De Testimonio. Sobre l'evidència en les decisions de l'ús d'intervencions terapèutiques (*Harveian Oration, 2008*)

Michael Rawlins

Traducció de G. Permanyer-Miralda i J. M .V. Pons Ràfols de l'article: Rawlins M. De Testimonio: on the evidence for decisions about the use of therapeutic interventions. Clinical Medicine. 2008;8:579-88.

William Harvey (1578-1657) fou un dels "filòsofs naturals" del segle XVII que ja no estaven disposats a acceptar l'autoritat d'Aristòtil, Plató o Galè com un fonament segur per entendre el món natural. Com va dir el propi Harvey: "És indigne rebre instruccions dels comentaris d'altri sense examinar els objectes mateixos, sobretot quan el llibre de la natura roman tan obert i tan fàcil de consultar".

Encara que estaven units en la seva cerca per examinar el "llibre de la natura" per si mateixos, els "filòsofs naturals" de l'època tenien aspres divisions sobre com calia fer-ho. Tres-cents cinquanta anys després, aquesta disputa sobre la naturalesa de la ciència i el seu mètode encara persisteix, particularment pel que fa a les orientacions inductiva i deductiva per establir el coneixement científic.

Però les discussions no són enlloc tan apassionades, i a vegades violentes, com pel que fa a la naturalesa de l'evidència per donar suport a l'ús de les intervencions terapèutiques. Això ha estat posat de manifest especialment per l'aparició, en els darrers 30 anys, del que s'anomenen, de manera variada, "regles", "nivells" o "jerarquies" d'evidència. Se'n mostra un exemple a la Taula 1.

L'evidència, en el context actual, té només un objectiu. Representa la base per informar els decisors sobre l'ús apropiat de les intervencions terapèutiques en la pràctica clínica diària. Cal prendre aquestes decisions a diferents nivells però, invariablement, poden tenir conseqüències crítiques per als pacients, les famílies i les societats. Per exemple, els decisors determinen l'adequació dels tractaments oferts als pacients individuals, estableixen el ventall de medicaments que cal incloure en la farmàcia d'un hospital i se'ls pot encarregar que avaluïn si determinades intervencions són prou segures i efectives —i cost-efectives— per a l'ús de tot un sistema de salut. Els errors en la presa de decisions poden tenir repercussions dramàtiques a tots els nivells.

Les jerarquies situen els assajos clínics aleatoritzats (ACA) al cim, amb diferents tipus d'estudis observacionals arrecerats en els graons inferiors. Es fan servir, a manera d'esquema, per donar una noció sobre la "força" de l'evidència corresponent i, especialment, per part dels autors de guies, per tal de "graduar" les recomanacions terapèutiques en funció de la força observada.

TAULA 1. **Una jerarquia de l'evidència**

Nivell	Criteris
1++	Metanàlisis d'alta qualitat, revisions sistemàtiques d'ACA o ACA amb risc molt baix de biaix
1+	Metanàlisis ben executades, revisions sistemàtiques d'ACA o ACA amb baix risc de biaix
1 -	Metanàlisis, revisions sistemàtiques d'ACA o ACA amb alt risc de biaix
2++	Revisions sistemàtiques d'alta qualitat d'estudis de casos i controls o de cohorts, o estudis de casos i controls o de cohorts d'alta qualitat amb un risc molt baix de biaix, confusió o atzar
2+	Estudis de casos i controls o de cohorts ben executats, amb un risc baix de biaix, confusió o atzar
2 -	Estudis de casos i controls o de cohorts amb un risc alt de biaix, confusió o atzar
3	Estudis no analítics (p. ex. comunicació o estudis de casos)
4	Opinió d'experts

ACA: assajos clínics aleatoritzats

El concepte que l'evidència es pot distribuir de manera fiable en jerarquies és il·lusori. Aquestes jerarquies posen en un pedestal no gaire còmode els ACA: aquesta tècnica té avantatges, però té també inconvenients. Els estudis observacionals tenen defectes, però tenen també mèrits. Els decisors han d'avaluar i tenir en compte tota l'evidència que hi hagi, independentment de si procedeix d'ACA o d'estudis observacionals i, si se n'han d'obtenir conclusions raonables i dignes de confiança, cal entendre les fortaleses i debilitats de cadascun. I tampoc, per assolir aquestes conclusions, representa cap vergonya acceptar que cal fer judicis sobre si els components de les fonts d'evidència són "apropiats per al seu objectiu". Al contrari, els judicis són un ingredient fonamental de la majoria d'aspectes del procés de presa de decisions.

Assaigs clínics aleatoritzats

La introducció dels ACA, a la meitat del segle xx, ha tingut un impacte profund en la pràctica de la medicina i se'n coneixen bé les principals característiques. Representen la comparació d'una (o vàries) intervencions que han estat distribuïdes a l'atzar en grups de pacients tractats simultàniament.

No hi ha dubte que els ACA amb cegament doble, quan s'executen i analitzen correctament, donen confiança en la validesa interna dels seus resultats pel que fa als beneficis de la intervenció; més encara, si són replicats en estudis ulteriors. Per tant, els ACA són sovint qualificats com el "patró or" per demostrar (o refutar) els beneficis d'una intervenció determinada. Això no obstant, la tècnica té limitacions importants, quatre de les quals són especialment pertorbadores: la hipòtesi nul·la, la probabilitat, la generalitzabilitat i les implicacions sobre els recursos.

La hipòtesi nul·la

Tradicionalment, l'anàlisi d'un ACA s'ha basat en la hipòtesi nul·la, que pressuposa que no existeix diferència entre els tractaments i que es comprova per a una estimació de la probabilitat (o freqüència) d'obtenir un resultat tant o més extrem que l'observat si la hipòtesi nul·la fos certa. Si aquesta probabilitat és menor que un valor arbitrari —habitualment menys d'1 entre 20 (és a dir, $p < 0,05$)— aleshores es rebutja la hipòtesi nul·la. Aquesta és l'orientació "freqüentista" del disseny i anàlisi dels ACA i té un indubtable atractiu: els càlculs estadístics són relativament senzills. Se n'ha acceptat àmpliament la metodologia i els criteris de "significació" estan ben establerts.

