# 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*