A klinikai gyógyszervizsgálatok alapjai

Ferenci Tamás

2025. június 19.

Tartalomjegyzék

Előszó

A klinikai gyógyszervizsgálatok gyakran képezik közbeszéd tárgyát. Amikor kipróbálnak egy teljesen új szert, amikor megjelenik egy új vizsgálati eredmény, sokszor kerülnek gyógyszervizsgálati eredmények a címlapokra. Sajnos nem ritkán alanyai félreértéseknek, rosszabb esetben szándékos félrevezetéseknek is. Gyógyszergyárak anyagilag érdekeltek abban, hogy feltupírozzák az eredményeket, bizonyos csoportok inkább ideológiai semmint tudományos alapon utasítanak el vagy épp magasztalnak fel készítményeket, de a hamis állítások többsége egyszerűen abból fakad, hogy a megszólalók, véleményt nyilvánítók egy része maga sem ismeri alaposan a gyógyszervizsgálatokat. A helyzetet sajnos sokszor a szakemberek is rontják, amikor misztifikálják a területet, és megszólalásaikban csak úgy repkednek a szakkifejezések négyes fázistól a randomizált placebo-kontrolláltig.

Jelen írásnak nem célja, hogy az ilyen félreértéseket, félrevezetéseket célirányosan tárgyalja. Nem mintha ez nem lenne fontos, ellenkezőleg, nagyon is az, de előtte még van egy lépés: az alapgondolatok, az alapfogalmak bemutatása. A kettő ráadásul nagyon szorosan összefügg: annál kevésbé fogékony valaki a valótlan információkra, minél jobban ismeri a területet magát.

Sajnos azonban magyar nyelven nem áll rendelkezésre olyan, az alapkoncepciókat bemutató anyag, ami nem szakemberekhez szól, de egyidejűleg kellően részletes is. Jelen összefoglaló célja e hiány pótlása. Fontos rögzíteni, hogy a gyógyszervizsgálatokhoz az itt tárgyalt területeken túl még rengeteg további tudományág kapcsolódik (Lakner, Renczes, és Antal 2009; Gachályi, Lakner, és Borvendég 2003), biokémiától gyógyszertanon át akár egészen a közgazdaságig. Jelen összefoglalóban igyekeztem azokat a területeket kihagyni, melyek a gyógyszerfejlesztésben résztvevők számára fontosak, és azokra a kérdésekre fókuszálni, melyek a gyógyszervizsgálatok eredményeit olvasók számára lényegesek.

Nem szeretnék zsákbamacskát árulni: ez nem egy 5 perces olvasmány. Nem tudom megígérni, hogy nem igényel energiát végigolvasni, de meg tudom ígérni, hogy ezen kívül mást nem igényel: nem kell hozzá orvosi ismeret, nem kell hozzá statisztikatudás, nem lesznek képletek és szakkifejezések.

Reménykedem abban, hogy emiatt hasznos olvasmány lesz mindazoknak, akik szeretnék megérteni, hogy a gyógyszervizsgálati eredmények mit jelentenek, és hogyan kell őket értelmezni. Legyenek érdeklődő laikusok, akik szeretnék megérteni, hogy mit olvastak az újságban, nem ezzel a területtel foglalkozó egészségügyisek, akiknek már homályosak az idevágó iskolai emlékei,

vagy éppen újságírók, akik szeretnének hitelesen beszámolni az olvasóknak. Ezek az eredmények sokszor nagyon is közvetlen hatással vannak életünkre, egészségünkre, így rendkívül fontos, hogy minél többen értsék, hogy mit jelentenek a számok, fogalmak, következtetések.

Minden észrevételt, javaslatot, kritikát örömmel várok a tamas.ferenci@medstat.hu email-címen!

A kézirathoz fűzött észrevételekért, gondos átolvasásért és javításokért szeretném köszönetemet kifejezni (alfabetikus sorrendben) Gombos Tímeának, Lakos Andrásnak, Molnár Márknak és Singer Júliának. Természetesen minden megmaradt hiba engem terhel.

1 A klinikai gyógyszervizsgálatok célja és módszerei

Arra a kérdésre, hogy egy gyógyszer „jó”-e, lehetetlen egyszerűen válaszolni, hiszen ez egy többdimenziós kérdés: számít, hogy biztonságos-e, számít, hogy hatásos-e, számít, hogy megfelelő, állandó minőségben gyártható-e, manapság egyre több esetben számít az ára is. Ezek a kérdések ráadásul sok esetben nem függetlenek sem egymástól, sem a kezelni kívánt betegség jellegétől: egy sok mellékhatást okozó gyógyszer is lehet jó, ha egy gyógyszer nélkül nagyon rossz kilátású betegségben sokat segít, egy kevés mellékhatást okozó gyógyszer is lehet rossz, ha egy gyógyszer nélkül sem túl súlyos betegségen csak kicsit javít. A gyógyszer adására vonatkozó döntés mindig **kockázat-haszon mérlegelés** eredménye; fontos feladat azonban, hogy ehhez a mérlegeléshez a valóságnak megfelelő kiindulási információink legyenek.

Ez az összefoglaló ebből a feladatból most két, az orvoslásban és a közbeszédben is talán legfontosabb kérdésre fog fókuszálni: a hatásosság és a biztonság megítélésére.

1.1. Gyógyszerek empirikus vizsgálata

Egy gyógyszer(jelölt)nek mind a hatásossága, mind a biztonságossága megvizsgálható számos módon. Használhatunk – és használnak is – állatkísérleteket, matematikai modelleket, biológiai megfontolásokat és így tovább, ám a folyamat végén, ha a gyógyszer még mindig megfelelőnek tűnik, manapság szinte kivétel nélkül meg kell vizsgálni a gyógyszer hatását – mind fő-, mind mellékhatásait – **empirikusan** is, azaz embereknek ténylegesen beadva a gyógyszert, és való életbeli adatokat gyűjtve az ezután történtekről. Ennek során megpróbáljuk a „hat-e a gyógyszer” (ha igen, pontosan mire, mennyire, esetleg milyen gyógyszerrel vagy egyéb kezeléssel együtt alkalmazva) és az „okoz-e mellékhatást a gyógyszer” (ha igen, milyet, milyen súlyosat, mennyire gyakran) kérdéseket megválaszolni. Vegyük észre, hogy a két kérdés nagyon hasonló: mindkét esetben arra vagyunk kíváncsiak, hogy valamilyen esemény előfordulásának a valószínűségét megnöveli-e a gyógyszeres kezelés (csak épp az egyik esetben ez egy pozitív, a másikban egy negatív jellegű esemény). Azt gondolhatnánk, hogy a kérdés nem nehéz, hiszen az előfordulási valószínűséget tudjuk közelíteni azzal, hogy milyen gyakran fordul elő az esemény, így egyszerűen annyit kell tenni, hogy gyűjtünk betegekből egy nagyobb vizsgálati csoportot aki kap gyógyszert és egy másikat, aki nem, majd összehasonlítjuk az adott esemény – akár pozitív, akár negatív – gyakoriságát. A valóságban sajnos ennél bonyolultabb a helyzet.

1.2. A confounding problémája

Egy gondolatkísérlettel könnyen illusztrálhatjuk, hogy elvileg[1] mi az *ideális* vizsgálati módszer: fogunk egy alanyt, beadjuk neki a gyógyszert, felírjuk mi történik, utána beülünk egy időgépbe, visszamegyünk, *nem* adjuk be a gyógyszert, majd így is felírjuk mi történik. A kettő *különbsége* lesz a gyógyszer *hatása* az adott embernél (akár negatív, akár pozitív hatásról van szó). Az időgép fontos volt, hiszen emiatt mondhatjuk, hogy *minden* körülmény *pontosan* ugyanaz volt, *kizárólag* a kezelés tényében volt eltérés – ez viszont garantálja, hogy ha a kimenetben találunk eltérést, akkor az *biztosan* a kezelés miatt kellett legyen. Tényleg megtaláltuk tehát a kezelés valódi, okozati hatását.

Időgép híján viszont gondban leszünk a viszonyítással. Ha az alanynak adunk gyógyszert, akkor megtudjuk az egyik érdekes kimenetet, hogy mi történik vele kezelés mellett, de mi a helyzet a másikkal? Honnan tudjuk, hogy mi történt *volna* vele kezelés nélkül? (Hiszen ezt a kettőt kell összevetni.) Sajnos – ha nincs időgépünk – sehonnan! Abban a pillanatban, hogy beadjuk a gyógyszert, megsemmisül az a kimenet, hogy mi történt volna gyógyszer nélkül. Ha nem adunk neki gyógyszert, akkor meg pont fordított a helyzet, tudjuk mi történik gyógyszer nélkül, de nem tudjuk – és nem tudhatjuk – mi történt volna gyógyszerrel.

Az egyik megoldás, hogy két beteget nézünk: az egyik kap gyógyszert (kezelt beteg), a másik nem (kontroll beteg), és az ő eredményeiket hasonlítjuk *egymáshoz*. Vegyük azonban rögtön észre, hogy ez a megközelítés csak és kizárólag akkor működik, ha igaz az a háttérfeltételezés, hogy a kontroll beteggel történtek mutatják, hogy mi történt *volna* a kezelt beteggel, ha nem kapott *volna* gyógyszert. Ez azonban nem nyilvánvaló, hiszen egy másik emberről van szó! Ő pedig *más* jellemzőiben is eltér(het) azon túl, hogy nem kap gyógyszert, amely jellemzők között lehet olyan, ami hat a vizsgált kimenetre. Mi van akkor, ha a kezelt beteg túlél és a kontroll meghal, de a kezelt beteg fiatalabb volt? Lehet, hogy valójában a kezelésnek semmi szerepe nem volt, egyszerűen a kezelt beteg alacsonyabb életkora a valódi oka a túlélésének! Még az is lehet, hogy a kezelés kimondottan ront a túlélési kilátásain, csak annyival fiatalabb volt, hogy az többet számított.

A legtöbb ember ezt hallva azt mondja, hogy jó, de ez a probléma egyszerűen megoldható: ne egy-egy embert nézzünk, hanem nagyobb csoportokat. Ez valóban fontos lesz, de más miatt! A mostani problémánkon azonban nem segít.

