Per què la majoria dels resultats de la recerca són falsos

John P. A. Ioannidis

Traducció de Joan M. V. Pons Ràfols i Gaietà Permanyer Miralda de l'article: Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124.

Nota: Per qüestions d'espai no es reprodueix la bibliografia de l'article original, per consultar-la podeu clicar aquí.

Algunes vegades, els resultats de la recerca que es publiquen són refutats per l'evidència posterior, amb la conseqüent confusió i decepció. La refutació i la controvèrsia estan presents en tota la gamma de dissenys d'investigació, des dels estudis clínics i estudis epidemiològics tradicionals fins a la més moderna recerca molecular. Hi ha una preocupació creixent que en la recerca moderna els resultats falsos puguin ser la majoria o, fins i tot, que ho siguin la gran majoria de les assercions que es publiquen. Tanmateix, això no hauria de sorprendre. Pot provar-se que la majoria dels resultats de la recerca són falsos. En aquest article examinaré els factors que influeixen en aquest problema i alguns del seus corol·laris.

Modelització del marc conceptual per uns resultats positius falsos

Diferents metodòlegs han assenyalat que l'alta taxa de no replicació (manca de confirmació) de les troballes de la recerca és una conseqüència de l'estratègia, còmoda però mal fonamentada, de reivindicar uns resultats concloents de la recerca a partir d'un únic estudi valorat per la significació estadística formal, normalment un valor de p menor de 0,05. La recerca no és representada i sintetitzada de la manera més apropiada amb els valors reals de p però, dissortadament, hi ha una noció generalitzada que els articles de recerca mèdica haurien d'interpretar-se solament d'acord amb els valors de p. Definim aquí els resultats de la recerca com a qualsevol associació que assoleix una significació estadística formal, siguin intervencions efectives, variables predictores, factors de risc o associacions. També és molt útil la recerca "negativa". "Negativa" és de fet un nom inadequat i la seva mala interpretació és generalitzada. Tanmateix, aquí ens referirem a les relacions que els investigadors afirmen que existeixen, més que als resultats nuls.

Com s'ha mostrat anteriorment, la probabilitat que el resultat d'una recerca sigui en efecte verdader depèn de la probabilitat *a priori* que ho sigui veritablement (abans de realitzar l'estudi), del poder estadístic de l'estudi i del nivell de significació estadística. Considerem una taula 2 x 2 en què els resultats de la recerca es comparen amb l'estàndard d'or d'una relació real en un camp científic. En un camp de recerca poden formular-se hipòtesis verdaderes i falses quant a la presència d'associacions. Anomenem R la raó entre el nombre de "relacions verdaderes" i l'"absència de relacions" entre aquelles variables examinades. R característicament pot variar força, depenent de si el camp apunta a relacions molt probables o si se cerquen tan sols una o unes poques associacions veritables entre milers i milions d'hipòtesis que poden postular-se. Considerem també, per simplicitat computacional, camps circumscrits en els quals o bé hi ha una sola relació veritable (entre moltes que poden ser hipotetitzades) o bé la potència (estadística) per trobar qualsevol de les diferents relacions veritables existents és semblant. La probabilitat preestudi que la relació sigui veritable és $R/(R + 1)$. La probabilitat de l'estudi de trobar una relació verdadera reflecteix el poder $1 - \beta$ (un menys la taxa d'error tipus II). La probabilitat d'afirmar que hi ha una relació quan veritablement no n'hi

TAULA 1. **Resultats de la recerca i relacions veritables**

Resultats de la recerca	Relacions veritables		
	SÍ	NO	Total
SÍ	$c(1 - \beta)R/(R + 1)$	$c\alpha/(R + 1)$	$c(R + \alpha - \beta R)/(R + 1)$
NO	$c\beta R(R + 1)$	$c(1 - \alpha)/(R + 1)$	$c(1 - \alpha + \beta R)/(R + 1)$
Total	$\boldsymbol{cR/(R + 1)}$	$\boldsymbol{c/(R + 1)}$	$\boldsymbol{c}$

ha cap reflecteix la taxa d'error tipus I o α. Assumint que s'estan examinant *c* relacions en aquest camp, els valors esperats en una taula 2 x 2 es mostren en la Taula 1.

Després que s'ha afirmat el resultat d'una recerca basat en una significació estadística formal, la probabilitat postestudi que sigui verdadera és el valor predictiu positiu (VPP). El VPP és la probabilitat complementària del que Wacholder i col·ls. han anomenat la probabilitat de referir falsos positius. D'acord amb la taula 2 x 2, hom obté VPP = $(1 - \beta) R/(R - \beta R + \alpha)$. Aleshores, un resultat de la recerca és més probablement verdader que fals si $(1 - \beta)R > \alpha$. Donat que la majoria d'investigadors depenen d'$\alpha = 0{,}05$, això significa que la troballa de la recerca és més probablement verdadera que falsa si $(1 - \beta)R > 0{,}05$.