Però la hipòtesi nul·la pot ser irrellevant si existeixen estudis previs que demostren que un tractament determinat té beneficis. Això pot succeir durant el desenvolupament d'un nou fàrmac, quan l'evidència preliminar de demostració d'efecte en estudis de fase 2 és investigada en fase 3 en grups més grans de pacients. En aquest moment, basar l'anàlisi dels resultats d'estudis de fase 3 en la hipòtesi nul·la sembla un contrasentit. Així mateix, la hipòtesi nul·la no és adequada quan s'ha demostrat benefici en estudis previs. Tot i això, hi ha revisions dels darrers 10 anys que mostren que el 73% dels ACA publicats en les principals revistes no fan en absolut cap intent sistemàtic de situar els seus resultats en el context de recerques prèvies.

La hipòtesi nul·la és encara més incòmoda en assajos que busquen si no hi ha diferència (equivalència), si hi ha un benefici no menor (no inferioritat) o si hi ha una diferència preespecificada (futilitat) entre els grups de tractament. Tots ells exigeixen suposicions prèvies sobre fins a quin punt les diferències entre els tractaments podrien ser rellevants o importants. Certament, la hipòtesi nul·la pot ser coherent metodològicament amb l'orientació deductiva de la ciència; però, tal com va observar Rothman, "sustentar la hipòtesi nul·la universal equival a suspendre la creença en el món real i, per tant, posar en dubte les premisses de l'empirisme".

Probabilitat

El valor de p. Segons l'orientació freqüentista, si el valor de p és prou petit, vol dir que la hipòtesi nul·la és falsa o que ha esdevingut alguna cosa molt inhabitual. De manera convencional, generalment es fa servir una probabilitat de menys del 5% (és a dir, $p < 0,05$) per distingir entre aquestes dues possibilitats. Però un valor de p major o menor de 0,05 no prova ni refuta (respectivament) la hipòtesi nul·la. Alguns, però no tots, dels problemes relacionats amb els valors de p es poden evitar expressant els resultats en forma d'intervals de confiança, que indiquen el grau d'incertesa, o manca de precisió, de l'estimació d'interès. No obstant, els valors de p i els intervals de confiança estan estretament interrelacionats.

Multiplicitat. Les dificultats per interpretar els valors freqüentistes de p encara es compliquen més quan es vol decidir, durant un assaig clínic, si s'ha d'acabar l'estudi prematurament o com avaluar els resultats en subgrups de pacients i si cal fer-ho una vegada conclòs l'estudi. Un problema semblant té lloc durant les anàlisis de seguretat d'un ACA. En tots aquests casos, les proves repetides de significació estadística —adoptant el valor convencional de $p < 0,05$— tenen cada cop major probabilitat d'obtenir un o més resultats "significatius". Per exemple, si s'analitzen 10 suposicions independents, hi ha una probabilitat del 40% que una sigui aparentment significativa (al nivell de $p < 0,05$). Això es coneix com "el problema de la multiplicitat". Però els estadístics tenen punts de vista molt discordants sobre com superar aquestes dificultats quan s'estableixen regles d'interrupció d'estudis (*stopping rules*) o anàlisis de subgrups.

Hi ha un desig natural entre els investigadors, durant el curs d'un ACA, de dur a terme anàlisis preliminars de les dades que van apareixent, per tal de decidir si l'estudi es continua o si s'interromp abans d'hora. Es pot justificar la finalització prematura segons el criteri que l'estudi ja ha assolit els seus objectius finals o a causa de la preocupació per la seguretat en un dels seus grups. Però la decisió d'acabar prematurament un assaig i de com fer-ho comporta nombrosos perills. Si una anàlisi preliminar mostra un benefici inesperat, pot resultar difícil diferenciar un efecte real d'una casualitat (el que s'anomena un "alt valor aleatori").

Per resoldre el problema de la multiplicitat s'han elaborat diferents mètodes. Molts es basen en canviar el nivell de significació estadística (és a dir, el valor de p) a cada anàlisi preliminar que es duu a terme, de manera que en exàmens prematurs de les dades es necessiti un valor més baix de p per refusar la hipòtesi nul·la. Però no hi ha con-

sens entre els estadístics pel que fa a quines han de ser les regles d'interrupció. Per això, la solució del problema ha esdevingut urgent. Ja que interrompre prematurament els estudis degut a un benefici pot sobreestimar sistemàticament els seus efectes, hi ha un perill autèntic que algunes declaracions d'eficàcia —especialment en oncologia— puguin ser injustificades.

Les anàlisis dels efectes d'una intervenció en subgrups de pacients poden ser importants per tal d'establir si diferents tipus de persones responen de manera diferent. La solució més habitual dels problemes de multiplicitat en les anàlisis de subgrups és acceptar com a dignes de confiança només un nombre limitat de grups, biològicament o clínicament plausibles, que hagin estat preestablerts durant la fase de disseny. En termes generals, no s'ofereix una definició de què s'ha de considerar "limitat". Les opinions sobre l'avaluació de subgrups una vegada l'assaig ha finalitzat varien. Algunes rebutgen completament totes les anàlisis *post hoc*, mentre que altres suggereixen cautelosos ajustos estadístics del valor de p.

L'orientació bayesiana. Hi ha cada cop més estadístics que opinen que la solució de moltes de les dificultats inherents a l'orientació freqüentista del disseny, anàlisi i interpretació dels ACA rau en un ús més gran de l'estadística bayesiana. L'orientació bayesiana de la probabilitat deu el seu nom a Thomas Bayes (1701-1761), que fou sacerdot no conformista a Tunbridge Wells. Aquesta noció de la probabilitat —la probabilitat subjectiva o inversa— és la versemblança d'una hipòtesi a partir de certes dades. Així, mentre que l'orientació freqüentista es refereix a la probabilitat d'unes dades condicionades a una hipòtesi específica (habitualment la hipòtesi nul·la), l'orientació bayesiana és la inversa (probabilitat d'una hipòtesi condicionada a unes dades).

El teorema de Bayes relaciona les probabilitats a partir d'allò que se sap abans (*a priori)* d'un experiment, tal com un ACA, amb les probabilitats recalculades després de l'experiment (*a posteriori*). L'enllaç entre la probabilitat "prèvia" i la "posterior" és el propi experiment. La probabilitat "posterior" representa una estimació de la probabilitat d'una hipòtesi condicionada a les dades observades, però tenint en compte allò que ja se sabia (la "prèvia") abans de realitzar l'experiment.