Nézhetünk akár 1000 kezelt beteget is, ha ők fiatalabbak – nem feltétlenül mind, akár csak átlagosan is – mint az 1000 kontroll betegünk, akkor pontosan ugyanazzal a problémával fogunk szembesülni: innentől kezdve egyszerre *két* eltérés van a csoportok között, a vizsgált (egyik kap gyógyszert, másik nem), és egy vele együttjáró másik (egyik fiatalabb, másik idősebb).

[1] Az itt bemutatott felfogás, az ún. counterfactual megközelítés valójában csak az egyik lehetséges megközelítése az okozatiság fogalmának. Ez a kérdés azonban részben már a filozófia tárgyköréhez tartozik; a gyakorlati gyógyszervizsgálati munkához, illetve a gyógyszervizsgálati eredmények megértéséhez elégséges az itt bemutatott gondolkodási keret (Greenland és Robins 1986; Maldonado és Greenland 2002; Maldonado 2013, 2016; Dawid 2000).

Innentől kezdve, ha találunk is különbséget a túlélők arányában, nem tudhatjuk, hogy az mi miatt van: a vizsgált eltérés miatt, a vele együttjáró másik eltérés miatt (vagy a kettő valamilyen keveréke miatt). Ez első ránézésre meglepő, de drámai problémákhoz vezet: akár még az is előfordulhat, hogy a kezelés kimondottan ártalmas, *és* a túlélők aránya a kezelt csoportban nagyobb! (Ha a kezeltek annyival fiatalabbak, hogy a kisebb életkor jótékony hatása nagyobb, mint a kezelés káros hatása). Ezt a problémát hívjuk angol szóval **confoundingnak**. Nem igazán terjedt el rá magyar megfelelő, de angolul nagyon szemléletes: szó szerint „egybemosódást" jelent, és csakugyan, arról van szó, hogy a vizsgált tényezőbeli eltérés (a becsülni kívánt hatás) összemosódott más eltéréssel vagy eltérésekkel.

Az ilyen változókat, amik ezt a galibát okozzák, tehát amelyekre egyszerre igaz, hogy eltérnek a vizsgált csoportok között *és* hatnak a kimenetre (ha bármelyik nem áll fenn, akkor nincsen baj, gondoljuk végig), szokás confoundernek, vagy magyarul zavaró változónak nevezni.

1.3. A randomizáció szerepe, megfigyeléses és kísérletes vizsgálatok

Mit tudunk ez ellen tenni? Olyan confoundereknél, amelyekről van információnk (ismerjük az értéküket az egyes alanyokra), rendelkezésre állnak statisztikai megoldások, amikkel segíthetünk a helyzeten. Megtehetjük például, hogy külön vizsgáljuk a gyógyszer hatását csak a fiatalokban és csak az idősekben (rétegzett elemzés). Erre vannak kifinomultabb módszerek is, de bármilyen kifinomultan járunk is el, biztosan nem kezeljük jól a helyzetet abban az esetben, ha eszünkbe sem jut valamelyik változóról, hogy az confounder lehet.

Az ember azt gondolhatná, hogy ez ellen a probléma ellen nincs védelem, de ez nem így van! Igaz, hogy csak egy, de egy módszer van, ami akkor is működik, ha nem is tudjuk, hogy mi(k) a confounder(ek).

Ez a módszer pedig nem más, mint az, ha az alanyokat véletlenszerűen – például pénzfeldobással – sorsoljuk a kezelt és kontrollcsoportok egyikébe! Ezt szokás **randomizációnak** hívni. Ebben az esetben ugyanis a fiatalok és idősek aránya között nem lesz semmilyen szisztematikus[2] eltérés a két csoport között. De ennél jóval fontosabb, hogy *semmi másban sem* lesz szisztematikus eltérés! Azaz: abban sem, amiről eszünkbe sem jut, hogy confounder! Ha véletlenül a zöld szeműek inkább kapnak gyógyszert, és a zöld szemszín valamiért javítja a túlélést, az bizonyosan elrontana minden vizsgálatot, mert nincs olyan gyógyszervizsgálat a világon, ahol a szemszínt egyáltalán felírnák – de vegyük észre, hogy a randomizálás *még ekkor is* működni fog! Hiszen ha pénzfeldobással sorsoltuk az alanyokat két csoportba, akkor szemszínben

[2]Érdemes picit pontosítani, hogy mit értünk „szisztematikus" alatt, hiszen valaki azt mondhatná (joggal), hogy éppenséggel a pénzfeldobásnál is előfordulhat, hogy az összes fiatal az egyik csoportba kerül, míg az összes idős a másikba. A dolog úgy nyer értelmet, ha hosszú távon szemléljük a kérdést, mintha képzeletben újra meg újra megismételnénk a randomizálást. Lehet, hogy előfordulhat az előbbi helyzet is, de az ellenkezője is, és *átlagosan* kiegyensúlyozott lesz a két csoport. Belátható, hogy a klinikai vizsgálat szempontjából ez fog számítani, ugyanis a kiértékelésnél megadjuk az eredmények bizonytalanságát is – ezt nemsokára látni is fogjuk – és ez épp úgy van számolva, hogy figyelembe veszi ezt az ingadozást (Senn 2013).

sem lesz szisztematikus eltérés a csoportok között, azaz nem baj, hogy hat a túlélésre. Úgy is fogalmazhatunk, hogy a randomizáció – ha jól végzik – automatikusan és teljesen megoldja a confounding problémáját, mert az összes potenciális confoundert kiszűri, azokat is, amikről nincs információnk, sőt, azokat is, amikről eszünkbe sem jut, hogy confounderek.

Ha ez ilyen egyszerű, és ezzel teljeskörűen megoldhatjuk a problémát, akkor adódik a kérdés, hogy miért nem csinálunk mindig randomizált, kontrollált vizsgálatot? Egyrészt, bizonyos helyzetekben ezt fizikailag vagy etikailag lehetetlen elvégezni. Ha érdekel minket, hogy a vöröshús-fogyasztás rákot okoz-e, nem lehet embereket randomizáltan etetni vörös hússal, ha érdekel, hogy a császármetszéssel születés tényleg cukorbetegséget okoz-e, nem lehet szülő nőket randomizáltan császármetszésre küldeni (függetlenül attól, hogy szükségük van-e rá…) stb. A példák is mutatják, hogy ez inkább a betegségek okait, terjedését kutató tudományágban, az epidemiológiában tipikus helyzet. Mi a probléma gyökere? Az, hogy a randomizáció elvégzéséhez szükséges, hogy a vizsgált tényezőt mi határozzuk meg, mi „adagoljuk" az alanyoknak. A probléma tehát az, hogy erre sokszor nincs lehetőség.

Azokat a kutatásokat, ahol a vizsgált tényezőt nem a kutatók irányítják, csak passzíve feljegyzik, hogy az alanyokkal mi történt vagy történik, anélkül, hogy bármivel is befolyásolnák az eseményeket, **megfigyeléses vizsgálatnak** nevezzük. Ha a kutatók aktívan befolyásolják a vizsgált tényezőt, **kísérletes vizsgálatról** beszélünk.

Míg az epidemiológiában legtöbbször csak megfigyeléses vizsgálatot lehet végezni, addig a gyógyszerkutatásban nagyon sokszor van mód kísérlet végzésére – randomizált, kontrollált kísérletre – is.

A kísérletek hatalmas előnye tehát, hogy elvileg mentesek tudnak lenni a confoundingtól, anélkül, hogy ehhez bármit kellene tennünk a randomizáláson túl. (Ez nem jelenti azt, hogy egy klinikai kísérletben, ha van információnk egy kimenetet befolyásoló változóról, akkor azt más okokból ne lenne célszerű felhasználni, de ez már technikai kérdés.) Megfigyeléses vizsgálatnál, bármilyen ügyesen is járunk el, mindig a fejünk felett fog lebegni, hogy mi van, ha valamilyen confounderről megfeledkeztünk. A kísérletek egy további előnye, hogy lehet maszkolni a kezelés tényét, de erről majd kicsit később.

Hozzá kell tenni, hogy a klinikai kísérleteknek hátrányaik is vannak a megfigyeléses vizsgálatokkal szemben (azon túl persze, hogy nem mindig végezhetőek el). Az egyik, hogy a kísérletekbe bevont alanyok száma jellemzően jóval kisebb, az utánkövetésük hossza pedig általában csak jóval rövidebb tud lenni, mint egy megfigyeléses vizsgálatnál. Ez az oka annak, hogy az utóbbi időben egyre többször végeznek megfigyeléses vizsgálatot a gyógyszerkutatásban is.

2 A véletlen ingadozás kérdésköre

2.1. A véletlen ingadozás szerepe, a gyógyszer hatásának statisztikai szignifikanciája

Térjünk most vissza arra a – korábban függőben hagyott – kérdésre, hogy miért kell nagyobb létszámú csoportokat képeznünk. A helyzet megértéséhez képzeljük el, hogy egy bizonyos betegségben kezelés nélkül az alanyok 20%-a hal meg adott időn belül. Hány ember fog egy 100 fős betegcsoportból meghalni? Ha a betegek halála független egymástól, akkor ez úgy fogható fel, mintha lenne egy nagy urnánk, benne a golyók 20%-a piros, 80%-a fehér, mi pedig véletlenszerűen, azaz alapos megkeverés után kihúznánk belőle 100 golyót és megszámolnánk, hogy hány pirosat húztunk. Az ember érzése, hogy a legvalószínűbb, hogy 20-at, és ez a benyomás igaz is. De ha magunk elé képzeljük ezt a példát, akkor egy még fontosabb dolgot is látni fogunk: hogy ez csak a legvalószínűbb, de nem biztos! Húzhatunk 19 piros golyót, vagy 21-et is, pontosan úgy, ahogy ez szabályos pénzérmét 100-szor feldobva sem biztos, hogy *pont* 50 fejet és 50 írást dobunk. Lehet, hogy 19-en halnak meg, lehet, hogy 21-en, az is lehet, hogy 15-en, vagy 25-en, sőt, az is lehet, hogy senki, vagy mind a 100. Ez utóbbi esetek valószínűsége hihetetlenül kicsi – de nem nulla.

