Metodološki pogovori:

Pristranosti v epidemiologiji

Ana MIHOR, Aleš KOROŠEC
Nacionalni inštitut za javno zdravje

Povzetek: Pri delu v javnem zdravju se mladi raziskovalci na začetku svoje kariere soočajo s težavami pri interpretaciji strokovne in znanstvene literature, ki je prvi korak pri spoznavanju posameznih področij in oblikovanju strokovnega mnenja. Med temi težavami z metodološkega vidika v ospredje postavljamo pristranosti v epidemioloških raziskavah ter jih z uporabo pristopa pogovora med raziskovalcem in metodologom orišemo, podrobno pojasnimo ali pa bralca usmerimo k nadaljnjemu branju. Poleg glavnih treh skupin pristranosti (pristranosti izbire, merjenja in motenja) se nekoliko posvetimo tudi statistični interpretaciji ter bolj poglobljeno naslovimo pristranosti v anketnih raziskavah in povzemanju literature.

Navajajte kot: MIHOR A, KOROŠEC A. Metodološki pogovori: Pristranosti v epidemiologiji. Javno zdravje 2019; 08: 1–15.
Prispelo: 1. 7. 2019 Sprejeto: 1. 8. 2019 Korespondenca: ana.mihor@nijz.si, ales.korosec@nijz.si

Poznavanje pristranosti v epidemioloških raziskavah je pomembno, ko si javnozdravstveniki gradimo podlago znanja in oblikujemo veljavno strokovno mnenje, ki ga lahko suvereno zagovarjamo. Svoje strokovno mnenje moramo namreč znati zagovarjati tudi z vidika morebitnih pomanjkljivosti v obstoječih dokazih. V zadnjih desetletjih se vse bolj poudarja, da je veliko objavljenih epidemioloških raziskav tako ali drugače pristranih oziroma so njihove ugotovitve napačne ter tudi sistematični pregledi in metaanalize niso imune za pristranosti (1). V prispevku bomo uporabili didaktični pristop pogovora med mlajšo raziskovalko in metodologom, da sistematično predstavimo veljavnost in pristranosti v epidemiologiji ter razjasnimo najpogostejša vprašanja, povezana s temi koncepti. Ker gre za izjemno obsežno področje, smo nekatere vrste pristranosti samo omenili, malo več prostora pa smo namenili anketnim raziskavam in povzetnim člankom, kot sta sistematični pregled in metaanaliza. Kritično branje literature poleg poznavanja pristranosti zahteva tudi kritično oceno rezultatov uporabljenih statističnih testov, zato del prispevka naslavlja še pomanjkljivosti statističnega preverjanja hipotez. Ob tem smo na enem mestu zbrali kar nekaj orodij za oceno kakovosti študij, virov za metodološke pristope pri sistematični in kvantitativni analizi pristranosti ter predstavili uporabo programa za grafični prikaz pristranosti.

Raziskovalka: »Na izbranem področju v javnem zdravju bi rada izvedla pregled obstoječih dokazov in ugotovitev iz epidemioloških študij, vendar me kot mlado strokovnjakinjo, ki šele začenja pot v javnem zdravju in nima veliko izkušenj, skrbi, da nimam dovolj znanja o mogočih pristranostih in ne znam kritično oceniti posameznih epidemioloških študij, kaj šele celotnega raziskovalnega področja. Ali mi lahko, prosim, razložiš, kaj so pristranosti?«

Metodolog: »**Pristranost** (angl. bias) v najširšem pomenu besede pomeni tiste zmote v sklepanju o dejstvih na podlagi rezultatov raziskav, ki so posledica sistematičnih napak v zbiranju in analizi podatkov. Temu nasprotne so napake v sklepanju zaradi naključja, kjer gre za naravno variabilnost praktično katere koli merljive lastnosti (spremenljivke), bodisi zaradi naključne variabilnosti njenih vrednosti ali naključne variabilnosti pri uporabi metod za njeno merjenje. Gre pravzaprav za naključno variabilnost vzorčenja. Odstopanja pri naključnih napakah se načeloma z enako verjetnostjo pojavljajo v pozitivno in negativno smer (v primerih simetričnih porazdelitev vrednosti v populaciji, kar pa ne velja za nesimetrične porazdelitve), pri čemer lahko rečemo, da rezultati **niso natančni** (angl. precision), natančnost pa lahko izboljšamo z večanjem ponovitev ali števila preiskovancev. Nasprotno pri sistematičnih napakah nenaključno variiranje popači dejstva vedno v isto smer, čemur pravimo, da rezultati **niso veljavni** (angl. validity). Veljavnosti ni mogoče ‚izboljšati' z večanjem študije. Oboje sicer pomeni, da posamezne ugotovitve **niso pravilne oziroma točne** (angl. accuracy), vendar nas upravičeno bolj skrbijo sistematične napake, pri katerih je potencial za velike zmote večji in pri katerih z večanjem števila to zmoto samo še bolj utrjujemo, saj hkrati povečujemo natančnost (2, 3). Naj poudarim, da se ti izrazi v različnih disciplinah uporabljajo različno, tako da veljavnost v psihometriji pomeni nekaj drugega kot v klasični epidemiologiji, na katero se bova osredotočila.«

Raziskovalka: »Ali lahko navedeš primer za naključno in sistematično napako v epidemiologiji?«

Metodolog: »Pogost didaktični primer za poenostavitev tega koncepta je merjenje obsega (npr. pasu) z merilnim trakom (4). Če z našim merilnim instrumentom ni nič narobe, bomo isti obseg v večkratnih poskusih zaradi variiranja v tem, kako merilni trak držimo, izmerili z manjšimi odstopanji – enkrat malo več od dejanske vrednosti, enkrat malo manj, v povprečju pa približno toliko, kot meri v resnici. Če pa obstaja neka sistematična napaka – na primer napaka v lestvici merilnega traku (recimo, da smo merilni trak oprali na visoki temperaturi in se je ta skrčil, mi pa tega ne vemo) –, bomo hkrati z naključno napako naredili še sistematično napako, saj bo izmerjeni obseg vsakokrat večji, kot je v resnici, ker merimo s pomanjšano lestvico.«

Raziskovalka: »Aha, torej pristranost je, če nekaj sistematično narobe delamo in vedno znova dobimo napačen rezultat.«

Metodolog: »Da, težava je, če ne vemo, da nekaj delamo/merimo narobe. Kadar vemo, da nekaj sistematično delamo narobe, pa gre za zavestno pristranost in neetično ravnanje. Kljub temu se včasih izplača uporabiti merilni instrument, ki je pristran, a natančen, kot pa neki drugi instrument, ki je veljaven, a nenatančen. To je odvisno od okoliščin oz. če točno vemo, koliko je merske napake, lahko le-to potem naknadno odštejemo.«

Raziskovalka: »Kakšne so sistematične napake in s tem pristranosti v epidemioloških raziskavah?«

Metodolog: »Preden se pogovoriva o pristranostih v epidemiologiji, morava najprej razumeti, kakšen je namen epidemiološkega raziskovanja na splošno. Z raziskovanjem želimo bodisi znane (širše) zakonitosti uporabiti v specifičnih primerih ali pa želimo iz specifičnih ugotovitev sklepati na (širše) zakonitosti. V prvem primeru govorimo o aplikativni, v drugem pa znanstveni epidemiologiji (2). Ta delitev je pomembna, da razumemo, kakšne zahteve imamo, ko pride do vrednotenja rezultatov oziroma veljavnosti študije.«

Raziskovalka: »O veljavnosti raziskav sem že nekaj prebrala, vendar ne razumem, kako je veljavnost povezana s posploševanjem in sklepanjem o dejstvih, zakonitostih?«

Metodolog: »Zagotovo si že slišala, da poznamo **notranjo in zunanjo veljavnost** (angl. internal and external validity). Notranja veljavnost se po nekaterih avtorjih nanaša na zahtevo, da lahko ugotovitve iz našega vzorca posplošimo na **izvorno populacijo** (angl. source population), iz katere smo vzorčili in o kateri želimo nekaj povedati. Temu bi lahko rekli tudi statistično posploševanje, naš vzorec pa mora nujno biti reprezentativen za izvorno populacijo. Pri zunanji veljavnosti pa nas zanima, ali lahko ugotovitve posplošimo tudi na populacije, katerih z našo študijo sploh nismo proučevali – **tarčne populacije** (angl. target population) (2, 5). Različni avtorji ta dva koncepta populacij drugače poimenujejo, kar je povzročilo precej zmede v terminologiji, vendar je pomen podoben (6). V slednjem primeru gre za manj formalno obliko posploševanja, pri čemer izključno statistično sklepanje ni dovolj, cilj pa je oblikovati splošnejšo znanstveno teorijo, ki ni vezana na katero koli populacijo in upošteva tudi ugotovitve ostalih ved (biologija, kemija, sociologija itn.). Temu nekateri pravijo znanstveno posploševanje in v tem primeru ni potrebno, da je vzorec reprezentativen za tarčno populacijo. Nasprotno, včasih je bolje, da na primer vzorčimo tako, da so kategorije, med katerimi iščemo morebitne razlike, po številu izenačene, kljub temu da v populaciji razporeditev ni enakomerna (2).«