La Figura 1 mostra l'aplicació de l'orientació bayesiana a l'anàlisi d'un ACA. L'assaig GREAT va ser dissenyat per avaluar la hipòtesi que el tractament trombolític precoç, ja al domicili del pacient, en l'infart de miocardi podia ser millor que el tractament més tardà a l'hospital. Per tant, els investigadors començaren un ACA que comparava l'efectivitat de la trombòlisi administrada pels metges generals en el mateix domicili del pacient amb el tractament administrat quan havia arribat a l'hospital local. Al cap de tres mesos, la reducció relativa de mortalitat de qualsevol causa fou del 49% (p = 0,04) en els pacients tractats a casa

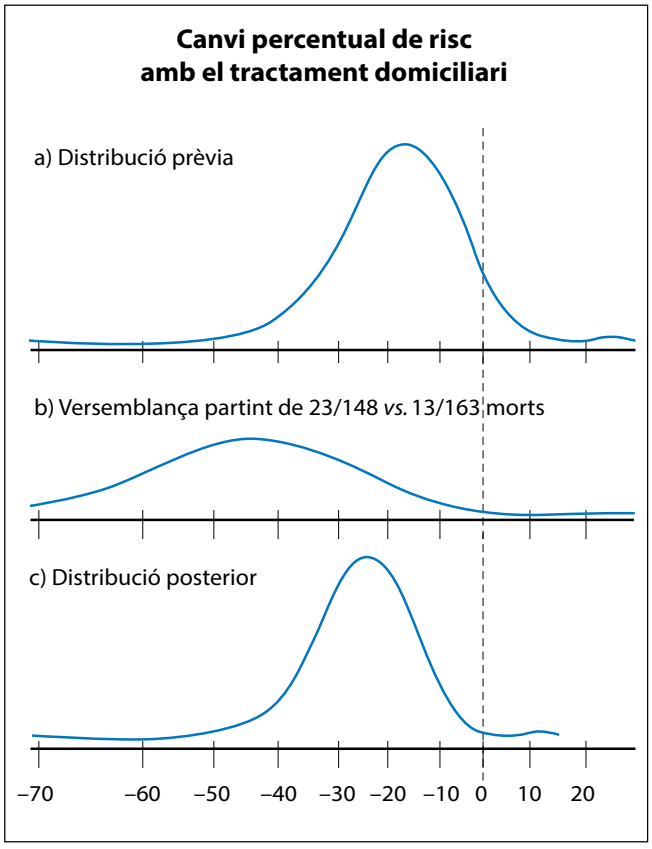

FIGURA 1. **Reanàlisi bayesiana de l'assaig GREAT. Mostren el canvi (reducció) en la mortalitat amb la teràpia trombolítica domiciliària en comparació amb el tractament hospitalari: a) distribució de la probabilitat prèvia del tractament domiciliari, b) distribució de la probabilitat de versemblança de l'assaig GREAT, c) distribució de la probabilitat posterior del tractament domiciliari utilitzant el teorema de Bayes**

comparats amb aquells que només reberen tractament en arribar a l'hospital. Encara que és ben possible que la trombòlisi més precoç pogués representar una millor supervivència, una reducció de la mortalitat de gairebé el 50% va semblar poc plausible ja que la trombòlisi hospitalària ja la redueix un 25%.

Per tant, Pocock i Spiegelhalter varen dur a terme una reanàlisi bayesiana (Figura. 1). Varen elaborar una distribució "prèvia" (Figura 1a) que es basava en resultats anteriors d'ACA sobre trombòlisi hospitalària, però manifestant la seva opinió que una reducció de mortalitat del 15% al 20% era molt plausible i que els extrems, tant d'una manca d'efecte com d'una reducció major del 40%, eren tots dos improbables. La Figura 1b mostra la distribució de probabilitat de versemblança dels resultats de l'estudi GREAT. La Figura 1c representa la distribució posterior obtinguda multiplicant la "prèvia" i la versemblança. Aquesta figura il·lustra com la versemblança derivada de l'anàlisi original de l'estudi GREAT ha estat "enretirada" i dona una representació formal de l'opinió que els resultats originals eren "massa bons per ser certs".

A més de per evitar l'ús indiscriminat de la hipòtesi nul·la, es defensa que l'orientació bayesiana resol problemes associats amb el disseny i anàlisi dels ACA, i també amb qüestions de multiplicitat. Per què no es fan servir més els mètodes bayesians? Sembla que hi ha cinc raons principals.

Primer, encara que l'orientació subjectiva de l'estudi de la probabilitat data ja del segle XVIII, alguns (especialment els que tenen un marc mental freqüentista) miren amb disgust aquesta interpretació —pel seu element d'opinió o judici personal—. Prefereixen la seguretat aparent (però il·lusòria) d'una definició clara d'allò que constitueix un resultat extrem quan se'l contrasta amb la hipòtesi nul·la i són reticents a acceptar que l'opinió o el judici personals puguin intervenir en la presa de decisions.

Segon, hi ha hagut controvèrsies substancials sobre l'obtenció de la probabilitat prèvia. Quan existeix evidència a partir d'estudis anteriors es pot tenir fàcilment allò que s'anomena una "prèvia clínica". Quan no hi ha una "prèvia clínica" es fa ús de "prèvies per defecte". S'ha exagerat massa les suposades dificultats d'utilitzar-les i, en qualsevol cas, els bayesians acostumen a utilitzar un nombre de "prèvies per defecte" en absència, i fins i tot a vegades en presència, de "prèvies clíniques", com a part de les seves anàlisis de sensibilitat.

En tercer lloc, les anàlisis bayesianes són de computació complexa i més exigents que els mètodes que habitualment s'utilitzen en la majoria d'anàlisis freqüentistes.

En quart lloc, alguns estadístics —però cada cop menys— no estan familiaritzats amb les tècniques d'anàlisi bayesiana i no volen (o no poden) adaptar-s'hi. Alguns atribueixen aquesta heterogeneïtat en les competències al tipus original d'universitat on s'havia format cada estadístic. Altres, menys amablement, pensen que és una qüestió generacional. Tal com em va dir un bayesià: "Habitualment, els estadístics que varen ser formats en l'ús de llibres de registre i regles de càlcul no saben fer estadística bayesiana".