El que s'aprecia menys bé és que el biaix i el nombre de proves independents repetides per diferents equips investigadors arreu del món poden distorsionar encara més aquesta imatge i poden donar lloc a probabilitats encara més petites que els resultats de la recerca siguin verdaders. Provarem de modelitzar aquests dos factors en el context d'una taula 2 x 2 semblant.

Biaix

Primer de tot, definim biaix com aquella combinació de diferents factors de disseny, dades, anàlisi i presentació que tendeixen a produir uns resultats de la recerca quan no s'haurien de produir. Anomenem *u* la proporció d'anàlisis provades que no haurien de ser "resultats de la recerca" però que, no obstant, acaben presentades i referides com a tals. El biaix no s'hauria de confondre amb la variabilitat per atzar, que dona lloc a què alguns resultats siguin negatius per atzar malgrat que el disseny de l'estudi, les dades, l'anàlisi i la presentació siguin perfectes. El biaix pot comportar manipulació en l'anàlisi o en la publicació dels resultats. La publicació selectiva o distorsionada és una forma típica d'aquest biaix. Podem assumir que *u* no depèn de si hi ha o no una relació veritable. No és una assumpció poc raonable, ja que habitualment és impossible conèixer quines relacions són realment certes. Davant la presència de biaix (Taula 2), hom té VPP = $([1 - \beta] R + u\beta R)/(R + \alpha - \beta R + u - u\alpha + u\beta R)$, i el VPP disminueix amb l'augment d'*u*, a menys que $1 - \beta$ sigui igual o menor que, és a dir, $1 - \beta \leq 0{,}05$ per la major part de situacions. Per tant, amb l'increment del biaix, les possibilitats que els

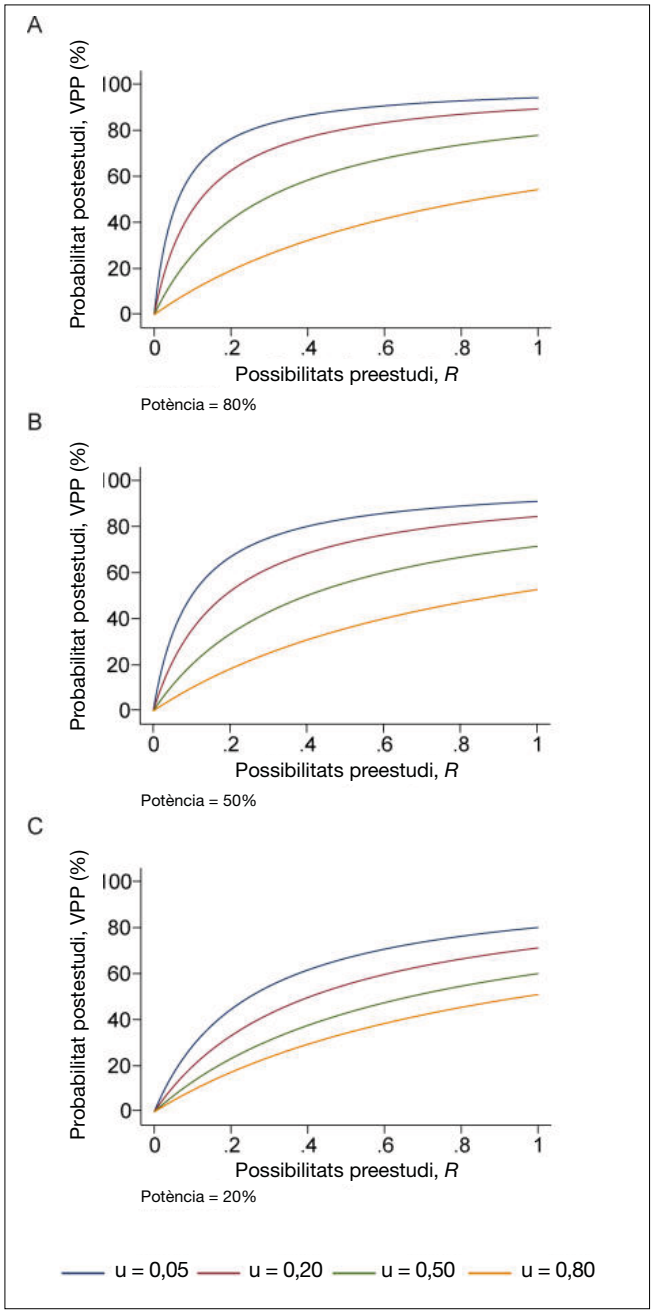

FIGURA 1. **Valor predictiu positiu (VPP) (probabilitat que un resultat de recerca sigui veritable) com a funció de les possibilitats preestudi a diferents nivells de bsiaix [u]**

Els tres panells corresponen a valors de potència 0,20, 0,50 i 0,80.