Raziskovalka: »Katera je pomembnejša, zunanja ali notranja veljavnost?«

Metodolog: »To je odvisno od tega, ali ima raziskava bolj aplikativno vprašanje ali bolj znanstveno vprašanje. V aplikativni epidemiologiji (npr. kadar želimo na podlagi vzorca oceniti prevalenco bolezni v naši izvorni populaciji za potrebe načrtovanja storitev, pri čemer je izvorna populacija tista, iz katere črpamo naš vzorec, npr. prebivalci Ljubljane leta 2016) nas zunanja veljavnost ne zanima, a poudariti moram, da mora naš vzorec biti reprezentativen za našo izvorno populacijo. V znanstveni epidemiologiji pa je cilj posplošitev za oblikovanje splošnih zakonitosti (npr. cigaretni dim pri ljudeh povzroča raka na pljučih, ne glede na časovno obdobje, kraj in proučevano populacijo) in v tem primeru je zunanja veljavnost pomembna. Vendar je za zunanjo veljavnost notranja vedno predpogoj, medtem ko obratno ne velja (2, 5).«

Raziskovalka: »Torej dober pristop bi lahko bil, da vedno najprej ocenim notranjo veljavnost posameznih študij, ki jih bom proučevala. Kadar želim ugotovitve tako identificiranih ‚kakovostnih' študij uporabiti še za oblikovanje splošnejših trditev, me dodatno zanima tudi njihova zunanja veljavnost.«

Metodolog: » Načeloma drži.«

Raziskovalka: »Kako torej vem, ali je pri posamezni raziskavi kršena notranja veljavnost?«

Metodolog: »V večini epidemioloških učbenikov je poudarek na treh vrstah ‚napak', ki resno ogrozijo notranjo veljavnost epidemiološke raziskave oziroma povzročijo sistematično napako. To so **pristranost izbire** (angl. selection bias), **pristranost merjenja, razporeditve ali klasifikacije** (angl. measurement/classification/information bias) in **pristranost zaradi motečih dejavnikov** (angl. confounding), med katerimi pa meje niso vedno jasne (2, 4). Poskusov klasifikacije vseh mogočih vrst in podvrst pristranosti je bilo že veliko, a se nobena ni izkazala za popolno zaradi tega, ker se mehanizmi, ki vodijo do pristranosti, velikokrat prekrivajo in jih je nemogoče preprosto klasificirati v jasne razrede (6). Drugi najpogostejši pristop klasifikacije je glede na to, kdaj v poteku raziskave se pojavi: 1) zbiranje in pregled literature, 2) načrtovanje študije, 3) izvedba študije, 4) zbiranje podatkov, 5) analiza podatkov, 6) interpretacija rezultatov in 7) objava ter uporaba rezultatov (7). Vendar se mi zdi osnovna delitev v tri skupine najprimernejša.«

Raziskovalka: »Ali mi lahko, prosim, vsako od teh treh vrst napak malo opišeš?«

Metodolog: »Pristranost izbire nastane pri tistih postopkih in procesih v raziskavi, ki določajo, kdo bo in kdo ne bo na koncu vključen v študijo. Pojavi se lahko tako v fazi načrtovanja, izvedbe kot analize. Posledica pristranosti izbire je, da se povezanost med proučevanim izidom in dejavnikom razlikuje od dejanske povezanosti v izvorni skupini, ki naj bi jo vzorec predstavljal. Vključeni se torej sistematično razlikujejo od izvorne populacije, ki je predmet raziskave (oziroma kontrolna in preiskovana skupina nista primerljivi); povezanost, ugotovljena v selekcionirani podskupini, pa je lahko bodisi realna za to podskupino (a je ne moremo posplošiti na izvorno populacijo) ali pa popolnoma lažna (4).«

Raziskovalka: »Zakaj v osnovi pride do teh razlik?«

Metodolog: »V grobem razloge za pristranost izbire lahko razdelimo v nekaj skupin: samoizbira (npr. tisti, ki se sami ponudijo, da bodo sodelovali v študiji, so morda bolj ozaveščeni in zdravi), neodgovor (npr. tisti, ki zavrnejo sodelovanje, so drugačni od sodelujočih), izguba iz sledenja (npr. tisti, ki predčasno izstopijo iz raziskave, so lahko bolj bolni), vzorci uporabe zdravstvenih storitev (npr. tisti, ki uporabljajo urgentne storitve, imajo drugačen socioekonomski status – SES), opredeljevanje oz. diagnoza (npr. pri tistih, ki imajo določen dejavnik tveganja za neko bolezen, je diagnoza bolj verjetno postavljena, ker so deležni natančnejše obravnave), izbira neprimerne kontrolne skupine (npr. kontrola bolnikom v bolnišnični študiji ne sme biti splošna populacija, ampak bolnišnična), izguba posameznikov v analizi (npr. kadar uporabimo multivariatni model, ki v analizo vključi samo primere, ki nimajo manjkajočih vrednosti v nobeni od vključenih spremenljivk). Obstaja veliko poimenovanj za specifične tipe pristranosti izbire, naj jih naštejem samo nekaj: pristranost samoizbire (angl. self-selection bias), pristranost neodgovora (angl. non-response bias), pristranost nepokritja (angl. non-coverage bias), pristranost zdravega ali krhkega uporabnika (angl. healthy/frail user bias), Berksonova pristranost, pristranost zaradi različnega dostopa do zdravstvenih storitev, učinek zdravega delavca (angl. healthy worker effect), Neymanova pristranost selektivnega preživetja, pristranost zaznave bolezni (angl. detection bias), pristranost zaradi izgube iz sledenja (angl. loss to follow up bias). Več o tem si lahko prebereš v preglednem članku (8).«

Raziskovalka: »Ali lahko, prosim, katero od teh ilustriraš na primeru?«

Metodolog: »**Pristranost zdravega ali krhkega uporabnika** je velik problem v kohortnih intervencijskih študijah, kadar je verjetnost, da bo posameznik prejel intervencijo, večja za tiste uporabnike, ki so bolj zdravi/bolni. Gre torej za problem, ki nastane zaradi tega, ker pri kohortni študiji (v primerjavi z randomiziranim kontroliranim poskusom) ni randomizacije preiskovanega dejavnika. Primer takih pristranosti so nekatere raziskave uspešnosti raznih intervencij v realnem okolju pri zmanjševanju obolevnosti in umrljivosti, zlasti pri starejših. Če ni randomizacije starejših med prejemnike in neprejemnike zdravljenja, ampak preprosto primerjamo ‚naravni' skupini zdravljenih in nezdravljenih, se v podrobnejši analizi lahko pokaže, da so tisti, ki prejmejo zdravljenje, v povprečju bodisi bolj ali manj zdravi od tistih, ki ga ne prejmejo. Ugotovljena (ne)uspešnost zdravljenja je tako lahko (vsaj do neke mere) posledica pristranosti zdravega/krhkega uporabnika, čeprav je učinkovitost zdravljenja že bila dokazana v idealnih, kontroliranih pogojih (9).

Izpostavil bi še **pristranost zdravega delavca**, ki jo moramo imeti v mislih, kadar beremo raziskave o vplivu izpostavljenosti zdravju škodljivim dejavnikom na delovnem mestu, ki rezultate poročajo kot razmerje v neki epidemiološki meri med delavci in splošno populacijo (npr. standardiziran količnik umrljivosti). Splošna populacija v takih raziskavah ni primerna kontrolna skupina, saj je za poklice, pri katerih je veliko tveganj za zdravje ali pri katerih je potrebna dobra fizična pripravljenost, značilno, da postopek izbire za zaposlitev pomembno vpliva na lastnosti delavcev – delavci v takih sektorjih so boljšega zdravja kot splošna populacija. To lahko, v primeru, da delovno mesto ne vpliva škodljivo na zdravje delavcev, pokaže paradoksalno manjšo zbolevnost ali umrljivost izpostavljenih delavcev v primerjavi z neizpostavljeno populacijo, v primeru škodljivih učinkov pa lahko le-te zabriše ali vsaj močno zmanjša (10).«

Raziskovalka: »Kot drugo skupino napak si omenil pristranost merjenja. Ali gre tu za takšne napake, kot si jih omenil na začetku in ilustriral s prispodobo skrčenega merilnega traku?«

Metodolog: »Ne čisto. Pristranost merjenja se pojavi v fazi zbiranja podatkov in je lahko posledica napak v **definiciji, merjenju ali klasifikaciji** katere koli od preiskovanih spremenljivk, bodisi odvisne ali neodvisne (11). Vzroki so lahko na strani merilnega instrumenta, vira podatkov, napak tistega, ki meri, ali subjekta merjenja in napak pri ravnanju s podatki. Poudaril bi, da čeprav smo pristranost v širšem smislu opredelili kot sistematično napako, moraš vedeti, da pravzaprav tudi naključna napaka lahko vodi do pristranosti. To se zgodi zlasti, kadar naključno napačno določimo status izpostavljenosti ali bolezni. Najpreprosteje ti razložim ta pojav na nekem primeru. Recimo, da z uporabo nekega orodja klasificiramo ljudi na tiste, ki so izpostavljeni, in tiste, ki niso. Ker praktično nobeno orodje nima 100-odstotne občutljivosti in specifičnosti, bomo čisto po naključju nekatere neizpostavljene napačno ocenili kot izpostavljene in obratno. Temu pravimo z angleškim izrazom ‚non-differential misclassification' in bi ga lahko prevedli kot **naključna napačna klasifikacija**. Bolj kot je merilni instrument nezanesljiv, večja je takšna napaka (npr. vprašanja o pretekli izpostavljenosti – težko se je natančno spomniti, kaj si, recimo, jedel pred enim tednom). Kadar je spremenljivka dihotomna, bomo vedno dobili rezultate, ki kažejo manjši učinek izpostavljenosti (angl. bias towards the null) od realnega, to pa ne drži za politomne. Hujše posledice pa nastanejo, kadar je **napačna klasifikacija sistematična** (angl. differential misclassification) – kadar je verjetnost napačne klasifikacije bolj verjetna v eni skupini, ker je povezana bodisi z izpostavljenostjo, boleznijo ali drugim dejavnikom (4).«

