 001 vagy 9999? Másik oldalról, ha ezt a szerzők csak így meg tudták határozni, bármilyen módon is, akkor meg egyáltalán minek kellett a G-Power szoftver, a szépen felírt paraméterek, az ábra, meg a 760-as eredmény...?

Ekkora mintanagyság mellett a szokásos nullhipotézis-szignifikanciateszt típusú vizsgálatok tökéletesen értelmetlenek

A szerzők által használt nullhipotézisek „pont null" jellegűek, tehát azt kötik ki, hogy a paraméter (Spearman-féle rangkorreláció) valódi értéke egyetlen konkrét szám (0). A „pont null" jellegű nullhipotézisek a valóságban szinte soha nem teljesülnek egzaktan egy orvosbiológiai helyzetben, tehát mindig cáfolhatóak, ha kellően nagy mintát gyűjtünk. Pontosan emiatt *teljesen felesleges* arról beszélni, hogy mennyi a *p*-érték, hiszen ekkora mintanagyságnál előre megmondható, hogy minden *p* szinte 0 lesz. Ami egyedül érdekes, az a hatásnagyság, jelen esetben a korreláció nagysága: a *klinikai* szignifikancia (relevancia) érdekes, hiszen a *statisztikai* szignifikancia praktikusan automatikusan teljesül.

Ha már mindenképp frekventista statisztikát használunk, akkor legalább annyit meg kell tenni ilyenkor, hogy *p*-érték helyett inkább konfidenciaintervallumokat közlünk (bár jelen esetben ennek se lett volna sok értelme, hiszen ekkora mintanagyságnál azok szélessége is közel nulla lenne).

A fentiek vonatkoznak az eloszlásvizsgálatokra, ráadásul azok ettől függetlenül is feleslegesek

A normalitás lényegében egy „pont null" típusú nullhipotézis (csak nem paraméteresen), ezért ugyanúgy vonatkoznak rá a fentiek: ekkora mintanagyság mellett teljesen felesleges letesztelni bármilyen orvosbiológiai változó normalitását, mert mindenféle tesztelés nélkül is megmondható, hogy szinte nulla lesz a *p*-érték. (Hogy időt spóroljunk a szerzőknek, előre megmondjuk, hogy ugyanezzel az eredménnyel fog járni a Johnson-, Pearson- és minden egyéb eloszláscsaládhoz való illesztés, melynek lehetőségét felvetették.)

Tehát itt is kizárólag a normalitástól való eltérés *mértékének* van jelentősége; ebben egyébként nagyon jól beváltak a grafikus eljárások is (például QQ-ábra).

Megjegyzendő, hogy az eloszlás vizsgálata, hacsak nem maga az eloszlás a kutatás tárgya, a fentiektől teljesen függetlenül is felesleges. Ha ugyanis *más* vizsgálat igényel normalitást, akkor annak fennállását *úgysem* lehet *ugyanazon* mintán végrehajtott normalitásvizsgálattal megítélni. Ha *előre tudjuk*, hogy fennáll a normalitás, akkor válasszunk arra építő próbát, de ha nem, akkor ne használjunk ilyen próbát, és kész, történet vége, a normalitás (ugyanazon) mintából statisztikai teszteléssel történő eldöntésének nincs helye *(6–8)*!

Ne használjunk hisztogramot ekkora mintanagyság mellett

Ekkora mintanagyság mellett egy magfüggvényes sűrűségbecslő olyan pontosan le tudja tapogatni az eloszlás alakját, hogy annak várható szóródása a valódi sűrűségfüggvény körül valószínűleg a nyomdatechnikailag előállítható pont mérete alatt marad... Teljesen felesleges bármi mást használni, a magfüggvényes becslő a fenti értelemben tökéletes megoldás (ekkora mintanagyságnál).

Szinte fizikailag fájdalmas látni a 2. ábrát, ahol a szerzők ugyanezt hisztogrammal próbálták megoldani a KNIME szoftverrel (érdekes módon ezt még az ábrafeliratban is fontosnak tartották külön közölni, mintha valami specialitásról lenne szó, és az R, amivel egyébként számoltak, ne tudna hisztogramot rajzolni...), de még azt sem vették észre, hogy a program már gyakorlatilag képtelen volt kirajzolni a hisztogramot: az iszonyatos mintanagyság miatt olyan nagyszámú osztályt vett fel, hogy az oszlopok lényegében tüskékké mentek össze, és ezt láthatólag ábrázolni sem tudta jól, mert kis csíkok maradtak az „oszlopok” között.