TAULA 2. **Resultats de la recerca i relacions veritables en presència de biaixos**

Resultats de la recerca	Relacions veritables		
	SÍ	NO	Total
SÍ	$(c[1 - \beta]R + uc\beta R)/(R + 1)$	$c\alpha + uc(1 - \alpha)/(R + 1)$	$c(R + \alpha - \beta R + u - u\alpha + u\beta R)/(R + 1)$
NO	$(1 - u)c\beta R/(R + 1)$	$(1 - u)c(1 - \alpha)/(R + 1)$	$c(1 - u)(1 - \alpha + \beta R)/(R + 1)$
Total	$\boldsymbol{cR/(R + 1)}$	$\boldsymbol{c/(R + 1)}$	$\boldsymbol{c}$

resultats de la recerca siguin veritables disminueixen considerablement. Això es mostra en la Figura 1 per a diferents nivells de potència estadística i per a diferents possibilitats (*odds*)[1] preestudi.

A la inversa, hi ha resultats de recerca verdaders que poden veure's anul·lats de tant en tant pel biaix invers. Per exemple, amb grans errors de mesura, les relacions es perden amb el soroll, o bé els investigadors utilitzen dades de manera ineficient, o bé no s'adonen de relacions estadísticament significatives, o bé hi ha conflictes d'interès que tendeixen a "enterrar" resultats significatius. No hi ha evidència empírica a gran escala de amb quina freqüència es presenta el biaix invers en diversos camps de recerca. Tanmateix, probablement és just dir que el biaix invers no és tan comú. A més, els errors de mesura i la utilització de dades de manera ineficient s'estan convertint probablement en problemes menys freqüents, ja que els errors de mesura s'han reduït amb els avenços tecnològics en l'era molecular i els investigadors són cada cop més sofisticats en les seves dades. Malgrat això, el biaix invers pot ser modelat de la mateixa manera que el biaix anterior. El biaix invers tampoc s'ha de confondre amb la variabilitat per atzar que pot donar lloc a la desaparició d'una relació veritable per atzar.

Proves realitzades per diferents equips investigadors

Una mateixa pregunta de recerca pot ser formulada per diferents equips d'investigadors. En la mesura que els esforços de la recerca es globalitzen, ha passat a ser pràcticament una regla que diversos equips de recerca, sovint desenes d'ells, miren de provar qüestions idèntiques o semblants. En algunes àrees, malauradament, la mentalitat prevalent fins ara ha estat centrar-se en descobertes aïllades fetes per equips individuals i interpretar els experiments de recerca aïlladament. Un nombre creixent de preguntes tenen almenys un estudi que afirma un resultat de la recerca i que rep atenció unilateral. La probabilitat que almenys un estudi, entre diversos fets sobre la mateixa qüestió, presenti uns resultats estadísticament significatius és fàcil d'estimar. Sigui n els estudis independents amb la mateixa potència (la taula 2 x 2 es mostra en la Taula 3): $VPP = R(1 - \beta^n)/(R + 1 - [1 - \alpha]^n - R\beta^n)$ (sense considerar biaix). Amb l'increment del nombre d'estudis independents, el VPP tendeix a disminuir tret que $1 - \beta < \alpha$, és a dir