Raziskovalka: »Mislim, da razumem napako naključne klasifikacije. Če je v skupini izpostavljenih tudi veliko takih, ki v resnici niso izpostavljeni, bomo zaradi njih dobili nižje vrednosti nekega negativnega izida, ravno obratno pa pri neizpostavljenih. Tako se razlika v izidu med tema dvema skupinama in s tem tudi razmerje zmanjša in učinek izpostavljenosti ‚zvodeni oz. se razredči'. Ne znam si pa predstavljati, v kakšnem primeru bi v obeh skupinah različno pogosto napačno klasificirali izpostavljenost.«

Metodolog: »Klasičen primer tega je **pristranost spominjanja** (angl. recall bias) v študijah primerov s kontrolami, kjer se primeri (tisti, ki imajo bolezen) ‚bolje' spomnijo izpostavljenosti v preteklosti kot zdrave kontrole, saj so bolj motivirani in so več razmišljali o vzrokih svoje bolezni. Tako dobimo močno precenjene povezave med izpostavljenostjo in boleznijo. Sorodna napaka je **pristranost tistega, ki meri** (angl. observer/interviewer bias) – če ve, kakšen je status osebe, lahko ‚bolje' izmeri njeno izpostavljenost (npr. tisti, ki intervjuva, natančneje sprašuje obolele). Povezanost pri sistematični merski napaki je lahko pod- ali precenjena (angl. bias toward or away from the null). Kadar je le mogoče, je zato nujno, da je merjenje spremenljivk slepo (4, 11).«

Raziskovalka: »V javnem zdravju se veliko ukvarjamo z rutinsko zbranimi podatki, ki niso vedno najbolj kakovostni, saj v večini primerov niso bili zbrani za namene (našega) raziskovanja. Je tudi to vir pristranosti merjenja?«

Metodolog: »Lahko je, da. Prednost rutinskih podatkov je zagotovo njihova številčnost (več deset tisoč podatkov), ki pa je po drugi strani tudi slabost, saj je nadzor nad zbiranjem podatkov, ki je v osnovi bolj heterogeno (poročanje iz bolnišnic), manjši. Razlika je tudi v namenu zbiranja podatkov. Podatki, zbrani za statistični namen, so lahko bolj kakovostni, a manj pravočasni kot tisti, ki so zbrani za namene beleženja in obračunavanja storitev. Poleg tega raziskovalec dela s podatki, ki jih ni sam zbral in jih brez pomoči nosilcev zbiranja ne more pravilno interpretirati, ker ne pozna vseh posebnosti kodiranja, poročanja, zbiranja in obdelave v podatkovni bazi. Pri bazi rutinskih podatkih nas sicer zanimajo kakovost osnovnih podatkov (s kakšno natančnostjo postavimo osnovno diagnozo/stanje), popolnost (registracija manj nevarnih bolezni je slaba, zato se maligni melanom v primerjavi z ostalimi kožnimi raki poroča popolneje), dvojniki (ali lahko razlikujemo med prvimi pojavi in ponovitvami), pravočasnost (za bolezni, pri katerih potrditev lahko traja dlje časa, je popolnost boljša, če analiziramo nekaj let stare podatke v primerjavi z najnovejšimi), natančnost pri zbiranju in agregaciji (naključne napake pri vnosu) ter kakovost kodiranja (zlasti pri osnovnem vzroku smrti). Naštete značilnosti se včasih močno razlikujejo v času (npr. izrazita izboljšava diagnostične občutljivosti) in kraju (npr. razlike v praksah kodiranja), kar lahko onemogoči nepristrane ocene trendov in primerjave med regijami ali državami (12).«

Raziskovalka: »Kako pa so opredeljene pristranosti zaradi motečih dejavnikov?«

Metodolog: »Najpreprosteje lahko motenje razložimo kot zmotno pripisovanje učinka nekemu dejavniku. Pojavi se, kadar je moteči dejavnik

povezan tako z izpostavljenostjo kot izidom in je hkrati neenakomerno porazdeljen v skupinah, ki jih primerjamo. Rothman postavi naslednja pogoja, ki ju mora izpolnjevati dejavnik, da postane moteč: 1) je povezan z opazovanim izidom, vendar ni njegova posledica (lahko je bodisi vzrok za izid ali le posredni kazalnik drugega vzročnega dejavnika za izid); 2) je povezan z izpostavljenostjo, vendar ni posledica izpostavljenosti (ni vmesni korak na vzročni poti med izpostavljenostjo in učinkom) (4).«

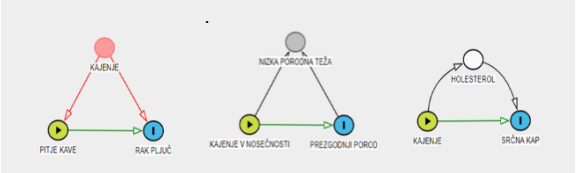
Raziskovalka: »Ne razumem dobro – ali to pomeni, da je moteči dejavnik vzročni dejavnik tako za izid kot izpostavljenost?«

Metodolog: »Da. Slikovit primer je zmotna povezava med nekim dejavnikom, ki je povezan s kajenjem, in pljučnim rakom. Zamislimo si študijo primerov obolelih za pljučnim rakom z zdravimi kontrolami, pri kateri merimo izpostavljenost pitju kave v obeh skupinah. Kajenje je v tem primeru moteči dejavnik, saj 1) povzroča raka in 2) ‚povzroča' pitje kave (tisti, ki kadijo, bolj verjetno tudi pijejo več kave). Ker kajenje povzroča raka, bo zagotovo tudi neenakomerno razporejeno med skupinama obolelih in zdravih. Izpolnjeni so vsi pogoji in analiza brez upoštevanja kajenja bi zmotno pokazala, da je pitje kave povezano s pljučnim rakom (13). Če bi v opisanem primeru veljalo obratno, namreč da pitje kave vodi v (povzroča) kajenje, potem kajenje ne bi bilo moteči dejavnik, ampak vzročni korak na poti od pitja kave do raka – pitje kave bi v tem primeru lahko šteli kot delni vzrok pljučnega raka, mediiran prek vpliva na kajenje, saj bi lahko zmanjšali zbolevanje, če bi z ukrepi preprečili pitje kave.

Posebna oblika motenja nastane zaradi ustaljenih kliničnih praks (angl. confounding by indication) (8). Kadar je odločitev, da bo oseba zdravljena z določenim zdravilom ali postopkom (npr. zaviralci protonske črpalke), povezana z določenim zdravstvenim stanjem (gastroezofagealni refluks), ki je hkrati vzročno povezano z izidom (rak požiralnika), lahko napačno interpretiramo, da zdravljenje povzroča ta izid. Podobno lahko podcenimo učinek intervencije, če je klinična praksa takšna, da le najbolj bolni prejmejo sicer učinkovito zdravljenje.«

Raziskovalec: »Med najinim pogovorom vse bolj opažam, da so si vse tri vrste pristranosti v nekih točkah zelo podobne – končni rezultat je, da primerjamo dve ali več skupin med seboj ali s populacijo, vendar tega ne bi smeli, ker so zaradi nekega vzroka različne, in torej s primerjanjem dobimo izkrivljene rezultate.«

Metodolog: »Poenostavljeno rečeno, drži. Obstaja tudi način, kako z grafičnim prikazom lahko bolje razumemo mehanizme, ki vodijo v pristranost. Imenujejo se **usmerjeni aciklični grafi** (angl. directed acyclic graphs ali DAG) (14, 15). V teh grafih so posamezni dejavniki med seboj povezani z enosmernimi puščicami, kadar so v vzročni povezavi, in iz njih lahko lažje razberemo, kateri dejavniki bi lahko vodili v pristranosti. Za risanje takih grafov ti priporočam prosto dostopen program **DAGitty** (16), ki nam glede na povezave med dejavniki tudi svetuje, katere je smiselno v analizi nadzorovati. Naj ti pokažem nekaj primerov, ki sem jih pripravil.