A változók normalitása és a kapcsolatuk linearitása két külön dolog

A szerzők azt írják, hogy „[t]ekintettel arra, hogy a normalitás feltételei nem teljesültek, a Pearson-féle korrelációs együtthatókat nem számítottuk ki”. A Pearson-féle korrelációs együttható a változók közti lineáris kapcsolatot méri, a fenti mondat tehát azt sugallja, mintha a szerzők azt gondolnák, hogy ha két változó nem normális eloszlású, akkor köztük szükségképp nem lineáris az összefüggés. Ez teljesen nyilvánvalóan nem igaz, ellenpéldaként generáljanak a szerzők számokat egy *tetszőleges* eloszlásból, lehet bármilyen távol a normálistól, majd pedig szorozzák be őket kettővel, adjanak hozzájuk hármat. Az így kapott két változó akármilyen messze lehet a normalitástól, a kapcsolatuk mégis *tökéletesen* lineáris lesz.

Ettől függetlenül abban teljesen igazuk van a szerzőknek, hogy mielőtt *bármilyen* korrelációs metrikát kiszámolunk, érdemes előbb tisztázni a kapcsolat jellegét, csakhogy ezt ők sem teszik meg! Mert a Spearman-féle rangkorreláció *ugyanúgy* nem árul el erről semmit; érdemes lett volna a szóródási diagramokat megvizsgálni.

Konklúzió

Volt a szerzőknek egy nagy adatbázisa, és állításuk szerint feltáró elemzést végeztek, olyat, amely páratlannak bizonyult az általuk végzett PubMed-keresés alapján. (A fenti megjegyzések talán segítik az olvasót annak megértésében, hogy miért annyira páratlan ez az elemzés...) Sajnos a nagy adatbázis önmagában nem biztosítja, hogy orvosilag érvényes, hasznosítható eredményeket kapjunk, az „exploratív” jelző kitétele pedig nem oldja meg a jelzett problémákat. Reméljük azonban, hogy írásunk egyúttal a kutatók, érdeklődő orvosok szélesebb köre számára is támpontokat ad az ilyen és ehhez hasonló kutatások kritikus értékeléséhez.

Ferenci Tamás
klinikai biostatisztikus, habilitált egyetemi docens,
Óbudai Egyetem, Élettani Szabályozások Kutatóközpont, Budapest

Singer Júlia
biostatisztikus, Chief Scientific Officer,
Accelsiors Kft., Budapest

Irodalom

1. *Brys Z, Nagy E, Magyar G, Molnár DL, Kis JT.* Nagyszámú laboratóriumi vérvizsgálati eredmény exploratív jellegű vizsgálata rangkorrelációval. *Lege Artis Med 2019;29(12): 609-14.* *https://doi.org/10.33616/lam.29.057*
2. *Goodwin L, Leech N.* Understanding Correlation: Factors That Affect the Size of r. *J Exp Educ 2006;74(3):251-66.*
3. *Wood SN, Li Z, Shaddick G, Augustin NH.* Generalized additive models for gigadata: modeling the UK black smoke network daily data. *J Am Stat Assoc 2017;112(519): 1199-210.* *https://doi.org/10.1080/01621459.2016.1195744*
4. *Bland JM, Altman DG.* Calculating correlation coefficients with repeated observations: Part 1 – correlation within subjects. *Br Med J 1995;310:446.* *https://doi.org/10.1136/bmj.310.6977.446*
5. *Bland JM, Altman DG.* Calculating correlation coefficients with repeated observations: Part 2 – correlation between subjects. *Br Med J 1995;310:633.* *https://doi.org/10.1136/bmj.310.6980.633*
6. *Rasch D, Kubinger KD, Moder K.* The two-sample t test: pre-testing its assumptions does not pay off. *Stat Pap 2011;52(1):219-31.* *https://doi.org/10.1007/s00362-009-0224-x*
7. *Rochon J, Kieser M.* A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *Br J Math Stat Psychol 2010;64:410-26.* *https://doi.org/10.1348/2044-8317.002003*
8. *Schoder V, Himmelmann A, Wilhelm KP.* Preliminary testing for normality: some statistical aspects of a common concept. *Clin Exp Dermatol 2006;31:757-61. https://doi.org/10.1111/j.1365-2230.2006.02206.x*

