VIDENSKAB OG PRAKSIS | SEKUNDÆRPUBLIKATION

Antaget: 16. april 2007
Interessekonflikter: Ingen angivet

Taksigelse: Forfatterne har modtaget økonomisk støtte fra Syddansk Universitet.

This article is based on a study first reported in BMJ 2006;333:1095.

Litteratur

1. Herskind AM, Basso O, Olsen J et al. Is the natural twinning rate still declining? Epidemiology 2005;16:591-2.
2. Hack M, Taylor HG, Drotar D et al. Chronic conditions, functional limitations, and special health care needs of school-aged children born with extremely low-birth-weight in the 1990s. JAMA 2005;294:318-25.
3. Shenkin SD, Starr JM, Deary IJ. Birth weight and cognitive ability in childhood: a systematic review. Psychol Bull 2004;130:989-1013.
4. Record RG, McKeown T, Edwards JH. An investigation of the difference in measured intelligence between twins and single births. Ann Hum Genet 1970;34:11-20.
5. Ronalds GA, de Stavola BL, Leon DA. The cognitive cost of being a twin: evidence from comparisons within families in the Aberdeen children of the 1950s cohort study. BMJ 2005;331:1306.
6. Petersen J K. The Danish Demographic Database: Longitudinal Data for Advanced Demographic Methods. Research Report. 15. Odense: Syddansk Universitet, Institut for Statistik og Demografi, 2000.
7. Andersen TF, Madsen M, Jørgensen J et al. The Danish National Hospital Register: a valuable source for modern health sciences. Dan Med Bull 1999;46:263-8.
8. Skytthe A, Kyvik K, Holm NV et al. The Danish Twin Registry: 127 birth cohorts of twins. Twin Res 2002;5:352-7.
9. Undervisningsministeriet. Prøver, evaluering, undervisning – en samlet evaluering af folkeskolens afsluttende prøver maj-juni 2004. København: Undervisningsministeriet, 2004.
10. Naglieri JA, Bornstein BT. Intelligence and achievement: just how correlated are they? J Psychoeduc Assess 2003;21:244-60.

Kvaliteten af kvalitetsmål – sekundærpublikation

Ph.d. Sune Lehmann, professor Andrew D. Jackson & professor Benny E. Lautrup

Danmarks Tekniske Universitet,
Institut for Informatik og Matematisk Modellering, og
Københavns Universitet, Niels Bohr Instituttet

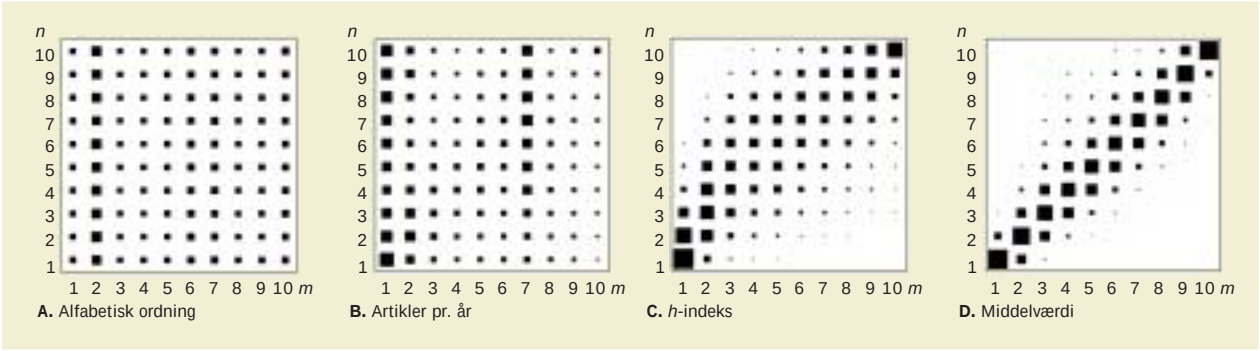
Resume

Er visse måder at måle videnskabelig kvalitet på bedre end andre? Vi analyserer pålideligheden og præcisionen af forskellige metoder, der typisk benyttes til at rangordne lister af videnskabelige publikationer. Vi påviser, at flere almindeligt anvendte kvalitetsmåls manglende pålidelighed gør dem ubrugelige i praksis.