Slika 1: Usmerjeni aciklični grafi za prikaz odnosa a) motečega dejavnika, b) trkalnika in c) mediatorja

Slika 1a prikazuje tipičen primer motečega dejavnika (rdeči oval). Le-ta je povezan tako z izpostavljenostjo (zeleni ▶ oval) kot izidom (modri ▮ oval) v obliki razvejišča (puščici kažeta proti obema), zato ga moramo upoštevati v analizi. Slika 1b pa kaže primer ‚trkalnika' (angl. collider, sivi oval), pri čemer je majhna porodna teža posledica tako izpostavljenosti kot izida. Tukaj pa bi bilo narobe, če bi v analizi upoštevali še nizko porodno težo, saj bi, tudi če v resnici kajenje ne bi imelo nikakršnega vpliva na prezgodnji porod, tega pokazali, pri čemer bi imele nekadilke paradoksno večje tveganje za prezgodnji porod. To je posledica pogojne verjetnosti – če vemo, da je ženska rodila otroka z nizko porodno težo, pa ni kadilka, je verjetnost, da je prezgodaj rodila, večja. Na ta način s postopki analize ustvarimo nov problem – povzročimo pristranost izbire.«

Raziskovalka: »Če prav razumem, bi torej, glede na to, da vemo, da kajenje vpliva na prezgodnji porod, z nadzorovanjem porodne teže zabrisali realni učinek kajenja.«

Metodolog: »Tako je. Problemi stratifikacije po spremenljivkah ali vključevanje teh v multivariatno analizo, ki so ‚trkalniki', so že dolgo znani v neonatalni epidemiologiji. Problemi so zelo veliki pri

raziskovanju neposrednega vpliva prenatalne izpostavljenosti škodljivim dejavnikom na neonatalno umrljivost, kadar bi želeli nadzorovati vpliv tega dejavnika na umrljivost prek prezgodnjega poroda (17). Dejavnik je namreč lahko hkrati moteči dejavnik in ‚trkalnik' – v teh primerih kontrola motenja prvega lahko povzroči pristranost izbire na račun drugega.«

Raziskovalka: »Ali je mogoče, da so tudi drugi opaženi zdravstveni paradoksi posledica pristranosti zaradi ‚trkalnika'?«

Metodolog: »Seveda. Potrjeno je že, da je ta mehanizem (vsaj deloma) kriv za t. i. debelostni paradoks, pri katerem imajo kronični bolniki z višjim indeksom telesne mase nižjo umrljivost kot tisti z nižjim (18).«

Raziskovalka: »Zelo poučno. Kaj pa kaže slika 1c?«

Metodolog: »Na sliki 1c vidimo mediator (beli oval; holesterol), prek katerega se kaže del vpliva kajenja na pojav srčne kapi. Učinek kajenja na tej shemi je dvojen: neposreden in posreden prek vpliva na holesterol. Če bi želeli oceniti celokupni vpliv kajenja, je napačno, da nadzorujemo v analizi raven holesterola – s tem dobimo oceno, ki ne vključuje posrednega vpliva, ampak samo neposrednega, kar pa je upravičeno le takrat, kadar nas zanima izključno ta. Prikazal sem ti le najosnovnejše primere, če bi se želela poglobiti v metodologijo neposrednih acikličnih grafov, pa si preberi navedene vire (14, 15, 19–21).«

Raziskovalka: »Zdaj imam pregled nad mogočimi pristranostmi, vendar me zanima, pri katerih raziskavah moram biti nanje še posebej pozorna – ali so določene vrste raziskav še posebej dovzetne?«

Metodolog: »Najprej bi poudaril, da nobena zasnova ni popolnoma imuna za pristranosti. Sicer velja, da je zlati standard slepa randomizirana intervencijska študija (RCT), ki naj bi v teoriji (če je bila randomizacija popolna) povsem odpravila tretjo vrsto pristranosti – moteče dejavnike. Kljub temu da vseh motečih dejavnikov ne poznamo in jih ne merimo, sklepamo, da z randomizacijo dobimo dve identični skupini, ki imata enako porazdelitev vseh znanih in neznanih motečih dejavnikov. A kot pri merjenju tudi tukaj lahko povsem po naključju randomizacija ni popolna, kar je bolj verjetno pri majhnih vzorcih, in pristranost zaradi motečih dejavnikov ni odpravljena, rezultat pa je neveljaven. Vendar je v tem primeru napaka kljub vsemu naključna, saj z večanjem vzorca lahko zmanjšamo verjetnost njenega pojava. Poleg tega so RCT občutljive zlasti za pristranost zaradi izgube iz sledenja ter nenaključne napake pri randomizaciji in slepljenju (angl. blinding). Veliko dovzetnejše za pristranosti pa so opazovalne raziskave, med katerimi je bila največ kritik deležna študija primerov s kontrolami. V teh se pojavlja večina znanih opisanih oblik pristranosti (4).«

Raziskovalka: »Kako lahko raziskovalci obvladujejo pristranosti?«

Metodolog: »Na mnogo načinov. Gre za zelo široko področje ter obstaja veliko knjig in člankov na to temo, zato ti bom samo orisal, katere možnosti so na voljo. Metode, ki jih največ uporabljamo za preprečevanje pristranosti, so randomizacija (naključno in enakomerno razporejanje motečih dejavnikov), restrikcija (omejitev raziskave na skupino, ki ima enake vrednosti motečega dejavnika – le ženske, le kadilci ipd.) in usklajevanje (angl. matching; takšno načrtno usklajevanje skupin, da imajo enako porazdelitev motečih dejavnikov). S prvo metodo pokrijemo tako znane kot neznane moteče dejavnike, s preostalima dvema pa smo omejeni samo na tiste, ki jih že poznamo. Za obvladovanje v raziskavah, kjer ti postopki niso mogoči ali smiselni, pa moramo prirediti analizo tako, da upoštevamo tudi moteče dejavnike – bodisi naredimo stratifikacijo ali jih vstavimo v regresijske modele. Največja pomanjkljivost stratifikacije je problem majhnih frekvenc v stratificiranih skupinah, kadar bi želeli obvladovati veliko motečih dejavnikov, zato se v glavnem uporablja regresija (2). Kljub temu bi bilo tako za avtorja kot bralca koristno, da tabelarično in/ali grafično prikaže porazdelitev po najpomembnejših motečih dejavnikih z uporabo stratifikacije. S tem dobimo boljši vpogled v podatke. Omenil bi še metodo, ki se uporablja zlasti za pristranosti, povezane s klinično prakso, pa tudi za pristranosti izbire, in sicer gre za oblikovanje takšne kontrolne skupine, ki ima na podlagi modela s ključnimi napovednimi spremenljivkami podobno verjetnost, da bi bila deležna intervencije, kot skupina, ki je dejansko prejela intervencijo (angl. propensity score matching). Ta verjetnost se lahko uporabi tudi za stratifikacijo, uteževanje in kot moteči dejavnik v regresijskem modelu (22). Obstajajo še metode, s katerimi preverjamo trdnost rezultatov modela tako, da spreminjamo model in s tem predpostavke. Temu pravimo **občutljivostna analiza** (angl. sensitivity analysis). Največ se

uporablja za oceno pristranosti zaradi nemerjenih motečih dejavnikov (23).«

Raziskovalka: »Ali bi se lahko zdaj malce poglobila v presečne anketne študije, ki so temelj dela v javnem zdravju? Katere vrste pristranosti jih pestijo? Sklepam, da se najpogosteje pri takih zasnovah pojavi problem reprezentativnosti vzorca za izvorno populacijo.«

Metodolog: »Imaš prav. Najpogostejše ‚napake' v teh raziskavah so povezane z izbiro: pristranost vzorčenja, samoizbire, nepokritja in neodgovora. **Pristranost neodgovora**, ki je lahko posledica zavrnitve, nezmožnosti ali nedostopnosti izbrane osebe, je pogosta pri študijah, ki temeljijo na prostovoljnem odzivu preiskovancev, na primer pri zdravstvenih anketah, ki jih omenjaš in na katerih temelji veliko našega dela. Znano je, da se tisti, ki se odzovejo na povabilo in izpolnijo vprašalnik, razlikujejo od nerespondentov v mnogih lastnostih – po spolu, starosti, izobrazbi, SES itn., zato domnevamo, da se razlikujejo tudi v odnosu do zdravja in vzorcih vedenj, povezanih z zdravjem. Če je ta lastnost povezana s proučevanim izidom, bomo dobili rezultate, ki ne izražajo stanja v izvorni populaciji, ki jo želimo proučevati, velikost pristranosti pa je odvisna od stopnje neodgovora in tega, kako močno se respondenti in nerespondenti razlikujejo v merjeni lastnosti (24). Kot dodaten primer naj navedem večino mednarodno usklajenih zdravstvenih anket, ko govorimo, da so podatki reprezentativni za osebe, stare 25–64 let, ali osebe, stare 15 ali več let, pozabimo pa na institucionalizirane osebe, do katerih praktično ni mogoče priti.«

Raziskovalka: »Kaj pa, kadar se povabljena oseba odzove, vendar na nekatera vprašanja ne odgovori? Je to tudi pristranost neodgovora?«

Metodolog: »Da. To, o čemer sem malo prej govoril, je pravzaprav neodgovor posameznika (angl. unit non-response bias). Pri manjkajočih podatkih posameznih vprašanj pa gre za neodgovor spremenljivke (angl. item non-response bias) (24, 25). Pristranost se pojavi, kadar je spremenljivka, ki jo vprašanje meri, povezana z neko lastnostjo respondentov, ki na to vprašanje ne odgovorijo (na primer, kako odvisniki odgovarjajo na vprašanja o uporabi drog).«

Raziskovalka: »Kako pa raziskovalci lahko povečajo reprezentativnost rezultatov anketne študije?«

