A kis „n”, nagy „P” probléma a neuropszichofarmakológiában, avagy hogyan kontrolláljuk a hamis felfedezések arányát

Petschner Péter[1,2], Bagdy György[1,2] és Tóthfalusi László[1]

[1] *Semmelweis Egyetem, Gyógyszerhatástani Intézet, Budapest*
[2] *MTA-SE Neuropszichofarmakológiai és Neurokémiai Kutatócsoport, Budapest*

Számos korszerű neuropszichofarmakológiai vizsgálati módszer jellegzetessége, hogy aránylag kevés vizsgálati egyénről (n) nagyon sok adatot (paramétert, P) gyűjt. Példaképpen említhetjük a képalkotó módszereket (pl. funkcionális mágneses rezonancia és egyéb képalkotó eljárásokat), az elektroenkefalográfiát (EEG), vagy a genomikai vizsgálatokat. Egyetlen microarray chip például több ezer próbát tartalmazhat, azaz a P ezres nagyságrendekkel haladhatja meg az n-t. Az ilyen elrendezésű vizsgálatok elemzése komoly statisztikai problémákat vet fel, amit a statisztikai szakirodalomban kis „n” nagy „P” problémának neveznek. A többszörös tesztelés problémája akkor lép fel, ha két vagy több csoportba tartozó egyéneket hasonlítunk össze a mért P számú jellemző alapján. Amennyiben az összehasonlítás az egyes jellemzők alapján történik, akkor akár több ezer statisztikai hipotézisvizsgálat elvégzése is szükségessé válhat. Amennyiben a többszörös tesztelés okozta megnövekedett klasszifikációs hibát nem vesszük figyelembe, akkor számos statisztikailag szignifikáns különbséget fedezhetünk fel a vizsgálati csoportok között. Azonban ezeknek a felfedezéseknek egy része valójában a véletlen műve és ezek a kísérleti eredmények általában nem reprodukálhatóak. A problémára több megoldás is született. Ezek közül cikkünkben a klaszter szintű összehasonlítást, valamint a hamis találati arányon alapuló statisztikai tesztet mutatjuk be.

(Neuropsychopharmacol Hung 2015; 17(1): 023–030)

Kulcsszavak: funkcionális mágneses rezonancia képalkotó vizsgálatok, microarray, hamis találati arány, permutációs teszt, gene set enrichment analysis, fMRI, statisztika

A bevezető jellegű orvosi statisztikai tankönyvek kedvenc példája, hogy adott egy vizsgálat, ahol a betegek egy része placebot kap, a másik csoport meg egy új vérnyomáscsökkentőt, és a vizsgálat végén megmérjük mindkét csoportban a vérnyomást. Minden beteghez egyetlen egy vizsgálati érték tartozik. A két csoport értékeit a jól ismert kétmintás t-próbával hasonlítjuk össze, és az eredményt szignifikánsnak minősítjük, ha a t statisztikához tartozó p valószínűségi érték kisebb, mint 0,05. Ebben az esetben, az „n” a vizsgálatba bevont betegek száma és a paraméterek száma (P) egyenlő 1-gyel (hiszen csak egy paramétert mértünk, a vérnyomást).

A valós élethez jóval közelebb áll, amikor egy egyénről, vizsgálati objektumról nem egy, hanem több adatot gyűjtünk, például öt különböző módon mérjük a gyógyszer hatását a betegre. Ekkor a P egyenlő 5-tel. A naiv elképzelés az lenne, hogy az öt paramétert egyenként hasonlítjuk össze, és ha egyetlen teszt p-értéke is kisebb, mint 0,05, akkor a gyógyszer hatása szignifikáns $p<0,05$ szinten. Ez a megközelítés azonban hibás, mert a statisztikai próbák számának növekedésével a döntési hibák összeadódnak, és az összesített döntési hiba végül lényegesen meghaladhatja a 0,05 értéket.

Több megoldás ismeretes annak érdekében, hogy a statisztikai döntések számával az összesített hiba ne növekedjen (Neuhauser, 2006). A gyógyszerek törzskönyvezéséhez megkövetelt „döntő” (pivotal) klinikai vizsgálatokban használják a legegyszerűbb módszert. Csak egy elsődleges végpont van, és a döntésnél csak ezt veszik figyelembe. Akárhány „másodlagos” végpontban jön ki pozitív eredmény, csakis a protokollban rögzített elsődleges végpont számít. Így csak egy döntés van, és a többszörös tesztelés okozta hibák összeadódása elkerülhető. A tudományban azonban nehéz előre megjósolni, mi lesz érdekes megfigyelés, és nem ésszerű egy felfedezést azért nem figyelembe

venni, mert nem szerepelt a protokollban. Ezért olyan statisztikai módszereket használnak, amelyek az összesített döntési hibát egy konstans, tipikusan 0,05 alatt tartják. Az angol nyelvű szakirodalom az összesített döntések hibájára mint „family-wise error rate" (FWER) hivatkozik (van den Oord, 2008). A FWER kontrollálásra számos statisztikai teszt létezik különböző szerzői nevek alatt (Cleophas és Zwinderman, 2006; Neuhauser, 2006).

