Öppna data kan öka kunskapsmassan och motverka fusk

EU, VETENSKAPSRÅDET, FINANSIÄRER OCH VETENSKAPLIGA TIDSKRIFTER STÄLLER NU KRAV PÅ ATT FORSKNINGSDATA GÖRS FRITT TILLGÄNGLIGA

Gustav Nilsonne, med dr, forskare, Karolinska institutet; Stockholms universitet
● gustav.nilsonne@ki.se

Rebecca Willén, fil dr, forskare, IGDORE (Globally Distributed Institute for Open Research and Education)

Det är god sed att bevara forskningsdata och att dela dem med andra forskare på begäran, om nödvändigt efter anonymisering. Universitet och högskolor har också en juridisk skyldighet att arkivera forskningsdata.

I praktiken är dock tillgången till data ofta beroende av den enskilde forskarens arkivreda och beredvillighet att dela med sig. Idealfallet är att forskaren kan förklara var data finns, hur filerna kan öppnas och hur variablerna kan identifieras samt tillhandahålla metadata som beskriver hur datasetet tagits fram.

Men dessvärre är det inte ovanligt att forskare har bytt dator eller arbetsplats, inte längre minns var data finns sparade eller har glömt bort vad variabelnamnen betyder eller hur data genererats och bearbetats [1, 2]. Det som hade kunnat upptäckas genom ny- eller omanalyser förblir då dolt och oupptäckt.

Varje gång ett dataset går ad mortem och blir funktionellt otillgängligt blir det svårare att få en korrekt överblick av fältets kunskapsläge. Risken för snedvridning (bias) ökar när andelen tillgängliga data minskar och effektskattningar blir mindre säkra. En vetenskaplig litteratur med ett växande antal påståenden om bevisade effekter men för vilka data inte längre kan granskas riskerar att få ett sjunkande samlat bevisvärde över tid.

Öppna data för att utvinna ny kunskap

Rapporter av kliniska prövningar och andra experiment innehåller ofta endast sammanfattningar av data och effektskattningar från statistiska modeller. Det fullständiga datasetet är vanligen mycket rikare och innehåller dessutom ofta värdefull information som inte analyserats och rapporterats.

Att förutsäga exakt vilka frågor som ett dataset kan besvara är svårt. Metaanalyser med aggregerade data ger bättre effektskattningar än enskilda studier, men metaanalyser med individdata ger ännu bättre effektskattningar. Detta beror bl a på att mer varians kan förklaras i statistiska modeller som använder variabler som har registrerats på individnivå. Heterogenitet som minskar skattningens säkerhet kan uppstå exempelvis om olika studier har analyserat data på skilda sätt, t ex genom att använda olika kovariat i regressionsmodeller. Om individdata är tillgängliga kan sådan heterogenitet övervinnas.

Tillgången på data påverkar också starkt risken för snedvridning (bias) i en metaanalys. Snedvridning uppstår exempelvis när vissa resultat inte har publicerats på grund av att fynden gick i en viss riktning eller saknade statistisk signifikans. Ju mer data som saknas, desto större blir risken för snedvridning. Ju mer aggregerade data är, desto sämre går det att hantera förväxlingsfaktorer [3].

För att en vetenskaplig rapport ska få sitt största värde som byggsten för kunskap är det därför angeläget att data görs tillgängliga, helst i så fullständig och obearbetad form som möjligt. I dag finns mycket goda möjligheter att publicera öppna data genom olika fältspecifika eller allmänna arkiv, ofta utan kostnad för den enskilde forskaren. En bra utgångspunkt är den lista över öppna arkiv som hålls av tidskriften Scientific Data (http://www.nature.com/sdata/policies/repositories#general).

Tvivelaktiga forskningspraktiker och fusk

Tvivelaktiga forskningspraktiker minskar forskningens tillförlitlighet. Ett exempel är utfallsväxling i kliniska prövningar: om det från början avsedda utfallsmåttet (t ex mortalitet) inte uppvisar några tydliga resultat, kan det hända att forskarna i stället lyfter fram ett annat utfallsmått (t ex radiologisk tumörprogress). Problemet med detta är förstås att om man använder tillräckligt många utfallsmått är det alltid något som visar en statistiskt signifikant effekt av ren slump.