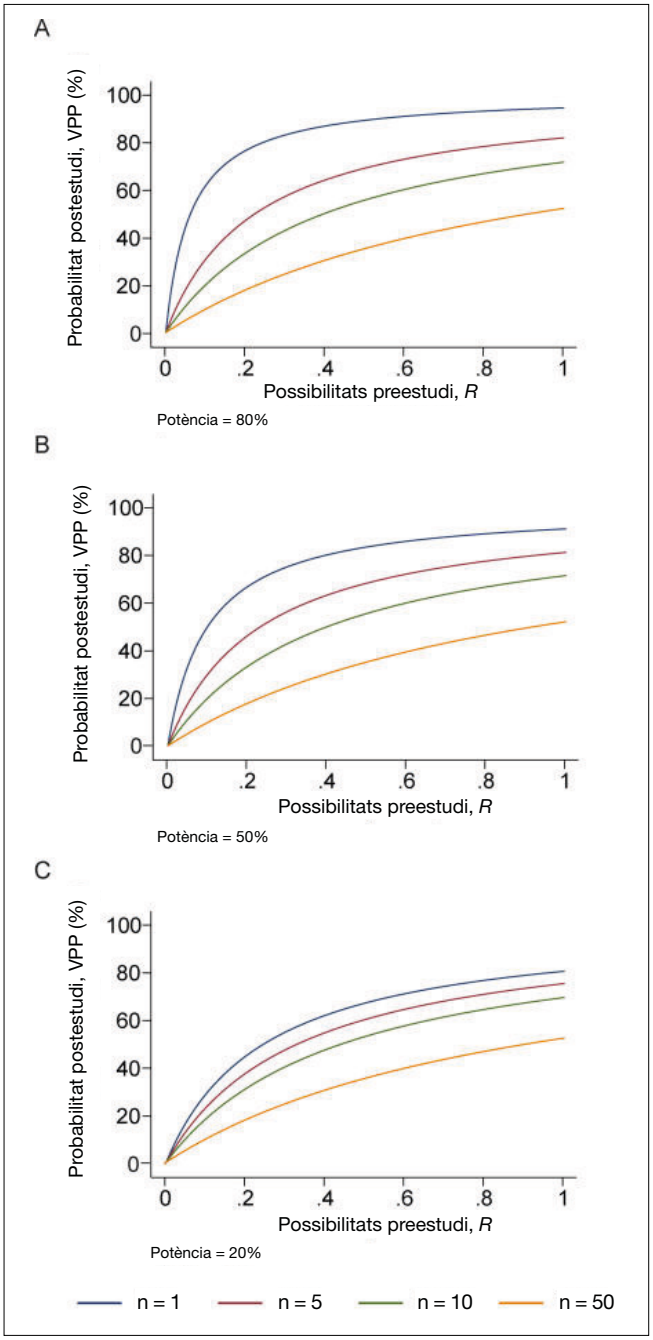

FIGURA 2. **Valor predictiu positiu (VPP) (probabilitat que un resultat de recerca sigui veritable) com a funció de les possibilitats preestudi per a diferents nombres d'estudis duts a terme**

Els tres panells corresponen a valors de potència 0,20, 0,50 i 0,80.

TAULA 3. **Resultats de la recerca i relacions veritables en presència de múltiples estudis**

Resultats de la recerca	Relacions veritables		
	SÍ	NO	Total
SÍ	$cR(1 - \beta n)/(R + 1)$	$c(1 - [1 - \alpha]n)/(R + 1)$	$c(R + 1 - [1 - \alpha]n - R\beta n)/(R + 1)$
NO	$cR\beta n/(R + 1)$	$c(1 - \alpha)n/(R + 1)$	$c([1 - \alpha]n + R\beta n)/(R + 1)$
Total	$\boldsymbol{cR/(R + 1)}$	$\boldsymbol{c/(R + 1)}$	$\boldsymbol{c}$

que, típicament 1 - β sigui menor de 0,05. A la Figura 2 es mostra per a diferents nivells de potència estadística i per a diferents possibilitats preestudi. Per a *n* estudis amb diferent potència, el terme β^n es veu substituït pel producte dels termes β_i per $i = 1$ a n, però les inferències són semblants.

Corol·laris

Un exemple pràctic es presenta en el Requadre 1. A partir de les consideracions anteriors, hom pot deduir alguns corol·laris interessants sobre la probabilitat que un resultat de recerca sigui de fet veritat.

Corol·lari 1. Quant més petits són els estudis, menys probable és que els resultats de la recerca siguin veritables.

Una grandària petita de la mostra significa menys potència i, d'acord amb les funcions exposades abans, el VPP per a un resultat verdader decreix en la mesura que la potència disminueix cap a $1 - \beta = 0,05$. Per tant, essent iguals els altres factors, els resultats de la recerca són més probablement veritables en un camp científic que porta a terme grans estudis, com és el cas dels estudis comparatius i aleatoritzats en cardiologia (alguns d'ells amb milers d'individus aleatoritzats), que en aquells camps que fan petits estudis, com és la major part de la recerca sobre predictors moleculars (amb grandàries de la mostra 100 vegades més petites).

Corol·lari 2. Quant més petita sigui la grandària de l'efecte en un camp científic, menys probable és que el resultat de la recerca sigui verdader.

La potència es relaciona també amb la grandària de l'efecte de la intervenció que s'examina. Per tant, el resultat de la recerca és més probablement veritable en aquells camps científics amb efectes grans, com el cas de l'impacte del tabaquisme en el càncer o la malaltia cardiovascular (riscos relatius 3-20), que en camps científics en què els efectes postulats són petits, com són els factors de risc genètics per a malalties multigenètiques (riscos relatius 1,1-1,5). L'epidemiologia moderna està obligada de manera creixent a avaluar petites grandàries d'efecte. En conseqüència, la proporció de resultats de recerca verdaders s'espera que es redueixi. En la mateixa línia, si la grandària d'efecte en un camp científic és molt petita, és probable que aquest camp estigui contaminat per afirmacions falses positives gairebé omnipresents. Per exemple, si la majoria de deter-

REQUADRE 1. **Exemple: ciència amb possibilitats (*odds*) preestudi baixes**

Suposem un equip d'investigadors que realitzen un estudi d'associació del genoma complet per provar si alguns dels 100.000 polimorfismes genètics està associat amb una susceptibilitat a l'esquizofrènia. A partir del que coneixem sobre el grau d'heretabilitat de la malaltia és raonable esperar que probablement al voltant de 10 polimorfismes genètics entre els que s'examinin mostraran estar associats veritablement amb l'esquizofrènia, amb unes raons de possibilitats (*odds ratios*) semblants al voltant d'1,3 per als més o menys 10 polimorfismes i amb una potència força semblant per a identificar-ne qualsevol d'ells. Aleshores $R = 10/100.000 = 10^{-4}$, i la probabilitat preestudi que qualsevol polimorfisme estigui associat amb l'esquizofrènia és també $R / (R + 1) = 10^{-4}$. Suposem també que l'estudi té una potència del 60% per trobar una associació amb una raó de possibilitats d'1,3 per $\alpha = 0,05$. Pot estimar-se aleshores que, si es troba una associació estadísticament significativa amb un valor *p* que amb prou feines creua el llindar de 0,05, la probabilitat postestudi que sigui veritable augmenta aproximadament 12 vegades en comparació amb la probabilitat preestudi, però és encara sols 12×10^{-4}.

