doi: 10.21860/medflum2017_187369

Slučajne pogreške u istraživanjima i važnost veličine uzorka

Random errors in research and the importance of sample size

Goran Poropat*, Sandra Milić

Zavod za gastroenterologiju, Klinika za internu medicinu, Klinički bolnički centar Rijeka, Rijeka

Sažetak. Slučajne pogreške sastavni su dio svakog eksperimenta, odnosno kliničkog istraživanja, i posljedica su varijabilnosti uzorkovanja. Rezultati istraživanja trebaju biti primjenjivi na čitavu populaciju, zbog čega su reprezentativnost i adekvatna veličina ispitivanog uzorka izuzetno bitne. Provođenjem određenog mjerenja na uzorku populacije nastojimo utvrditi s kojom vjerojatnošću je opažena razlika između mjerenih varijabli posljedica slučajnosti, odnosno postojanja stvarne razlike u populaciji. Navedena vjerojatnost izražava se P-vrijednošću. Slučajnost se očituje u istraživanjima kao pogreška tipa I i pogreška tipa II. Rizik slučajne pogreške u istraživanjima umanjuje se povećanjem broja mjerenja, odnosno veličine istraživanog uzorka. Zbog toga je u fazi planiranja istraživanja nužno izračunati i definirati potreban broj ispitanika na temelju prethodno definiranog primarnog ishoda istraživanja. Time se omogućuje provođenje vjerodostojnih i metodološki čvrstih istraživanja koja rezultiraju solidnim dokazima.

Ključne riječi: P-vrijednost; pogreške tipa I i II; slučajne pogreške, snaga; veličina uzorka

Abstract. random errors are essential parts of every experiment and clinical research resulting from sampling variability. Research results should be applicable to the entire population, which is why the representativeness and adequate sample size are of crucial importance. By conducting a specific measurement on a population sample, we try to determine the probability of the observed difference between measured variables being the result of chance or acutal difference in the population. This probability is known as the P-value. The chance in research manifests itself as an error type I and error type II. The risk of random error diminishes by increasing the number of measurements or enlarging the sample size. Therefore, it is necessary to calculate and define the required sample size based on a priorly defined primary outcome. This enables the implementation of credible and methodologically strong research resulting in solid evidence.

Key words: error type I and II; P-value; random errors; sample size

***Dopisni autor:**
doc. dr. sc. Goran Poropat, dr. med.
Zavod za gastroenterologiju,
Klinika za internu medicinu
Klinički bolnički centar Rijeka
Krešimirova 42, 51 000 Rijeka, Hrvatska
e-mail: gporopat8@gmail.com

http://hrcak.srce.hr/medicina

UVOD

Istraživanja u biomedicini u osnovi predstavljaju pokuse kojima se nastoji utvrditi učinkovitosti, odnosno eventualne štetnosti određenih dijagnostičkih ili terapijskih postupaka u kontekstu specifične indikacije. Navedeni pokusi provode se u umjetno stvorenim uvjetima, odnosno, neovisno o tome je li riječ o laboratorijskim ili kliničkim istraživanjima, navedeni uvjeti definirani su dizajnom studije, tj. definiranim uključnim i isključnim kriterijima, protokolom istraživanja, primarnim i sekundarnim ishodima od interesa itd. Upravo zbog takve prirode istraživanja ne predstavljaju stvarni život, jer navedeni uvjeti i okolnosti u stvarnom životu ne mogu biti jasno definirani. No, istraživanja bi trebala u najvećoj mogućoj mjeri simulirati stvarni život te je zbog toga vrlo važno definirati uzorak ispitanika na kojem se provodi istraživanje, koji bi bio reprezentativan za opću populaciju. Osim samog dizajna istraživanja i protokola provođenja eksperimenata, reprezentativnost uzorka u najvećoj mjeri ovisi o njegovoj veličini, odnosno broju uključenih ispitanika. Važnost veličine uzorka u današnjoj biomedicinskoj znanosti potvrđena je i sve većim brojem časopisa koji zahtijevaju da randomizirana klinička istraživanja (engl. *Randomised Clinical Trials;* RCT) jasno istaknu metodologiju izračuna i određivanja veličine uzorka u skladu s CONSORT izvješćem (engl. *Consolidated Standards of Reporting Trials*)[1].