Metodolog: »Nereprezentativnost najučinkoviteje preprečujemo že v fazi načrtovanja in izvajanja raziskave – tako da izberemo pravilno vzorčenje (npr. stratificirano, da zajamemo vse skupine; vzorčimo naključno, da zmanjšamo pristranost vzorčenja [angl. sampling bias] in samoizbire), primerno obliko izvajanja (takšno, ki bo dosegla čim več ljudi in ne bo nekaterih sistematično izključila, kar bi povzročilo pristranost nepokritja), oblikujemo vprašanja in cel vprašalnik na način, da so vprašanja postavljena nedvoumno ter razumljiva in tako zmanjšamo število vprašanj brez odgovora, po potrebi prevedemo vprašalnik za ciljno marginalno skupino, pošiljamo opomnike ali uporabimo druge načine motiviranja ljudi, da se odzovejo (npr. nagradne igre, pozivi v medijih k sodelovanju v raziskavi, možnost odgovarjanja v spletni anketi). Te metode so najpomembnejše, da dobimo res reprezentativne odgovore (24, 25).«

Raziskovalka: »Kaj (če sploh kaj) pa je mogoče narediti, če se raziskovalec vsega tega drži, vendar na koncu dobi nereprezentativne rezultate?«

Metodolog: »Najbolj razširjen pristop za post hoc odpravo nereprezentativnosti je uteževanje podatkov. Poznamo tri korake uteževanja: 1) glede na verjetnost vzorčenja, kadar le-to poznamo (z utežmi, ki so inverzne verjetnosti izbora v vzorec, ta pa je odvisna od protokola in načina vzorčenja); 2) glede na neodzivnost (osnovno utež delimo z odstotkom odzivnosti po določenih skupinah); 3) glede na celotno populacijo po izbranih sociodemografskih lastnostih. Med temi je za tretji tip zelo razširjena metoda **poststratifikacije**, pri kateri uteži dobimo tako, da celotno populacijo ter vzorec stratificiramo na primer po vseh kombinacijah starosti, spola, regije in podobno, nato pa v posamezni poststratifikacijski celici delimo delež populacije z deležem v vzorcu. Dobljeni količnik predstavlja utež za posameznega respondenta, s katero dobimo v vzorcu enako porazdelitev glede na postratifikacijske spremenljivke, kot jo ima izvorna populacija. Metoda se uporablja, če je mogoče preprosto pridobiti število prebivalcev v vsaki celici. Ostale sorodne metode so uteževanje z linearno ali logistično regresijo ter metoda ‚grabljenja' (angl. raking/iterative proportional fitting), ki je iterativni računalniški postopek prilagajanja vzorca na znano populacijo (24, 25).«

Raziskovalka: »Kako uspešno je uteževanje – ali lahko zaupam, da je v raziskavi, ki je uporabila uveljavljen način uteževanja (recimo po protokolu

za raziskave, ki jih predpisuje Evropska unija), nepristranost odgovora popolnoma odpravljena?«

Metodolog: »Popolnoma pristranosti ne moremo odstraniti z nobeno post hoc metodo. Uteževanje ima tudi svoje pomanjkljivosti. Uspešnost klasičnega uteževanja je v glavnem odvisna od tega, v kolikšni meri so v študiji zajeti predstavniki skupine, ki se sicer redkeje odzove na povabilo, reprezentativni za to celotno skupino. Za marsikatero skupino (npr. mladi moški z nizkim SES) se je izkazalo, da to ne drži, in pokazali so, da je na ta račun lahko prevalenca zdravju škodljivih navad močno podcenjena (26). Dopolnilno zbiranje podatkov o lastnostih nerespondentov z uporabo namenskega vprašalnika za nerespondente ali povezovanje z registrskimi podatki (če je dovoljeno in tehnično izvedljivo) oziroma iskanje teh podatkov v literaturi sicer omogoča razvoj sodobnih naprednih metod, s katerimi lahko še učinkoviteje zmanjšamo posledice neodzivnosti, kot so imputacijske metode, pri katerih na podlagi dostopnih podatkov vstavimo povprečne vrednosti za specifično skupino ali določimo manjkajoče vrednosti z regresijskim modelom v multipli imputaciji. Potem naj omenim še uteževanje z metodo nagnjena (angl. inverse propensity score weighting), pri kateri utež dobimo na podlagi logističnega regresijskega modela iz zanesljivega referenčnega vzorca in predstavlja verjetnost, da oseba z določenimi lastnosti sodeluje v anketi. Princip je podoben kot tisti, ki sem ga omenjal pri motečih dejavnikih. Ponovno pa se moramo zavedati, da lahko ima vsak postopek tudi veliko pomanjkljivosti in temelji na predpostavkah, ki morda niso izpolnjene, tako da so lahko tudi metode za popravljanje pristranosti pristrane. Največja slabost vstavljanja povprečnih vrednosti je zmanjšanje variance imputirane spremenljivke ter posledično preozki intervali zaupanja v statističnih testih, v ostalih primerih pa povečanje in nezanesljivost ocene variance ter s tem nezanesljiva ocena natančnosti ugotovitev ali pa neobstoječa ali pomanjkljiva podpora inferenčnim metodam za delo z imputiranimi podatki v običajno uporabljenih programih za analizo podatkov (24, 25).«

Raziskovalka: »Kaj nas mora najbolj skrbeti glede pristranosti merjenja v anketnih študijah?«

Metodolog: »Poleg pristranosti spominjanja, ki sva jo že omenila, velike težave predstavlja **pristranost družbene zaželenosti** (angl. social desirabilty bias), ki se prikrade v raziskavo, ker smo ljudje nagnjeni k temu, da odgovarjamo tako, kot mislimo, da je družbeno sprejemljivo – predvsem pri občutljivih temah (alkohol, droge, kajenje, spolnost, prihodek, telesna masa, prehrana otrok, kot o njej poročajo starši, ipd.) radi podcenimo negativne vzorce vedenja. Tudi to je eden izmed razlogov, zakaj je potrebna validacija vprašalnikov, pri čemer je nujno, da je validacija potekala na vzorcu, ki je reprezentativen za našo preiskovano populacijo. Za identifikacijo in merjenje te pristranosti lahko uporabimo **Marlowe-Crownovo lestvico socialne zaželenosti** ali **Martin-Larsenov točkovalnik** (11).«

Raziskovalka: »Zdaj pa bi želela vprašati še nekaj o analizi podatkov. Ali bi morala biti tudi tu pozorna na napake, ki ogrozijo veljavnost študije?«

Metodolog: »To, po čemer sprašuješ, so ‚napake', ki se redkeje omenjajo – pristranost (statistične) analize in interpretacije podatkov. Rezultati ne bodo veljavni, če na primer izberemo napačen model (npr. linearni model, pri katerem povezanost ni linearna), neprimerno statistično metodo (npr. napačna korelacijska metoda glede na naravo spremenljivke – Pearson namesto Spearman), pride do odpovedi modela (npr. zaradi majhnega števila podatkov – angl. sparse data bias) ali rezultate napačno interpretiramo.«

Raziskovalka: »Osnove statistike razumem in mislim, da znam interpretirati rezultate statističnih testov. Vem, da se lahko na rezultate bolj zanesem, če so statistično značilni oziroma je manjša vrednost p. Vrednost p nam pove, kakšna je verjetnost, da bi dobili tako veliko testno statistiko ali večjo, ob predpostavki, da velja ničelna hipoteza. Zavedam se tudi, da je meja značilnosti pri vsakem statističnem testu določena arbitrarno in da vedno obstaja možnost za napako bodisi prve vrste (lažno pozitiven rezultat z verjetnostjo alfa), pri čemer zavržemo ničelno hipotezo, čeprav je le-ta pravilna, bodisi druge vrste (lažno negativen rezultat z verjetnostjo beta), pri čemer ne uspemo zavreči ničelne hipoteze, kadar ta ne velja.«

Metodolog: »Vse, kar si povedala, drži le deloma. Prvič je zelo narobe, če vrednotimo verjetnost hipoteze samo na podlagi statistične značilnosti in velikosti vrednosti p. Bolj kot to sta ključni velikost opazovanega učinka ob upoštevanju širšega konteksta raziskovalnega vprašanja in naša kritična (strokovna) presoja. Vedeti moramo, da je vsako statistično sklepanje odvisno od predpostavk poljubnega statističnega modela, ki so obsežnejše, kot se jih zavedamo, in jim je v resnici izjemno težko

zadostiti, zato v veliki večini primerov preprosto ne moremo vedeti, ali so raziskovalci res preverili in potrdili vse predpostavke modela. Marsikatere statistične metode in klasični testi so bili razviti v drugi polovici 20. stoletja, ko računalniki še niso bilo tako zmogljivi in so zahtevali določene predpostavke ali poenostavitve, ki pa se danes lahko z uporabo novejših metod obidejo oz. niso več omejitev. Potem pa je še problem, da veliko ljudi napačno dojema statistično značilnost in vrednost p, ki je manj zanesljiva in objektivna mera, kot se zavedamo. Mnogi raziskovalci se še vedno preveč naslanjajo in interpretirajo le rezultate, pri katerih je $p < 0,05$ (27).«

Raziskovalka: »Kaj misliš s tem, ko praviš, da napačno dojemamo statistično značilnost?«