Assumim ara que els investigadors manipulen el disseny, les anàlisis i la publicació per tal de fer que més associacions creuin el llindar $p = 0,05$, tot i que aquest llindar no es creuaria amb una adhesió perfecta al disseny, l'anàlisi i la publicació completa dels resultats, estrictament segons el pla original d'estudi. Aquesta manipulació podria fer-se, per exemple, amb la inclusió o exclusió fortuïta d'alguns malalts o controls, l'anàlisi de subgrups *post-hoc*, investigant contrasts genètics no especificats inicialment, canvis en les definicions de malaltia o control i diverses combinacions de la publicació selectiva o distorsionada dels resultats.

Alguns programes de "mineria de dades"[2] comercialitzats actualment estan orgullosos de la seva capacitat per donar resultats estadísticament significatius mitjançant la tortura de dades. Davant la presència de biaix amb $u = 0,10$, la probabilitat postestudi que el resultat sigui veritable és sols $4,4 \times 10^{-4}$. A més, fins i tot en l'absència de qualsevol biaix, quan 10 equips independents d'investigadors realitzen experiments semblants arreu del món, si un d'ells troba una associació estadísticament significativa, la probabilitat que aquest resultat sigui veritable és sols $1,5 \times 10^{-4}$, gairebé no més alta que la probabilitat que teníem abans que es portés a terme aquesta costosa recerca.

minants genètics o nutricionals veritables mostressin un risc relatiu menor a 1,05, l'epidemiologia genètica i nutricional representaria una tasca utòpica.

Corol·lari 3. Quant més gran sigui el nombre i menor la selecció de relacions analitzades en un camp científic, menys probable és que els resultats de la recerca siguin veritables.

Com s'ha mostrat anteriorment, la probabilitat postestudi que un resultat sigui veritable (VPP) depèn en gran mesura de les possibilitats preestudi (*R*). Per tant, uns resultats de la recerca són més probablement veritables en dissenys confirmatoris, com és el cas dels grans estudis comparatius i aleatoritzats fase III o en les seves metanàlisis que en els experiments generadors d'hipòtesis. Els camps considerats altament informatius i creatius, donada la riquesa d'informació reunida i avaluada, com és el cas dels *microarrays* (bioxips d'ADN o ARN) i altres recerques orientades de seqüenciació d'alt rendiment[3], forçosament han de tenir VPP extremadament baixos.

Corol·lari 4. Quant major sigui la flexibilitat en els dissenys, definicions, resultats i models analítics en un camp científic, és menys probable que el resultat de la recerca sigui veritable.

La flexibilitat augmenta el potencial de transformació del que seria un resultat "negatiu" en un resultat "positiu", és a dir, biaix *u*. Per a diversos dissenys de recerca, per exemple, els estudis comparatius i aleatoritzats, hi ha hagut esforços per estandarditzar la seva realització i publicació. L'adhesió a estàndards comuns és probable que incrementi la proporció de resultats veritables. Això mateix fa referència als resultats. Els resultats veritables són més freqüents quan els resultats són inequívocs i acordats de manera universal (per exemple, mort) que quan s'elaboren resultats múltiples (per exemple, escales per resultats en esquizofrènia). De manera semblant, camps que utilitzen mètodes analítics estereotipats i acordats (per exemple, mètode de Kaplan-Meier i prova log-rang en corbes de supervivència) donen lloc a una proporció major de resultats verdaders que no en camps on els mètodes analítics estan encara sota experimentació (per exemple, mètodes d'intel·ligència artificial) i sols es publiquen els "millors" resultats. Malgrat això, fins i tot en els dissenys de recerca més estrictes, els biaixos semblen ser un problema major. Per exemple, hi ha evidència sòlida que la publicació selectiva de resultats i la manipulació dels resultats i de les anàlisis publicades són un problema comú fins i tot en els assaigs aleatoritzats. Aquest problema no desapareixerà abolint simplement la publicació selectiva.

Corol·lari 5. Quant majors siguin els interessos financers o d'altra mena i els prejudicis en un camp científic, menys probable és que el resultat de la recerca sigui verdader.

Els prejudicis i els conflictes d'interès poden incrementar el biaix *u*. Els conflictes d'interès són molt comuns en la recerca biomèdica i es reporten, habitualment, de manera escassa i inadequada. El prejudici no ha de tenir necessàriament arrels financeres. Els científics que treballen en un camp determinat poden tenir prejudicis simplement per la seva creença en una teoria científica o el compromís amb els seus resultats. Molts estudis, que d'altra banda semblen independents i de caràcter acadèmic, poden no estar realitzats per cap altra raó que donar als metges i als investigadors qualificacions per a la seva promoció. Aquests conflictes no financers també poden donar lloc a la publicació i la interpretació de resultats distorsionats. Investigadors de prestigi poden suprimir per la via de la revisió d'experts l'aparició i disseminació de resultats que refuten els resultats propis, condemnant així el seu camp a un dogma fals i perpetu. L'evidència empírica quant a l'opinió d'experts mostra que aquesta és extremadament poc fiable.