Tegyük a kérdést izgalmasabbá! A gyógyszeres kezelés ezt a valószínűséget 10%-ra csökkenti, azaz a gyógyszer hatásos. Ez azt jelenti, hogy most két urnánk van, egy „kontroll" és egy „kezelt" feliratú, az előbbiben a golyók 20%-a piros, az utóbbiban csak 10%. Mit jelent egy klinikai vizsgálat a hatásosság eldöntésére? Azt, hogy mindkét urnából – véletlenszerűen, alapos keverés után – kihúzunk 100 golyót. Ha egy randomizált kísérletről van szó, semmilyen más hibaforrás nincs, akkor természetes lenne azt mondani, hogy a kezelés akkor tekinthető hatásosnak, ha a kezeltek között kisebb a halálozások aránya. De az urnás példa mutatja, hogy mi a probléma: *attól még*, mert az urnában 20%-10% az arány, *véletlenül* húzhatunk mindkettőből 15 pirosat! Az sem lehetetlen, hogy a 20%-os urnából 10, a 10%-osból 20 pirosat húzunk. Még egyszer: pusztán a véletlen szeszélye folytán!

Látható tehát, hogy kétirányú a probléma: ha a kezelt csoportban kevesebb halálozás lesz (kevesebb pirosat húzunk abból az urnából), akkor sem biztos, hogy hat a gyógyszer (kisebb a pirosak aránya az egész urnában), de ha ugyanannyit vagy épp többet húzunk, akkor sem lehetetlen, hogy mégis hat a gyógyszer. Ez a **véletlen ingadozás problémája**.

Fontos észrevétel, hogy a problémára tökéletes megoldás nincs: ha a kezeltek közül mindenki meggyógyul, és a kontrollcsoportból mindenki meghal, *elvileg* akkor is elképzelhető, hogy nem

hat a gyógyszer. (Mindkét urnában a golyók fele piros, de az egyikből pont 100 pirosat, a másikból pont 100 fehéret húztunk – pusztán a véletlen szeszélye folytán. Extrém valószínűtlen, de nem lehetetlen.) Ebben a helyzetben tehát *biztos* döntést nem tudunk hozni; ha ragaszkodnánk a biztos válaszhoz, akkor a kutatás minden eredménye esetén azt kellene mondani, hogy „nem tudjuk, hogy hat-e a gyógyszer". Ahhoz tehát, hogy egyáltalán válaszolni tudjunk a gyógyszer hatásosságára vonatkozó kérdésre, *muszáj* megengednünk valamennyi hibázást.

Valahol határt kell húznunk: már akkor mondjuk, hogy hat a gyógyszer, ha 20% helyett 19% hal meg? Vagy csak akkor, ha 15? Vagy még szigorúbbak vagyunk, és csak akkor, ha 5? Ezzel a döntéssel a kétféle hibázás között egyensúlyozunk: ha nagyon szigorúak vagyunk, akkor ritkán fogunk egy hatástalan gyógyszert tévesen hatásosnak minősíteni, de sokszor fogjuk a hatásosakra is azt mondani, hogy hatástalanok. Ha enyhék vagyunk, akkor pont fordítva.

Az orvosi vizsgálatok kiértékelését ma domináló statisztikai iskolában, amit frekventista iskolának szoktak nevezni, úgy húzzuk meg ezt a határt, hogy igaz legyen, hogy ha a gyógyszer nem hat, akkor csak adott, kicsi – tipikusan 5% - valószínűséggel mondjuk mégis, a véletlen szeszélyéből fakadó tévedés folytán azt, hogy hat. Ezt az 5%-ot szokás szignifikanciaszintnek nevezni. Ha a gyógyszer az így meghúzott küszöbnél nagyobb hatást mutat, akkor azt mondjuk: statisztikailag **szignifikáns a hatás**. Értsd: nem tudható be a véletlen ingadozásnak, olyan nagy, hogy azt kell feltételeznünk, hogy az nem pusztán a véletlen ingadozás miatt jött ki.

Természetesen nem kötelező 5%-ot használni. Választhatunk más értéket is, például 1%-os szignifikanciaszintet: ez azt jelenti, hogy szigorúbb küszöböt szabunk meg, hiszen 1%-ra kell csökkenteni annak valószínűségét, hogy egy valójában nem hatásos gyógyszer pusztán a véletlen ingadozás miatt megugorja ezt a küszöböt. Azaz csak a vizsgálatunkban erősebben ható gyógyszerekre mondjuk azt, hogy „ez vélhetően nem a véletlen ingadozás miatt jött ki, miközben a valóságban nem hat a gyógyszer".

Kézenfekvő lenne a fenti, fordított logikával szemben a természetes kérdést feltenni, tehát azt, hogy „mennyire valószínű, hogy hat a gyógyszer". A fenti gondolkodási keretben azonban erre a kérdésre nem lehet válaszolni; ez már átvezet minket egy másik iskolába, melyet bayesi statisztikának szokás nevezni. Sok érv szólna ennek használata mellett, de addig is, amíg elterjed, marad a „szignifikáns" szó használata, és az, hogy ne feledkezzünk meg erről a fordított logikáról: ha szignifikáns a gyógyszer hatása, az nem azt jelenti, hogy valószínű, hogy hat a gyógyszer, hanem azt, hogy ha nem hatna, akkor valószínűtlen lenne, hogy azt kapnánk a vizsgálatban, amit ténylegesen kaptunk is. Természetesen egyáltalán nem észszerűtlen ilyenkor azt mondani, hogy akkor „minden bizonnyal" hat a gyógyszer, de ettől még a dolog nem fordítható egyszerűen meg.

2.2. A kutatás ereje és mintanagysága

Ha egyszer a szignifikanciaszint egy hibázás valószínűségét fejezi ki, és az értékét mi határozzuk meg, akkor miért 5%-ot használunk? Miért nem 1%-ot? (Így jóval kisebb lenne ennek a hibázásnak a valószínűsége!) Miért nem 0,1%-ot? Miért nem 0,01%-ot?

Hogy ezt megértsük, gondoljuk végig, hogy a szignifikanciaszint csökkentése mit jelentene: azt, hogy csak azokra a gyógyszerekre mondjuk, hogy hatnak, amelyek rendkívül hatásosnak bizonyulnak a vizsgálatban. Már egy közepes mértékű hatásra is azt mondjuk, hogy „ez még belefér a véletlen ingadozásba". Ez tényleg vitathatatlanul csökkenti annak a valószínűségét, hogy a hatástalan gyógyszereket tévedésből hatásosnak mondjuk, eddig rendben van, csak épp lenne egy kellemetlen következménye is: az, hogy a hatásos, nem extrém hatásos, de azért hatásos gyógyszerek egy nagy részére is azt mondanánk, hogy hatástalan! Hiszen nem tudják megugrani a nagyon magasra tett lécet.

A probléma tehát az, hogy a szignifikanciaszinttel az *egyik* típusú hibázás valószínűségéről nyilatkoztunk, arról, hogy ha nem hat a gyógyszer, mekkora valószínűséggel minősítjük mégis, pusztán a véletlen ingadozásból fakadó hiba folytán hatásosnak, de ezzel szemben áll egy másik típusú hibázás is: a hatásos gyógyszer téves hatástalannak minősítése. Mi a helyzet ennek a valószínűségével? Ez keményebb dió, hiszen ennek valószínűsége attól is függ, hogy *mennyire* hat a gyógyszer (amit mi magunk sem tudhatunk!). Minél hatásosabb, annál kisebb lesz ez a valószínűség.

Egy **kutatás erejének** nevezzük annak a valószínűségét, hogy a hatásos gyógyszert *helyesen* hatásosnak minősítjük. Az előbbiekből tehát látható, hogy ez függ a gyógyszer valódi hatásának mértékétől.

Nem kevésbé fontos, hogy függ a bevont betegek számától, a mintanagyságtól is: minél *több* betegünk van, a vizsgálatból kapott érték (a kihúzott golyók között a pirosak aránya) annál *kevésbé* ingadozik a valódi érték (a pirosak aránya az urnában) körül. Ez azért fontos, mert a korábban látott okfejtés azt mondja, hogy a szignifikancia azon múlik, hogy mennyire tudjuk igazolni, hogy a hatás nem fakadhatott véletlen ingadozásból – ez pedig természetesen annál könnyebb, minél kisebb a véletlen ingadozás mértéke. Azaz: minél nagyobb a mintanagyság, annál nagyobb az erő.

A gyakorlatban erre az összefüggésre általában fordítva néznek rá: nem azt nézik, hogy adott mintanagyságnál mekkora az erő, hanem megkövetelnek egy adott erőt (tipikusan 80 vagy 90%), és azt számolják ki, hogy ennek eléréséhez mennyi beteg kell. Ez tehát lehetővé teszi egy vizsgálat mintanagyságának racionális megtervezését!

Természetesen, mint az az egész fenti okfejtésből is látszik, mindez függ attól, hogy mekkora a gyógyszer hatása, amit mi sem tudunk, ezért az egész mintanagyság-tervezés feltételes: *ha* a hatás nagysága ekkora, *akkor* ennyi beteg kell annak kimutatásához (adott erővel, adott szignifikanciaszinten). Az tehát egy fontos kérdés, hogy mit feltételezünk hatásnagyságnak,

vagy pontosabb lenne úgy fogalmazni: mekkora hatás kimutatására szeretnénk képesek lenni (például az alapján, hogy mi klinikailag lényeges – erre a kérdésre még visszatérünk).

Első ránézésre úgy tűnhet, hogy a kutatás erejének ügye a „gyógyszergyártó saját problémája": megcsinálja a kísérletet, még hat is a gyógyszere, aztán mégis azt kapja, hogy nem hatásos. (Néha szokták is a szignifikanciaszintre azt mondani, hogy a fogyasztó kockázata – hatástalan gyógyszert kapok – míg erre a hibára azt, hogy gyártó kockázata – hat a gyógyszer, és mégsem tudja forgalomba hozni.) Ez első közelítésként rendben van, de valójában ennél kicsit bonyolultabb a helyzet. Ha túl kicsi a vizsgálat ereje (általánosan elterjedt angol szóval: underpowered a kutatás), tehát túl kevés beteget vontak be, akkor a fals hatás kimutatásának a *valószínűsége* tényleg ugyanúgy 5%, mintha közepesen sok, nagyon sok, vagy akármennyi beteget vontak volna be. Igen ám, de *ha* fals hatást mutatunk ki, akkor a kimutatott fals hatás *mértéke* nagyobb lesz akkor, ha kicsi volt a vizsgálat ereje! Ez matematikailag bebizonyítható, de talán intuitíve is érzékeltethető: ha kicsi a mintanagyság, akkor az 5% tartása *épp azért* lesz lehetséges, mert csak nagyon hatásos szerekre mondjuk azt, hogy tényleg hatnak (hiszen ilyenkor nagy az ingadozás, a nem hatásos szerek is könnyebben bizonyulhatnak erősebben hatásosnak pusztán a véletlen ingadozás miatt). Így abban az 5%-ban viszont nagy lesz a kimutatott hatás. Pontosan emiatt valójában kárt okoz az orvosi megismerésnek az is, ha nagy számban készülnek nem megfelelő erejű vizsgálatok (Button és mtsai. 2013).