P-vrijednost predstavlja mogućnost da je dobiveni rezultat, odnosno razlika između mjerenih varijabli posljedica slučajnosti.

TESTIRANJE HIPOTEZE

Klinička istraživanja, posebno ona prospektivna, randomizirana i placebo-kontrolirana započinju postavljanjem jedne ili više hipoteza koje je istraživanjem potrebno testirati. Hipoteze predstavljaju pretpostavke o međusobnoj povezanosti ili odnosu dviju ili više varijabli[1]. Uobičajeno je navedena pretpostavka postavljena tako da između promatranih ili mjerenih varijabli nema povezanosti, odnosno nema razlike, zbog čega se takav oblik hipoteze zove nul-hipoteza. Provođenjem statističke analize rezultata vrši se testiranje hipoteze, koja se na temelju dobivenih rezultata može prihvatiti ili odbaciti. Međutim, ne postoji varijabla koja se može mjeriti s apsolutnom sigurnošću, jer su sva mjerenja podložna pogrešci. Zbog toga se rezultati u sklopu istraživanja izražavaju i interpretiraju kao vjerojatnost. Vjerojatnost nekog rezultata izražava se P-vrijednošću (engl. *probability*). Primjenom statističkog testa istraživači nastoje utvrditi mogućnost da je dobiveni rezultat, odnosno razlika između mjerenih varijabli posljedica slučajnosti, a ne stvarne razlike između mjerenih varijabli. Odnosno, P-vrijednost predstavlja mogućnost da mjerenjem utvrdimo određenu razliku između mjerenih varijabli pri istinitoj nul-hipotezi, odnosno kad navedena razlika realno ne postoji. P-vrijednost je najčešće dogovorno definirana na razini 0,05. Navedena vrijednost izražava se kao razina statističke značajnosti. Međutim, nerijetko čitatelji znanstvenih radova, ali i sami istraživači, nepravilno interpretiraju razinu statističke značajnosti, odnosno, ako neki test rezultira P-vrijednošću < 0,05, rezultat se definira kao statistički značajan, ali se nepravilno interpretira kao povoljan ili nepovoljan za grupu u kojoj je određen. Ako nije utvrđena statistički značajna razlika između mjerenih varijabli, odnosno P-vrijednost iznosi > 0,05, rezultat se često definira kao negativan, pa čak i čitavo istraživanje. Navedeni termin pogrešno implicira da razlike između mjerenih varijabli nema, ali zapravo jedino što je moguće zaključiti jest da nema dokaza postojanja razlike[2]. Ako je, primjerice, P-vrijednost < 0,05, pravilno je interpretirati rezultat tako da je mogućnost utvrđene razlike između mjerenih varijabli utvrđena posljedica slučajnosti manja od 5 %. Ako statistički značajne razlike između promatranih varijabli nema, nul-hipoteza se prihvaća. Ostvareni rezultati ne pružaju dovoljno dokaza da uspješno odbacimo nul-hipotezu, odnosno postoji velika vjerojatnost da je razlika između mjerenih varijabli slučajna i posljedica samog mjerenja. U slučaju utvrđene statistički značajne razlike između mjerenih varijabli, nul-hipotezu odbacujemo i prihvaćamo alternativnu hipotezu, jer za to postoje dovoljni dokazi, odnosno mala je

vjerojatnost da je utvrđena razlika posljedica slučajnosti[3].