Det er alment accepteret, at antallet af en videnskabelig artikels citationer – set i forhold til andre artikler i samme felt – udgør et kvantitativt mål for artiklens videnskabelige værdi. Hvis en artikel citeres ofte, betyder det, at artiklens indhold er en aktiv del af mange andres videnskabelige arbejde. Sammenhængen mellem citationer og forskeres videnskabelige værd er imidlertid mere kompliceret, idet de fleste forskere producerer mere end kun en artikel. Spørgsmålet om, hvordan citationerne af en række artikler i en sådan citationsliste kombineres til et mål for personens videnskabelige formåen, er ikke trivielt, og det er emnet for denne undersøgelse.

Vores analyse er baseret på alle artikler, der er publiceret i teoretisk højenergifysik og samlet i databasen SPIRES, men vi forventer kvalitativt lignende resultater for andre natur- og sundhedsvidenskabelige felter, hvor publikation af videnskabelige artikler spiller en central rolle i formidlingen af forskningsresultater. En af de primære grunde til, at »gode« videnskabelige kvalitetsmål er vanskelige at konstruere, er, at fordelingen af citationer for enkelte artikler er skævvredet med en tung asymptotisk hale af højt citerede artikler [1]. For SPIRES udgør antallet af ikkeciterede artikler 29% af alle publikationer i databasen. Halvdelen af artiklerne har to eller færre citationer (medianen), mens middelantallet af citationer er 12,6. Dette skyldes, at en lille del af artiklerne får næsten alle citationerne, mens størsteparten af artiklerne er lavt citerede: 4,3% af alle artikler genererer 50% af alle citationerne, mens de 50% mindst citerede artikler kun står for 2,1% af citationerne. I forbindelse med vurderingen af kvalitetsmål har den skæve fordeling af citationer den vigtige konsekvens, at når vi udregner gennemsnittet eller medianen af en forfatters citationsliste, fortæller svarene os om vidt forskellige aspekter af forfatternes publikationshistorie.

Målet med enhver kvantitativ evaluering af forskere er at etablere en rangorden. En rangorden kan kun konstrueres ved at reducere citationslisten til et enkelt tal. Men der er mange måder at danne en sådan rangorden på, og forskellige metoder resulterer i mål med forskellige statistiske egenskaber. Det er på baggrund af hvert enkelt måls anvendelighed til at skabe en præcis og akkurat rangordning, at vi taler om »kvaliteten« af et kvalitetsmål. Set fra dette perspektiv er det bedste mål det, som minimerer usikkerheden af de tildelte værdier og derfor maksimerer muligheden for at skelne mellem forfattere. Vi analyserede tre anvendte kvalitetsmål: 1) antallet af publikationer pr. år, 2) Hirsch-indekset h [2] og 3) citationslistens middelværdi. Siden det blev foreslået i 2005, er h-indekset i stigende grad taget i brug som mål for kvalitet – også på danske universiteter. En forsker har Hirsch-index h, hvis h af vedkommendes N artikler har mindst h citationer og de tilbageværende (N–h) artikler har færre end h citationer. Hirsch-



Figur 1. Sammenligning af kvalitetsmål. Hver række, indekseret ved *n*, viser de gennemsnitlige sandsynligheder for, at en forfatter, der som udgangspunkt var placeret i den givne decil *n*, rent faktisk tilhører decil *m*. Sandsynlighederne er proportionale med størrelsen af kvadraternes areal.

indekset er ekstensivt (vokser med antallet af publikationer), men da vi er interesserede i intensive kvalitetsmål (konstante over tid), analyserer vi den tilsvarende normaliserede *h*-værdi, *h/N*, jf. [2, 3]. Endelig inkluderer vi et mål som burde være ukorreleret med videnskabelig kvalitet, nemlig en rangordning baseret på forfatternes forbogstaver.

Analysen indledes med, at man udregner et af kvalitetsmålene for alle forfattere i SPIRES og derefter inddeler dem i decilgrupper på baggrund af den udregnede værdi. Gruppe 1 består af de 10% lavest rangerede forskere, gruppe 2 af de 10% næstlaveste og så videre indtil gruppe 10, der består af de 10% højest rangerede forfattere. Derefter udregner vi den betingede sandsynlighed for, at en artikel skrevet af en forfatter i gruppe *n* har *k* citationer. En citationsliste er en statistisk størrelse, som ville blive anderledes, hvis forfatteren gennemløb sit liv igen, ligeledes er forskerens kvalitetsmål statistisk og kunne også være anderledes inden for visse grænser sat af kvaliteten af målet. Ud fra de betingede sandsynligheder og under anvendelse af Bayes' sætning kan vi bestemme den gennemsnitlige sandsynlighed for, at en forfatter, som oprindeligt blev placeret i gruppe *n*, burde have været placeret i gruppe *m*. Dette resultat opnås ved at udregne sandsynligheden for, at citationslisterne for forskerne i gruppe *n* er opnået ved tilfældige træk på betinget sandsynlighedsfordeling for gruppe *m* (se [3] for en formel matematisk udledning). Da udregningen af *m* er baseret på citationslisten i samspil med alle tilgængelige data i SPIRES, er denne gruppeinddeling mere pålidelig end inddelingen, der er baseret på det rå kvalitetsmål *n*. Proceduren gentages for alle kvalitetsmål og alle decilgrupper.