A legegyszerűbb FWER-t kontrolláló teszt az úgynevezett Bonferroni teszt, ami azt mondja, hogy használjunk tetszésünk szerinti számú tesztet, de ha P számú tesztet végzünk, akkor mindegyik tesztet p/P szignifikanciaszinten végezzük el. Például legyen egy klinikai vizsgálatban 3 paraméter, amelyek jellemezhetik az eltérést a kezelt- és a kezeletlen beteg csoportok közt. A két csoport összehasonlítására, Bonferroni szerint, három hipotézisvizsgálatot kell végezni az egyes paraméterek szerint, de az egyes tesztek esetén a szignifikancia 0,05/3=0,016 kell, hogy legyen. Ebben az esetben az összesített hiba kisebb lesz, mint 0,05. A Bonferroni eljárás és a hasonló elvek szerint működő más próbák elfogadható hatékonysággal (statisztikai erővel) működnek, ha 5-10 összehasonlítás végzünk, de efölötti számú statisztikai vizsgálat felett már gondok adódhatnak.

Tételezzük most fel, hogy a két csoportot nem 3, hanem 100 paraméter alapján hasonlítjuk össze. Ekkor az egyenkénti tesztek szignifikanciaszint-határait 5×10^{-4} szinten kéne meghúzni, holott például p<0,001 esetén már minden eredményt „nagyon erősen" szignifikánsnak szoktunk hívni. Bonferroni szerint így nem mondhatnánk, hogy p<0,001 szinten a két csoport eltér, még akkor se, ha a két csoport mind a száz paraméterben különbözik egymástól p<0,001 szinten. A pszichofarmakológiában használt korszerű vizsgáló módszerek esetén azonban nem ritka, hogy tízezres nagyságrendben (vagy még annál is nagyobb számban) gyűjtünk adatokat egyetlen egyedről. Íme, két állatkísérletes példa Intézetünk gyakorlatából.

Patkány agyban transzkripciós microarray segítségével néztük, hogy ecstasy (3,4-metiléndioximetamfetamin, MDMA) vagy venlafaxin (VLX) kezelés hatására milyen gének mutatnak emelkedett vagy csökkent expressziót a kontrollhoz képest (Petschner et al., 2013; Tamasi et al., 2014). A vizsgálatokhoz Illumina RatRef-12 microarray chip-et használtunk. Egy chip 22523 próba szekvenciát tartalmazott. Minden vizsgált agyterület esetén 7 vagy 8 minta segítségével mértük az expresszióváltozás mértékét. Azaz, ebben a vizsgálatban agyterületenként n=14 (16) és P=22523.

Állatkísérletes modellben escitalopram hatását vizsgáltuk elektroenkefalográf (EEG) segítségével krónikus alvásmegvonást követően a kontroll kezeléshez viszonyítva (Kostyalik et al., 2014). Az EEG jeleket 4 másodperces időközökben értékeltük ki a 0,5-60 Hz frekvencia tartományban, fél Hertzes lépésközzel. Ennek megfelelően egy kétórás vizsgálat esetén összesen 119 (frekvencia-intervallumok teljes száma)x(3600/4) (az időintervallumok teljes száma)x2 (az órák száma)=214200 adatpont keletkezik. A vizsgálatban 6-6 állat vett részt. Azaz, ha a teljes EEG-spektrumot akarjuk összehasonlítani a két csoport között, akkor n=12 és P=214200.

Humán pszichofarmakológiai vizsgálatokban gyakran legalább ekkora számokkal találkozunk. Egy fMRI „kép" alapegysége a voxel, ami lényegileg egy pixel 3D-ben (Lange, 2003). A kérdés, hogy a voxel aktivitása nő vagy csökken a kontrollhoz képest. Egy fMRI pillanatfelvétel közel 100000 voxelt tartalmaz azaz P=100000.

Természetesen a pszichofarmakológián kívül számos egyéb tudományterület van ahol a kis „n" nagy „P" problémával a statisztikusok találkoztak (National Research Council (U.S.). Committee on Mathematical Sciences Research for DOE's Computational Biology, 2005). A probléma megoldására több javaslat született, amelyek közül mi a tömeges, egy-szempontos analízisen alapuló (massive univariate analysis) megközelítést ismertetjük (Groppe et al., 2011).

HALMAZOK (KLASZTEREK) ÖSSZEHASONLÍTÁSA

Egy microarray vizsgálatnál nem egy gén, hanem sokkal inkább egy szabályozási útvonalban történő változás az érdekes. Így tanulmányainkban mi azt vizsgáltuk, hogy mi a hatása az MDMA vagy a venlafaxin adagolásnak a különféle jelátviteli útvonalakra (Petschner et al., 2013; Tamasi et al., 2014). Ha egy ilyen útvonal aktivitása megnő, akkor feltehető, hogy nem egy, hanem számtalan gén átíródásában áll be változás. (Az egyszerűség kedvéért az aktivitás változás előjelétől most eltekintünk.) A géneket az irodalomból ismerjük. Jelöljük T_i-vel a kezelt csoportba, C_i-vel a kontroll csoportba tartozó géneket. A kérdés tehát az, hogy különbözik-e egymástól a két halmaz, nevezetesen $\{T_1, T_2 \ldots T_k\}$ és $\{C_1, C_2 \ldots C_k\}$ génjeinek aktivitása.