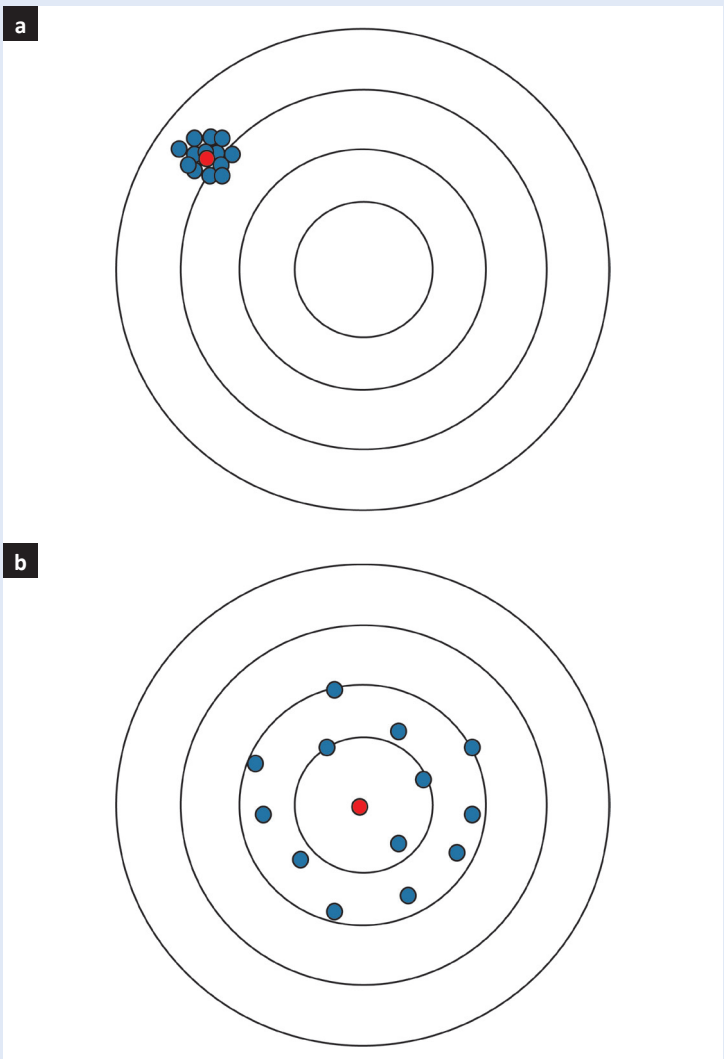
SLUČAJNE POGREŠKE U ISTRAŽIVANJIMA

Kao što je prethodno navedeno, sva mjerenja koja se provode u istraživanja podložna su pogrešci. Eksperimentalne pogreške ne odnose se na tehničke pogreške u procesu mjerenja, nesporazume ili zablude, već su nerazdvojivi dio procesa mjerenja. Eksperimentalne pogreške određene su točnošću mjerenja – koliko je izmjerena vrijednost blizu istinite ili prihvaćene vrijednosti i preciznošću – koliko su dvije ili više izmjerenih vrijednosti međusobno bliske (ponovljivost ili reproducilibnost rezultata). Pouzdanost rezultata istraživanja definiraju tri dimenzije: sustavne pogreške (engl. *bias*), pogreške u dizajnu istraživanja i slučajne pogreške (engl. *random errors*). Sustavne pogreške predstavljaju devijaciju od istine u vidu precjenjivanja ili podcjenjivanja stvarne veličine učinka. Sustavne pogreške ne utječu na preciznost mjerenja, već na točnost. Višestrukim ponavljanjem istog mjerenja ili iste studije postiže se pogrešan odgovor u prosjeku. Nepreciznosti mjerenja odnose se na slučajne pogreške, koje predstavljaju nepredvidive varijacije između opaženih, odnosno izmjerenih vrijednosti i stvarne istinite vrijednosti neke varijable. U ovom slučaju, višestrukim ponavljanjem istog mjerenja ili iste studije, unatoč nepreciznostima, postiže se točan odgovor u prosjeku (slika 1). Ako sustavna pogreška ne postoji, slučajne pogreške predstavljaju uzrok pogrešnog statističkog zaključivanja, koje dovodi do netočne procjene određenog učinka, odnosno veličine određene varijable[3].