Et perfekt kvalitetsmål ville – i et plot af sandsynligheden for at opnå klassifikationen *m* givet a priori-klassifikationen *n* – placere al vægt i diagonalen. Et mål uden nogen korrelation med kvalitet ville derimod fordele vægten jævnt ud over et sådant plot. **Figur 1** viser tydeligt, at både præcision og pålidelighed er afhængig af, hvilket mål for kvalitet, vi vælger. Som forventet er den alfabetiske rangorden ukorreleret med videnskabelig kvalitet, så forfatterne er jævnt fordelt mellem de forskellige decilgrupper (Figur 1A). Den kvadratiske (*rms*) middelusikkerhed for dette plot er maksimal, dvs. ±29%. Det nok mest udbredte kvalitetsmål i Danmark er det gennemsnitlige antal artikler publiceret pr. år. Dette mål giver en fordeling, som er sammenlignelig med resultaterne opnået ved ren alfabetisk rangordning af forfatterne. Det bedste, man kan sige om publikationsfrekevens, er, at man med den måler flittighed frem for dygtighed.

Hirschs indeks forsøger at skabe balance mellem produktivitet og kvalitet og derved tage højde for den store vægt, som de skæve citationsfordelinger lægger på et relativt lille antal højt citerede artikler. Som nævnt bestemmes *h* ved at rangordne artiklerne fra højest til lavest citerede, hvor artikel i har $C(i)$ citationer, og *h* findes ved at løse ligningen $h = C(h)$. Imidlertid er dette blot den enkleste version af den mere generelle ligning $h = A\ C(h)^{\kappa}$ med en faktor *A* og potens κ. *Hirschs* valg af $A = \kappa = 1$ er fuldstændig vilkårligt og ikke baseret på nogen form for indsigt eller data. Dog viser Figur 1C, at *h*-indekset klarer sig bedre end publikationsfrekvensen, formentlig fordi det faktisk gør brug af citationsdata. Specifikt overestimerer *h*-indekset den indledende *n*-gruppetildeling med otte procentpoint, hvilket afspejles i en øget sandsynlighedstæthed over diagonalen. Gruppeinddeling på baggrund af *h* har en *rms*-usikkerhed på ±16, hvilket kun er en faktor to bedre end ordning efter forbogstav. Dermed er det klart, at man med *Hirschs h* ikke kan inddele forskere i decilintervaller med sikkerhed. Figur 1 viser, at det gennemsnitlige antal citationer pr. artikel er et klart overlegent kvalitetsmål, både med hensyn til pålidelighed og præcision. Den gennemsnitlige inddeling har en fejl på kun 1,8 procentpoint, og *rms*-usikkerheden er ±9.

Ved hjælp af enkle argumenter kan vi vise, at *rms*-usikkerheden for ethvert kvalitetsmål falder eksponentielt, når det tilgængelige antal artikler for en forfatter vokser [3]. Ved at man benytter middelværdien som kvalitetsmål, kan en forfatter placeres i grupperne 2-3 og 8-9 med 90%'s sandsynlighed på baggrund af 50 artikler, mens gruppe 1 og gruppe 10 kræver færre artikler. Ethvert forsøg på inddeling af forfattere

i grupper på baggrund af signifikant færre artikler bør behandles med forsigtighed. Hvis citationsdata for et specifikt forskningsfelt er tilgængelige, kan den her beskrevne metode anvendes. Vor metode gør det muligt at sammenligne forskere fra forskellige videnskabelige områder ved at antage, at forskere, som tilhører samme percentil af deres respektive felter, befinder sig på samme videnskabelige niveau. Metoderne kan endda benyttes til at sammenligne forskere, som arbejder tværvidenskabeligt og publicerer deres arbejde i mere end et felt. Desværre er citationsdatabaser med tilstrækkelig homogenitet og omfang ikke offentligt tilgængelige på nuværende tidspunkt.