Érdemes végezetül megjegyezni, hogy az erő még egy szempontból fontos: a hatásosnak és a hatástalannak minősítés nem szimmetrikus. Ha a 100 kezeltből mindenki meggyógyul és a 100 kontrollból mindenki meghal, akkor van értelme azt mondani, hogy ez erős bizonyíték a hatásra, de ha pont 50 gyógyul meg itt is meg ott is, az még nem erős bizonyíték arra, hogy nem hat a gyógyszer. Ez azt jelenti, hogy nem tudtunk kimutatni hatást, de ezen állítás nem értelmezhető anélkül, hogy tudnánk, hogy mekkora hatást tudtunk volna egyáltalán kimutatni! Ha kicsi a mintanagyság, kicsi az erő, akkor simán lehet, hogy akár egy nagyobb hatást sem tudunk észrevenni. Gondoljunk arra, hogy ha mindkét csoportban a betegek fele gyógyul meg, az nagyon nem ugyanazt jelenti csoportonként 10, és csoportonként 10 ezer beteggel! Jobban szeretjük azt mondani, hogy „nem tudtuk kimutatni, hogy lenne hatás", ahelyett, hogy „kimutattuk, hogy nincs hatás". Ez utóbbi igazolása sem lehetetlen (természetesen csak olyan értelemben, hogy adott megbízhatósággal kijelenthető módon egy adott küszöbnél kisebb a hatás!), de ez messze többet igényel – kellően nagy beteglétszámú vizsgálatot – mint egyszerűen azt, hogy a vizsgálat eredménye nem szignifikáns lett. Szokták néha azt is mondani: a „bizonyíték hiánya nem a hiány bizonyítéka".

2.3. A *p*-érték fogalma

Ha csak annyit tudunk, hogy egy vizsgálat szignifikáns 5%-on, akkor még sok eset lehetséges: elképzelhető, hogy éppen csak nagyobb lett a hatása, mint az 5%-on meghúzott küszöb, de az is elképzelhető, hogy sokkal nagyobb lett. Ha megmondjuk, hogy 5%-on szignifikáns, de

1%-on már nem, akkor már többet tudunk: valahol az 5 és 1%-os küszöbök között van. Ha megmondjuk, hogy 3%-on szignifikáns, de 2%-on nem, akkor még többet tudunk.

Ezt a logikát tovább folytatva eljutunk a ***p-érték*** fogalmához: ez a legkisebb szignifikanciaszint, aminél még éppen hatásosnak minősítenénk a gyógyszert. Azaz: ha ennél nagyobb szignifikanciaszintet választunk, akkor hatásosnak minősítjük a gyógyszert, ha ennél kisebbet, akkor nem. Ha a p-érték mondjuk 3,5% (általában így szokták írni: $p = 0{,}035$), akkor 5%-ot választva szignifikanciaszintnek hatásosnak minősítjük a gyógyszert, 4%-on szintén, de 3%-on már nem.

A p-érték tehát egyrészt számszerűen mutatja, hogy „mennyire" szignifikáns a gyógyszer, másrészt lehetővé teszi, hogy tetszőleges szignifikanciaszinten döntsünk a hatásosságról.

Ha valaki úgy érzi, hogy ez így elég zavaros, az nem véletlen: valójában itt két statisztikai iskola keveredik. Az egyik iskola azt mondja, hogy két lehetőség van, hat a gyógyszer és hatástalan a gyógyszer, ebből adódóan kétféle hibázás van. Meghatározzuk a kétféle hibázás súlyát, ebből matematikailag levezethetjük, hogy mi az optimális szignifikanciaszint (minél nagyobb baj a hatástalan gyógyszer tévesen hatásosnak minősítése, értelemszerűen annál kisebb), és ez alapján döntenünk a két lehetőség között. Itt tehát két lehetőségünk van, egy „igen-nem" típusú döntést hozunk a végén, és nincsen semmilyen p-érték. A másik iskola azt mondja, hogy csak egyféle lehetőség van, hogy hatástalan a gyógyszer, mi azt nézzük, hogy ennek mennyire mondanak ellent a vizsgálati eredmények, és ezen ellentmondás mértékét mérjük a p-értékkel. Itt tehát egy lehetőség van, ebből adódóan az erő fogalma sem létezik és „igen-nem" típusú döntés helyett egy folytonos, számszerű mérőszámunk lesz, a p-érték. Mint láthatjuk, a manapság használt módszer az orvosi vizsgálatok kiértékelésére e két iskola inkonzisztens elegye: van p-érték, de mégis két lehetőséget nézünk és „igen-nem" típusú döntést hozunk. Ez az egyik fontos oka annak, hogy nagyon sok vita övezi ennek a módszernek a használatát az orvosi vizsgálatok kiértékelésére, de egyelőre megingathatatlan az uralma (Goodman 1999; Wasserstein és Lazar 2016; Greenland és mtsai. 2016).

2.4. A konfidenciaintervallum fogalma

A p-érték nagy problémája, hogy nem mond semmi szemléleteset sem a hatás nagyságáról, sem az annak becslésében rejlő, véletlen ingadozásból fakadó bizonytalanságról: a p-érték lehet ugyanannyi egy olyan esetben, ahol pontosan (nagy mintán) kimértük, hogy a gyógyszer kicsit hat, és egy olyan esetben, ahol pontatlanul (kis mintán) kimértük, hogy nagyon hat. Márpedig ez a kettő nagyon nem ugyanaz, úgyhogy kell valamilyen eszköz, ami erről is számot tud adni: ez lesz a **konfidenciaintervallum**.

A konfidenciaintervallum épp a becsült értékben lévő, véletlen ingadozásból fakadó bizonytalanságot próbálja megragadni. Mit jelent a véletlen ingadozás? Az, hogy ha a valódi érték 15%, a mintámban kaphatok 10-et vagy épp 20-at is. Ugyanez fordítva elmondva: ha 10%-ot kaptam, a valódi érték lehetett 15 (de éppenséggel 5 is). A kérdés tehát az, hogy ha 10%-ot

kaptam, akkor mi lehetett a valódi érték? Ahogy korábban is láttuk, erre nem tudunk válaszolni, de a fordított kérdésre igen: milyen értékekre igaz, hogy ha ennyi lenne a valóság, akkor kijöhetett 10% pusztán a véletlen ingadozás miatt. Melyek azok az értékek, amelyek ilyen értelemben „hihetőek", mint valóság, ha a mintában megfigyelt eredményekre támaszkodunk csupán – ezt adja meg a konfidenciaintervallum. Ha például a 10% konfidenciaintervalluma a 8%, 12% tartomány, az azt jelenti, hogy ha 8 és 12% között lenne a valódi érték, akkor még nem elhanyagolható valószínűséggel kijöhet 10% pusztán a véletlen ingadozás miatt, de ha 8-nál kisebb, vagy 12-nél nagyobb lenne a valódi érték, akkor már nagyon valószínűtlen lenne, hogy pusztán a véletlen ingadozás miatt 10-et kapjunk. Azt, hogy mit értünk „nem elhanyagolható" valószínűségen, meg „nagyon valószínűtlen" alatt az ún. megbízhatósági szint adja meg, a tipikus értéke 95%. Ami ebben benne van, azoktól az értékektől nem különbözik szignifikánsan az eredményünk, ha 5%-os szignifikanciaszintet használunk.

2.5. Statisztikai szignifikancia és klinikai relevancia

Az előbbiekben végig a statisztikai értelemben vett szignifikanciáról volt szó: arról, hogy egy vizsgálatban talált hatás van-e olyan nagy, hogy azt gondoljuk róla, hogy valódi hatás van-e a hátterében, vagy pedig el tudjuk fogadni, hogy nincs valódi hatás, és a vizsgálatban tapasztalt hatás kijöhetett pusztán a véletlen ingadozás miatt. A „szignifikáns" szónak azonban van egy teljesen más értelme is, az, hogy a hatás nagysága akkora, ami az adott tárgyterületen szakmailag lényeges: „a betegnek tegnap óta szignifikánsan csökkent a vérnyomása" (mert 180/140-ről lement 120/80-ra). Ez egy teljesen értelmes és érthető mondat, de az is világos, hogy az itt használt szignifikanciának semmi köze az előbbi értelemhez, úgyhogy megkülönböztetésül ezt nevezzük **klinikai relevanciának**. Azonban a mindennapi szóhasználatban a kétféle jelentés gyakran keveredik.

A probléma, hogy egyik jelentésből sem következik a másik: egy hatás lehet statisztikailag szignifikáns, de klinikailag nem, ám lehet klinikailag szignifikáns is úgy, hogy közben meg statisztikailag nem az. Hogy hogyan? A választ az előző pont adja meg: a dolog a hatás becslésének véletlen ingadozásból fakadó bizonytalanságán, például a mintanagyságon múlik! Lehet, hogy a gyógyszer klinikailag nagyon is szignifikánsan, 20 Hgmm-rel csökkenti a vérnyomást, de ez statisztikailag inszignifikáns (mert 4 betegen mérték ki), és lehet, hogy klinikailag teljesen inszignifikánsan, 0,2 Hgmm-rel csökkenti a vérnyomást, de ez mégis szignifikáns statisztikailag (mert 4000 betegen mérték ki).

Pontosan ezért oda kell figyelni arra, hogy ha a szignifikáns szót látjuk, akkor azt milyen értelemben használják: attól még mert valami statisztikailag szignifikáns, a hatás mérete lehet bármilyen kicsi, adott esetben minden klinikai jelentőség nélküli is.