Utfallsväxling kan därför skapa en missvisande bild av hur effektiv en behandling är. Med öppna data är det möjligt att omanalysera ett dataset, exempelvis i enlighet med den plan som angivits i ett preregistrerat studieprotokoll, för den händelse att den slutliga vetenskapliga rapporten avviker från protokollet utan övertygande skäl.

»Varje gång ett dataset går ad mortem och blir funktionellt otillgängligt blir det svårare att få en korrekt överblick av fältets kunskapsläge.«

I ljuset av avslöjanden om oredlighet på svenska lärosäten är det angeläget att fundera över hur fusk och tvivelaktiga praktiker kan förebyggas. Centrala etikprövningsnämndens expertgrupp för oredlighet i forskning beslöt nyligen i ett fall med duplicerade fotomikrografier att det var oredligt att de rätta bilderna inte kunde visas upp på begäran [4]. Om de foton för vilka de publicerade bilderna uppgavs vara representativa hade publicerats öppet i ett digitalt arkiv, skulle det inte ha varit möjligt att på detta sätt tappa bort data.

Det är dessutom troligt att en oavsiktlig förväxling aldrig hade skett om man öppet arkiverat dem, eftersom arkiveringen medför att man noga går igenom och annoterar data. En eventuell avsiktlig för-

HUVUDBUDSKAP

- Öppna data från kliniska prövningar och andra studier gör det möjligt att utvinna mer kunskap, minskar risken för snedvridning (bias) och kan motverka fusk.
- Krav på öppna data ställs nu av EU liksom av flera forskningsfinansiärer och vetenskapliga tidskrifter.
- Meritsättning av öppna data är en delvis olöst utmaning.

Illustration: Colourbox

Öppna och fritt tillgängliga forskningsdata ger både möjligheter och utmaningar.

växling hade lättare kunnat upptäckas av referentgranskare om de hade haft tillgång till hela materialet.

Etisk skyldighet mot forskningspersonerna

När forskningspersoner deltar i kliniska prövningar eller andra studier underkastar de sig risker och intrång för att bidra till kunskapsutvecklingen. Den etiska bedömningen av en studie grundar sig i en avvägning mellan risker och förväntad nytta. Kunskapsnyttan är beroende av att studiens resultat blir tillgängliga. Om data inte i någon form kan användas har försökspersonernas risker varit förgäves. Det ankommer på varje ägare av forskningsdata, särskilt från människor och djur, att se till att största möjliga kunskapsvärde kan utvinnas och helst återföras till den population från vilken data hämtats.

Data från människor måste anonymiseras på ett tillfredsställande sätt innan de publiceras. Kunskapsvärdet av att publicera data öppet måste balanseras mot risken för identifikation och uppgifternas känslighet. I de flesta fall kan risken för identifikation hanteras genom att variabler med unika värden (t ex ålder och kroppslängd) kategoriseras eller stryks.

Vissa typer av data, exempelvis radiologiska bilder, kan inte anonymiseras helt. Man måste då överväga vilka riskmodeller som kan vara aktuella, hur sannolika de är, och om ytterligare åtgärder kan vidtas för att minska risken, såsom att beskära bilderna. Exempelvis är det möjligt vid publicering av hjärnavbildningsdata att inte ta med ansiktsregionen.

Meritsättning av öppna data

Det är en utmaning för utvecklingen av öppna data att flytta det vetenskapliga meritvärdet från artefakten (publikationen) till själva innehållet (data, analyskod, tolkningar). På sätt och vis är detta bara en aspekt av det välkända problemet att författarlistor inte på något tydligt sätt avspeglar författarnas respektive bidrag till publikationen.