A különbséget vizsgálhatjuk tagonként is, mint T_1 versus C_1 és T_2 versus C_2, ugyanakkor könnyen előfordulhat, hogy minden gén expressziója csak kis mértékben változik, így az egyéni hatás kicsi marad. Ha minden génexpresszió csak húsz százalékkal változik, akkor az egyenkénti hatás még alacsony, de mivel azonos útvonalban játszanak szerepet, a multiplikatív, azaz egymásra épülő összesített hatás (1,2x1,2x1,2…) óriási lehet.

A másik probléma maga a teszt. A Bonferroni (vagy bármely hasonló) FWER-t megőrző teszt különösen hatástalan (szakszóval konzervatív) lenne, mert ezek a tesztek mind azon alapulnak, hogy az elvégzendő „k" számú összehasonlítás független dolgokat hasonlít össze. Esetünkben azonban ez nem így van, hiszen egy adott útvonalon levő gének aktivitása egymástól szinte biztosan nem független. Még nyilvánvalóbb a probléma az fMRI esetén. Itt egy agyterület fMRI aktivitásának növekedése nem egy voxel, hanem egy, a térben összefüggő voxel csoport aktivitásnak növekedésével jár. A voxelenkénti összehasonlítás itt teljességgel irracionális, mivel nem feltételezhető, hogy az egymás melletti voxelek aktivitása független volna.

A kérdés megválaszolása, hogy hogyan lehet hatékonyan összehasonlítani két halmazt, hogy a fenti bukatókat elkerüljük, nem egyszerű. Léteznek komplex módszerek (pld. Hoteling t-teszt), de ezek a tesztek olyan feltevéseken alapulnak, amelyeket kis „n" esetén (a mi esetünkben a microarray chipek száma) képtelenség kielégíteni. Ennél célszerűbb a permutációs teszttel való megközelítés, amelyet a legtöbb bioinformatikai- és képalkotó szoftver széleskörűen használ.

Permutációs teszt

A permutációs teszt ötlete viszonylag egyszerű (Ludbrook, 1994). Tételezzük fel, hogy van egy kísérletünk, amiben van egy kontroll csoport (C) és egy kezelt csoport (T). Az egyszerűség kedvéért legyen mindkét csoportban az elemszám egyenlően n. A minket érdeklő kérdés az, hogy a két csoport átlagának különbsége (d) mennyiben tekinthető véletlenszerű különbségnek. Az ötlet az, hogy ha a kezelésnek nincs hatása, akkor teljesen mindegy, hogy a T vagy a C csoportban szerepel egy érték. Ezek alapján az algoritmus így néz ki:

1. Egyesítsd a T és C csoportokat, ez lesz az U csoport.
2. Véletlenszerűen válassz ki az egyesített U csoportból n értéket, mintha ez lenne egy új T' csoport. Ez valószínűleg tartalmazni fog az eredeti T és C csoportból is értékeket, de ha nincs különbség T és C között, akkor mindegy, honnan vettük az értéket.
3. Hasonlóképpen válassz n értéket a C' csoport részére és számold ki a T'-C' csoport különbségét. A kapott különbséget, d'-t, pedig raktározd el.
4. Ismételd meg a 2-3. lépést sokszor, tipikusan 1000-szer vagy ennél is többször.

Az elraktározott d' értékek adják a d null-eloszlását, azaz a különbség eloszlását, ha T és C között tényleg nincs lényegi különbség. Ha az eredeti különbségünk d, nagyobb vagy kisebb, mint a feljegyzett d' értékek 95 százaléka, akkor ez már kellően valószínűtlen eseménynek számít statisztikailag és azt mondhatjuk, hogy a T és C csoport szignifikánsan eltérnek egymástól ($p<0,05$ szinten).

A permutációs teszt nagyon szemléletes és számtalan előnye van. Alapvetően kevés matematikai feltevéstől függ. Ennek következtében a különbség helyett bármilyen más statisztikai változót is használhattunk volna a fenti példánkban, akkor is helyes választ kapunk. Például kíváncsiak lehettünk volna, hogy a két csoport (T és C) maximuma tér-e el egymástól szignifikánsan. Az sem probléma, ha az egy csoportban levő megfigyelések egymástól nem függetlenek. Ugyanakkor a permutáción alapuló tesztek hátránya, hogy alkalmazásuk kellően gyors számítógépet igényel, és ezért használatuk csak mostanában kezd terjedni. Az alábbiakban két példával illusztráljuk egymástól nem független megfigyelésekből álló csoport összehasonlítását permutációs teszttel.