Finalment, les autoritats reguladores han tingut dubtes a vegades en acceptar que l'orientació bayesiana pot tenir avantatges. No obstant, hi ha senyals d'un augment d'interès, especialment pel que fa a l'avaluació de dispositius. I els propis fabricants cada cop adopten més orientacions bayesianes en assajos en fase 2 i fase 3.

És probable que, en el futur, les orientacions bayesianes juguin un paper molt més gran. Eliminar la rigidesa del valor de p i resoldre alguna de les qüestions de la multiplicitat són premis que val la pena assegurar. Per damunt de tot, les orientacions bayesianes poden ajudar els decisors a arribar a conclusions més apropiades.

Generalitzabilitat

Per regla general, els ACA es duen a terme en poblacions seleccionades de pacients durant un període de temps delimitat i, habitualment, relativament curt. En la pràctica clínica, la intervenció s'aplicarà a una població més heterogènia de pacients —habitualment amb comorbiditats— i sovint durant períodes de temps molt més llargs. Conèixer fins a quin punt les troballes dels ACA tenen validesa externa i es poden extrapolar o generalitzar a poblacions més àmplies de pacients és una qüestió que s'ha fet cada cop més important. Els seus problemes més destacats s'exposen a la Taula 2. En un altre lloc (veure al final de

TAULA 2. **Influències adverses en la generalitzabilitat dels resultats dels assajos clínics aleatoritzats (ACA)**

Factors	Elements	Problemes potencials
Pacients	Edat	Efectivitat en pacients més joves o més grans
	Gènere	Efectivitat en general
	Gravetat de la malaltia	Efectivitat en formes més greus o més lleus de la malaltia
	Factors de ric	Efectivitat en pacients amb factors de risc per a la malaltia (p. ex. fumadors)
	Comorbiditats	Influència d'altres malalties en l'efectivitat
	Ètnia	Efectivitat en altres grups ètnics
	Estatus socioeconòmic	Efectivitat en pacients amb desavantatge
Tractament	Dosi	Dosi massa alta en els ACA
	Temporalitat	Influència de l'adherència (compliment) al règim de tractament
	Durada del tractament	Efectivitat durant l'ús a llarg termini
	Comedicació	Interaccions adverses
	Efectivitat comparativa	Efectivitat en comparació amb altres productes utilitzats per a la mateixa indicació
Context	Qualitat de l'atenció	Prescripció i supervisió per proveïdors sanitaris menys especialitzats o experts

l'article N. dels T.) comento amb més detalls el fet que la qüestió de la generalitzabilitat planteja autèntics dubtes. Per exemple, Bartlett i col·ls. varen revisar els criteris d'exclusió adoptats en ACA amb estatines (27 estudis) i antiinflamatoris no esteroidals (AINE) (25 estudis). Varen comprovar que hi estaven infrarepresentats les dones, els vells i les minories ètniques en comparació amb el seu ús en la població general. S'ha observat una infrarepresentació equivalent en ACA d'altres intervencions cardiovasculars.

Avaluació de beneficis. Per tant, hi ha incertesa sobre si els beneficis obtinguts pels pacients "mitjans" dels ACA es poden extrapolar als pacients "mitjans" que reben atenció clínica ordinària. Per exemple, importa realment la infrarepresentació d'alguns grups en els ACA? Alguns autors fan la suposició que els resultats dels ACA en poblacions ben seleccionades de pacients, si altres condicions són iguals, es poden extrapolar amb fiabilitat a la cura de pacients en general. Es defensa que, si la patogènia d'una malaltia és la mateixa en tots els subgrups, es poden esperar beneficis similars en poblacions de pacients més àmplies.

El problema d'aquesta posició és que hi ha poca evidència sistemàtica que la recolzi i n'hi ha alguna que la refuta. Indubtablement, hi ha estudis aïllats que demostren una concordança entre els efectes beneficiosos vistos als ACA i els que s'observen en el tractament mèdic convencional. En són un bon exemple els beneficis de l'anticoagulació en pacients amb fibril·lació auricular no valvular. Però el que roman incert és fins a quin punt les característiques diferencials dels pacients tractats en ACA, comparades amb les dels que reben tractament rutinari, tenen realment importància pel que fa als beneficis suposats. Certament, com els propis autors del grup CONSORT reconeixen, "la validesa externa és una qüestió de judici".

Avaluació de danys. Malgrat haver-hi optimisme pel que fa a la generalitzabilitat dels ACA en relació amb l'eficàcia (tot i que amb un bon gruix d'incertesa), l'experiència mostra que els ACA són febles per obtenir evidència rellevant en l'avaluació dels danys. Els ACA poden detectar qüestions "dramàtiques" de seguretat, però donen una orientació poc fiable.

Actualment és habitual que en els ACA es detectin i registrin tots els esdeveniments adversos que tenen lloc després de l'aleatorització. Això redueix la probabilitat d'un biaix de l'investigador en la interpretació de la naturalesa causal de qualsevol malaltia intercurrent, que alguns malalts desenvoluparan inevitablement en el curs d'un estudi. Els esdeveniments adversos inclouen símptomes i signes anormals, anomalies detectades amb anàlisis bioquímiques de rutina (recomptes hemàtics, urea i electròlits, proves de funció hepàtica, anàlisis d'orina, etc.) i els resultats d'estudis especials (per exemple, electrocardiograma o ecocardiograma). Teòricament, els esdeveniments adversos relacionats de manera causal amb la intervenció es poden identificar simplement comparant els grups. Encara que a primera vista aquesta orientació té atractius, planteja diferents problemes.

Els ACA estan dissenyats per assegurar que la potència estadística serà suficient per demostrar benefici clínic. Però els càlculs de potència generalment no tenen en consideració els danys. En conseqüència, encara que els ACA poden identificar les reaccions adverses més freqüents, de manera singular passen per alt les menys freqüents o les que tenen una latència més llarga (com ara els càncers). La majoria dels ACA, fins i tot per a intervencions que és probable que els usuaris utilitzin durant anys, només tenen una durada de sis a 24 mesos. I si es detecten efectes adversos a un nivell de significació estadística, és fàcil que siguin menystinguts com efectes de l'atzar més que a causa d'una diferència real entre els grups.