Dvije su vrste slučajne pogreške u istraživanjima, pogreška tipa I ili alfa pogreška (α) i pogreška tipa II ili beta (β) pogreška (slika 2). Pogreška tipa I nastaje kad odbacujemo nul-hipotezu, koja je zapravo istinita, odnosno pogrešno prihvaćamo postojanje razlike između dviju varijabli, koja zapravo ne postoji. Navedena pogrešno prihvaćena razlika, odnosno rezultat, predstavlja lažno pozitivni rezultat. Vjerojatnost da se učini pogreška tipa I jednaka je razini statističke značajnosti, odnosno P-vrijednosti. Ako se primjenjuje konvencionalna razina statističke značajnosti od 0,05, vjerojatnost pogreške tipa I iznosi 5 %, što znači da će zaključak baziran na P-vrijednosti < 0,05 biti pogrešan u 5 % slučajeva[3]. Što je niže postavljena razina statističke značajnosti, to je manja vjerojatnost pogreške tipa I, odnosno manja je vjerojatnost odbacivanja nul-hipoteze, koja je istinita.

Pogreška tipa II nastaje kad greškom prihvaćamo nul-hipotezu, koja je zapravo pogrešna. U tom slučaju pogrešno prihvaćamo da između mjerenih varijabli nema značajne razlike, iako ona zapravo postoji. Takav pogrešno prihvaćeni rezultat, odnosno prihvaćeni izostanak razlike između vari-

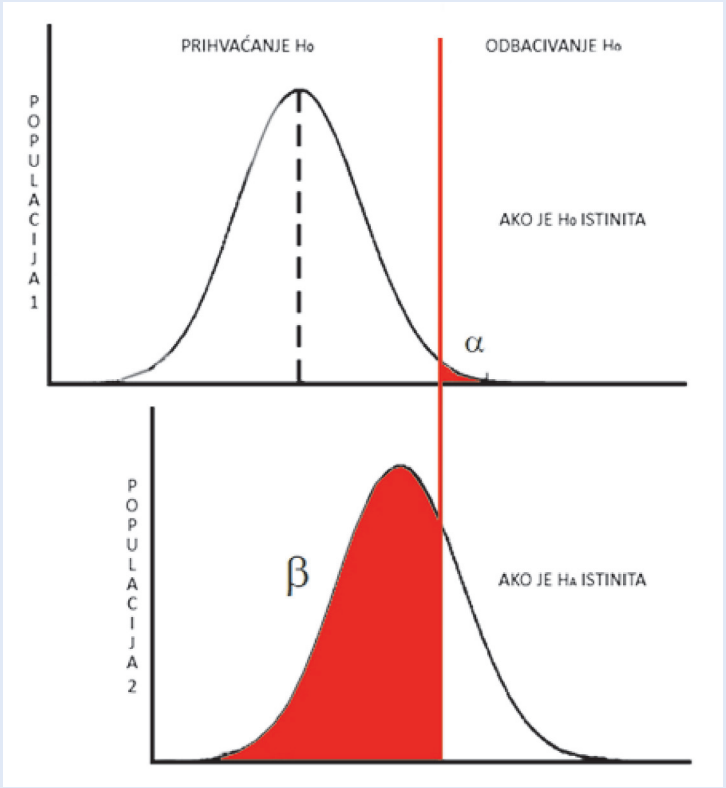
Rizik slučajne pogreške u istraživanjima umanjuje se povećanjem broja mjerenja, odnosno veličine istraživanog uzorka.



Slika 1. Pojednostavljeni grafički prikaz razlike između sustavne pogreške (a) i slučajne pogreške (b) u istraživanjima.

jabli kad ona zapravo postoji, predstavlja lažno negativni rezultat. Pogreška tipa II izravno je povezana sa snagom testa, odnosno snagom studije (engl. *power*). Snaga testa jednaka je 1 ↕ β. Pogreška tipa 2 konvencionalno se postavlja na 0,2, što znači da je uobičajena snaga testa 0,8, iako nerijetko može biti postavljena i na vrijednost 0,9, odnosno β pogreška iznosi 0,1. Ako je snaga testa 80 %, to zapravo znači da je istraživač voljan prihvatiti 20 % (1 : 5) vjerojatnosti da će zaključiti kako ne postoji realna razlika između mjerenih varijabli u populaciji, iako ona zapravo postoji[4]. Snaga testa je sposobnost, odnosno vjerojatnost testa da će točno detektirati razliku kada ona u populaciji realno i postoji. Također, snaga testa je vjerojatnost testa da će izbjeći lažno negativni rezultat.