Permutációs teszt és fMRI

Az fMRI esetén, mint említettük, az alap megfigyelési egység a voxel. A kérdés, hogy mikor mondhatjuk, hogy két, voxelekből álló, összefüggő terület aktivitása eltér egymástól. Például mérjük egy kép hatását n önkéntesen, és szeretnénk összehasonlítani, hogy a kép hatására egy kilenc voxelből álló, 3x3 elrendezésű terület aktivitása növekszik-e vagy sem. Nichols és Holmes (Nichols és Holmes, 2002) „super threshold" elnevezésű, a problémára megoldást nyújtó algoritmusa a következő:

1. Első lépésben voxelenként hasonlítsuk össze a beavatkozás előtti és utáni értékeket az n önkéntesben egymintás t-próbával. Ha kilenc voxelt hasonlítunk össze, akkor eredményül kilenc t-ér-

téket kaptunk. Válasszuk ki ezek közül az abszolút értékben legnagyobbat. Legyen ez t_{max}.
2. Permutációs lépés. Ahhoz, hogy eldöntsük, hogy a kapott t_{max} szignifikáns-e, határozzuk meg a t_{max} null-eloszlását a permutációs tesztnél leírt módon, tetszőlegesen hozzárendelve egy adott voxel értékéhez a stimulus vagy kontroll értéket. A további lépések megegyeznek a permutációs tesztnél leírtakkal.

A módszer előnye, hogy az így megállapított szignifikanciaszint figyelembe veszi mind a többszörös tesztelést, mind a válaszok korreláltságának hatását. Ráadásul könnyen módosítható, hogy ne csak azt nézze, hogy mennyivel változik az intenzitás egy adott területen, hanem azt, hogy mennyivel változik meg az adott intenzitású terület. Sőt, a két szempont akár együttesen is mérhető (Bullmore et al., 1999; Hayasaka és Nichols, 2003). A módszer természeten nem kötött az fMRI-hez, hasonlóan használható az agyterületek aktivitás változásának követésére EEG-vel (Groppe et al., 2011).

Permutációs teszt és microarray adatok

A szakirodalomban mára adottak a jelátviteli, metabolikus vagy más sejtfunkcióhoz kapcsolható génlisták. (Elvileg nem olyan könnyű eldönteni, hogy adott gén milyen folyamathoz tartozik, de ezt a folyamatot biológiai adatbázisok segítségével és bioinformatikai módszerek alkalmazásával mára lényegében automatizálták.) A microarray elemzés során a feladat úgy néz ki, hogy adott az S_1, S_2 … és S_n génlista és azt kell eldönteni, hogy egy S_i génlista (útvonal) mennyivel változik, mennyire mutat jellegzetes változást a kezelésre. Az általunk is használt és a legáltalánosabban elterjedt ún. gene set enrichment algorithm (GSEA) eljárás a következő (Subramanian et al., 2005):

1. Minden gént hasonlítsunk össze a kezelt és kezeletlen csoport között. Erre a legegyszerűbb módszer a kétmintás t-teszt. Ha n számú chip segítségével mérjük a kontroll csoport expresszióját, és szintén n számú chip segítségével a kezelt csoport génexpresszióját, akkor minden egyes gén esetén a kétmintás teszt ad egy t-értéket.
2. Rendezzük sorba az így kapott t-értékeket egy T-listába. Ha a chip például 20000 gén aktivitásának mérésére képes, akkor 20000 t-értéket rendezünk sorba. A lista i-edik elemét, tehát az i-edik t-értéket, jelölje T(i).
3. Definiáljunk egy, a T hosszúságával megegyező, tömböt. Legyen ennek neve S.
4. Fókuszáljunk csakis egy adott U útvonalra, azaz egy darab génlistára.
5. Menjünk végig a listán, és ha az U-hoz, vagyis az adott útvonalhoz tartozó gént látunk, akkor S(i) értéke legyen egyenlő S(i-1)+wT(i)i, ellenkező esetben S(i)=S(i-1)-wT(i). A w súly az U-hoz tartozó gének és a teljes génlistán szereplő gének számától függ.
6. Az U útvonalhoz így tartozik egy ún. „enrichment score" (ES), ahol ES=max(S).
7. Az ES értéke függ az U útvonalban szereplő gének számától, ezért további normalizációra van szükség. Ez lesz az ún. „normalized enrichment score" (NES).
8. Permutációs lépés. Annak megítélésére, hogy a kapott NES szignifikáns vagy nem, futassuk le az algoritmus 1–7. lépését, permutálva a kezelt és kontroll címkék (S) hozzárendelését a chipekhez. Az általános ajánlás, hogy legalább 1000 permutációt végezzünk.
9. Az így megkapott értékeket hasonlítsuk össze a valós NES értékkel. Ha a NES nagyobb, mint a permutációval kapott értékek 95%-a, akkor az U útvonalhoz tartozó gének megváltoztak, egyébként nem.
10. Végezzük el az 1–9 lépést minden U útvonalra.