3 A kontrollálás kérdései

3.1. A kontrollcsoport megválasztása, aktív kontroll

Az eddigi leírásban végig „kezelt” és „kontroll” csoportokat emlegettük, hallgatólagosan feltéve, hogy a kezelt csoport a vizsgált gyógyszert kapja, a kontrollcsoport pedig semmit. Ez csakugyan egy lehetséges elrendezés (noha ebben az esetben sem szerencsés, ha szó szerint semmit nem kap a kontrollcsoport – de erre a kérdésre a következő pontban részletesen kitérünk), ám van, hogy etikai okokból nem követhető ez az irány: ha a betegségre van ismerten hatásos és biztonságos gyógyszer, és annak megvonása lényeges és biztos kockázatot jelent, akkor nem engedhető meg, hogy a betegek egy része ne kapjon kezelést.

Abban az esetben, ha erre tekintettel nem a semmihez, hanem egy *másik* gyógyszerhez hasonlítják a vizsgált készítményt, **aktív kontrollról** szokás beszélni. Ez a fent vázolt etikai problémákat megoldja ugyan, de egy sor módszertani problémát teremt, kezdve azzal, hogy így valójában nem azt tudjuk meg, hogy a gyógyszer hat-e, hanem azt, hogy *legalább ugyanannyira* hat, mint a másik. Sok esetben ráadásul az nem is cél, hogy a gyógyszer jobb legyen, mint a másik, ezért igazából csak annyit tudunk és akarunk bizonyítani, hogy *ugyanannyira* hat, de ehhez meg statisztikai okokból kifolyólag általában lényegesen nagyobb mintanagyság is kell. További probléma, hogy azt, hogy a *másik* hat-e, nem tudjuk meg magából a vizsgálatból, külső feltevéseket kell elfogadnunk – különben akár még az is előfordulhat, hogy azért hat egyformán mindkét gyógyszer, mert valójában mindkettő hatástalan. Végezetül az is bonyolítja a helyzetet aktív kontroll esetén, hogy míg semmit csak egyféleképp lehet adni, addig valamit elég sokféleképp – azaz ilyenkor számítani fog az is, hogy konkrétan milyen gyógyszerhez viszonyítunk, akár még az is, hogy milyen adagolásban adva. Tehát ebben az esetben az eredmény egy újabb tényezőtől fog függeni, amit meg lehet – jó- vagy kevésbé jóhiszeműen – szerencsétlenül, vagy félrevezető módon is választani.

A fentiek miatt az aktív kontrollálás melleti döntés sokszor finom mérlegelést igényel. Megengedhető a nem aktív kontroll, ha a betegség amúgy sem feltétlenül igényel orvosi ellátást (férfias típusú kopaszodás), ha minimális kellemetlenséggel jár a kezeletlenül hagyás is (enyhe köhögés, átmeneti szédülés), de ha módszertanilag nyomós indoka van, akár nagyobb kockázat is elfogadható. Annak természetesen nem tehető ki egy beteg, hogy meghaljon, de ha a készítmény nem, vagy nem igazoltan csökkenti a halálozást, csak mondjuk a kórházban töltött időt rövidíti néhány nappal, akkor ez egy olyan kockázat, aminél teljes írásos felvilágosítás után egy felnőtt emberre rábízható, hogy eldöntse, vállalja-e annak érdekében, hogy az emberiség

orvosi tudása bővüljön. (Természetesen megfelelő korai leállítási szabályokkal arra az esetre, ha kiderülne, hogy a gyógyszer mégis csökkenti a halálozást is; erről később még szó lesz.)

Nagyon fontos ennek kapcsán hangsúlyozni, hogy a nem aktív kontroll sem azt jelenti, hogy a kontrollcsoport nem kap kezelést. A szokásos ellátást természetesen *ugyanúgy* megkapja mindkét csoport, pont emiatt ez nem okoz sem etikai, sem statisztikai problémát, mert a csoportok közti különbség továbbra is csak abban van, hogy ezen a szokásos ellátáson *felül* mit kapnak, vizsgált készítményt vagy semmit. Így, amennyiben randomizáltak, az eredmény továbbra is csak a kísérleti szernek – és a véletlen ingadozásnak – tulajdonítható.

3.2. Placebóhatás, kezeletlen és placebo kontroll, a kezelés maszkolása

Régóta ismert, hogy a kezelésben részesülés *tudata* önmagában képes hatást kiváltani, pozitív hatást, ha a kezelésben részesüléshez pozitív várakozások kapcsolódnak – ezt szokás **placebóhatásnak** nevezni (Price, Finniss, és Benedetti 2008). (Az máig vita tárgya, hogy ez mennyire érvényesül a szubjektívebb, tehát a beteg saját értékelésén alapuló, és mennyire az objektív, például műszerrel lemért végpontokon.) Ez azért fontos, mert ha úgy döntünk, hogy a kezelést a kezelés hiányához hasonlítjuk, akkor nem célszerű azt tenni, hogy a kontrollcsoportnak nem adunk semmit, hiszen ekkor azonnal megtudnák, hogy ők a kontrollcsoportba tartoznak, és hasonlóan a kezelt csoport is azonnal tudná, hogy ők a kezelt csoportba tartoznak, így a legtökéletesebb randomizálás mellett is létrejönne egy eltérés a két csoportban. Igaz, hogy „csak" a kezelésben részesülés tudatában, de a placebóhatás pont azt jelenti, hogy már ez is probléma lehet. Ehhez persze az is kell, hogy megállapodjunk abban, hogy a placebóhatást nem tekintjük a gyógyszernek betudható hatásnak (mivel az nem is a gyógyszer *specifikus* hatása, hanem magáé a gyógyszerszedésé, függetlenül attól, hogy mit szed az alany). Ezt az összehasonlítási alapot szokás **kezeletlen kontrollnak** nevezni.

A placebo hatás egyfajta „kivonását" jelenti, ha – amennyiben megoldható – a kontrollcsoportnak egy, a vizsgált gyógyszerrel mindenben azonos, tőle megkülönböztethetetlen szert kell kapnia, amely egyetlen dologban tér el a vizsgált készítménytől: abban, hogy nincs benne hatóanyag. Ezt szokás **placebónak** nevezni, és az így kontrollált vizsgálatot pedig **placebókontrollált kutatásnak** hívjuk.

Sokan azt gondolják, hogy a placebo adása pusztán a kontrollcsoport „ügye", de ez tévedés: a placebo adása *ugyanannyira* érdekes a kezelt csoport számára is, hiszen innentől *ők sem* tudhatják, hogy melyik csoportban vannak! Azaz megvalósul a kiindulási cél, hogy ti. ne legyen ebben (sem) eltérés a két csoport között.

Azt a helyzetet, ha a betegek nem tudják, hogy kezelt vagy kontrollcsoportba tartoznak, amely tehát csak placebókontrollal vagy aktív kontrollal érhető el, szokás **maszkolásnak** nevezni. Ha csak a beteg nem ismeri a csoportját, de a kezelőorvos igen, akkor **egyszeres maszkolásról** (**egyszeres vak próbáról**) szokás beszélni. Érdekes módon még ez sem tökéletes: ha az orvos

tudja a besorolást, az is baj lehet (akár teljesen szándékolatlanul, de sugallhatja a betegnek, viselkedésével, testbeszédével, hogy melyik csoportba tartozik), ezért a magasabb bizonyítóerő érdekében még jobb, ha az *orvos sem* tudja, hogy a beteg melyik csoportba tartozik. Ilyenkor beszélünk **kétszeres maszkolásról** (**kettős vak próbáról**). A nem maszkolt vizsgálatokat **nyílt elrendezésű** vizsgálatnak is szokták nevezni.

Aktív kontroll esetén kisebb problémát jelent a nyílt elrendezés, hiszen ahhoz vélhetően kisebb várakozások fűződnek, hogy két, kimondhatatlan nevű szer közül pontosan melyiket kapja a beteg, annál mindenesetre bizonyosabb kisebb, mint a kap/nem kap helyzetben. (De a „kisebb" sem feltétlenül nulla, jól ismert például, hogy az újabb, jobban reklámozott, drágább szerek jobb placebót jelentenek.)

Vannak esetek, amikor a kontrollálás technikai okokból nehezebben kivitelezhető, például ha nem egy gyógyszerről, hanem egy műtétről van szó, amit placebóval szeretnénk kontrollálni, vagy, ha aktív kontrollt választunk egy gyógyszernél, de a vizsgált és a kontroll gyógyszer formája eltérő, például az egyik inhaláló szer, a másik filmtabletta. Sokszor azonban még ezek a problémák is feloldhatóak, például megfelelő etikai feltételek teljesülése esetén „álműtét" is végezhető (Wartolowska és mtsai. 2014), amikor a beteg ugyanolyan műtét előtti és utáni ellátáson keresztül megy keresztül, mint egy igazi műtétnél, sőt, akár még a műtéti bemetszést és összevarrást is megkapja, csak épp a kettő között nem csinálnak semmit. Ha pedig a két gyógyszerforma eltérő, akkor megtehetjük, hogy mindenki inhaláló szert és filmtablettát *is* kap, de az egyik csoportban a betegek placebót inhalálnak és a tablettában lesz az aktív szer, a másikban fordítva (double dummy elrendezés).

3.3. Párhuzamos és keresztezett kontroll

Az eddigi leírásból úgy tűnhet, hogy a klinikai kísérlet úgy néz ki, hogy begyűjtjük a betegeket, véletlenszerűen két csoportba soroljuk őket, alkalmazzuk a gyógyszert, majd utánkövetjük az alanyokat és lemérjük a végpontot. A legtöbb esetben tényleg ez történik, azzal az apró kiegészítéssel, hogy valójában a randomizálás – és az azt követő kezelés – nem egyetlen időpontban történik az összes betegre, hanem folyamatosan, ahogy a betegek érkeznek a vizsgálatba; még az sem ritka, hogy valamelyik beteg teljesen befejezi az egész vizsgálatot, amikor más még csak épp randomizálásra kerül (Senn 2013). Ezzel együtt is, az ilyen kutatásokat **párhuzamosan kontrollált** vizsgálatnak szokás nevezni, utalva arra, hogy az előbbi technikai kérdéstől eltekintve valóban párhuzamosan fut a kezelt és kontroll csoport kezelése.