Szerzői válasz az „Egy nagymintás exploratív vizsgálat érvényessége a limitációi tükrében" című olvasói levélre

Az exploratív adatelemző vizsgálatok helye az orvosi tudás előállításában

Az „Egy nagymintás exploratív vizsgálat érvényessége a limitációi tükrében" *(1)* című olvasói levélben a „Nagyszámú laboratóriumi vérvizsgálati eredmény exploratív jellegű vizsgálata rangkorrelációval" *(2)* című közlemény érvényességének határairól szerepelnek gondolatok.

Kontextus

Az orvosok számára evidens, hogy a D-dimerek emelkedett szintje jellemző a mélyvénás trombózisban és a tüdőembóliában szenvedő betegekben, azonban a teszt pozitivitása más betegségekben is előfordulhat, tehát a pozitív D-dimerteszt nem bizonyítja a fenti betegségek meglétét. Egyéb tényezőket *(kontextus)* is vizsgálni kell a helyes diagnózis felállításához. A nyelvész számára evidens, hogy hiba keletkezhet abból, ha a szavakat önálló entitásként kezelik, kiragadva azokat a környezetükből *(3, 4)*. *Mark Twain* egy alkalommal azt a csípős megjegyzést tette, hogy az angol humort nehéz megérteni, ha az ember nincs felkészülve rá *(5)*. Az írás az adatelemzés különböző hagyományainak kontextusában mutatja be a kritizált feltáró vizsgálatot.

Történeti perspektíva

Az adatok összegyűjtésének, elemzésének és az eredmények értelmezésének, interpretációjának módszereit érdemes történeti perspektívából nézni, mivel ezen a területen az elmúlt kétszáz év során egymással párhuzamosan és egymásra hatva több, tudománytörténeti szempontból is fontos fejlődési mozzanatot láthatunk *(6)*. Az olvasó így bepillantást nyerhet az adatelemzések dinamikus tudománytörténeti folyamatába anélkül, hogy statikus, esetenként dogmatikusnak tűnő kinyilatkoztatások csapdájába esne.

Érvényesség (validitás)

Vizsgáljuk meg először az olvasói levél címében szereplő *érvényesség* jelentését és kontextusát. *Campbell* óta a nemzetközi szakirodalomban elfogadott megközelítés, hogy a validitás arra vonatkozik, hogy vajon azt mérjük-e, amit mérni vélünk vagy mérni szeretnénk *(7)*. Amennyiben a kutatás validitása nagy, akkor az eredmények összhangban vannak a fizikai, kémiai, biológiai vagy szociális világ valódi természetével, jellemző vonásaival és változékonyságával, továbbá a meglévő elméletekkel. Tekintettel arra, hogy az elméletek maguk is változhatnak, a validitás definíciója paradox módon magában hordozza a tudomány fejlődése előmozdításának a lehetőségét éppúgy, mint akadályozását.

A nagy validitás egyik indikátora a *megbízhatóság*, tehát az, hogy megismételt mérés esetén hasonló eredmények keletkeznek. Ha egy módszer nem megbízható, akkor feltehetőleg nem is valid. Az *oksági* kapcsolatok feltárásával kapcsolatos validitás két fontos aspektusa, oldala a *belső* és a *külső validitás*. A belső validitás a kísérletek tervezésével, míg a külső validitás az eredmények általánosíthatóságával van összefüggésben. Campbell és *Stanley* a *belső validitást* a következő kérdésre adandó válaszként definiálta: „A *kísérletben* alkalmazott *inger, stimulus* szignifikáns különbséget okozott-e a vizsgált szempont szerint?" A *külső validitást* a szerzők a következő kérdésre adandó válaszként definiálták: „A kísérletben kiváltott hatás milyen »populációra« (a hagyományos »frekventista« statisztikai gondolkodás alapfogalma a »populáció«, amely valós vagy elképzelt sokaság, amelyre a »mintából« következtetnek), kísérleti elrendezésre és változókra *általánosítható*?" A Campbell-cikk címében figyelemre méltó, hogy hangsúlyozottan *kísérletekről*, a később elkészült könyv címében pedig *kísérletekről és félig-kísérletekről* van szó *(8)*. Később a validitás egyes kategóriáit átminősítették és a veszélyeket is átsorolták *(9)*. *Shadish* és munkatársai hangsúlyozták, hogy az *érvényesség* a *„populációról"* levont következtetések igaz voltát jellemzi *(9)*, másutt viszont a validitást fenyegető tényezőket nem az igazság, hanem a *hatásnagyság* becslésének pontosságával hozták kapcsolatba *(10)*.