Megjegyezzük, hogy a GSEA több variánsa is ismeretes (Nam és Kim, 2008), és a fenti sémában, az egyszerűség kedvéért, feltettük, hogy a kezelés mindig növeli az aktivitást. A GSEA lényege azonban az, hogy nagymértékben csökkentettük a „P" méretét, mert kiválasztottuk a számunkra érdekes gén halmazokat az összes gén közül. Az összes többi elhagyásával ezek az analízisben nem vesznek részt és így a nagy „P" értéke drámaian csökken. Statisztikai szempontból éppen ez az egyik hasonlóság az fMRI-nél leírt „super threshold" és a GSEA eljárás között. Nem az egyes elemeket, hanem azokból képzett halmazokat hasonlítunk össze egy megfelelően választott statisztika segítségével és a lényegtelen halmazokat a további elemzésből kizárjuk. Ez jó ötletnek tűnik, a többszörös tesztelés problémát nagymértékben leegyszerűsíti, de végleg nem oldja meg.

Példaképpen válasszunk ki a „super threshold" algoritmus segítségével a 100000 voxelből 1000 klasztert, amelyiknél a beavatkozás szignifikáns ($p<0,05$) változást okozott. Azonban a talált 1000 klaszter közül 50 téves, ún. „false positive", mert az egyszeri hipotézisvizsgálatok hibája összeadódik.

Hasonlóan tételezzük fel, hogy az ezer génútvonal közül a GSEA száz klasztert azonosít szignifikánsnak p<0,05 szinten. De ebből a százból öt átlagosan téves. Ráadásul arra is kíváncsiak lehetünk, hogy egy adott génútvonalon melyek azok a gének, amelyek átírása tényleg nagymértékben változott és melyeké nem. Ha egy útvonalon van 100 gén (az általános ajánlás, hogy egy útvonalon minimum 15 és maximum 500 gén legyen (Petschner, Tamasi et al. 2013)), akkor megint megoldhatatlan tesztelési problémába ütközünk. A probléma megoldása a klasszikus statisztikai hipotézisvizsgálat alapvető feltevéseinek kiterjesztése.

A HAMIS TALÁLATI ARÁNY FOGALMA ÉS AZON ALAPULÓ TESZTEK

Benjamini és Hochberg (Benjamini és Hochberg, 1995) voltak az elsők, akik a hamis találati arány, angol rövidítéssel FDR (false discovery rate), fogalmát bevezették és az azon alapuló statisztikai hipotézisvizsgálati eljárást kidolgozták. Ahhoz, hogy megértsük ezt az eljárást, célszerű megismételni, hogy mit jelent a statisztikai teszt p-értéke. Ez közelítőleg azt jelenti, hogy mennyi annak a valószínűsége, hogy egy állításra tévesen mondom azt, hogy a két csoport eltér, mivel valójában azonosak. Ha százszor végezzük el az összehasonlítást, akkor maximum ötször tévedhetünk, tehát ötször állíthatjuk tévesen, hogy szignifikáns különbég van köztük, mikor nincs (feltéve, hogy p=0,05). Ha több tesztet végzünk a klasszikus FWER-t kontrolláló tesztek (pl. Bonferroni-teszt) segítségével, ez akkor is így van. Azaz ha százszor végzünk el egy olyan tesztsorozatot, amelynek minden lépésében 100 statisztikai összehasonlítást végzünk, akkor is maximum öt téves állítást engedünk meg, de most 100x100=10000 hipotézisvizsgálatból. Ugyanakkor azzal nem foglakozunk, hogy hány esetben állítottuk tévesen azt, hogy az eltérés nem szignifikáns, holott valójában az. A klasszikus eljárások csak arra fókuszálnak, hogy alacsony limit alatt tartsák a tévesen szignifikánsnak mondható eltérések számát és felteszik, hogy az összes hiba attól függ, hogy hány statisztikai összehasonlítást teszünk.

Benjamini és Hochberg (Benjamini és Hochberg, 1995) úgy érvelt, hogy a kutatókat nem az érdekli, hogy az összes elvégzett tesztből hány a hibás, hanem az, hogy hány százalék a statisztikailag szignifikánsnak talált eredmények között a hibás, azaz mennyi az FDR. Benjamini és Hochberg rögtön eljárást is adott, hogy hogyan kell többszörös tesztelést végrehajtani úgy, hogy az FDR-t ellenőrzés alatt tartjuk. Ez a Benjamini-Hochberg (B-H) teszt. Az eljárás elég egyszerű. Végezzünk el bármilyen módon m számú összehasonlítást, például t-tesztet. Az így kapott $p_1..p_m$ értékek lesznek a bemeneti értékek. Adjunk meg egy FDR értéket, legyen például FDR=0,1. Akkor a B-H teszt megmondja, hogy mely állítások szignifikánsak az adott p szinten úgy, hogy az összes FDR hiba kisebb, mint az általunk választott FDR limit. Ha például az FDR=0,1, akkor lehetséges, hogy akár tíz százaléka is a B-H teszt szerinti szignifikáns állításoknak téves.