Izračun veličine uzorka predstavlja određivanje uzorka čija je veličina dovoljna za detekciju dobrobiti određene intervencije ako ona postoji, ali pritom nije veća od potrebne.



Slika 2. Ilustracija slučajne pogreške tipa I i II. Pretpostavimo da postoje dvije populacije s različitom srednjom vrijednošću mjerene varijable, ali čije se normalne krivulje raspodjele međusobno djelomično preklapaju. Ako je nul-hipoteza (H_0) istinita, odnosno nema značajne razlike između uspoređivanih populacija, tada postoji mala vjerojatnost da će stvarna srednja vrijednost ispitivane varijable biti desno od vertikalne linije, odnosno u zacrnjenom području (pogreška tipa I ili α), tj. postoji mala vjerojatnost pogrešnog odbacivanja istinite H_0. Ako je alternativna hipoteza (H_A) istinita, tada postoji više od 50 % vjerojatnosti da će srednja vrijednost biti lijevo od vertikalne linije i H_0 neće biti odbačena, iako bi trebala, što predstavlja pogrešku tipa II ili β. Ako je β = 0,67, tada je vjerojatnost da ćemo točno utvrditi postojanje stvarne razlike između mjerene varijable dviju populacija 1 ↕ β = 0,33, odnosno 33 % (prilagođeno prema ref. 12).

Rizik slučajne pogreške u istraživanjima umanjuje se povećanjem broja mjerenja, odnosno veličine istraživanog uzorka. Samo dovoljan broj uključenih ispitanika, odnosno provedenih istraživanja, osigurat će prihvatljiv rizik slučajne pogreške.

VELIČINA UZORKA

Iz dosada izloženog vidljivo je da adekvatna provedba istraživanja, odnosno nastojanje da istraživanjem ostvarimo vjerodostojne rezultate temeljem kojih možemo donijeti pouzdane zaključke uvelike ovisi o broju ispitanika, odnosno provedenih mjerenja u samom istraživanju, odnosno o veličini uzorka. Izračun je, odnosno definiranje potrebne veličine uzorka u istraživanju, ključno učiniti u fazi planiranja istraživanja. Glavni cilj prilikom definiranja potrebnog broja ispitanika odnosi se na određivanje uzorka čija je veličina dovoljna za detekciju dobrobiti određene intervencije, ako ona postoji, ali pritom nije veća od potrebne[5]. Izračun adekvatne veličine uzorka ovisi o nekoliko čimbenika: razine pogreške tipa I, razine pogreške tipa II ili statističke snage, najmanje klinički relevantne veličine učinka i naravi podataka koji se obrađuju[3]. Kao što je ranije spomenuto, dogovorno se najčešće pogreška tipa I ili α definira na 5 %, dok se pogreška tipa II ili β definira na 20 %, iz čega proizlazi da je snaga 80 %.

Razlika u veličini učinka ispitivane intervencije ili mjerene srednje vrijednosti ispitivane varijable u odnosu na kontrolu zapravo je nepoznata, te je istraživač procjenjuje na najbolji mogući način. Navedena je procjena vrlo bitna u izračunu veličine uzorka, ali sam postupak je nerijetko zahtjevan i subjektivan. Procijeniti razliku veličine učinka neke intervencije (engl. *intervention effect size*) možemo na temelju dostupnih podataka o njoj u pokusnoj studiji, ako ona postoji, ili ranijim, već provedenim istraživanjima. Navedeno može