A *következtetések érvényessége* azzal kapcsolatos, hogy a kvalitatív vagy kvantitatív jellegű kutatásból vagy kísérletből levont következtetések mennyire *észszerűek*. A következtetések érvényessége azzal a kérdéssel kapcsolatos, hogy vajon „az adatok alapján van-e kapcsolat vagy nincs a vizsgált tényezők között." A *statisztikai következtetések érvényessége* (SCV) értelemszerűen a számításokon alapuló összefüggések keresésére vonatkozik. Fontos észrevenni, hogy a validitás fogalmát és teljesülési feltételeit elsősorban az *oksági kapcsolatok* megragadását célzó kísérletekkel és félig-kísérletekkel kapcsolatban vizsgálták.

Az érvényességgel kapcsolatos kutatások vezettek ahhoz, hogy az egyes vizsgálati elrendezések között fontossági *hierarchiát* állapítsanak meg. Eszerint az *oksági kapcsolatok feltárásához* a legkevésbé alkalmasak a *megfigyeléses vizsgálatok* [különböző módszerekkel *gyűjtött adatok*, a *survey* típusú vizsgálatok (szociológiai felmérések), „hivatalos" adatok *(11)*, internetes adatok stb.]. Ezeknél jobban alkalmasak az oksági kapcsolatok feltárására a *félig-kísérletek* („csak utána", „előtte-utána" vizsgálatok, „megszakított idősoros elrendezések" stb.), elsősorban a lélektan, a társadalomtudományok és az orvostudomány területén. Az oksági kapcsolatok feltárására leginkább a *valódi kísérletek* alkalmasak, azok is elsősorban a fizika, kémia és biológia területén, amelyek között számos kísérleti elrendezés (factorial, fractional factorial, split-plot design stb.) található. A klinikai vizsgálatokat is a félig-kísérletek közé sorolják, amelyekkel a *véletlen besorolás* (random allocation) miatt egyfelől okságinak *tűnő* kapcsolatok igazolhatók, azonban a *mintakiválasztás önkényes* jellege (site) miatt több vizsgálat együttes elemzésére lehet szükség az eredmények általánosíthatóságához (metaanalízis). Mindezt azért is érdemes felidézni, mivel az alább tárgyalandó *exploratív* adatelemzések célja *nem* az oksági kapcsolatok, hanem sokkal inkább az adatok megismerése, a *lehetséges* összefüggések kezdeti vizsgálata, ezek vizualizációja, esetleg orvosi hipotézisek generálása. Az *érvényesség* fogalma tehát elsősorban az *oksági* viszonyok feltárására irányuló *kísérletes és félig-kísérleti vizsgálatok* ellenőrzésére szolgáló eszközrendszer, nem alkalmazható feltáró, exploratív vizsgálatokra.

Exploratív és konfirmatív kutatás

Történetileg a fenti folyamatokkal párhuzamosan, részben azokkal összefüggésben két, egymást kiegészítő megközelítés is kialakult. Az egyik a már említett *exploratív* (feltáró jellegű, megismerési célzatú), a másik a *konfirmatív* (megerősítő, oksági összefüggések bizonyítására törekvő) kutatás és a hozzájuk kapcsolódó statisztikai eljárások.