A B-H tesztet Storey fejlesztette tovább (Storey és Tibshirani, 2003). Storey algoritmusa segítségével lényegében iteratív módon lehet kontrollálni az FDR értéket többszörös tesztelés esetén. Egy, a Storey algoritmust használó többszörös összehasonlítást végző eljárás két érteket rendel minden egyes összehasonlításhoz. Az i-edik génhez rendelt p(i) érték lényegében megfelel a hagyományos p-értéknek és azt mutatja, hogy mi annak az evidenciája, hogy a kezelés hatására az adott génpár a kezelt és kezeletlen csoportokban eltér. A Storey-féle q(i) ugyanakkor azt mutatja meg, hogy mennyi az összesített hiba, ha az i-dik gént és az összes olyan gént, ahol a p-érték kisebb, mint p(i), szignifikánsnak fogadom el. A q(i) hiba lényegében egy, az i-edik génhez becsült FDR érték.

Egy, a Storey algoritmust használó többszörös összehasonlítást végző eljárás m számú génpár (kezelt – nem kezelt) összehasonlítására körülbelül így néz ki:

1. Add meg a p szignifikancia határt és azt a maximális kritikus q FDR-értéket, amely még elfogadható.
2. Végezd el a kívánt m számú összehasonlítást valamilyen teszt pld. t-teszt segítségével. A kapott m számú p-értéket rendezd sorba. Legyenek ezek p(1) < p(2) …… < p(m).
3. Számold ki alulról felfelé az adjusztált p-értékeket a B-H teszt segítségével. Legyen ez $p_{adj}(i)$. Minden $p_{adj}(i)$-hez számold ki a megfelelő FDR értéket Storey algoritmusával. Ez lesz a q(i).
4. Ha $p_{adj}(i)>p$ vagy $q(i)>q$, azaz nagyobbak, mint a limitként megadott értékek akkor az eljárás véget ért. Az első i-1 génekről fogjuk állítani, hogy szignifikánsan eltérnek a kezelt és kezeletlen csoportokban.

Megjegyezzük, hogy az eredeti B-H tesztnek számos módosítása létezik, de maga az FDR koncepció ezekben a tesztekben is változatlan maradt (Pounds, 2006).

B-H teszt alkalmazása az idegtudományi kutatásokban

Az FDR-en alapuló, többszörös összehasonlítási eljárások standard alkotóelemei a statisztikai szoftvereknek (Statistica, Graphpad, R, Stata), kivéve az SPSS-t. Az fMRI adatok feldolgozására használatos legnépszerűbb szoftverek (AFNI, SPM és BrainVoyager) mindegyikében elérhető az FDR opció. A teszt használata nagyobb biztonsággal teszi lehetővé a funkcionálisan aktív és inaktív klaszterek elkülönítését. A témáról ezen a téren az első összefoglaló közleményt Genovese (Genovese et al., 2002) publikálta. Habár a módszer elterjedőben van, jelenleg még gyakori, hogy két klasztert akkor is különbözőnek tekintenek, ha az összehasonlító teszt p-értéke egy előre definiált p-értéknél (pld. p<0,01) kisebb. Ez azonban megkérdőjelezi az ilyen módon kiértékelt vizsgálatok reprodukálhatóságát (Woo et al., 2014).

A sokszoros tesztelés a többcsatornás EEG adatok elemzése esetén is fontos probléma. A klaszter alapú összehasonlítási feladat lényegében megegyezik az fMRI módszerrel kapcsoltban leírtakkal. A kérdésről Hemmelmann írt áttekintő közleményt (Hemmelmann et al., 2005). Az FDR opció itt is szerepel a használatos programok (EEGLAB, FieldTrip) statisztikai eszköztárában.

A microarray adatok jelentették a legnagyobb inspirációt Benjamini és Hochberg számára ahhoz, hogy közel egy évszázad után lényegileg változtassák meg a statisztikai tesztek filozófiáját. Az általunk is idézett publikációnak (Benjamini és Hochberg, 1995) a Google Scholar szerint 25550 citációja volt e sorok írásakor. Ezen citációk döntő többsége microarray vagy ún. genome-wide association study (GWAS) és ezen a téren az FDR-t kontrolálló többszörös hipotézis vizsgálat lett a standard eljárás (Reiner et al., 2003). A bioinformatikában az R programozási környezet a legelterjedtebb és e nyílt programozási környezet filozófiájából adódóan bármely eljárás szabadon kombinálható az eredeti B-H teszttel vagy annak valamely módosított változatával.

Értelemszerűen az összes hasonló pszichiátriai vagy neurológiai vonatkozású vizsgálat, legyen az microarray vagy GWAS (van den Oord, 2008) is tipikusan ezt a módszert használja. Nem véletlenül, hiszen ha ilyen módon hasonlítjuk össze az eredményeket, akkor lényegesen több esetben mondhatjuk statisztikailag korrekt módon, hogy szignifikáns eltérést találtunk.