L'anàlisi de danys a partir dels ACA planteja un altre problema no resolt de multiplicitat. És gairebé inevitable que, en estudis a gran escala i a llarg termini, s'observi algun efecte estadísticament significatiu. Diferenciar els que són iatrogènics dels que són intercurrents i no causals, o un simple error per atzar, és tant un art com una ciència. Quan els esdeveniments són clarament iatrogènics (per exemple anafilaxi, erupcions morbil·liformes, necròlisi epidèrmica tòxica) se'n pot deduir una relació causal. Igualment, també pot suposar-se aquesta causalitat si els efectes adversos són biològicament plausibles (per exemple, càncer de mama després de tractament hormonal substitutiu [THS]). Quan no es tracta d'aquests factors, poden aparèixer dificultats d'interpretació. Certament, els assajos clínics ben executats i analitzats poden aportar informació important sobre efectes adversos. Entre els exemples hi ha els ACA de tractament antiarítmic profilàctic amb fàrmacs de classe I després d'un infart de miocardi i els de THS en dones postmenopàusiques. Però aquests són excepcions.

Recursos

Els costos dels ACA són substancials, en diners, temps i energies. La Figura 2 mostra el ventall de costos de 153 ACA que es van realitzar els anys 2005 i 2006. En aquestes dades es combinen els costos d'assajos finançats pel National Institute for Health Research i el Medical Research Council amb els corresponents a tres grans companyies farmacèutiques en els seus assajos en fase 2 i 3. La mediana dels costos va ser de 3.202.000 lliures, amb un interval interquartílic de 1.929.000 a 6.568.000. Aquestes dades no són exhaustives ni necessàriament representatives dels ACA en general, però il·lustren que els ACA poden ser una activitat molt cara. I sembla que els seus costos pugen. Segons estimació d'un fabricant, el cost mitjà per pacient inclòs en assajos ha augmentat de 6.300 lliures el 2005 fins a 7.300 el 2006 i 9.000 el 2007.

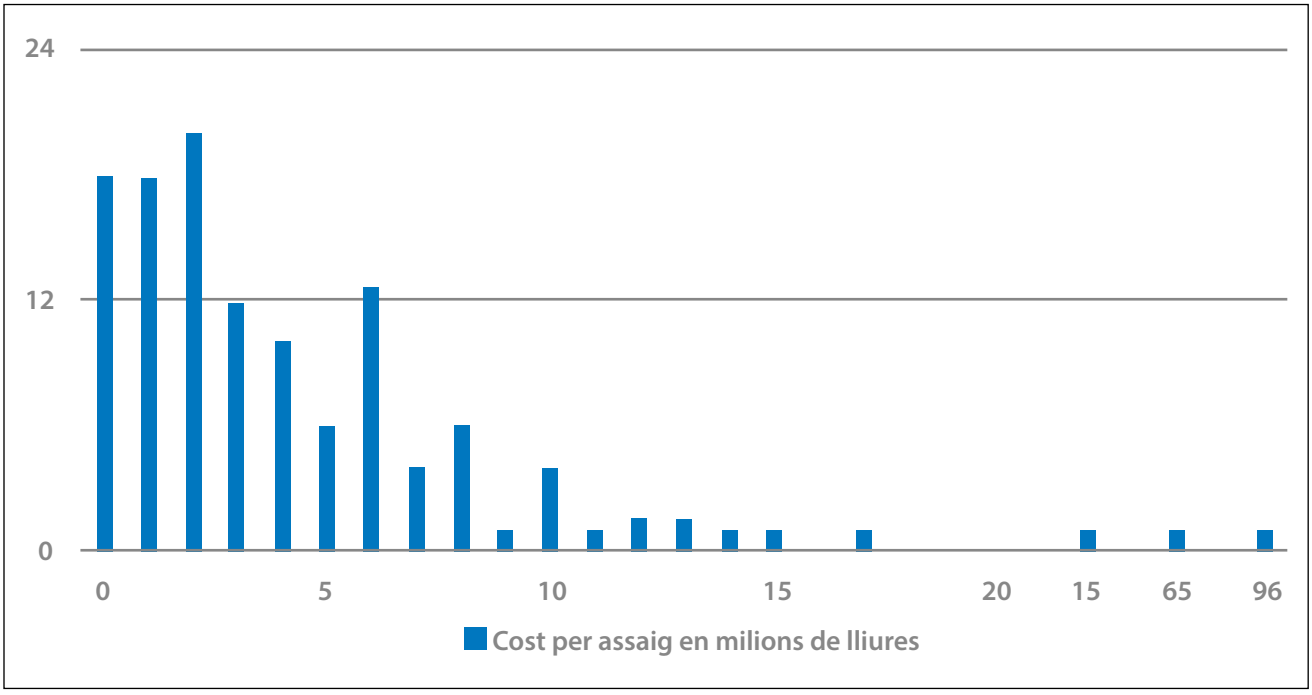

FIGURA 2. **Rangs de costos (milions de lliures) dels assajos comparatius i aleatoritzats (farmacèutics)**

Bona part d'aquest augment de costos es deu a les exigències reguladores i d'altres tipus que, en els darrers anys, han estat imposades als assajos amb finançament públic i privat. Cadascuna d'aquestes mesures ha estat introduïda amb la millor de les intencions. Entre aquestes hi ha el desig de protegir el pacient dels investigadors i finançadors poc escrupolosos, assegurar la recollida i comunicació puntual dels esdeveniments adversos durant l'assaig, auditar els documents de comunicació de casos individuals i evitar així les conseqüències de falsejament per part dels investigadors, i així successivament. Però fins i tot estudis senzills, amb productes dels que es disposa ja des de fa molts anys, ara imposen un repte burocràtic massiu a finançadors i investigadors potencials, tant si pertanyen a universitats com al sector privat.

Hi ha propostes recents dels propis investigadors clínics acadèmics que indiquen que els costos dels assajos clínics es podrien reduir entre un 40% i un 60% sense detriment de la seva qualitat. Mesures senzilles per reduir la càrrega burocràtica, com la captura electrònica de dades, la reducció de l'extensió dels quaderns de recollida de dades i modificacions de les activitats de maneig en cada centre, podrien reduir els costos de manera substancial.