Az *exploratív* statisztikai elemzés fogalmát *Tukey (12)* vezette be úgy, hogy hangsúlyozta egyebek mellett az *adatok* vizuális megjelenítésének a fontosságát, újfajta, beszédes statisztikai ábrák elkészítését *(13, 14)*, mindezt az *adatok* jobb megértése érdekében. Az exploratív jellegű kutatások elsődleges célja az adatok jobb megismerése, megértése, vizualizálása, a későbbi adatgyűjtések és adatfeldolgozások módjának meghatározásához támogatás nyújtása, a *lehetséges* összefüggések előzetes felderítése, és esetleg hipotézisek generálása, amelyek általában a kutatás kezdeti szakaszában lehetnek hasznosak. Az exploratív kutatások eredményeképpen, elsődlegesen az orvosszakmai szempontok érvényesülésére tekintettel, *az orvosokkal szoros szakmai együttműködésben*, a *lehetséges* oksági összefüggésekről *előzetes* hipotézisek fogalmazhatók meg, amelyek később megalapozott konfirmatív kutatásban verifikálhatók vagy falszifikálhatóak.

A *konfirmatív* kutatások során általában célzott adatgyűjtések, félig-kísérletek vagy kísérletek alapján a pontosan megfogalmazott elgondolásokat statisztikai hipotézisvizsgálatoknak vetik alá. A *Neyman* és *Pearson* által kidolgozott konfirmatív statisztikai hipotézisvizsgálatok elméletét *(15)* eleinte rendkívül hevesen támadták. A kezdeti óriási ellenállás után a nullhipotézisre alapozott szignifikanciavizsgálat elmélete (NHST) gyakorlatilag dogmává merevedett és az eljárás alkalmazása rutinszerűvé vált. A NHST és a p-érték kiszámítási módja körül azonban időközben számos problémát azonosítottak *(16)*, amelyeket nagyrészt *(17)*, de nem teljes mértékben sikerült megoldani. Részben ezzel összefüggésben manapság egyre népszerűbbek a *bayesi* statisztikai megoldások, amelyek a hagyományos frekventista megközelítéstől alapvetően különböző filozófiai alapokon nyugszanak, a kapott eredmények azonban sokszor számszerűleg rendkívül hasonlóak lehetnek a hagyományos statisztikai módszerekkel kapott eredményekhez *(18)*.

Adatbányászat és gépi tanulás

Az adatbányászat a konfirmatív statisztikai módszerektől eltérő szemléletet használ: az adatbányászok *hipotézisek megtalálásának módjaira* fókuszálnak és azokat tesztelik a hipotézisalkotásba nem bevont adatrészen *(19)*. A hagyományos adatbányászathoz tartozó eszközök a klaszterezés, az osztályozás, a nem klasszikus idősoros elemzések stb. Matematikai és számítástudományi ismeretek felhasználásával rengeteg új és hasznos adatfeldolgozási technikát fejlesztettek ki, ezeket összefoglaló néven gépi tanulásnak nevezhetjük. Három nagy csoportját különböztethetjük meg: *felügyelt tanulás, nem felügyelt tanulás és a megerősítéses tanulás*. A felügyelt tanulás egyik alcsoportja a mesterséges mély neuronhálózati modellek – amelyek egy részét egyebek mellett a prediktív modellezéseknél is rutinszerűen alkalmazzák, például SAS szoftver környezetben *(20–22)*. A régi és új adatfeldolgozási, adatbányászati és gépi tanulási technikák bonyolult csalásfelderítési *(23)*, kép-

feldolgozási, diagnosztikai, lingvisztikai *(24)* és számos egyéb problémák megoldására is hasznosnak bizonyultak *(25, 26)*.

„Big data”

Időközben megjelentek a többféle értelemben vett nagyobb adatállományok is, amelyeket „big data” néven emlegetnek. Az eredeti közleményben idézőjelben szereplő „big data” általában az angolul három V betűvel jellemzett adatokat (Volume: méret, Velocity: sebesség, Variety: változatosság) jelenti *(27, 28)*. A három V utalást tartalmaz, egyebek mellett a gyorsan nagy adattömegeket létrehozó *high throughput*, *next generation sequencing* *(29–33)* és más hasonló eljárásokkal keletkezett adatokra. A három V mellett újabb V-ket is javasoltak, ilyen például a V: veracity (igazságnak megfelelés). Mindazonáltal a „big data” kifejezés a mai napig nem pontosan definiált fogalom.