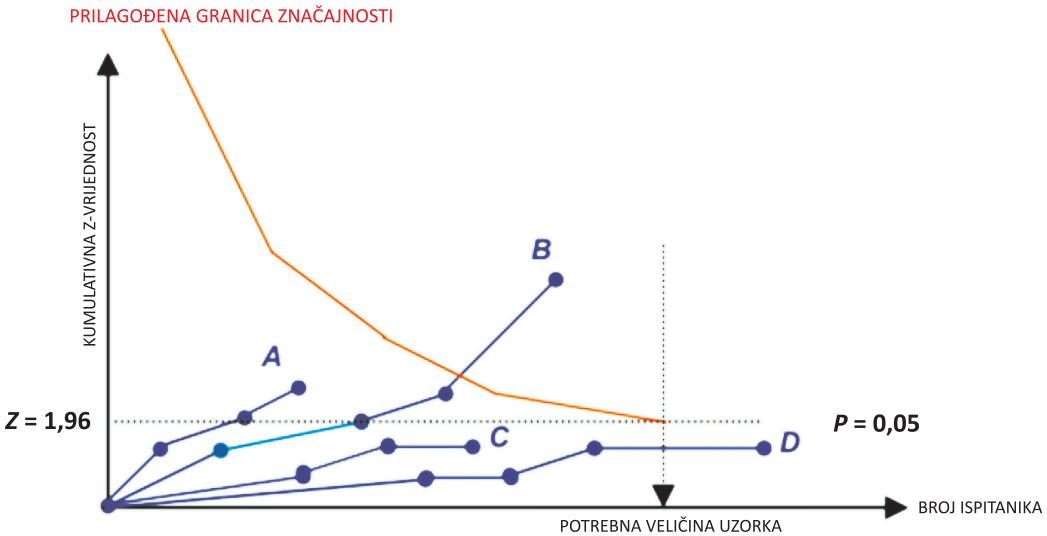
biti korisno, ali može isto tako dovesti do zablude i pogrešnog tumačenja učinka neke intervencije. Naime, procjena razlike veličine učinka neke intervencije ili srednje vrijednosti neke varijable od interesa prilikom izračuna potrebne veličine uzorka predstavlja procjenu koja se odnosi na čitavu populaciju. Razlika koja je utvrđena u pokusnoj studiji ili nekom drugom istraživanju u pravilu nije utvrđena na odgovarajućem broju ispitanika te predstavlja sama po sebi procjenu na koju utječe slučajna pogreška[6]. Zbog navedenoga je najbolje procijeniti razliku veličine učinka intervencije u odnosu na kontrolu tako da se istraživanje planira s ciljem detekcije minimalne klinički relevantne razlike veličine učinka. Navedena se razlika ponekad opisuje kao najmanja razlika u nekoj domeni od interesa, koju pacijenti doživljavaju dobrobitnom, a koja u odsutnosti bitnijih nuspojava i izrazitih troškova zahtijeva promjenu u pristupu liječenja pacijenata[7]. U tom slučaju mogu se uzeti u obzir razlike u veličini učinka koje su u prethodnim slučajevima dovele do uvođenja novih terapija, kao i provođenje upitnika među stručnjacima za određeno područje, odnosno specifičnim pacijentima. Važno je naglasiti da, što je manja klinički relevantna razlika, to će veći broj ispitanika biti potrebno uključiti da se ona detektira.

Narav podataka odnosi se na vrstu varijabli ili ishoda koje mjerimo, pa tako je u slučaju kontinuiranih podataka bitna njihova varijabilnost, koja je definirana standardnom devijacijom. Potrebna veličina uzorka raste što je veća varijabilnost podataka. Analogno standardnoj devijaciji često se u izražavanju varijabilnosti podataka, odnosno varijabilnosti uzorkovanja koristi interval pouzdanosti (engl. *confidence interval;* CI). Interval pouzdanosti predstavlja raspon unutar kojeg se s određenom vjerojatnošću, odnosno određenom razinom sigurnosti (najčešće 95 %) nalazi stvarna vrijednost neke varijable u populaciji. Široki interval pouzdanosti ukazuje na veliku varijabilnost podataka, odnosno na mali uzorak, što ne znači da su rezultati pogrešni, već da su podaci konzistentni sa širokim rasponom mogućih hipoteza, odnosno mogućih ishoda[2]. Uski interval pouzdanosti ukazuje na malu varijabilnost podataka, odnosno velik uzorak, što ukazuje da bismo analizom drugog uzorka prilično sigurno ostvarili slične rezultate[2].