Estudis observacionals

La nomenclatura que descriu els estudis observacionals (no aleatoritzats) és confusa. Evito la distinció entre estudis "comparatius" i "no comparatius" perquè tots els estudis observacionals comporten alguna forma de comparació implícita (informal) o explícita (formal). Tampoc trobo que els termes "estudis de cohorts" o "estudis quasi-experimentals" siguin gaire il·luminadors. El primer agrupa estudis que, en realitat, són entitats diferents. El segon és un terme que mai no he vist definit de manera adequada, ni tan sols coherent. Els tipus d'estudis observacionals que han estat i continuen essent usats per obtenir evidència sobre el beneficis i danys d'intervencions terapèutiques es presenten a la Taula 3.

La gran fortalesa dels ACA és que l'assignació dels tractaments es fa a l'atzar, de manera que els grups comparats són similars en els factors de base. Això pot no ser així en els estudis observacionals comparatius, on hi ha un autèntic risc de biaix de selecció i confusió. Certament, existeix una abundant (i a vegades encesa) literatura que compara els mèrits i els demèrits dels estudis aleatoritzats i els observacionals de l'efectivitat de les intervencions terapèutiques.

No obstant, els intents de revisió sistemàtica de les comparacions publicades entre les dues orientacions han estat emmetzinats per dos problemes. En primer lloc, hi ha les dificultats per identificar els estudis rellevants. Ja que molts estudis observacionals no han estat "etiquetats" de manera coherent en les bases de dades bibliogràfiques electròniques, és difícil assegurar que les tècniques de recerca convencional els han identificat de manera no esbiaixada. Per tant, molts revisors s'han basat en recollides personals d'articles, en els seus arxius propis (o d'altri) o en estudis identificats en revisions sistemàtiques prèvies. Per tant, no és pas petita la possibilitat d'un "biaix del revisor". La segona dificultat és que ben poques d'aquestes

TAULA 3. **Tipus d'estudis observacionals**

Estudis amb controls històrics
Estudis dels efectes d'una intervenció en un grup de pacients tractats amb una intervenció comparats retrospectivament amb un grup que prèviament havia rebut tractament convencional (la millor teràpia de suport inclosa)
Estudis controlats contemporàniament no aleatoritzats
Comparació dels resultats de pacients sotmesos a un tractament comparats amb un altre grup de pacients, no tractats o tractats amb una intervenció alternativa durant el mateix període de temps
Estudis de casos i controls
Comparació de l'ús d'una intervenció en grups de malalts amb i sense una malaltia o condició particular
Dissenys abans-després
Observacions en grups de pacients abans i després de ser tractats amb una intervenció. Per tant, els pacients actuen com els seus propis controls. Sovint, aquesta tècnica ha estat utilitzada en treballs amb controls històrics implícits on la història natural de la malaltia o condició està ben establerta i és previsible
Sèries de casos
Resultats d'un grup (sèrie) de pacients tractats amb una intervenció durant la pràctica clínica rutinària. Encara que no hi ha un grup formal de control, invariablement es fan comparacions implícites o explícites
Comunicacions de casos
Comunicacions de casos (anècdotes) de danys en pacients individuals, fetes tant a la literatura com a una agència central (com la Medicines Healthcare Products Regulatory Agency al Regne Unit o la Food and Drug Administration als EUA)

revisions han tingut en compte les diferències entre els diferents tipus de dissenys observacionals.

Hi ha un acord general que els efectes "espectaculars" es poden observar sense necessitat d'ACA. En canvi, hi ha molt menys consens sobre el paper dels estudis observacionals per precisar el benefici quan la magnitud de l'efecte és més modesta. Pot passar que hi hagi una tendència que els estudis observacionals mostrin efectes del tractament més grans que els ACA. No obstant, això no ha estat una observació invariable. De fet, en alguns casos s'han observat tant subestimacions com sobreestimacions. La magnitud de les diferències entre les dades dels ACA i els estudis observacionals també pot variar segons els tipus específics de disseny utilitzats en aquests. En altres treballs s'analitzen les estratègies analítiques per reduir els efectes dels biaixos de selecció i confusió en estudis observacionals.

Aquí considerem amb detall dues varietats d'estudi observacional perquè han estat especialment importants per obtenir evidència sobre els beneficis i danys d'intervencions terapèutiques. En una altra publicació (veure al final de l'article N. dels T.) es pot trobar una anàlisi més completa de les altres varietats, que també han fet contribucions extremadament importants.

Assajos amb controls històrics

A la Taula 4 es mostren exemples d'intervencions de benefici indubtable, demostrat en estudis amb control històric, on es feia la comparació entre una nova intervenció i l'experiència anterior amb la malaltia.

En temps passats, l'ús de controls històrics havia estat objecte de fortes crítiques. No obstant, a finals dels anys 1980, els investigadors clínics es varen tornar menys hostils cap a aquest concepte. Estimulats per la

TAULA 4. **Algunes intervencions amb efectivitat establerta a partir d'assajos amb controls històrics**

Intervenció (any)	Indicació
Tiroxina (1891)	Mixedema
Insulina (1922)	Cetoacidosi diabètica
Vitamina B_{12} (1926)	Anèmia perniciosa
Sulfamides (1937)	Sèpsia puerperal
Desfibril·lació (1948)	Fibril·lació ventricular
Estreptomicina (1948)	Meningitis tuberculosa
Blocadors ganglionars (1959)	Hipertensió maligna
Maniobra de Heimlich (1975)	Obstrucció laríngia per un cos estrany
Cisplatí més vinblastina i bleomicina (1977)	Càncer testicular disseminat
N-acetilcisteïna (1979)	Intoxicació per paracetamol
Ganciclovir (1986)	Retinitis per citomegalovirus
Tractament amb làser (2000)	Taques de vi de Porto
Imatinib (2002)	Leucèmia mieloide crònica

creixent epidèmia de sida, varen acceptar que "alguna de les orientacions tradicionals de l'assaig clínic eren innecessàriament rígides". Aquests autors proposaren que, per recolzar l'eficàcia d'un nou fàrmac per tractar la sida, els assajos amb controls històrics haurien de complir les següents exigències: 1) no hi ha d'haver un altre tractament apropiat per servir de control; 2) hi ha d'haver experiència suficient per assegurar que els pacients no tractats tenen un pronòstic uniformement dolent; 3) cal esperar que el tractament no tingui efectes colaterals que comprometin el benefici potencial per al pacient; 4) ha d'haver-hi una expectativa justificada que el benefici potencial per al malalt serà prou gran com per fer que la interpretació d'un estudi no aleatoritzat no sigui ambigua; i 5) el fonament científic del tractament ha de ser prou robust com per ser àmpliament acceptat un resultat positiu.