Statisztika, biostatisztika

A világot először számokká alakítjuk át. A számok nem beszélnek önmagukért. Mi beszélünk a számokról. Mi látjuk el a számokat jelentéssel *(34)*, amelynek egyik eszköze a statisztika. Az egyik definíció szerint a biostatisztika olyan tudomány, amely lehetővé teszi a bemutatott tények alapján következtetések levonását a bennünket körülvevő világ jobb megértése, megismerése érdekében *(35)*. *McElreath* szerint a statisztikai módszerek, eljárások a gólemekhez, robotszerű agyagszobrokhoz, kis robotokhoz hasonlíthatók *(36)*. McElreath szerint a statisztika *nem* matematika. Sőt, szerinte a statisztika nem is tudomány. Szerinte a statisztika művelői leginkább a gépészmérnökökhöz hasonlíthatók, akik utasítások végrehajtására képes absztrakt gépeket, algoritmusokat terveznek, hoznak létre és működtetnek. Ebben a megközelítésben a statisztikai módszerek különböző, ember által konstruált és folyton változó eszközök gyűjteménye. Ismét mások szerint a statisztika az emberi megismerést (tudás-előállítást) és kommunikációt elősegítő eszközrendszer.

Adatelemzés, adattudomány

A történetileg alig néhány évtizedes múltra tehető *data science* lényeges eleme a tudás, jelen esetben az *orvosi tudás* és az *adatok* értelmezése és kommunikációja között a kapcsolat erősítése, és amelynek fő lépései az orvosokkal szoros együttműködésben az adatok importja, rendezése, integrálása *(data integration)*, transzformációja, megjelenítése, modellezése és kommunikációja. Nem statikus, hanem iteratív, körkörös visszacsatolásokra épülő elemzési folyamatról van szó.

Az adatelemzés, adattudomány és a matematikai statisztika összehasonlítása, szembeállítása kapcsán *Hayashi* a következőket írta: *„...a matematikai statisztika művelői hajlamosak arra, hogy elszakadjanak a valóságtól. Ezzel szemben az adatelemzés módszereinek kialakításakor nem fordítottak túl sok figyelmet a matematikai statisztikára, mégis hasznos eszközöket hoztak létre bonyolult problémák megoldásához, amelyek a hagyományos értelemben nem mindig statisztikai következtetések, hanem gyakran leíró jellegűek” (37)*. A döntési fák, a véletlen erdő (random forest) és egyéb módszerek kidolgozásához köthető *Breiman* neve, aki a következőket írta: *„Az adatok alapján a következtetések levonásához két kultúra létezik a statisztikai modellezésben. Az egyik szerint feltételezik, hogy az adatokat valamilyen ismert sztochasztikus, véletlen folyamat generálta. A másik kultúra képviselői algoritmikus modelleket használnak és az adatok keletkezési mechanizmusát ismeretlennek tekintik. A statisztikusok közössége többnyire az első kultúrához sorolja magát. Ez a fajta elköteleződés irreleváns elméletekhez, megkérdőjelezhető következtetésekhez vezetett és távol tartotta őket egy sor érdekes és fontos jelenleg meglévő probléma megoldásától” (38)*.

Következtetés

A validitás fogalmának ismeretében az olvasói levélben szereplő *érvényesség* alig alkalmazható olyan vizsgálatra, amely bevallottan és egyértelműen *exploratív, leíró, tapogatózó* jellegű volt, hiszen *nem kísérleti, nem félig-kísérleti* adatokkal foglalkozott, nem volt szó semmilyen kísérleti elrendezésről, a statisztikusok által használt értelemben nem szerepelt benne semmilyen „populáció”, nem szerepelt „minta”, az adatok keletkezési mechanizmusát ismeretlennek tekintettük, nem szerepelt benne semmilyen inger vagy stimulus. Kizárólag leírni, ábrázolni, megismerni kívánt adatok voltak, és a vizsgálat alapján a szerzők nem kívántak semmilyen oksági kapcsolatot bizonyítani, sőt, a kapott eredményeket nem kívánták általánosítani sem. A vizsgálat során nem történt semmilyen, a hagyományos *konfirmatív statisztikai értelemben* vett *hipotézisvizsgálat*, és a kapott p-értékek kiértékelése kizárólag *exploratív* jellegű volt. Az iparban széles körben elfogadott és alkalmazott *„exploratív p-érték-számítás”* azt jelenti, hogy a kapott p-értékek csupán előzetes, tájékozódási célra használható iránymutatásnak tekinthetők, és ezekre p-érték-korrekció (Bonferroni-

korrekció, Holm-eljárás stb.) elvégzése sem szükséges (a gyógyszeriparban gyakran *secondary endpoints*). Ugyanez vonatkozik a mindössze érzékeltetésképpen elvégzett erőelemzésre is. A *konfirmatív* statisztikai elemzésekhez szokott statisztikusok szemszögéből nézve az alkalmazott módszerek és a kapott eredmények szokatlannak tűnhetnek, azonban az *exploratív* adatelemzések és az orvosi tudás szemszögéből nézve a kapott tájékoztató jellegű eredmények ismeretében több lehetőség kínálkozik később a célzott vizsgálatok megtervezéséhez.