U slučaju dihotomnih podataka bitna je učestalost ili pojavnost ishoda od interesa u kontrolnoj skupini[3]. Važno je naglasiti da, što je manja klinički relevantna razlika između pojavnosti određenog ishoda u kontrolnoj i ispitivanoj skupini, to će veći broj ispitanika biti potrebno uključiti da se ona detektira. Primjerice, ako procjenjujemo da je učestalost nekog događaja u kontrolnoj skupini 10 % i procjenjujemo da je veličina učinka intervencije od interesa redukcija tog istog događaja od 40 %, te postavimo α = 0,05 i snaga = 0,9, izračunat ćemo potrebnu veličinu uzorka od ukupno 965 ispitanika. No ako pod istim parametrima našu procjenu veličine učinka intervencije smanjimo na redukciju navedenog događaja od 20 %, izračunat ćemo četiri puta veći uzorak, odnosno ukupno 4.301 ispitanika, jer je za detekciju manje klinički relevantne razlike potreban veći broj mjerenja, odnosno veći broj ispitanika[4].

Trial Sequential analiza kombinira unaprijed definiranu veličinu uzorka i granice značajnosti prilagođene s obzirom na akumulirajuće podatke u metaanalizama.

Veličina uzorka izrazito je bitna u provođenju vjerodostojnih i metodološki čvrstih istraživanja koja rezultiraju solidnim dokazima. Studije na nedovoljnim uzorcima najvjerojatnije neće biti u mogućnosti detektirati potencijalnu stvarnu razliku u ishodima među grupama, što može predstavljati nepotrebno financijsko opterećenje[3]. Potencijalno su takva istraživanja i neetična, jer izlažemo pacijenta eksperimentalnim intervencijama, a da prethodno nismo s metodološkog aspekta osigurali konzistentnost i adekvatnost istraživanja. Izračun potrebne veličine uzorka uvijek je potrebno vršiti na temelju prethodno definiranog primarnog ishoda istraživanja, koji bi u pravilu trebao biti klinički najrelevantniji ishod. Tijek izračuna i procijenjene parametre potrebno je jasno i unaprijed prikazati u metodološkoj sekciji istraživanja te time omogućiti čitatelju primjerenu interpretaciju rezultata i zaključaka istraživanja. Potrebno je također naglasiti da se prilikom analize podgrupa ispitanika s određenim ishodom ili karakteristikama unutar istraživanja u pravilu vrše na broju koji je manji od ukupnog broja ispitanika



Slika 3. Grafički prikaz sekvencijske analize studija (engl. *Trial sequential Analysis;* TSA). Četiri plave krivulje (A-D) ili Z-krivulje predstavljaju rezultate četiriju zasebnih metaanaliza, a točke na krivuljama rezultate zasebnih studija uključenih u pojedine metaanalize. Rezultati metaanalize A prelaze konvencionalnu razinu statističke značajnosti *P* = 0,05 (horizontalna točkasta linija), ali ne i granicu značajnosti prilagođenu kumulativnoj metaanalizi (crvena krivulja), što znači da su potrebna dodatna istraživanja kako bi se adekvatno procijenio stvarni učinak intervencije uz prihvatljivu razinu slučajne pogreške. Rezultati metaanalize B prelaze obje navedene granice te se unatoč tome što nije postignuta izračunata veličina uzorka (vertikalna točkasta linija), može uz prihvatljivu razinu slučajne pogreške zaključiti da ispitivana intervencija ima značajni učinak te daljnja istraživanja nisu potrebna. Rezultati metaanalize C ne prelaze ni jednu od navedenih granica značajnosti, ali nije postignuta ni odgovarajuća veličina uzorka. Stoga je indicirano nastaviti istraživanja dok rezultati metaanalize D uključuju broj ispitanika veći od definirane veličine uzorka te pritom ne prelazi navedene granice značajnost, zbog čega je moguće zaključiti da ispitivana intervencija nema značajnog učinka i nije je potrebno dalje istraživati (prilagođeno prema ref. 13).

te da stoga takve analize najčešće nemaju dovoljnu snagu, jer je snaga samog istraživanja temeljena na ukupnom broju uključenih ispitanika[3]. Studije s malim uzorcima također je potrebno objavljivati kako bi njihovi rezultati bili uključeni u eventualne metaanalize te na taj način pridonijeli stjecanju sveukupne slike neke intervencije, odnosno varijable od interesa.