Bizonyos esetekben azonban van egy alternatív lehetőség: az, hogy *ugyanazok* a betegek alkotják a kezelt és kontrollcsoportot, akik így nem egymással párhuzamosan, hanem egymás *után* kerülnek kezelésre. Magyarán: a bevont alanyokat véletlenszerűen kettéosztjuk, egyik részük előbb kezelést kap, utána placebót vagy a másik gyógyszert, a másik része pedig pont fordítva. Az ember első ránézésre azt is gondolhatja, hogy ezzel igazából megoldottuk a felvezetőben említett problémáját az időgép hiányának, hiszen így mégis *ugyanazon* emberen mértük le a kezelés és a kezeletlenség hatását is. Valójában ez sem teljesen igaz, hiszen elképzelhető, hogy

nem mindegy, hogy melyik időszakban kapja az alany a kezelést (ha valamiért a kezelés hatása változik időben), az sem feltétlenül mindegy, hogy úgy kapja a kontrollt, hogy előtte kapott kezelést (lehet, hogy a kezelésnek van időben elnyúló hatása), de a legnagyobb probléma, hogy ez a fajta elrendezés sokszor fizikailag megvalósíthatatlan: nem lehet ilyet csinálni akkor, ha a beteg teljesen meggyógyul a kezeléstől, hiszen utána már hiába adunk neki mást. Akkor sem valósítható meg ez az elrendezés, ha a beteg állapota súlyosbodik az első kezelés alatt vagy után (szélsőséges esetben ha meghal, hiszen akkor megint csak hiába akarnánk később mást adni). Valójában tehát krónikus, stabil betegségeknél van mód rá, ahol a kezelés „csak" az állapoton javít, és ez az esetleges javulás nem túl hosszú kezelés után előáll.

Ha meg lehet valósítani ezt az elrendezést, amit **keresztezett vizsgálatnak** szokás nevezni, akkor az viszont sok előnnyel bír, abból fakadóan, hogy minden alany lényegében a saját maga kontrollja, így egy sor egyénre jellemző tulajdonság (életkor, nem, milliónyi genetikai tényező) *automatikusan* ki van egyensúlyozva. Az ilyen vizsgálatok ezért általában hatásosabbak, azaz kisebb a véletlen ingadozás bennük (ugyanazzal a mintanagysággal pontosabb becslés lehetséges, avagy ugyanolyan pontossághoz kisebb mintanagyság is elég), és lehetővé teszik a betegek közti – nem csak a kezelések közti – ingadozás becslését is (Senn 2003).

4 A végpontok megválasztásának és mérésének kérdései

4.1. A végpont fogalma, elsődleges és másodlagos végpontok

A **végpont (hatásmutató)** a vizsgálat klinikailag érdekes kimenetének lemérése, annak számszerűsítése, hogy mi történt a beteggel akár a gyógyszer hatásosságát, akár a biztonságát tekintve. Végpont lehet, hogy a beteg meghalt-e, hogy mennyi lett a vérnyomása, hogy mennyi idő után épült fel, felszívódott-e egyáltalán a gyógyszer, vagy az, hogy egy bizonyos mellékhatás fellépett-e nála. Az külön kérdéskör, hogy ezekből az eredményekből hogyan lehet egyetlen számba sűrítve kifejezni a gyógyszer hatását, de ezzel most nem foglalkozunk.

Látható, hogy végpontból több is lehet, de ezeket általában rangsorolni szokták: elsődleges végpontnak nevezzük azt, ami a legfontosabb kimenet a vizsgálat kérdése szempontjából, és ami alapján a kísérletbe bevont betegek számát tervezik. Hiszen láttuk, hogy egy bizonyos méretű hatás – adott erővel és adott szignifikanciaszinten történő – kimutatásához szükséges beteglétszám előre meghatározható statisztikai módszerekkel. Na de melyik hatásról beszélünk? Arról, hogy a gyógyszer hogyan változtatja a halálozások számát? Arról, hogy hogyan változtatja az infarktus előfordulását? Ezek nagyon különböző beteglétszámokat tehetnek szükségessé, hiszen halálozásból tipikusan kevesebb van, mint infarktusból, így annak csökkentése kisebb méretű hatás, ezért a kimutatásához több betegre lesz szükség. A vizsgálatban mindig ki kell választani az[1] elsődleges végpontot, arra meghatározni a kimutatni tervezett hatást, és ez alapján tervezni mintanagyságot. Az összes többi, ún. másodlagos végpontra nincs méretezve a kutatás ereje.

4.2. Kemény és helyettesítő végpontok

Az előbbi gondolatnak tulajdonképpen még messzebbre vezető következményei is vannak. Gondoljunk bele a következőbe: miért adunk a betegeknek vérnyomáscsökkentő gyógyszert? Azért, hogy ritkábban kapjanak infarktust, stroke-ot, vagy még általánosabban fogalmazva: hogy tovább legyen jó az egészséggel összefüggő életminőségük, tovább legyenek munkaképesek, később haljanak meg. Ezért ha egy új vérnyomáscsökkentő gyógyszerjelöltet vizsgálunk, akkor *elvileg*

[1] Előfordulhat, hogy egynél több elsődleges végpont van, ebben az esetben azonban a többszörös összehasonlítások pontban leírt kérdésekre oda kell figyelni a vizsgálat tervezése és kiértékelése során.

ezeket kell végpontnak megtenni: infarktus előfordulása, stroke előfordulása, halálozás. Az ilyen, beteg számára igazán releváns végpontot szokták **kemény végpontnak** nevezni, és ha mód van rá, ilyet célszerű mérni. Igen ám, de a vérnyomáscsökkentős példánkban belefutunk egy problémába: ha csak nem nagyon rossz állapotú alanyokról van szó, akkor ezek az események elég ritkák lesznek, illetve sokat kell várni a kialakulásukra. Ami persze általában jó hír, de most, kutatástervezési szempontból nem az: a ritkaság statisztikai szempontból azt jelenti, hogy rengeteg alanyra, a lassan kialakuló jelleg pedig azt, hogy nagyon hosszú utánkövetési időre lesz szükség, és ezek egy ponton túl kivitelezhetetlenek.

Mit csináljunk tehát a vérnyomáscsökkentő esetében? Azt fogjuk lemérni, hogy a vérnyomást csökkenti-e! Hiszen ez mindenkinél lemérhető, és rövid időn belül lemérhető (ráadásul könnyen, veszélytelenül és olcsón). De bármennyire is kézenfekvőnek tűnik ez, vegyük észre, hogy itt van egy háttérfeltételezés: az, hogy a vérnyomás csökkentése *tényleg* maga után vonja a kemény végpontok, például az infarktus vagy a halálozás csökkenését is! Ezt muszáj biztosan tudnunk, különben az egész vizsgálattal nem megyünk semmire. Az ilyen végpontokat szokás **helyettesítő végpontnak** nevezni. A helyettesítő végpontok sok esetben hasznosak, sőt, néha ezek teszik egyáltalán lehetővé kutatások elvégzését, de másrészről a használatuk „veszélyes üzem”, mert az előbbi feltétel teljesülését alaposan meg kell vizsgálni. Ezzel kapcsolatban itt csak arra hívjuk fel a figyelmet, hogy *önmagában* az, hogy valami együttmozog a végponttal (a magasabb vérnyomású betegek körében gyakoribb az infarktus) még *nem* bizonyítja, hogy jó helyettesítő végpont! Az, hogy a helyettesítő végpont javítása maga után vonja a kemény végpont javulását is, egy sokkal erősebb feltétel, mint az, hogy a kettő együtt jár; ha erről megfeledkeznénk, akkor egy fogfehérítő szerről kimutathatnánk, hogy hatásos a tüdőrák ellen (gondoljuk végig!).

5 Néhány további témakör a gyógyszervizsgálatok témájából

5.1. Több végpont használata, többszörös összehasonlítások helyzete

A statisztikai szignifikanciánál láttuk, hogy a szokásos 5%-os szignifikanciaszint használata azt jelenti, hogy ha nincsen hatás, mi 5% valószínűséggel fogunk mégis, a véletlen ingadozás miatti pech miatt, tévesen hatást kimutatni. (És láttuk, hogy ezt a szintet csökkenteni sem feltétlenül érdemes, hiszen akkor meg a valódi hatások detektálását is megnehezítjük.) Ez teljesen rendben is van akkor, ha *egyféle* hatást vizsgálunk meg. De mi történik akkor, ha több összehasonlítást végzünk? Ennek talán legkézenfekvőbb példája, amikor több végpontot vizsgálunk, például azt is megnézzük, hogy a gyógyszer az infarktus előfordulását csökkenti-e, és azt is, hogy a stroke előfordulását csökkenti-e.

Külön-külön vizsgálódva nincsen semmi különös: ha a gyógyszer egyikre sem hat, akkor 5% valószínűséggel fogunk mégis hatást kimutatni mindkettő esetében. Ha azonban azt kérdezzük, hogy mekkora valószínűséggel fogunk *legalább az egyiknél* tévesen hatást kimutatni, az máris rögtön nagyobb lesz, mint 5%! Gondoljunk bele, a dolog úgy is felfogható, hogy van egy cinkelt pénzérménk, ami 5% valószínűséggel mutat fejet feldobásnál. Ekkor annak a valószínűsége, hogy egy dobásból fejet kapunk 5%, de annak a valószínűsége, hogy két dobásból legalább egy fejet kapunk, nyilván jóval nagyobb. Az meg pláne, hogy 10-ből dobunk legalább egy fejet. Ha tehát „vagylagosan" vizsgálódunk, azaz akkor mondjuk, hogy találtunk valamit, ha *bármelyik* összehasonlításnál (pl. bármelyik végponton) találunk hatást, akkor óvatosan kell eljárni.

Ez egy példa a **többszörös összehasonlítások helyzetének** nevezett általános jelenségre, ami akkor lép fel, ha a fenti értelemben „vagylagosan" végzünk több vizsgálatot (Schulz és Grimes 2005). Az érzékeltetés kedvéért, a következő ábra mutatja pontosan kiszámolva[1] annak valószínűségét, hogy valahol (legalább egy összehasonlításban) találunk hatást, ha valójában sehol nincs hatás, annak függvényében, hogy hány összehasonlítást végeztünk:

[1] Egyedül annyit kellett a számításhoz feltételeznünk, hogy az egyes összehasonlítások egymástól függetlenek.