Konstruktív kritikák

Megemlíthető, hogy az eredeti kéziratra összesen négy anonim lektori vélemény érkezett, ezekből kettő biostatisztikai jellegű kéréseket tartalmazott. Az anonim lektorok úgy találták, hogy exploratív adatelemzési szempontból elfogadható ez az előzetes vizsgálat.

A már megjelent közleményre konstruktív kritikák is érkeztek orvosoktól. Az egyik klinikus azt javasolta, hogy az érzékeltetésképpen leírt erőelemzés helyett szimulálhatóak változónkként hasonló eloszlású véletlen adatok és az ott számított korrelációk felhasználhatóak összehasonlításra. A másik orvos pedig azt jelezte, hogy néhány helyen tévedésből minta és mintanagyság szerepel a cikk szövegében – exploratív vizsgálat esetében mintáról és érvényességéről általában nem lehet beszélni, csak adatról és adatvizualizációról. Orvoskonferencián pedig felmerült, hogy a vérképadatokban az összefüggések exploratív feltárására használhatóak lennének a gépi tanulás egyes elemei (support vector machine, mesterséges neuralis hálók stb.), illetve a vizualizációra pedig az *Epskamp* által leírt módszer.

Az oksági viszony megállapítására *nem* törekvő *feltáró* vizsgálat *(1)* igazán meglepő eredménye az, hogy az ismétlődés, az adathalmazba való bekerülés valószínűségeinek ismeretlensége és további (széles értelemben vett) „zajok” ellenére az alkalmazott nagyon egyszerű, robusztus távolságmetrika (adatgeometriai szempontból egy sokdimenziós térben vektorok között bezárt szögek értékeinek számítása) az orvosi tudás egy részét megragadta, leírta. Ez akár azt is jelentheti, hogy a nem kutatási célból gyűjtött zajos vérképadatok jobban hasznosíthatóak, mint eddig gondoltuk.