KONKLÚZIÓK

Az utóbbi 15-20 év technikai fejlődése robbanásszerű változást idézett elő a kísérleti adatok mennyiségében. Egyetlen vizsgálatban több tízezer, akár több százezer adatot (számot) is gyűjthetünk. A technikai fejlődést lehetővé tevő eszközök látványosak és közismertek. Jóval kevésbé ismert, hogy a roppant adattömeg új kihívást jelentett egy olyan klasszikus tudományágnak, mint a statisztika, és számos területen statisztikai paradigmaváltozást indukált. A paradigmaváltásban természetesen a statisztikusoknak is nagy segítség volt a gyors és olcsó számítógépek megjelenése. E két hatás kombinációjának jó példája a napjainkban terjedő, permutáción alapuló teszt, amelynek elve közel száz éve ismert volt, de a gyakorlatban senki sem tudta kiszámolni. Számos más eljárás is kifejlesztésre került a kis „n" nagy „P" problémával kapcsolatban, de e közlemény keretén belül csak két igen általánosan használt megközelítést ismertettünk.

A kis „n" nagy „P" probléma egyik lehetséges megoldása a „P" csökkentése, az adatok klaszterba való gyűjtése, aggregálása, valamint az azt követő szelektív tesztelés. A klaszterértelmezésre két példát is adtunk, egyfelől mint térben közeli pontok (voxelek) az fMRI esetén, másfelől mint egy adott útvonalon levő gének halmaza a microarray adatok esetén. Matematikai szempontból a klaszterek összehasonlításának problémája mindkét esetben hasonló. Ezért nem meglepő, hogy két teljesen eltérő területen használt algoritmus, az fMRI adatok feldolgozásánál használt „super threshold" és a microarray adatok esetén használt GSEA is azonos sémát követ. Közleményünkben a klaszterek összehasonlítására, szakszóval a klaszterek statisztikai inferenciájára helyeztük a hangsúlyt, és nem magára a klasztert létrehozó algoritmusra. Azáltal, hogy el tudtuk dönteni, hogy melyik klaszter szignifikáns és melyik nem, lehetségessé vált, hogy a nem szignifikáns klaszterbe sorolt adatokat ne vegyük figyelembe a további analízis során, így az összehasonlítandó adatok mennyisége nagymértékben csökkent. Az aggregálásra, majd azt követő szelektív tesztelésre a GSEA volt jó példa. Itt először eldöntöttük, hogy melyek az érdekes útvonalak, majd az érdekes útvonalakon belül vizsgáltuk, melyek az érdekes gének. Ezt a többlépcsős tesztelési eljárást hierarchikus tesztelési eljárásnak hívják.

A klaszterezés ugyan csökkenti a probléma méretét, az elvégzendő tesztek számát, de nem oldja meg többszörös tesztelés problémáját. Ennél radikálisabb megoldásra volt szükség, amelynek keretében került sor az FDR fogalmának bevezetésére. Az FDR, nem

teljes matematikai precizitással, lényegében egy limit, ami mutatja, hogy egy közlemény szignifikáns állításainak maximum mekkora százaléka lehet hibás. Vizsgálatainkban (Petschner et al., 2013; Tamasi et al., 2014) ez a limit 25% volt, ami a microarray analízisben nem járatos olvasónak nagynak tűnhet, valójában azonban ez a standard hibahatár ilyen vizsgálatok esetén. Microarray, fMRI és egyéb nagy adatmennyiséget generáló vizsgálatok gyakran voltak bírálat tárgyai, mert az eredményeket más kutatók nem tudják (tudták) reprodukálni. Az új megközelítés legalább lehetővé teszi, hogy kvantitatív módon megbecsüljük, mi annak az esélye, hogy két független vizsgálat megerősíti egymás állításait. Nem mellesleg, hiszen tulajdonképpen ez a statisztikai próbák elsődleges célja. A közleményünkben vázolt kvantitatív megközelítésre nemcsak a vizsgálati eredmények értékelésekor, hanem a vizsgálat tervezéskor is szükség van. Az alapkérdés mindig az elemszám, azaz hogy minimálisan hány állat vagy önkéntes kell a kísérlet statisztikailag is korrekt kivitelezéséhez, értékelhető adatokhoz. A kérdés megválaszolása általában komolyabb matematikai ismereteket és szakember segítségét igényli. Ugyanakkor a szakemberrel való konzultáció során olyan kérdésekre is fel kell készülni, hogy mennyi lehet maximálisan az FDR, a tévesen szignifikánsnak nyilvánított gének aránya. Ez a kérdés a terület statisztikai metodológiáját nem ismerő kutató számára meglepő lehet. Így közleményünk célja elsősorban az volt, hogy egy hangsúlyozottan nem matematikai jellegű bevezetést adjon a kis „n” nagy „P” probléma világába és az alapvető szempontok körvonalazásával elősegítse a pszichofarmakológusok és a statisztikusok közti párbeszéd hatékonyságát.