Corol·lari 6. Quant més "calent" sigui un camp científic (més equips d'investigadors implicats), menys probable és que els resultats de la recerca siguin verdaders.

Aquest corol·lari aparentment paradoxal sorgeix perquè, com s'ha dit anteriorment, el VPP d'un resultat aïllat disminueix quan molts equips investigadors estan implicats en el mateix camp. Això pot explicar per què veiem, ocasionalment i en camps que criden àmpliament l'atenció, una gran excitació seguida ràpidament per una greu decepció. Amb molts equips treballant en el mateix camp i amb la producció de dades experimentals massiva, el temps és fonamental per vèncer en la competició. Per tant, cada equip pot prioritzar la persecució i disseminació del resultat "positiu" més impressionant. Els resultats "negatius" podran ser atractius per a la seva difusió solament si algun altre equip ha trobat una associació "positiva" en la mateixa qüestió. En aquest cas pot ser atractiu refutar un afirmació feta en alguna revista de prestigi. S'ha encunyat el terme "fenomen Proteu" per descriure aquest fet d'alternar afirmacions extremes resultants de la recerca i refutacions extremadament oposades. L'evidència empírica suggereix que aquesta seqüència d'oposicions extremes és molt comuna en la genètica molecular.

Aquests corol·laris consideren cada factor separadament, però sovint aquests factors s'influeixen entre ells. Per exemple, els investigadors que treballen en camps en què les grandàries d'efecte verdaderes es consideren petites, poden estar més predisposats a realitzar grans estudis que els investigadors que treballen en camps en què les grandàries d'efecte verdaderes són considerades com a grans. El prejudici pot prevaldre en un camp científic calent, soscavant encara més el valor predictiu dels resultats de la recerca. Grups d'interès amb molts prejudicis poden

fins i tot creuar una barrera que avorta els esforços per obtenir i disseminar resultats oposats. A la inversa, el fet que un camp sigui "calent" o que hi hagi forts interessos involucrats, algun cop pot promoure grans estudis i millorar els estàndards de la recerca, potenciant el valor predictiu dels seus resultats. Les proves massives orientades al descobriment poden donar lloc a un nombre tan gran de relacions significatives que els investigadors en tenen prou per informar i cercar més informació i, per tant, abstenir-se de la tortura i la manipulació de dades.

La major part dels resultats de la recerca són falsos per a la major part de dissenys i camps d'estudi

En el marc referit, un VPP superior al 50% és força difícil d'assolir. La Taula 4 mostra els resultats de simulacions utilitzant les fórmules desenvolupades segons la influència de la potència, la raó entre relacions verdaderes i no verdaderes, i el biaix, per a diferents tipus de situacions que poden ser característics d'uns dissenys d'estudi i contextos concrets. Un resultat d'un assaig clínic comparatiu i aleatoritzat ben realitzat i amb la potència apropiada, iniciat amb una probabilitat preestudi del 50% que la intervenció és efectiva, pot ser finalment veritat el 85% de les vegades.

És esperable un bon funcionament semblant d'una metaanàlisi confirmatòria d'estudis clínics aleatoritzats de bona qualitat: probablement s'incrementi el biaix potencial, però la potència i la probabilitat pretest són majors en comparació a un únic estudi aleatoritzat. I a la inversa, un resultat metaanalític d'estudis no concloents on la sumació de resultats s'utilitza per "corregir" la baixa potència dels estudis individuals, és probablement fals si $R \leq 1{:}3$. Els resultats de la recerca feta amb baixa potència —és el cas d'estudis clínics en fases primerenques— seria veritable en una de cada quatre vegades o encara menys si hi ha biaix. Estudis epidemiològics exploratoris (generadors d'hipòtesis) funcionen encara pitjor, especialment si són de baixa potència; però, amb bona potència, els estudis epidemiològics poden tenir una de cinc probabilitats de ser verdaders si $R = 1{:}10$.

Finalment, en recerca orientada a la descoberta amb un nombre massiu de proves, on les relacions testades superen les veritables una entre mil vegades (per exemple, quan s'examinen 30.000 gens, dels quals 30 poden ser responsables verdaders), el VPP per a cadascuna de les associacions observades és extremadament baix, fins i tot amb una estandardització considerable en el laboratori, els mètodes estadístics, els resultats i la publicació, per tal de minimitzar el biaix.