Metodolog: »Že dolgo časa ugledni epidemiologi in statistiki svarijo pred nekritično uporabo in interpretacijo statističnih testov in vrednosti p. Zelo problematično je, da zreduciramo informacijo v arbitrarno dihotomno trditev o statistični (ne)značilnosti: $p = 0,049$ je značilen rezultat, $p = 0,051$ pa nič več. S tem, ko preverimo samo ničelno hipotezo (ni učinka), poleg tega zanemarimo preverjanje vseh ostalih hipotez (vse ostale velikosti učinka). Na podlagi vrednosti p za izključno eno hipotezo namreč zelo površno opišemo informacijo, ki se skriva v podatkih. Informacija sicer narašča na naslednji način: najmanj nam pove samo trditev o značilnosti (dihotomna arbitrarna odločitev), nekoliko več pove točna vrednost p (vendar je nikakor ne smemo enačiti z močjo povezanosti, bolj kot le-to izraža namreč natančnost merjenja povezanosti) in še več velikost učinka v obliki intervala zaupanja. Če bi teoretično testirali vse mogoče hipoteze o velikosti učinka, bi dobili razporeditev vrednosti p, ki bi imela stožcu podobno obliko, pri čemer bi najvišji $p = 1$ označeval ocenjeno vrednost, vrednosti p pa bi na vsako stran padale asimptotično proti 0. Vrednost p tako lahko razumemo kot moč ujemanja med posamezno vrednostjo učinka in našimi podatki – večji kot je p, bolj se ta vrednost ujema s podatki. Interval zaupanja je skupek mogočih velikosti učinka, izhajajoč iz podatkov, kjer vrednosti p najprej naraščajo od 0 za spodnjo mejo do vrednosti 1 za ocenjeno vrednost (načeloma sredina intervala), nato pa spet padajo proti 0 za zgornjo mejo. Širina tega intervala izraža natančnost merjenja (če se naveževa nazaj na napake, gre za velikost naključne napake), oddaljenost ocenjene vrednosti od nične pa velikost učinka. Smiselno je interpretirati oboje – zavedati se moraš, da lahko izjemno velika in natančna študija ‚statistično značilno' dokaže tudi tako majhen učinek, da je praktično zanemarljiv, kar naredi ugotovitve pravzaprav nepomembne; in obratno lahko nenatančna študija da statistično neznačilen rezultat, a je ocenjena velikost učinka velika. Taka študija pa ni zanemarljiva, saj nakazuje potencial velikega učinka (4, 28). Dobro se je držati načela: odsotnosti statističnega dokaza za neki učinek ne gre enačiti s prisotnostjo dokaza, da učinka ni (29)!«

Raziskovalka: »Mislim, da razumem. Praviš, da moramo vedno imeti pred seboj širšo sliko vseh mogočih učinkov in oceniti njihovo pomembnost glede na velikosti. Prej si omenil, da pri vsakem statističnem testu naredimo veliko predpostavk. Ali jih lahko nekaj našteješ?«

Metodolog: »Klasične predpostavke, ki se jih večina zaveda, da morajo držati, so v ožjem smislu tiste, ki se neposredno tičejo statističnih modelov. Te so na primer normalna (ali druga) porazdelitev in homogenost variance pri parametričnih testih, linearna povezanost, normalna porazdelitev ostankov in homoskedastičnost pri linearni regresiji, neinformativno krnjenje pri analizi preživetja itn. V širšem smislu pa ob uporabi vsakega statističnega testa predpostavljamo tudi o stvareh, ki lahko imajo veliko večji vpliv na izkrivljanje rezultatov: naključno vzorčenje, vnaprejšnja izbira raziskovalnega vprašanja, testa in spremenljivk (in ne post hoc, ko že imamo zbrane podatke; **pristranost analize** [angl. analysis bias]) ali poročanje samo o značilnih rezultatih (**pristranost poročanja** [angl. reporting bias]). ‚Napake' v analizi so namreč lahko tudi namerne, če je želja po pozitivnih rezultatih premočna. Kot primer naj navedem prilagajanje nastavitev v statističnih modelih, da vrnejo bolj pričakovane rezultate, uporabo bistveno prevelikega števila enot v testih, kjer potem že minimalna, klinično nepomembna razlika med skupinami pomeni statistično značilno razliko, ali uporabo več analiz s ciljem dobiti statistično značilnost (t. i. p-heking oz. angl. data dredging). Statistična značilnost in vrednost p nam prav tako nič ne povesta o verjetnosti prej naštetih pristranosti ali motečih dejavnikov, saj pri testiranju predpostavljamo, da deluje izključno naključje – torej, ko uporabimo statistične teste za preverjanje hipotez, hkrati predpostavljamo, da smo pristranosti in moteče dejavnike bodisi v zasnovi ali analizi popolnoma odpravili, kar pa je praktično nemogoče (27). Dodaten problem so multipli testi, pri katerih moramo biti zelo previdni, saj v množici

testiranj skoraj zagotovo dobimo kakšen nizek p, zato so potrebni popravki, npr. popravek Bonferroni (30).«

Raziskovalka: »Torej praviš, da vsem predpostavkam v bistvu nikoli ni zadoščeno in statistični test tega ne razkrije. Sklepam pa, da to nikakor ne pomeni, da so vse študije glede tega enako kakovostne. Zagotovo obstaja razpon v tem, kako dobro se raziskava približa najboljši mogoči zadostitvi predpostavk.«

Metodolog: »Seveda, to je ključno in lahko v grobem ocenimo s tem, koliko pozornosti so raziskovalci namenili identifikaciji, obvladovanju in razpravi o vseh mogočih razlogih, ki bi vodili do kršenja predpostavk, torej ne samo statističnih predpostavk, ampak tudi predpostavk o tem, kako so rezultati generirani in predstavljeni – skratka, predpostavk o notranji veljavnosti.«

Raziskovalka: »Koliko pa se lahko zanesem na ugotovitve študije, ki so se čim bolj približale predpostavkam?«

Metodolog: »Morda je zastrašujoče, da so nekateri pokazali, da tudi v (malo verjetnem) primeru, ko je vsem predpostavkam zadoščeno, lahko po analogiji specifičnosti in občutljivosti diagnostičnega testa izračunana pozitivna napovedna vrednost statistično značilnega rezultata znaša tudi manj kot 50 % (31). Gre za t. i. krizo replikacije – večina študij s ‚statistično značilnimi rezultati' jih v ponovni raziskavi ne bi potrdila, kar se kaže v tem, da veliko preliminarnih ugotovitev s poznejšimi raziskavami ovržemo (32). Stališče laične javnosti glede pogosto popolnoma nasprotujočih dokazov je izjemno sarkastično in škodi ugledu znanosti ter ruši zaupanje ljudi v stroko. Zato nekateri zagovarjajo, da moramo povečati povprečno moč študij in odpraviti arbitrarne meje vrednosti p ali pa jih vsaj močno znižati (33). Želja mnogih statistikov je, da bi odpravili statistično sklepanje in poimenovanje rezultatov kot statistično značilne le na podlagi $p < 0,05$.«

Raziskovalka: »Kakšna je pravzaprav alternativa statističnemu preverjanju ničelne hipoteze, če praviš, da je tako nezanesljiva?«

Metodolog: »Alternativ do zdaj na žalost ni bilo prav veliko, nedavno pa so novim idejam na tem področju namenili cel suplement revije The American Statistician (34). Zanimiv predlog je sprememba načina objave rezultatov: avtorji bi najprej izvedli eno ali dve eksplorativni študiji, ki bi jih usmerili pri oblikovanju končne študije, ki bi bila vnaprej registrirana z natančno določenim protokolom zbiranja in analize podatkov. Še en mogoč pristop je, da raziskovalci na istih podatkih uporabijo različne metode analize in ugotovijo, ali se rezultati spreminjajo. Torej gre spet za neke vrste občutljivostno analizo. Tisti, ki se zavedajo, da se vrednosti p še dolgo časa ne bomo znebili, zahtevajo od avtorjev, da če že poročajo vrednosti p, izpišejo natančne vrednosti. Še bolj povedno pa je, če avtorji izkažejo, da njihovi zaključki ne temeljijo (izključno) na statističnih testih, ampak točkovnih ocenah parametrov (velikostih učinka) z intervali zaupanja in njihovem kritičnem vrednotenju (s poudarkom na analizi pristranosti) ob upoštevanju informacij zunaj lastnih podatkov in perspektiv. Tako oblikovanim zaključkom lahko bolj ‚verjamemo'. Nazorno razlago interpretacije dobljene vrednosti p v odvisnosti od tega, kako (biološko, psihološko, fizikalno …) smiselna je pravzaprav osnovna hipoteza, si lahko prebereš v članku Regine Nuzzo (35).«

Raziskovalka: »Za konec bi vprašala še, kaj misliš, kako zaskrbljujoča je tako imenovana **pristranost objave** (angl. publication bias)? Vemo, da so objavljeni članki vir, iz katerega (vse pogosteje) črpajo odločevalci za oblikovanje na dokazih temelječih odločitev (angl. evidence-based decision making), in ne nazadnje je prav priprava podlage za ta namen cilj mojega dela. Kako velika je ta pristranost?«