La meva adaptació personal d'aquestes exigències per als assajos amb controls històrics està influenciada de manera general, i ho manifesto obertament, per les consideracions expressades per Bradford Hill per distingir associacions causals i no causals en els estudis epidemiològics. Per tant, considero que, generalment, els assajos amb controls històrics haurien de ser acceptats com a evidència d'efectivitat a condició de complir totes les següents condicions:

1. El tractament ha de tenir una base biològicament plausible. Tots els tractaments de la Taula 4 compleixen aquest requisit.
2. No hi ha d'haver cap tractament apropiat que pugui ser raonablement utilitzat com a control. El terme "apropiat" exclouria, per exemple, l'ús de transplantament de medul·la òssia com a control de tractament enzimàtic substitutiu per a la malaltia de Gaucher.
3. La malaltia ha de tenir una història natural ben establerta i previsible. Prefereixo aquesta expressió més que no pas "mal pronòstic". Malalties com les "taques de vi de Porto" poden afectar prou la qualitat de vida dels pacients sense afectar la seva esperança de vida.
4. No s'ha d'esperar que el tractament tingui efectes indesitjables que comprometrien els seus beneficis. Aquesta condició ha de ser *sine qua non*.
5. Hi ha d'haver una esperança raonable que l'efecte del tractament serà prou gran com per fer que la interpretació del benefici no sigui ambigua. Un quocient "senyal/soroll" de 10 o més sembla intensament suggestiu d'un efecte terapèutic genuí. No obstant, la magnitud d'un quocient "senyal/soroll" indicatiu d'un efecte "espectacular" (és a dir, de 10 vegades) està basada en una impressió i, fins ara, no està fonamentada en cap evidència empírica sòlida.

En el futur, hi haurà circumstàncies en què haurem d'estar preparats per acceptar l'evidència de benefici a partir d'assajos amb controls històrics. Entre les intervencions d'aquest tipus hi podria haver, per exemple, tractaments que aturessin completament la neurodegeneració progressiva que es veu en les malalties de Creutzfeldt-Jakob o de Huntington. En aquestes dues entitats hi ha mesures objectives, i també subjectives, per confirmar (o refutar) que s'ha aturat la seva progressió. El fet que, en el moment actual, hi ha investigadors al Regne Unit, Europa i els EUA que acumulen cohorts de pacients amb aquestes malalties —amb el propòsit específic de tenir controls històrics en estudis futurs— em fa sentir optimista.

Estudis de casos i controls

Els estudis de casos i controls comparen l'ús d'una intervenció en grups de persones amb i sense una malaltia o condició particular. Igual que altres dissenys observacionals, aporten informació sobre una *associació* entre l'exposició a una intervenció particular, però no indiquen necessàriament que la relació sigui *causal*. Els problemes del biaix de selecció i confusió no són pas menys importants en la interpretació dels estudis de casos i controls que els que ho són en altres dissenys observacionals controlats. No obstant, poden ser minimitzats per la cura en el disseny i l'anàlisi.

Avaluació de *beneficis*. Els estudis de casos i controls s'han utilitzat, però amb resultats variables, per recolzar la demostració del benefici d'intervencions.

Durant els anys 1980, alguns estudis observacionals (sobretot de casos i controls) varen suggerir que l'ús de THS a llarg termini s'associava a una reducció significativa de la cardiopatia isquèmica. Les revisions quantitatives fetes a començaments de la dècada de 1990 indicaven que el risc relatiu en les usuàries, comparat amb el de les no usuàries, podia estar associat amb una reducció de fins al 50%. En base a això, el THS va representar el tractament farmacològic amb major prescripció als EUA.

Actualment se sap, a partir de diferents ACA grans i ben executats, que el THS no té cap efecte beneficiós en la cardiopatia isquèmica i que pot augmentar el risc d'ictus. Les discrepàncies entre els resultats dels estudis observacionals i els ACA sobre els beneficis que se suposaven del THS eren en bona part deguts a un biaix de selecció. Si els estudis observacionals haguessin tingut en compte l'edat, el nivell socioeconòmic, l'hàbit tabàquic i la durada del tractament, molts dels seus suposats avantatges (però no tots) haurien desaparegut. I algunes dones han pagat un preu elevat per aquest error.

No obstant, hi ha hagut circumstàncies en què els estudis de casos i controls han aportat indicacions significatives dels beneficis de les intervencions. Entre elles hi ha els efectes protectors de l'aspirina sobre l'infart agut de miocardi, la prevenció dels defectes del tub neural amb àcid fòlic, la relació entre la posició en el son i la síndrome de la mort sobtada en el lactant i l'efecte

protector del càncer colorectal dels AINE. En el futur ens caldrà desenvolupar orientacions que ens permetin tenir confiança en els estudis observacionals de manera general; en particular, els estudis de casos i controls poden donar informacions que permetin fer suposicions raonables sobre la validesa interna. Les noves tècniques, com ara l'aleatorització mendeliana, poden ser una bona ajuda. Calen més recursos, temps i energies fent recerca metodològica si es vol basar la causalitat de manera més segura en l'evidència observacional.

Avaluació de danys. Contrastant amb les dificultats de valorar els beneficis de les intervencions amb disseny de casos i controls, aquest mètode ha estat extremadament important per identificar relacions causals entre intervencions específiques i els seus efectes adversos. A la Taula 5 se'n mostren alguns exemples. Els estudis de casos i controls també han estat útils per tranquil·litzar, quan han demostrat que alguns suposats efectes indesitjables "descoberts" per mètodes de declaració espontanis no són problemàtics. Entre els seus exemples hi ha la sospita d'associació entre els bifosfonats i la fibril·lació auricular o entre els broncodilatadors simpaticomimètics i un excés de morts per asma.