I New England Journal of Medicine har öppna data debatterats under året. Det började med att Dan Longo och Jeff Drazen, två av redaktörerna, i en ledare beklagade sig över risken för att »dataparasiter« ska snylta på andras arbete och publicera resultat från data som de själva inte samlat in [5]. I sociala medier väckte ledaren muntert löje.

Ändå pekar Longo och Drazen på en öm punkt: Till dess att vi har hittat ett sätt att meritvärdera öppna data är varje dataset en investering för den enskilde forskaren. Här vilar ett ansvar på forskningsfinansiärerna att utveckla riktlinjer som tydligt värdesätter forskningens innehåll i stället för potentiellt missvisande indikatorer såsom antalet publicerade artiklar.

En åtgärd som visat sig öka publicering av öppna data är att tidskriften sätter en digital stämpel (badge) på artikeln om den innehåller öppna data. När stämplar för öppna praktiker infördes i tidskriften Psychological Science ökade andelen artiklar med öppna data från mindre än 5 procent till över 40 procent [6]. Någon motsvarande ökning kunde inte observeras i jämförbara tidskrifter under samma tid.

Åtgärder för öppna data

Internationella tidskrifter och anslagsgivare har börjat kräva att forskningsdata och forskningsmaterial ska göras fritt tillgängliga, inte bara för andra forskare utan också för allmänheten (öppen tillgång eller »open access«). Vetenskapsrådet har föreslagit riktlinjer som innebär att forskningens resultat både i form av rapporter och data ska göras fritt tillgängliga senast från år 2025 [7].

Även EU är i färd med att införa krav på öppen tillgång, och det återstår att se hur dessa kommer att implementeras i Sverige och andra medlemsstater [8].

Vi ser fram emot att regeringen tar ställning till Vetenskapsrådets föreslagna riktlinjer. Forskningsfinansiärer borde väga in öppna data som kvalitetsmarkör vid beslut om anslag, och lärosäten borde göra det vid beslut om anställning och befordran.

Lärosäten och finansiärer borde också agera för att rädda viktiga dataset som riskerar att bli funktionellt otillgängliga. Prioritet bör ges åt sådana dataset som informerar klinisk och annan praktik, t ex data från kliniska prövningar, och åt dataset som är unika och svåra att återskapa.

Till sist vilar ansvaret för att värdesätta praktiker som stödjer reproducerbarhet på oss alla som är aktiva som forskare och tillsammans bygger upp vår vetenskapliga kultur. ○

● Potentiella bindningar eller jävsförhållanden: Inga uppgivna.

Citera som: Läkartidningen. 2016;113:ELCU

REFERENSER

1. Vines TH, Albert AYK, Andrew RL, et al. The availability of research data declines rapidly with article age. Curr Biol. 2014;24(1):94-7.
2. Krawczyk M, Reuben E. (Un)available upon request: field experiment on researchers' willingness to share supplementary materials. Account Res. 2012;19(3):175-86.
3. Lakens D, Hilgard J, Staaks J. On the reproducibility of meta-analyses: six practical recommendations. BMC Psychol. 2016;4(1):24.
4. Centrala etikprövningsnämnden (CEPN). Expertgruppen för oredlighet i forskning. Yttrande 2016-09-08 [citerat 6 okt 2016]. Dnr O 2-2016. http://www.epn.se/media/2377/o-2-2016-expertgruppens-yttrande-160908.pdf
5. Longo DL, Drazen JM. Data sharing. N Engl J Med. 2016;374(3):276-7.
6. Kidwell MC, Lazarević LB, Baranski E, et al. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. PLoS Biol. 2016;14(5):e1002456.
7. Vetenskapsrådet. Nationella riktlinjer för öppen tillgång till vetenskaplig information 27 jan 2016 [citerat 6 okt 2016]. http://www.vr.se/omvetenskapsradet/regeringsuppdrag/avrapporterade2015/avrapporterade2015/nationellariktlinjerforoppentillgangtillvetenskapliginformation.4.7e727b6e141e9ed702b1307e.html
8. The Netherlands EU Presidency 2016. Amsterdam Call for Action on Open Science. 7 apr 2016 [citerat 6 okt 2016]. https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science