Metodolog: »Pristranost objave je mnogo širši pojem, kot pove ime samo. Pod njegovo okrilje spadajo namreč različne oblike pristranosti, ki bi jim s skupnim imenom pravilneje rekli **pristranost diseminacije** (angl. dissemination bias), saj vključujejo poleg pristranosti objave v ožjem smislu tudi pristranost jezika, pristranost citiranja, pristranost indeksiranja v podatkovnih bazah, pristranost sive literature, pristranost medijskega poročanja ipd. Kot pri vsaki pristranosti je tudi pri pristranosti objave jasno, da če se objavljene ali upoštevane študije sistematično razlikujejo od tistih, ki niso objavljene ali so manj upoštevane/vidne, imamo velik problem pri ocenjevanju veljavnosti celotne baze ugotovitev (36). Mnogi tudi opozarjajo, da se je, odkar je statistična analiza postala hitra in preprosta zaradi uporabe računalnikov, kakovost študij zmanjšala, ob tem pa so se izjemno pomnogoterile. Tako se po eni strani utapljamo v informacijah, po drugi pa nimamo pravega uvida zaradi selekcije izključno

pozitivnih (ali negativnih v primeru, kadar gre za nasprotje interesov) ali lahko dostopnih rezultatov (37). Ocenjujejo, da okoli polovica vseh raziskav nikoli ni objavljenih, pri čemer je med objavljenimi nesorazmerno veliko lažno pozitivih rezultatov. Selekcija rezultatov nastane na strani raziskovalcev, ki v objavo pošiljajo le pozitivne rezultate, na strani pokroviteljev, ki naročajo študije z ‚vnaprej želenimi' rezultati, včasih pa celo onemogočijo objavo negativnih rezultatov, na strani urednikov in recenzentov, ki v večji meri objavijo pozitivne študije, in na strani uporabnikov, ki pristrano iščejo, berejo in citirajo literaturo (36). Na tem mestu je treba omeniti tudi povsem realno dilemo raziskovalca: če v raziskavi določen javnozdravstveni ukrep, ki ga je predlagal in uvedel, ne pokaže nedvoumno (statistično) pomembnega izboljšanja nekega z zdravjem povezanega vedenja ciljne populacije in je zaradi tega ogroženo nadaljnje financiranje ukrepa, so namigi k izboru ‚pravilnih' metod in interpretacij v okviru analize podatkov lahko nezanemarljivi.«

Raziskovalka: »Kako pa pristranost objave in druge pristranosti vplivajo na sicer najzanesljivejše oblike dokazov, ki so najvišje v piramidi hierarhije dokazov – sistematični pregledi in metaanalize? Običajno tovrstnim študijam namenim največ pozornosti in jih štejem kot najpomembnejše. Velikokrat celo začnem raziskovati z iskanjem in prebiranjem pregledov.«

Metodolog: »Logično je, da če je baza, iz katere črpamo za povzemanje, pristrana, potem bodo pristrani tudi povzetki, zato se pristranost objave že dolgo šteje kot ena od največjih pomanjkljivosti sistematičnih pristopov povzemanja ugotovitev.«

Raziskovalka: »Ali je mogoče pristranosti v sistematičnih pregledih in metaanalizah na neki način preprečiti?«

Metodolog: »Z vidika transparentnosti in zmanjševanja pristranosti je pomembno, da se raziskovalci natančno držijo protokolov (**PRISMA-P** (38) – protokol za sistematične preglede in **QUOROM** (39) – protokol za metaanalize RCT), ki urednikom in bralcem pomagajo pri oceni kakovosti. Zaželena je vnaprejšnja registracija v mednarodni register **PROSPERO** (40). Vendar želim poudariti, da pristranosti ne moremo popolnoma preprečiti, lahko jih le zmanjšamo.«

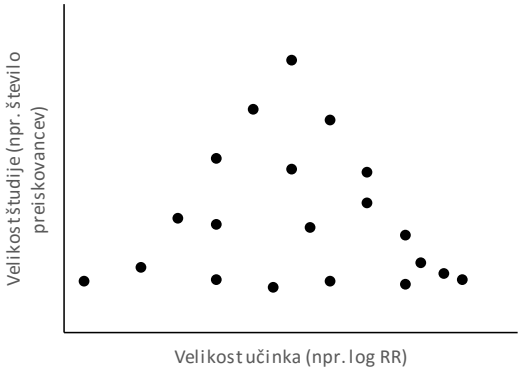
Raziskovalec: »Praviš torej, da ima vsak sistematični pregled in metaanaliza problem pristranosti – kako pa jih lahko prepoznam?«

Metodolog: »Pri ocenjevanju sistematičnih pregledov lahko poleg zgoraj omenjenih protokolov uporabimo preprosto orodje **AMSTAR** (41), ki je sestavljeno iz 16 vprašanj za oceno metodološke kakovosti. Med temi naj z vidika pristranosti objave na primer omenim, da ena od postavk kakovosti zahteva, da so avtorji naredili izčrpno iskanje literature (angl. comprehensive literature search). To pomeni, da so iskali ne samo v podatkovnih bazah objavljene članke, ampak tudi sivo literaturo, raziskovalne registre, strani pomembnih ustanov in organizacij, sezname literature iz vključenih študij ter kontaktirali eksperte na področju za pridobitev njihovega mnenja ali neobjavljenih raziskav. Še posebej je takšno iskanje pomembno, kadar raziskujemo v javnem zdravju, kjer je precej velik delež literature težje dostopen (v obliki poročil, ki niso bila objavljena v indeksiranih revijah in pogosto niso napisana v angleškem jeziku), posledice pristranosti pa so lahko velike, saj vodijo v neprimerne odločitve o ukrepih in politikah (42). Podobno orodje kot AMSTAR je tudi **ROBIS** (43), ki pa je bolj osredotočeno na oceno tveganja za pristranosti v sistematičnih pregledih kliničnih študij. Ker je veljavnost sistematičnega pregleda močno odvisna tudi od veljavnosti vključenih študij, bi vsak tak pregled moral oceniti tudi metodološko kakovost in tveganje za pristranosti vseh vključenih študij (pozorni smo lahko, ali so uporabili katero od orodij, kot so **ROBINS-I** (44), **Newcastle-Ottawa** (45) ali **STROBE** (46) za študije primerov s kontrolami in kohortne študije, **Cochranovo orodje RoB2** (47), program **RevMan** (48), **EPHPP** (49), **Downs in Black** (50) orodje za oceno pristranosti in kakovosti randomiziranih kontrolnih in drugih študij, **Hoyevo orodje** za prevalenčne študije (51)). V zadnjem desetletju se tovrstna orodja vse bolj razvijajo, tako da je vedno smiselno poiskati najnovejše in najbolj uveljavljeno orodje, pri tem pa ne pozabiti, da nobeno orodje ni popolno. Ključno je, da so avtorji sistematičnega pregleda izločili metodološko sporne študije po merilih, ki so jih jasno opredelili!«

Raziskovalka: »Ali ta orodja pridejo v poštev tudi pri metaanalizi?«

Metodolog: »Seveda, vse, kar smo uporabili pri sistematičnih pregledih, lahko prav tako uporabimo za metaanalize. Vendar moramo biti pozorni še na to, ali so avtorji metaanalize poskusili kvantificirati

tveganje za pristranost objave. Pri metaanalizah imamo namreč poleg zgoraj naštetega na voljo še nekatere kvantitativne metode, ki pa so uporabne le takrat, ko imamo vključeno veliko število izvirnih raziskav. Najpreprostejša je uporaba **lijakastih diagramov** (angl. funnel plot; slika 2) (52).«



Slika 1: Lijakasti diagram za prikaz (a)simetrije v metaanalizo vključenih študij

Teorija v ozadju teh je, da kadar študije narišemo z razsevnim diagramom, pri čemer na absciso nanašamo velikost ugotovljenega učinka (na primer vrednost ocenjenega relativnega tveganja za izid), na ordinato pa mero velikosti študij (na primer število enot), bi v primeru, če pristranost objave ni vplivala na vključene študije, morali dobiti obliko lijaka. Razlog za tako obliko je v tem, da zaradi naraščanja težav v izvedljivosti (stroški, organizacija) število študij pada glede na velikost študij, ob tem pa se zaradi večje natančnosti manjša tudi heterogenost rezultatov. Kadar deluje pristranost objave, se to kaže v asimetričnosti lijaka – predvsem obstaja praznina, kjer bi morale biti razpršene manjše študije s (skoraj) ničelnimi rezultati. Lijakasti diagram lahko razkrije tudi pristranost poročanja, če manjkajo študije, ki ne kažejo koristnosti intervencij (npr. zdravljenja), in je lijak tako po eni strani prazen. Preprosta metoda za odpravo takšne asimetrije je t. i. trim and fill, pri katerem asimetrični del preslikamo na praznino in ponovno izračunamo skupni učinek ter ga primerjamo z nepopravljenim (53). Nekoliko drugačen pristop je t. i. Rosenthalov fail-safe N (54). Tu v primeru, da smo z metaanalizo pokazali neki skupni učinek, dodajamo vključenim študijam namišljene študije z ničelnim rezultatom in preverjamo, koliko takšnih študij moramo vključiti, da postanejo tudi rezultati metaanalize ničelni. Če je to število zelo majhno, je tveganje za pristranost objave veliko. Obstajajo še druge, zapletenejše metode, na primer Eggerjev statistični test, o katerih si lahko več prebereš v knjigi *Publication bias in meta-analysis* (55).«

Raziskovalka: »Če odmislimo pristranost objave, ali so sistematični pregledi in metaanalize dovzetni še za druge pristranosti – glede na to, da je postopek vključevanja študij odvisen tudi od raziskovalčeve presoje?«