Però el biaix de selecció i la confusió per indicació també poden tenir lloc en estudis de casos i controls dissenyats per investigar danys. Per exemple, el 1974, tres estudis de casos i controls publicats simultàniament suggerien una associació entre la reserpina, fàrmac per al tractament de la hipertensió, i el desenvolupament ulterior de càncer de mama. En altres estudis, publicats més tard, no es va confirmar l'associació descrita, que sembla haver estat conseqüència d'haver exclòs, com a controls, les pacients amb malaltia cardiovascular. En aquest cas, una forma subtil de biaix de selecció (el biaix d'exclusió) va ser probablement el responsable de les conclusions errònies a què es va arribar primer.

Jerarquies d'evidència

La primera jerarquia d'evidència es va publicar a finals dels anys 70. Des d'aleshores n'han aparegut moltes de similars en la literatura, amb rebuscament i complexitat creixents. L'any 2002, una revisió va identificar 40 d'aquests sistemes i un estudi més recent, el 2006, en trobà 20 més.

En la jerarquia de la Taula 1, com en altres, els ACA estan situats al nivell més alt, amb una posició més baixa per a la informació basada en estudis observacionals. Aquesta aproximació jeràrquica a l'evidència ha estat no tan sols adoptada per molts integrants del moviment de la medicina basada en l'evidència i en l'avaluació de tecnologies, sinó que ha arribat a dominar el desenvolupament de les guies de pràctica clínica. No obstant, donar tal preeminència als resultats dels ACA és poc raonable. Tal com va dir pertinentment Bradford Hill, l'arquitecte de l'ACA: "Qualsevol creença que l'assaig comparatiu és l'únic mètode no voldria pas dir que el pèndol ha anat massa enllà, sinó que s'ha desenganxat del tot".

Tal com he comentat, els ACA són especialment febles pel que fa a la generalitzabilitat i, molt especialment, a l'avaluació de danys. Encara que, certament, els ACA poden identificar aquells efectes adversos que tenen lloc relativament sovint i durant la breu escala temporal de la seva realització, romanen limitacions significatives. Contràriament a una afirmació recent, només els estudis observacionals poden proporcionar l'evidència que es necessita per avaluar danys menys freqüents o de major latència.

Encara més, les jerarquies no poden fer lloc a l'evidència que es basa en combinar els resultats dels ACA i els estudis observacionals. Combinar l'evidència derivada d'un ventall de dissenys d'estudis és una característica dels models d'anàlisi de decisions, igual que ho és dels camps emergents d'assajos amb teleanàlisi i assajos segons preferències dels pacients.

TAULA 5. **Alguns efectes adversos confirmats per estudis de casos i controls**

Intervenció (any de publicació)	Efecte advers
Anticonceptius orals (1967)	Tromboembolisme venós
Estilbestrol durant l'embaràs (1972)	Carcinoma del tracte genital (en dones joves)
Aspirina en nens (1985)	Síndrome de Reye
L-triptòfan	Síndrome eosinofília-miàlgia
Antiinflamatoris no esteroidals (1994)	Hemorràgia digestiva alta
Tractament hormonal substitutiu (1996)	Tromboembolisme venós
Tractament hormonal substitutiu (1997)	Càncer de mama
Inhibidors selectius de la recaptació de serotonina (1999)	Hemorràgia digestiva alta
Anticonvulsivants (1999)	Síndrome de Stevens-Johnson i necròlisi epidèrmica tòxica
Olanzapina (2002)	Diabetis
Fluoroquinolones (2002)	Trastorns del tendó d'Aquil·les

A part simplement del seu nombre, les incongruències entre jerarquies demostren la seva naturalesa insatisfactòria. Entre aquestes hi ha la variable prominència atorgada a les metanàlisis, que en algunes jerarquies estan situades per damunt d'ACA grans i de qualitat, mentre que altres les ignoren. També hi ha incongruències entre jerarquies en la seva graduació dels estudis observacionals: algunes puntuen més alt els estudis de cohorts que els de casos i controls, altres consideren que tots són iguals i altres inverteixen l'ordre.

Les jerarquies pretenen substituir el judici per una avaluació simplista i pseudoquantitativa de la qualitat de l'evidència existent. Els decisors han d'incorporar els judicis com a part de la seva avaluació de l'evidència per tal d'arribar a les seves conclusions. Aquests judicis fan referència al grau amb què cadascun dels elements de la base de l'evidència és "adequat per al seu objectiu". És fiable? És generalitzable? Compensen els beneficis de la intervenció els seus danys? I així successivament.

Consideracions finals

L'experimentació, l'observació i les matemàtiques —individualment i col·lectiva— tenen un paper crucial a jugar per obtenir la base d'evidència de la terapèutica moderna. Les disputes sobre la importància relativa de cadascuna són una distracció innecessària. Cal substituir les jerarquies de l'evidència per l'acceptació —i encara més, per l'adopció entusiasta— d'una varietat d'orientacions. Això no és un al·legat per abandonar els ACA i substituir-los per estudis observacionals. Tampoc és una declaració que l'orientació bayesiana del disseny i anàlisi de dades experimentals i no experimentals hauria de suplantar tots els altres mètodes estadístics. Al contrari, és un al·legat perquè els investigadors continuïn desenvolupant i millorant els seus mètodes, perquè els decisors evitin l'adopció de posicions aferrissades sobre la naturalesa de l'evidència i perquè tots acceptin que la interpretació de l'evidència demana judici.

A aquells que arrosseguen dubtes sobre la naturalesa de la pròpia evidència, els recordo que mentre Gregor Mendel (1822-1884) desenvolupava la teoria monogènica de l'herència a partir de l'experimentació, Charles Darwin (1809-1882) va concebre la teoria de l'evolució, com a resultat de l'observació atenta, i que la teoria de la relativitat d'Albert Einstein (1879-1955) era una descripció matemàtica d'alguns aspectes del món que ens envolta. El descobriment de la circulació de la sang per William Harvey —com ho va descriure a *De motu cordis*— es va basar en una síntesi elegant de les tres formes d'evidència.

N. dels T.: L'autor fa referència a la versió completa de la seva conferència: *De Testimonio. On the evidence for decisions about the use of therapeutic interventions. The Harveian Oration delivered before the Fellows of The Royal College of Physicians of London on Thursday 16 October 2008 by Professor Sir Michael David Rawlins*. Londres, Royal College of Physicians; 2008.