dr. Molnár D. László,
a vizsgálat orvos-biostatisztikus konzultánsa

Brys Zoltán,
a vizsgálat adatelemzési részének vezetője

Irodalom

1. *Singer J, Ferenczi I.* Egy nagymintás exploratív vizsgálat érvényessége a limitációi tükrében *Lege Artis Medicinae 2020;30(4-5):XXX-YYY.*
2. *Brys Z, Nagy E, Magyar G, Molnár DL, Kis JT.* Nagyszámú laboratóriumi vérvizsgálati eredmény exploratív jellegű vizsgálata rangkorrelációval. *Lege Artis Medicinae 2019;29 (12):609-14. https://doi.org/10.33616/lam.29.057*
3. *Goodwin C, Duranti A.* Rethinking context: An introduction. in: rethinking context: language as an interactive phenomenon. *Cambridge: Cambridge University Press; 1992.*
4. *Whitehead AN.* Philosophers do not think in a vacuum. dialogues of Alfred North Whitehead. Recorded by Lucien Price. David R. Godine. 2001.
5. *Twain M.* English know a joke, says Mark Twain. *The Evening World; 1907. p. 2.*
6. *Stigler SM.* Statistics on the table. The history of statistical concepts and methods. *Harvard University Press; 1999.*
7. *Campbell DT.* Factors relevant to validity of experiments in field settings. *Psychological Bulletin 1957;54:297-312. https://doi.org/10.1037/h0040950*
8. *Campbell DT, Stanley J.* Experimental and quasi-experimental designs for research. *Cengage Learning, 1963.*
9. *Shadish WR.* Revisiting field experiments: Field notes for the future. *Psychological Methods 2002;7:3-18. https://doi.org/10.1037/1082-989X.7.1.3*
10. *Reichardt CS.* Quasi-Experimentation. A guide to design and analysis. *The Guilford Press, 2019.*
11. *Slattery M.* Official Statistic. Tavistock. *London, New York; 1986.*
12. *Tukey JW.* Exploratory data analysis. *Addison-Wesley, 1977.*
13. *Chambers JM, Cleveland WS, Kleiner B, Tukey P.* Graphical methods for data analysis. *Woodsworth & Brooks/Cole, 1983.*
14. *Tufte ER.* Envisioning information. narratives of space and time. *Graphics Press, 1990.*
15. *Neyman J, Pearson ES.* On the problem of the most efficient tests of statistical hypotheses. *Phil Trans R Soc Ser A 1933;231:289-337. https://doi.org/10.1098/rsta. 1933.0009*
16. *Wagenmakers EJ.* A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review 2007;14(5):779-804. https://doi.org/10.3758/BF03194105*
17. *Mayo DG.* Statistical inference as severe testing. How to get beyond the statistics wars. *Cambridge, UK: University Printing House; 2018. https://doi.org/10.1017/9781107286184*
18. *Dinya E, Molnár DL, Mészáros J, Solymosi N.* Biometria a klinikumban. Feladatok bayesi megoldása. *Budapest: Medicina Kiadó; 2019.*
19. *Bodon F.* Adatbányászati algoritmusok. *Budapest, 2010. http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf (Letöltés ideje: 2020. január 2.)*
20. *Sarma KS.* Predictive modeling with SAS enterprise miner. Practical solutions for business applications. 2nd ed. *SAS Institute Inc. 2013.*
21. *Lewis ND.* Applied predictive modeling techniques in R. With step by step instructions on how to build them FAST! *AusCov, 2015.*
22. *Kuhn M, Johnson K.* Applied predictive modeling. *Springer, 2016.*

23. *Brown ILJ.* Developing credit risk models using SAS enterprise miner and SAS/STAT. Theory and applications. *SAS Institute Inc., 2014.*
24. *Géron A.* Hands-on machine learning with scikit-learn, Keras & TensorFlow. Concepts, tools, and techniques to build intelligent systems. *O'Reilly, 2019.*
25. *Fernandez G.* Statistical data mining using SAS applications, 2nd ed. *CRC Press, 2010. https://doi.org/10.1201/EBK1439810750*
26. *Williams G.* Data mining with rattle and R. The art of excavating data for knowledge discovery. *Springer, 2011. https://doi.org/10.1007/978-1-4419-9890-3_17*
27. *Laney D.* 3D data management: Controlling data volume, velocity and variety. *META Group Research Note 2001;6:70.*
28. *Baumer BS, Kaplan DT, Horton NH.* Modern data science with R. *CRC Press, 2017.*
29. Buffalo V. Bioinformatics Data Skills. Reproducible and robust research with open source tools. *O'reilly, 2015.*
30. *Datta S, Newletton D.* Statistical analysis of next generation sequencing data. *Springer, 2014. https://doi.org/10.1007/978-3-319-07212-8*
31. *Pevsner J.* Bioinformatics and functional genomics. 3rd ed. *Wiley Blackwell, 2015.*
32. *Bhadhadhara K.* Statistical analysis of genomic data using R and bioconductor packages. *LAMBERT Academic Publishing, 2017.*
33. *Hofmann A, Clokie S.* Wilson and Walker's principles and techniques of biochemistry and molecular biology. *Cambridge University Press, 2018. https://doi.org/10.1017/9781316677056*
34. *Silver N.* The Signal and the noise: Why so many predictions fail-but some don't. *Penguin Books; 2015.*
35. *Oliveira AG.* Biostatistics decoded. *Wiley, 2013. https://doi.org/10.1002/9781118670767*
36. *McElreath R.* Statistical Rethinking. A Bayesian course with examples in R and Stan. *CRC Press, 2016.*
37. *Hayashi C.* What is data science? Fundamental concepts and a heuristic example. In: *Hayashi K, Yajima HH, Bock N, Ohsumi Y, Tanaka, Baba Y* (eds.). Data science, classification, and related methods. studies in classification, data analysis, and knowledge organization. *Tokyo: Springer; 1998. https://doi.org/10.1007/978-4-431-65950-1_3*
38. *Breiman L.* Statistical modeling: The two cultures. *Statistical Science 2001;16(3):199-231.*