Metodolog: »Zaradi bolj rigoroznega in kvantitativnega pristopa so se sistematični pristopi kot (domnevno) boljša alternativna klasičnim ekspertnim pregledom (angl. narrative reviews) uveljavili ravno na podlagi pomislekov glede objektivnosti pri vključevanju in interpretaciji ugotovitev. Vendar se je izkazalo, da nimajo sistematični pristopi nič manj problemov s pristranostjo, morda celo več, saj so včasih preglede oblikovali izključno ljudje z ogromno izkušnjami z nekega področja in dostopom do objavljenih raziskav, zdaj pa naj bi skoraj vsak, največkrat mlajši raziskovalec, če se dosledno drži zapovedanega protokola, lahko izvedel kakovosten povzetek znanja (56). Tak pristop se mi zdi sicer veliko objektivnejši in primernejši za klinične raziskave, po drugi strani pa je lahko preveč redukcionističen in neprimeren, zlasti kadar gre za izjemno kompleksna, večdisciplinarna vprašanja – pogosto ravno v javnem zdravju. Vendar je kljub vsemu sistematično povzemanje velikokrat najboljša izbira, ki jo imamo na voljo, poleg tega se tudi vse bolj izboljšuje z raznimi metodološkimi pristopi za zmanjševanje pristranosti. **Naj zaključim z mislijo, da so ugotovitve, ki jim lahko najbolj zaupamo, zagotovo kombinacija ekspertnega znanja in sistematičnega povzemanja.**«

Raziskovalka: »Hvala za izčrpne odgovore. Zdaj sem opremljena z veliko več znanja o tem, kako z vidika pristranosti kritično vrednotiti epidemiološke študije pri spoznavanju nekega področja. To mi bo v veliko pomoč tudi pri prvih raziskovalnih korakih na moji strokovni poti.«

V pričujočem prispevku smo z didaktičnim pristopom pogovora med strokovnjakinjo in metodologom osvetlili najpomembnejše pristranosti in vire motenja, ki lahko ogrozijo bodisi veljavnost ene izvirne epidemiološke študije, veljavnost povzetne raziskave, kot sta sistematični pregled in metaanaliza, ali celo postavijo pod vprašaj veljavnost ugotovitev celotnega raziskovalnega področja. Za bralca smo na enem mestu zbrali veliko virov pristranosti, nasvetov in sodobnih orodij, ki so mu lahko v pomoč pri vrednotenju kakovosti študij, po drugi strani pa lahko o njih razmišlja tudi, kadar sam načrtuje raziskavo. Kjer bi bralec želel bolj poglobljeno znanje o posameznih pristranostih in pristopih, ga usmerimo k relevantnim virom. Osredotočili smo se sicer izključno na epidemiološke študije, zelo pomembno v javnem zdravju pa bi bilo raziskati in pojasniti tudi, kje in kako je veljavnost ogrožena v študijah, ki uporabljajo (bolj) kvalitativne pristope.

LITERATURA

1. Ioannidis JPA. Why most published research findings are false. PLOS Med 2005; 2(8): e124.
2. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.
3. Greenland S, VanderWeele TJ. Validity and bias in epidemiological research. In: Detels R, Gulliford M, Abdool Karim Q, Tan CC, editors. Oxford textbook of global public health. 6th ed. Oxford, UK: Oxford University Press, 2015: 569–90.
4. Rothman KJ. Epidemiology: an introduction. 2nd ed. New York, USA: Oxford University Press, 2012.
5. Rothman KJ, Greenland S. Validity and generalizability in epidemiologic studies. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. 2nd ed. Chichester, England: John Wiley & Sons, 2005.
6. Schwartz S, Campbell UB, Gatto NM, Gordon K. Toward a clarification of the taxonomy of "bias" in epidemiology textbooks. Epidemiology 2015; 26(2).
7. Choi BCK, Pak AWP. Bias, Overview [elektronski vir]. In: Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels JL, editors. Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, 2014. Dostopno 27. 6. 2019 na: https://doi.org/10.1002/9781118445112.stat05095
8. Delgado-Rodriguez M, Llorca J. Bias. J Epidemiol Community Heal 2004; 58(8): 635–41.
9. Soumerai SB, Starr D, Majumdar SR. How do you know which health care effectiveness research you can trust? A guide to study design for the perplexed. Prev Chronic Dis 2015; 12: E101–E101.
10. Chen W, Yu S, Zhu J, Chai H, He W, Wanga W. Personality characteristics of male sufferers of chronic tension-type and cervicogenic headache. J Clin Neurol 2012; 8(1): 69–74.
11. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016; 9: 211–7.
12. Swerdlow AJ. Data quality in vital and health statistics. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. 2nd ed. Chichester, England: John Wiley & Sons, 2005.
13. Sanikini H, Radoï L, Menvielle G, Guida F, Mattei F, Cénée S et al. Coffee consumption and risk of lung cancer: the ICARE study. Eur J Epidemiol 2015; 30(1): 81–5.
14. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2008: 183–211.
15. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol 2008; 8(1): 70.
16. Textor J, Hardt J, Knüppel S. DAGitty: A graphical tool for analyzing causal diagrams. Epidemiology 2011; 22(5).
17. Wilcox AJ, Weinberg CR, Basso O. On the pitfalls of adjusting for gestational age at birth. Am J Epidemiol 2011; 174(9): 1062–8.
18. Banack HR, Kaufman JS. The "Obesity paradox" explained. Epidemiology 2013; 24(3).
19. Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res 2018; 84(4): 487–93.
20. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15(5).
21. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes JM, Kaufman JS, editors. Methods in social epidemiology. San Francisco, CA: Jossey-Bass, 2006: 393–428.
22. Alan BM, Richard W, Bradley LJ, Til S. Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 2013; 6(5): 604–11.
23. Groenwold RHH, Nelson DB, Nichol KL, Hoes AW, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. Int J Epidemiol 2009; 39(1): 107–17.
24. De Leeuw ED, Hox J, Dillman D. International handbook of survey methodology. New York, US: Routledge, 2012.
25. Groves RM, Fowler Jr FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey methodology. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2011.
26. Gray L, McCartney G, White IR, Katikireddi SV, Rutherford L, Gorman E et al. Use of record-linkage to handle non-response and improve alcohol consumption estimates in health survey data: a study protocol. BMJ Open 2013; 3(3): e002647.

27. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN et al. Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016; 31(4): 337–50.
28. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. Jurimetrics 2011; 51(2): 113–29.
29. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311(7003): 485.
30. Victor A, Elsässer A, Hommel G, Blettner M. Judging a plethora of p-values: how to contend with the problem of multiple testing. Dtsch Arztebl Int 2010; 107(4): 50–6.
31. Sterne JAC, Smith GD. Sifting the evidence — what's wrong with significance tests? Phys Ther 2001; 81(8): 1464–9.
32. Peng R. The reproducibility crisis in science: a statistical counterattack. Significance 2015; 12(3): 30–2.
33. Ioannidis JPA. The proposal to lower p-value thresholds to .005. JAMA 2018; 319(14): 1429–30.
34. Jeske D, editor. Statistical Inference in the 21st Century: A World Beyond p < 0.05. The American Statistician 2019; 73 (Suppl 1).
35. Nuzzo R. Scientific method: statistical errors. Nat News 2014; 506(7487): 150.
36. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ et al. Dissemination and publication of research findings: an updated review of related biases. Heal Technol Assess 2010; 14(8): 1–193.
37. Song F, Hooper L, Loke Y. Publication bias: what is it? How do we measure it? How do we avoid it? Open Access Journal of Clinical Trials 2013; 2013(5): 71-81.
38. Moher D, Shamseer L, Clarke M, Ghersi D, Liberatî A, Petticrew M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst rev 2015; 4(1): 1.
39. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Oncol Res Treat 2000; 23(6): 597–602.
40. PROSPERO International prospective register of systematic reviews [elektronski vir]. Dostopno 27. 6. 2019 na: https://www.crd.york.ac.uk/prospero/
41. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007; 7(1): 10.
42. Howes F, Doyle J, Jackson N, Waters E. Evidence-based public health: the importance of finding 'difficult to locate' public health and health promotion intervention studies for systematic reviews. J Public Health (Bangkok) 2004; 26(1): 101–4.
43. Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016; 69: 225–34.
44. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355: i4919.
45. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses [elektronski vir]. Dostopno 27. 6. 2019 na: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp
46. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies STROBE statement. Ann Intern Med 2007; 147(8): 573–7.
47. Higgins JPT, Sterne JAC, Savovic J, Page MJ, Hróbjartsson A, Boutron I et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, Clarke M, McKenzie J, Boutron I, Welch V, editors. Cochrane Methods. Cochrane Database of Systematic Reviews, 2016: 29–31.
48. Review Manager (RevMan) [računalniški program]. Verzija 5.3 Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration; 2014. Dostopno 27. 6. 2019 na: https://community.cochrane.org/help/tools-and-software/revman-5
49. Effective Public Health Practice Project. Quality assessment tool for quantitative studies [elektronski vir]. Hamilton, ON: Effective Public Health Practice Project; 1998. Dostopno 27. 6. 2019 na: https://merst.ca/ephpp/
50. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998; 52(6): 377–84.
51. Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C et al. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. J Clin Epidemiol 2012; 65(9): 934–9.
52. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315(7109): 629–34.
53. Duval S, Tweedie R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000; 56(2): 455–63.
54. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull 1979; 86(3): 638.
55. Rothstein H, Sutton AJ, Borenstein M. Publication bias in meta-analysis: prevention, assessment and adjustments. Chichester, England: John Wiley & Sons, 2005.
56. Kemm J. The limitations of 'evidence-based' public health. J Eval Clin Pract 2006; 12(3): 319–24.