A ábrán piros vonallal jelölt 5% a szignifikanciaszint: *egy* összehasonlításnál tényleg rendben van minden, és ez lesz a fals találat valószínűsége, de látható, hogy az összehasonlítások számának emelkedésével – ahogy éreztük is – gyorsan nő ez a valószínűség.

A több végponton kívül van még pár eset, ami valójában ugyanerre a problémára vezet, például ha több betegcsoportban is megvizsgáljuk a gyógyszer hatását vagy ha több időpontban is megvizsgáljuk a gyógyszer hatását (ezekkel fogunk foglalkozni a következő pontokban), de az alapprobléma mindenhol ugyanaz: ha több összehasonlítást végzünk, megnő annak a valószínűsége, hogy tévesen hatást mutatunk ki. Ilyenkor szükség lehet arra, hogy megfelelő statisztikai korrekciókat tegyünk. Ezeknek számos módszere van, pusztán az érzékeltetés kedvéért, az egyik ilyen lehetőség, hogy az egyes vizsgálatok szignifikanciaszintjét csökkentjük. Hiszen a jelenség az volt, hogy a „fals találat" valószínűsége megnő ahhoz képest, mint az egyes összehasonlítások szignifikanciaszintje – akkor csökkentsük le ez utóbbit, hogy még a megnövekedés után is annyi legyen, ami számunkra elfogadható, például 5%!

5.2. Alcsoport-analízis

Bizonyos esetekben felmerül a lehetőség, hogy egy gyógyszer hatása valamilyen meghatározott csoportban, például adott nemben, életkorban, genetikai szempontból, valamilyen társbetegség mellett, a betegség súlyossága, vagy épp a terápia megkezdésének időpontja szerint eltér. Hogyan lehet ezt megvizsgálni?

Kézenfekvő lehetőség, hogy szűkítsük le az adatainkat a vizsgált csoportra, például vegyük ki csak a férfiakat a mintából, és számoljuk ki az eredményt az ő körükben. A gyakorlatban általában úgy járnak el, hogy ilyen esetben mindegyik csoportra – a példánál maradva: férfiakra és nőkre is – kiszámolják az eredményeket, és egymás mellett közlik. Természetesen több szempont szerinti csoportosítást is vizsgálhatunk. Ezt az eljárást szokás **alcsoport-analízisnek** hívni.

A módszer nagyon csábító, hiszen rendkívül egyszerű, és mégis megválaszolhatja sok kérdésünket arról, hogy egy bizonyos csoportban külön hogyan hat a gyógyszer – valójában azonban számos csapdája van (Sleight 2000; Schulz és Grimes 2005; Assmann és mtsai. 2000; Yusuf és mtsai. 1991).

Az még csak a kisebbik baj, hogy ha leszűkítjük a vizsgált betegek körét, akkor kisebb lesz a beteglétszám, így a véletlen ingadozás megnő, a becslések bizonytalanabbak lesznek, azaz kisebb lesz a statisztikai erő. Ennek egyik következménye, hogy fals módon mondhatjuk azt, hogy adott alcsoportban nincs hatás, holott akár még az is lehet, hogy a hatás pontosan ugyanakkora abban az alcsoportban is, mint az egész vizsgálatban, csak épp a kevesebb beteg miatt erről nem tudjuk igazolni, hogy szignifikáns. A nagyobb probléma, hogy alcsoportból nagyon sok tud lenni, és ez végképp igaz, ha kombináljuk is őket (pl. cukorbeteg, 60-70 éves férfiak alcsoportja). Azaz: az imént tárgyalt többszörös összehasonlítási helyzettel állunk szemben! Az összehasonlítások száma most az alcsoportok száma, és a következtetés ugyanaz: nagyon gyorsan megnő annak a valószínűsége, hogy *legalább egy* alcsoportban akkor is találunk hatást, ha sehol nincsen! Ez a két probléma ráadásul némiképp még fel is erősíti egymást, hiszen ahogy korábban láttuk, a túl kis erő pedig ahhoz fog vezetni, hogy ha téves hatást találunk, annak a mérete ráadásul még nagy is lesz.

Mindezeken némileg segít, ha limitált az alcsoportok száma, az pedig különösen fontos, hogy *előre* eldöntött legyen, hogy milyen alcsoportok kerülnek elemzésre, mert a legveszélyesebb az, ha ezt az adatok alapján, az eredményeket nézegetve találják ki a kutatók (így nagyon nagy lesz a valószínűsége, hogy olyan alcsoportot találnak, ami csak abban a konkrét mintában mutat eltérést). Tökéletes megoldást azonban ezek sem jelentenek.

Akkor mire jók mégis az alcsoport-analízisek? Mivel elképzelhető, hogy a gyógyszer hatása *tényleg* eltér bizonyos alcsoportokban, és ezt fontos is lenne felderíteni, így nem lehet egyszerűen figyelmen kívül hagyni ezt a témakört. Fontos azonban, hogy az alcsoport-analízisek eredményei mindig óvatosan kezelendőek, és jellegük szerint inkább csak *felvetik* a lehetőségét, hogy egy alcsoport máshogy viselkedik, de önmagukban nem bizonyítják. Ezért a valóban releváns feltevéseket később megerősítő jellegű vizsgálatokban célirányosan igazolni kell.

5.3. Interim elemzések, korai leállítás

A gyógyszervizsgálatokról eddig elmondottakból úgy tűnhet, hogy ha egy maszkolt vizsgálatról van szó, akkor egészen a vizsgálat lezárásáig senki nem tudhatja, hogy melyik beteg tartozott

a kezelt, és melyik a kontrollcsoportba. Ez elméletben tényleg így is van, de gyakorlatilag nem kivitelezhető, és nem csak azért, mert valakinek be kell csomagolnia a szereket kívülről nézve azonos csomagolásba. Bele kell nézni menet közben is az adatokba, minimum a biztonság miatt (ha valami bajt okoz a szer, azt muszáj hamar észrevenni), de akár a hatásosság miatt is: ha találunk egy nagyon hatásos készítményt, akkor sem szerencsés, ha hónapok, adott esetben évek múlva derül csak ki.

A megoldást az jelenti, hogy létezik egy, a kutatást végzőktől természetesen szigorúan független – hogy a maszkolás fennmaradjon – bizottság, amelynek tagjai belenézhetnek menet közben is az eredményekbe. Értékelhetik azokat, és ez alapján szükség esetén a független bizottság határozhat a vizsgálat **korai leállításáról**, de az egyéb módosításokról is. Az ilyen, menet közben végzett kiértékelést szokták **interim analízisnek** nevezni.

A fenti elmondásból az egész nagyon sima ügynek tűnhet, pedig valójában itt is vannak csapdák. A legfontosabb probléma, hogy minden egyes „belekukucskálás" az eredményekbe egy lehetőség, hogy szignifikáns hatást kapjunk akkor is, ha nincs hatás! Minél többször tesszük meg, annál nagyobb a valószínűsége, hogy fals módon hatást mutatunk ki – ez tehát pontosan ugyanaz a probléma, amit a többszörös összehasonlítások esetén láttunk; az összehasonlítások száma itt a menet közbeni kiértékelések száma.

Mivel azonban fontos, hogy mégis tudjunk interim analízist végezni, ezért számos módszert találtak ki, amit alkalmazni lehet, hogy korrekt eredményt kapjunk[2].

Amíg a biztonság miatti leállítás szerepét senki nem kérdőjelezi meg, addig a hatásosság miatti leállítás gyakran vált ki vitákat a szakirodalomban. A módszertani problémákon túl egy fontos etikai megfontolás is szerepet játszik ebben: bár „matematikailag" mindegy, hogy valaki azért hal meg, mert súlyos mellékhatást szenved el egy nem jó gyógyszertől, vagy pedig azért, mert nem jut hozzá időben egy jó gyógyszerhez, az orvosok számára ez a kettő nem ugyanaz – sokkal nagyobb súllyal esik latba az a kár, amit az aktív tevékenységükkel okoztak. Ezt fejezi ki igazából a „nil nocere" – „ne árts" – elve (nem azt, hogy semmit nem szabad tenni a beteggel, ami elvileg akár árthat is neki, mert akkor semmilyen orvosi beavatkozást nem lehetne soha senkin végezni).

5.4. A gyógyszervizsgálatok fázisai

Itt nem célunk a gyógyszervizsgálatok szervezésével kapcsolatos milliónyi – és a gyakorlatban fontos – technikai kérdésre kitérni, de egy dolgot megemlítünk, mert ez gyakran előkerül a közbeszédben is.

[2]Ezek helyenként egészen komplexek, de az alapgondolat általában az, amit a többszörös összehasonlításoknál is láttunk: belenézhetünk többször is, de akkor szigorúbbnak kell lennünk, hogy a végén összességében is megőrizzük az 5%-os fals hatás-kimutatás valószínűséget.

A gyógyszerfejlesztés folyamatának azon első lépéseit, amikor a gyógyszert még nem adják embernek, **preklinikai fázisnak** szokás nevezni. Ide tartoznak a számítógépes modellezések, a laboratóriumban sejtkultúrákon és szöveteken végzett vizsgálatok, az állatkísérletek stb. Ha egy gyógyszerjelölt e fázison átmegy és ígéretesnek tűnik, akkor jön az emberi kipróbálás szakasza, amit **klinikai fázisnak** szoktak nevezni. A preklinikai fázisban ugyan megpróbálják kizárni a kockázatokat (és maximalizálni annak a valószínűségét, hogy hatásos legyen a gyógyszer), de ez igazán biztosan csak akkor dől el, ha ténylegesen embernek adják a vizsgált készítményt. Ebből adódóan a klinikai fázis a gyógyszerfejlesztés legkritikusabb, és ezzel összhangban a legjobban szabályozott területe. Számos szomorú történelmi példa után ma már klinikai vizsgálatban kizárólag önkéntesek vehetnek részt.