VAŽNOST VELIČINE UZORKA U METAANALIZAMA

Slučajne pogreške mogu u velikoj mjeri utjecati na rezultate metaanaliza. Metaanalize se obično redovito ažuriraju objavljivanjem novih kliničkih istraživanja. Prilikom svakog ažuriranja provode se ponavljane analize na akumulirajućim podacima. Takve ponavljane analize i ponavljana testiranja značajnosti uvijek pri konvencionalnoj razini značajnosti od 5 % dovode do porasta rizika slučajne pogreške[8]. Usporedbe u sklopu sustavnih pregleda u pravilu dovode do multipliciranja. Multipliciranje predstavlja pojavu da, što se veći broj analiza učini, veća je vjerojatnost da će jedan od testova biti statistički značajan, unatoč tome što nema stvarnog učinka[9]. Ako se primjenjuje standardna razina statističke značajnosti od 5 %, očekivano će jedno od 20 testiranja značajnosti biti statistički značajno, čak i ako stvarne razlike među uspoređivanim intervencijama ne postoje[9]. Analogno tome, očekivano 1 od 20 nezavisnih 95 % intervala pouzdanosti neće obuhvaćati stvarnu vrijednost mjerenog parametra[9]. Kad ne postoji stvarni učinak intervencije, dodavanje novih podataka, odnosno novih studija u kumulativnu metaanalizu dovest će do postizanja razine statističke značajnosti i pogrešnog zaključka postojanja značajnog učinka kada on ne postoji[10]. Zbog navedene problematike Cochrane kolaboracija razvila je specifičnu metodu, sekvencijsku analizu studija (engl. *Trial Sequential Analysis;* TSA), koja kombinira unaprijed definiranu potrebnu veličine informacije metaanalize, odnosno

uzorak i prilagođene granice značajnosti ili granice praćenja (engl. *monitoring boundaries*) s obzirom na akumulirajuće podatke[8]. Veličina uzorka ili informacije u metaanalizi trebale bi biti barem toliko velike kao u pojedinačnoj studiji adekvatne snage[11]. Grafički primjer i objašnjenje TSA-a prikazani su na slici 3.

ZAKLJUČAK

Određivanje potrebnog broja ispitanika uključenih u istraživanje često predstavlja problem u njegovu planiranju, ali ujedno čini jednu od esencijalnih komponentni dobro dizajniranih kliničkih studija. Rezultati randomiziranih kliničkih istraživanja adekvatne snage praćeni su prihvatljivom razinom slučajne pogreške te su kao takvi pouzdani, dok zaključci dovode do pravilnih odgovora na istraživačka pitanja. Izračun potrebne veličine uzorka trebao bi unaprijed biti definiran i prikazan u tekstu istraživanja, a poznavanje i razumijevanje navedene materije omogućava čitateljima kritičko sagledavanje pojedinih studija i sustavnih pregleda, kao i adekvatnu procjenu njihove kvalitete.

Izjava o sukobu interesa: autori izjavljuju da ne postoji sukob interesa.

LITERATURA

1. Devane D, Begley CM, Clarke M. How many do I need? Basic principles of sample size estimation. J Adv Nurs 2004;47:297-302.
2. Slutsky DJ. Statistical errors in clinical studies. J Wrist Surg 2013;2:285-7.
3. Akobeng AK. Understanding type I and type II errors, statistical power and sample size. Acta Paediatr 2016; 105:605-9.
4. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365: 1348-53.
5. Beck RW. Sample size for a clinical trial: why do some trials need only 100 patients and others 1000 patients or more? Ophthalmology 2006;113:721-2.
6. Brasher PM, Brant RF. Sample size calculations in randomized trials: common pitfalls. Can J Anaesth 2007;54: 103-6.
7. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.
8. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive metaanalyses may be inconclusive – trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal metaanalyses. Int J Epidemiol 2009;38:287-98.
9. Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008;61:857-65.
10. Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random effects cumulative metaanalysis. Control Clin Trials 1996;17: 357-71.
11. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative metaanalysis. J Clin Epidemiol 2008;61: 64-75.
12. Hoffman J. Biostatistics for medical and biomedical practitioners. Tiburon, California, USA: Elsevier, 2015.
13. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many metaanalyses. J Clin Epidemiol 2008;61:763-9.