TAULA 4. **Valor predictiu positiu (VPP) de resultats de la recerca per a diferents combinacions de potència (1-β), raó de relacions verdaderes i no verdaderes (*R*) i biaix (*u*)**

1 - β	*R*	*u*	Exemple pràctic	VPP
0,80	1:1	0,10	Un estudi comparatiu i aleatoritzat amb potència apropiada, mínim biaix i possibilitats preestudi 1:1	0,85
0,95	2:1	0,30	Metaanàlisi confirmatòria d'estudis comparatius i aleatoritzats de bona qualitat	0,85
0,80	1:3	0,40	Metaanàlisi d'estudis petits no concloents	0,41
0,20	1:5	0,20	Estudi clínic comparatiu i aleatoritzat fase I/II ben realitzat, però amb baixa potència	0,23
0,20	1:5	0,80	Estudi clínic comparatiu i aleatoritzat fase I/II de mala qualitat i amb baixa potència	0,17
0,80	1:10	0,30	Estudi epidemiològic exploratori (generador d'hipòtesis) amb potència apropiada	0,20
0,20	1:10	0,30	Estudi epidemiològic exploratori amb baixa potència	0,12
0,20	1:1000	0,80	Estudi epidemiològic exploratori amb múltiples tests	0,0010
0,20	1:1000	0,20	El cas anterior d'estudi epidemiològic exploratori amb múltiples tests i biaixos més limitats (més estandarditzat)	0,0015

Els resultats afirmatius de la recerca poden ser simplement mesures acurades del biaix prevalent

La majoria de la recerca biomèdica moderna, com s'ha mostrat, té lloc en àrees amb molt baixa probabilitat, pre i postestudi, de resultats verdaders. Suposem que en un camp de la recerca no hi ha en absolut cap mena de resultat verdader per descobrir. La història de la ciència ens ensenya que sovint en el passat l'afany científic ha malbaratat esforços en camps sense cap rendiment d'informació científica veritable, si més no d'acord amb el nostre coneixement actual. En aquest "camp nul" hom esperaria que, idealment, totes les grandàries d'efecte observades variessin per atzar al voltant del valor 0 (nul) en l'absència de biaix. La mesura en què els resultats observats es desvien del que s'esperaria sols per atzar seria simplement una mesura del biaix prevalent.

Per exemple, suposem que no hi ha nutrients o patrons dietètics que siguin determinants de risc importants per desenvolupar un tumor determinat. Suposem també que la literatura científica ha examinat 60 nutrients i conclou que tots ells estan relacionats amb el risc de desenvolupar aquest tumor amb un risc relatiu en el rang entre 1,2 i 1,4 en la comparació entre el tercil superior i inferior d'ingesta. Aleshores, la grandària d'efecte referida està simplement mesurant el biaix net que ha envoltat la generació d'aquesta literatura científica. Les grandàries d'efecte referides són, de fet, l'estimació més acurada del biaix net. De tot això se segueix que, entre "camps nuls", els camps que afirmen efectes majors (sovint acompanyats d'afirmacions sobre la importància mèdica o de salut pública) són simplement aquells que han estat sotmesos als pitjors biaixos.

Per a camps amb un VPP molt baix, les poques associacions veritables no distorsionarien gaire aquesta imatge general. Encara que poques associacions siguin verdaderes, la forma de la distribució dels efectes observats encara donaria una mesura clara dels biaixos implicats en aquell camp. Aquest concepte inverteix completament la manera en què veiem els resultats científics. Els investigadors, tradicionalment, han vist els efectes grans i altament significatius amb excitació, com a signes d'una descoberta important. Efectes massa grans i massa altament significatius poden ser, de fet, més probablement signes d'un gran biaix en la major part de camps de la recerca moderna. Haurien de portar els investigadors a un pensament crític acurat sobre què pot haver anat malament amb les seves dades, anàlisis i resultats.

És clar que investigadors que treballin en qualsevol camp es resistiran probablement a acceptar que el camp complet en què han fet la seva carrera sigui un "camp nul". Tanmateix, altres línies d'evidència o avenços en la tecnologia i l'experimentació poden conduir a la llarga al desmantellament d'un camp científic. L'obtenció de mesures del biaix net en un camp pot ser també d'utilitat per visionar el que podria ser el rang de biaix present en altres camps on puguin estar presents mètodes analítics, tecnologies i conflictes semblants.

Com podem millorar aquesta situació?

És inevitable que la major part dels resultats de la recerca siguin falsos o podem millorar aquesta situació? Un gran problema és que és impossible conèixer amb el 100% de certesa què hi ha de veritat en qualsevol qüestió de recerca. Segons això, l'estàndard d'"or" pur és inassolible. Tanmateix, hi ha unes quantes aproximacions per millorar la probabilitat postestudi.