Rövidítések jegyzéke

FWER	–	family-wise error rate
FDR	–	false discovery rate
GSEA	–	gene set enrichment algorithm
NES	–	normalized enrichment score
ES	–	enrichment score

Köszönetnyilvánítás. Jelen tanulmány a Nemzeti Agykutatási Program (KTIA_NAP_13-1-2013-0001) támogatásával valósult meg (szerződés nyilvántartási száma: KTIA_13_NAP-A-II/14.).

Levelező szerző: Bagdy György, Semmelweis Egyetem, Gyógyszerhatástani Intézet, Budapest, Nagyvárad tér 4.
E-mail: bagdy.gyorgy@pharma.semmelweis-univ.hu

IRODALOM

1. Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B. 57: 289–300.
2. Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M.J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans Med Imaging 18(1): 32-42.
3. Cleophas, T.J., Zwinderman, A.H. (2006). Clinical trials are often false positive: a review of simple methods to control this problem. Curr Clin Pharmacol 1(1): 1-4.
4. Genovese, C.R., Lazar, N.A., Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15(4): 870-878.
5. Groppe, D.M., Urbach, T.P., Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48(12): 1711-1725.
6. Hayasaka, S., Nichols, T.E. (2003). Validating cluster size inference: random field and permutation methods. Neuroimage 20(4): 2343-2356.
7. Hemmelmann, C., Horn, M., Susse, T., Vollandt, R., Weiss, S. (2005). New concepts of multiple tests and their use for evaluating high-dimensional EEG data. J Neurosci Methods 142(2): 209-217.
8. Kostyalik, D., Vas, S., Katai, Z., Kitka, T., Gyertyan, I., Bagdy, G., Tothfalusi, L. (2014). Chronic escitalopram treatment attenuated the accelerated rapid eye movement sleep transitions after selective rapid eye movement sleep deprivation: a model-based analysis using Markov chains. BMC Neurosci 15(1): 120.
9. Lange, N. (2003). What can modern statistics offer imaging neuroscience? Stat Methods Med Res 12(5): 447-469.
10. Ludbrook, J. (1994). Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. Clin Exp Pharmacol Physiol 21(9): 673-686.
11. Nam, D., Kim, S.Y. (2008). Gene-set approach for expression pattern analysis. Brief Bioinform 9(3): 189-197.
12. National Research Council (U.S.). Committee on Mathematical Sciences Research for DOE's Computational Biology. Mathematics and 21st century biology. National Academies Press, Washington DC, 2005.
13. Neuhauser, M. (2006). How to deal with multiple endpoints in clinical trials. Fundam Clin Pharmacol 20(6): 515-523.
14. Nichols, T.E., Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15(1): 1-25.
15. Petschner, P., Tamasi, V., Adori, C., Kirilly, E., Ando, R.D., Tothfalusi, L., Bagdy, G. (2013). Gene expression analysis indicates CB1 receptor upregulation in the hippocampus and neurotoxic effects in the frontal cortex 3 weeks after single-dose MDMA administration in Dark Agouti rats. BMC Genomics 14: 930.
16. Pounds, S.B. (2006). Estimation and control of multiple testing error rates for microarray studies. Brief Bioinform 7(1): 25-36.
17. Reiner, A., Yekutieli, D., Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19(3): 368-375.
18. Storey, J.D., Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100(16): 9440-9445.
19. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L.,

Golub, T.R., Lander, E.S., Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43): 15545-15550.
20. Tamasi, V., Petschner, P., Adori, C., Kirilly, E., Ando, R.D., Tothfalusi, L., Juhasz, G., Bagdy, G. (2014). Transcriptional evidence for the role of chronic venlafaxine treatment in neurotrophic signaling and neuroplasticity including also glutatmatergic- and insulin-mediated neuronal processes. PLoS One 9(11): e113662.
21. van den Oord, E.J. (2008). Controlling false discoveries in genetic studies. Am J Med Genet B Neuropsychiatr Genet 147B(5): 637-644.
22. Woo, C.W., Krishnan, A., Wager, T.D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. Neuroimage 91: 412-419.

The problem of small „n" and big „P" in neuropsychopharmacology, or how to keep the rate of false discoveries under control

One of the characteristics of many methods used in neuropsychopharmacology is that a large number of parameters (P) are measured in relatively few subjects (n). Functional magnetic resonance imaging, electroencephalography (EEG) and genomic studies are typical examples. For example one microarray chip can contain thousands of probes. Therefore, in studies using microarray chips, P may be several thousand-fold larger than n. Statistical analysis of such studies is a challenging task and they are refereed to in the statistical literature such as the small "n" big "P" problem. The problem has many facets including the controversies associated with multiple hypothesis testing. A typical scenario in this context is, when two or more groups are compared by the individual attributes. If the increased classification error due to the multiple testing is neglected, then several highly significant differences will be discovered. But in reality, some of these significant differences are coincidental, not reproducible findings. Several methods were proposed to solve this problem. In this review we discuss two of the proposed solutions, algorithms to compare sets and statistical hypothesis tests controlling the false discovery rate.

Keywords: functional imaging studies, microarray, false discovery rate, permutation test, gene set enrichment analysis, fMRI, statistics