Her er det vigtigt at påpege, at fordelene ved samvittighedsfuld citationsanalyse overskygges af et stort potentiale for misbrug. Institutioner har ofte den fejlagtige opfattelse, at beslutninger truffet på baggrund af en algoritme pr. definition er retfærdige og fornuftige, hvilket ingenlunde er garanteret. Det forværres yderligere af manglen på indsigt i, hvad kvalitet er, og hvordan den måles. Da man er ude af stand til at måle det, man ønsker at maksimere, nemlig kvalitet, ender man i stedet med at maksimere det, man er i stand til at måle. Derfor vil beslutninger fremover fortsat blive truffet på baggrund af kvalitetsmål, som enten komplet ignorerer citationsdata (f.eks. publikationsfrekvensen), eller som hviler på ukomplette datasæt. Eksisterende databaser over videnskabelige publikationer kan aktivt hjælpe på denne situation ved at sammensætte feltspecifikke homogene underdatabaser i analogi med den database, vi har anvendt for teoretisk partikelfysik. Adgang til sådanne databaser vil give både institutioner og forskere mulighed at vurdere kvaliteten af enhver citationsliste i lyset af al offentlig tilgængelig information. Som forskere bør vi insistere på, at institutionerne åbent redegør for deres brug af citationsdata.

Indtil da er vi nok nødt til at gøre tingene på den gammeldags måde og rent faktisk læse artiklerne.

Korrespondance: *Sune Lehmann,* Institut for Informatik og Matematisk Modellering, Danmarks Tekniske Universitet, DK-2800 Lyngby.
E-mail: lehmann@nbi.dk

Antaget: 20. april 2007
Interessekonflikter: Ingen angivet

This article is based on a study first reported in Nature 2006;444:1003-4

Litteratur

1. Lehmann S, Lautrup BE, Jackson AD. Citation networks in high energy physics. Physical Review E 2003;68:026113.
2. Hirsch JE. An index to quantify an individual's scientific output. Proc Nat Acad Sci 2005;102:16569.
3. Lehmann S, Jackson AD, Lautrup BE. A quantitative analysis of measures of quality in science. arXiv: physics 2007:0701311.

Leptomeningeal karcinomatose – en usædvanlig årsag til høretab

Reservelæge Sidse Bregendahl & 1.reservelæge Jens Å. Lindberg

Herning Sygehus, Medicinsk Afdeling

Erhvervet høretab hos voksne kan have mange årsager. En ikke særlig kendt årsag er leptomeningeal karcinomatose (LC), som kan udløse et perceptivt høretab af retrokokleær type [1]. LC opstår hos omkring 5% af de patienter, der har cancer [2, 3], og ses hyppigst ved adenokarcinom fra lunger og mammae samt ved malignt melanom [1, 2, 4].

Isoleret bilateralt perceptivt høretab sekundært til LC er tidligere kun beskrevet i beskedent omfang [1, 3-5]. Vi ønsker derfor i det følgende at henlede opmærksomheden på denne problemstilling ved at beskrive en sygehistorie, hvor LC viser sig med høretab som hovedsymptom.

Sygehistorie

En 49-årig kvinde blev indlagt på en medicinsk afdeling. Hun havde haft en måned varende progredierende bilateralt høretab efterhånden ledsaget af svimmelhed, kvalme, opkastninger, tiltagende hovedpine og smerter i nakke med udstråling til ryggen. En audiologisk undersøgelse, der blev foretaget henholdsvis 18 dage og fire dage forud for indlæggelsen, vidnede om progredierende bilateralt høretab af perceptiv type med fald i *threshold carhart* (TC) fra 55 dB til 70 dB på højre øre og fra 35 dB til 40 dB på venstre øre. Skelnetabet (DL) blev målt til 20 og 24% på henholdsvis højre og venstre øre ved den seneste undersøgelse.

Patienten var et år tidligere blevet opereret for et lavt differentieret adenokarcinom i højre mamma med brystbevarende kirurgi efterfulgt af postoperativ stråle- og kemoterapi. En histologisk undersøgelse viste spredning til ti af de 22 fjernede aksillære lymfeknuder, anaplasigrad 3 og negativ østrogenreceptorimmunmarkering. Ved efterfølgende kontroller kunne der ikke afsløres recidiv.

Ved indlæggelsen var patienten i god almentilstand, og de eneste objektive fund var bilateral hørenedsættelse. Samtlige