Una evidència amb major potència —per exemple, estudis amb gran grandària mostral o metaanàlisis amb poc risc de biaix— pot ajudar en la mesura en què s'aproxima a l'estàndard d'"or" desconegut. Tanmateix, els grans estudis encara poden tenir biaixos i aquests han de ser reconeguts i evitats. A més, és impossible d'obtenir evidència a gran escala per a tots els milions o trilions de qüestions plantejades en la investigació actual. L'evidència a gran escala s'hauria de focalitzar en qüestions de recerca en què la probabilitat preestudi és ja considerablement alta, de manera que els resultats significatius de la investigació donin lloc a una probabilitat postprova que s'hauria de considerar força definitiva. L'evidència a gran escala està també particularment indicada quan es poden testar conceptes majors més que no pas qüestions específiques i estretes. Aleshores, un resultat negatiu pot refutar no sols una afirmació proposada específica, sinó tot un camp o una considerable porció d'aquest camp. Seleccionar la realització d'estudis amb gran grandària mostral a partir de criteris de ment estrets, com ara la promoció comercial d'un medicament concret, és una recerca en gran mesura desaprofitada. A més, hom ha de ser cautelós amb el fet que els estudis extremadament grans poden trobar més probablement una diferència estadística i formalment significativa per un efecte trivial que realment no difereix de manera significativa de 0.

En segon lloc, la major part de qüestions de recerca són adreçades per molts equips d'investigació i és enganyós emfatitzar els resultats estadísticament significatius de qualsevol equip aïllat. El que importa és la totalitat de l'evidència. També poden ser d'ajuda la disminució dels biaixos mitjançant l'increment dels estàndards de la recerca i la reducció de prejudicis. Tanmateix, això pot necessitar un canvi en la mentalitat científica que pot ser difícil d'assolir. Per a alguns dissenys de recerca, els esforços poden ser més exitosos amb el registre avançat dels estudis; per exemple, en els estudis aleatoritzats. El registre planteja un repte per a la recerca generadora d'hipòtesis. Alguna mena de registre o de xarxa de les col·leccions de dades o dels investiga-

dors dins d'un mateix camp pot ser més factible que el registre de cadascun i de tots els experiments generadors d'hipòtesis. Independentment d'això, si fins i tot no veiem un gran progrés amb el registre d'estudis en altres camps, els principis de desenvolupament i adhesió a un protocol, per l'estil dels que existeixen per als estudis controlats i aleatoritzats, es podrien aplicar més àmpliament.

Finalment, en lloc de perseguir la significació estadística, hauríem de millorar la nostra comprensió del rang de valors d'*R* —les possibilitats preestudi— existents en l'àmbit d'aquella investigació. Prèviament a la realització d'un experiment, els investigadors haurien de considerar quines creuen que són les probabilitats que estiguin testant una associació verdadera més que una no verdadera. Aleshores, algunes vegades es podrà especular quins valors elevats d'*R* es poden esperar. Com s'ha descrit anteriorment, sempre que sigui acceptable èticament, s'hauria de realitzar estudis grans amb mínims biaixos sobre resultats de la recerca que són considerats relativament establerts, per tal de veure amb quina freqüència es confirmen aquests resultats. Sospito que alguns dels "clàssics" establerts no superarien la prova.

No obstant, la major part de noves descobertes seguiran sorgint d'una recerca generadora d'hipòtesis amb unes possibilitats preestudi baixes o molt baixes. Haurien de reconèixer aleshores que una prova estadísticament significativa resultant d'un únic estudi mostra sols una imatge parcial, sense conèixer quants experiments més s'han realitzat més enllà de la publicació i en l'àmbit rellevant en general. Malgrat la molta literatura existent sobre correccions en cas de múltiples tests, generalment és impossible desxifrar fins a quin grau els autors de la publicació han torturat les dades o quants altres equips investigadors els han precedit en aquesta troballa. Encara que això fos factible de determinar, no ens informaria sobre les possibilitats preestudi. És, per tant, inevitable que hom faci assumpcions aproximades sobre quantes associacions s'espera que siguin verdaderes entre aquelles que han estat examinades en camps de recerca i dissenys de recerca rellevants. El camp més ampli pot mostrar una orientació per estimar aquesta probabilitat en un projecte de recerca aïllat. També seria útil aprofitar les experiències sobre els biaixos detectats en camp veïns. Malgrat que aquestes assumpcions serien considerablement subjectives, serien encara molt útils per a la interpretació de les afirmacions provinents de la recerca i per posar-les en context.

NOTES

1. *Odds* és el terme anglosaxó molt utilitzat en estadística i en apostes que expressa el quocient entre la probabilitat d'un esdeveniment dividit per la probabilitat inversa o contrària, és a dir, odds = p/(1-p). Creiem que traduir-ho per possibilitats és més entenedor que dir oportunitats i així s'utilitzarà en aquest text com en altres traduccions realitzades per als *Annals* en aquesta secció (N. dels T.).
2. D'acord amb el TERMCAT la mineria de dades és la tècnica informàtica que consisteix a analitzar un gran volum d'informació emmagatzemada en diferents bases de dades a fi de deduir patrons de coneixement que puguin generar aplicacions pràctiques (N. dels T.).
3. D'acord amb el TERMCAT, la seqüenciació d'alt rendiment és el procés de determinació de la seqüència d'aminoàcids d'una determinada proteïna o bé de la seqüència de nucleòtids d'una molècula de DNA o d'RNA efectuat amb un mètode que tracta una gran quantitat de dades de manera ràpida i eficaç (N. dels T.).