A klinikai szakaszt általában három fázisra szokták bontani. Az **1-es fázisú** vizsgálatok során néhány kivételtől eltekintve a gyógyszert kislétszámú (jellemzően néhány tucatnyi) és általában teljesen egészséges alanynak adják. Ebből is rögtön látható, hogy ekkor még nem kell gyógyítania a gyógyszernek! Akkor mi szükség van erre a fázisra? A válasz az, hogy itt ismerik meg a kutatók azt, hogy a gyógyszernek mi a sorsa az emberi szervezeten belül (hogyan szívódik fel, hogy oszlik szét a szervezetben, hogyan alakul át, hogyan távozik), hogy a szervezetben hogyan fejt ki hatást, ez milyen mutatókkal írható le, hogy ezek összefüggenek-e a szervezeten belüli viselkedés mutatóival. Általában az 1-es fázis során különböző dózisokat adnak be, hogy megtalálják, hogy melyik lehet az optimális. Bár hatásosságot itt még nem néznek, de biztonságot nagyon is: nyomon követik a mellékhatásokat, ez is segít annak kiválasztásában, hogy mely dózisok jöhetnek szóba a későbbi vizsgálatokhoz.

A **2-es fázis** kritikus lépés: itt kell először gyógyítania is a gyógyszernek. A 2-es fázisú vizsgálatokban a gyógyszerjelöltet már betegeknek adják, de egyébként jó állapotban lévő és nem túl nagy létszámú (jellemzően 100 körüli vagy alatti), rendkívül megválogatott és nagyon szorosan monitorozott csoportnak. Itt is rögzítik ugyanúgy a biztonsági végpontokat, de itt először vizsgálják a hatásosságot is. A cél még nem ennek a pontos felmérése, csak annak eldöntése, hogy egyáltalán a remény megvan-e arra, hogy a gyógyszer működjön, azaz van-e megfelelő dózis, aminél a kockázat-haszon mérleg kellő betegcsoport számára pozitív lehet. Szükség esetén tovább optimalizálják, és itt véglegesítik az adagolást is.

Ha igen, akkor a gyógyszerjelölt továbblép az utolsó, legnagyobb volumenű, és legkomplexebb szakaszba. A **3-as fázis** során a készítményt már nagy létszámú (akár több ezer fős) betegcsoportnak adják, nagyon gyakran számos vizsgálóhelyen, a betegeket jóval kevésbé megválogatva, ezzel is közelítve a való életbeli helyzethez. Amellett, hogy ugyanúgy nyomon követik itt is a biztonsági végpontokat, ez lesz a perdöntő információforrás a hatásosság megítéléséhez, a korábbi feltételezések megerősítéséhez. A gyógyszerjelölt akkor kaphat törzskönyvet, ha ezen a fázison is sikeresen átjut.

Manapság a 3-as fázisú vizsgálatok nagyon kevés, jól megindokolt kivételtől eltekintve mind randomizált, kontrollált kutatások.

Szoktak néha még **4-es fázisú** vizsgálatokról is beszélni. Ez a kifejezés némileg kakukktojás, mert ez már nem a gyógyszerfejlesztési folyamat része, hanem a törzskönyvezés utáni fázisé.

A kutatómunka ugyanis nem ér véget a törzskönyvezéssel és a forgalomba hozatallal: bármilyen nagy létszámú is volt a fázis 3 vizsgálat, gyakran előfordul, hogy vannak kérdések, amik eldöntéséhez még ez is kevés. Ahogy láttuk, a kis hatások megítéléséhez – legyen szó akár kis főhatásról, akár ritka mellékhatásról – nagy mintanagyság kell; nagyon kis hatáshoz akkora mintanagyság is szükséges lehet, ami a fázis 3 vizsgálatban sem érhető el. Ezeket csak akkor lehet tanulmányozni, amikor a gyógyszer már széleskörű alkalmazásba került; van rá példa, hogy ilyenkor már több százezer beteg bevonásával készítenek kutatásokat. Az ilyen vizsgálatok további célja, hogy információt gyűjtsön a való életbeli hatásról, hiszen még a 3-as fázis betegei is általában sok szempont (életkor, társbetegségek stb.) szerint válogatott körből származnak, a betegminta összetétele eltér az összes betegétől. Cserében a dolog kevésbé lehet szigorúan szabályozott, mint a 3-as fázisban; itt már nagyon sokszor kerül sor megfigyeléses vizsgálatok végzésére is.

5.5. A bizonyítékok összessége szemlélet

Az orvosi kérdések eldöntésekor – és ez alól természetesen a gyógyszerek hatásosságának a megítélése sem kivétel – soha nem az a kérdés, hogy *egy konkrét* vizsgálat mit mutat, hanem az, hogy mi a teljes tudásunk az adott témakörben. Ez adott esetben állhat egyetlen vizsgálatból, állhat többből is, ez utóbbi esetben a különböző kutatások bizonyító ereje eltérő is lehet.

Éppen ezért fontos, hogy egy kérdés megválaszolását lehetőleg a teljes, rendelkezésre álló tudásunkra alapozzuk. Ez a gyakorlatban két feladat megoldását igényli: a tudást tartalmazó közlemények valamilyen módon történő összegyűjtését (**review**), lehetőleg reprodukálható módon (**szisztematikus review**), és a közleményekből kiolvasható eredmények megfelelő statisztikai módszerrel történő aggregálását (**meta-analízis**). A statisztikai módszerekre azért van szükség, mert itt nem lehet egyszerűen átlagolni: minden részlet nélkül, ha egy 20 fős vizsgálat azt találta, hogy a halálozási arány 10%, egy 1000 fős pedig azt, hogy 20%, akkor érezhető, hogy nem lehet azt mondani, hogy együtt azt mondják, hogy a válasz 15%. Valahogy tehát a kutatások bizonytalanságát is figyelembe kell venni, erre szolgálnak a meta-analízis módszerei.

Fontos hangsúlyozni, hogy a meta-analízisre csak akkor van szükség, ha nem ismerjük a vizsgálatokban résztvevő betegek adatait személy szerint, egyénileg. Ha igen, akkor csak „össze kell önteni őket” (persze pár dologra még ekkor is figyelni kell, de ez már részletkérdés), de a gyakorlatban elég tipikus helyzet, hogy az egyéni eredményeket nem, csak az összesítést ismerjük. Például nem tudjuk az egyedi vérnyomás értékeket, csak az átlagukat és a szórásukat. Ez a tipikus helyzet akkor, ha a nyers adatokhoz nem férünk hozzá, csak a kutatásról készült publikációt tudjuk elolvasni. Ilyenkor van szükség a meta-analízis eszköztárára.

Végezetül nagyon fontos rögzíteni, hogy a meta-analízis sem jelent csodafegyvert: eleve limitálja, hogy csak aggregált adatok alapján tud dolgozni, és ha a bemenő adatok mind rossz minőségűek, akkor abból nem lesz varázslatosan jó minőségű eredmény semmilyen módszerrel sem.

Hivatkozások

Assmann, Susan F, Stuart J Pocock, Laura E Enos, és Linda E Kasten. 2000. „Subgroup analysis and other (mis)uses of baseline data in clinical trials". *The Lancet* 355 (9209): 1064–69. https://doi.org/10.1016/S0140-6736(00)02039-0.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, és Marcus R Munafò. 2013. „Power failure: why small sample size undermines the reliability of neuroscience". *Nature Reviews Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Dawid, A. P. 2000. „Causal Inference without Counterfactuals". *Journal of the American Statistical Association* 95 (450): 407–24. https://doi.org/10.1080/01621459.2000.10474210.

Gachályi, Béla, Géza Lakner, és János Borvendég. 2003. *Klinikai farmakológia a gyakorlatban*. Springer Tudományos Kiadó.

Goodman, Steven N. 1999. „Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy". *Annals of Internal Medicine* 130 (12): 995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008.

Greenland, Sander, és James M Robins. 1986. „Identifiability, Exchangeability, and Epidemiological Confounding". *International Journal of Epidemiology* 15 (3): 413–19. https://doi.org/10.1093/ije/15.3.413.

Greenland, Sander, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, és Douglas G Altman. 2016. „Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". *European journal of epidemiology* 31 (4): 337–50. https://doi.org/10.1007/s10654-016-0149-3.

Lakner, Géza, Gábor Renczes, és János Antal. 2009. *Klinikai vizsgálatok kézikönyve*. SpringMed Kiadó.

Maldonado, George. 2013. „Toward a clearer understanding of causal concepts in epidemiology". *Annals of Epidemiology* 23 (12): 743–49. https://doi.org/10.1016/j.annepidem.2013.09.001.

———. 2016. „The role of counterfactual theory in causal reasoning". *Annals of Epidemiology* 26 (10): 681–82. https://doi.org/10.1016/j.annepidem.2016.08.017.

Maldonado, George, és Sander Greenland. 2002. „Estimating causal effects". *International Journal of Epidemiology* 31 (2): 422–29. https://doi.org/10.1093/ije/31.2.422.

Price, Donald D., Damien G. Finniss, és Fabrizio Benedetti. 2008. „A Comprehensive Review of the Placebo Effect: Recent Advances and Current Thought". *Annual Review of Psychology* 59 (1): 565–90. https://doi.org/10.1146/annurev.psych.59.113006.095941.

Schulz, Kenneth F, és David A Grimes. 2005. „Multiplicity in randomised trials I: endpoints and treatments". *The Lancet* 365 (9470): 1591–95. https://doi.org/10.1016/S0140-

6736(05)66461-6.

Senn, Stephen. 2003. *Cross-over Trials in Clinical Research*. Wiley.

———. 2013. „Seven myths of randomisation in clinical trials". *Statistics in Medicine* 32 (9): 1439–50. https://doi.org/10.1002/sim.5713.

Sleight, Peter. 2000. „Debate: Subgroup analyses in clinical trials: fun to look at-but don't believe them!" *Trials* 1 (1): 1–3. https://doi.org/10.1186/cvm-1-1-025.

Wartolowska, Karolina, Andrew Judge, Sally Hopewell, Gary S Collins, Benjamin J F Dean, Ines Rombach, David Brindley, Julian Savulescu, David J Beard, és Andrew J Carr. 2014. „Use of placebo controls in the evaluation of surgery: systematic review". *British Medical Journal* 348. https://doi.org/10.1136/bmj.g3253.

Wasserstein, Ronald L., és Nicole A. Lazar. 2016. „The ASA Statement on p-Values: Context, Process, and Purpose". *The American Statistician* 70 (2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Yusuf, Salim, Janet Wittes, Jeffrey Probstfield, és Herman A. Tyroler. 1991. „Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials". *JAMA* 266 (1): 93–98. https://doi.org/10.1001/jama.1991.03470010097038.