Crowdsourcing er en ny metode til indsamling af data til videnskabelige undersøgelser

Josefine Stokholm Bækgaard[1] & Peter Hallas[2]

STATUSARTIKEL

1) Det Sundhedsvidenskabelige Fakultet, Københavns Universitet
2) Anæstesiafdelingen, Rigshospitalet

Ugeskr Læger 2015;177:V07140415

Crowdsourcing (CS) er en ny metode til indsamling af data til videnskabelige undersøgelser. CS er et voksende fænomen, hvor man indhenter information ved hjælp af mange mennesker, oftest via internettet. Inden for medicinens verden begynder der at dukke artikler op, hvor man benytter sig af CS-teknikker.

CS har bl.a. været anvendt til forudsigelse af influenzaudbrud [1, 2], søgning efter årsagen til *severe acute respiratory syndrome* (SARS) [3], overvågning af hiv-udbrud [4], undersøgelse af lithiums effekt hos amyotrofisk lateralsklerose (ALS)-ramte [5] og indsamling af DNA fra Parkinson-ramte [6]. I 2013 var der 124 artikler på PubMed [7] med søgeordet *crowdsourcing* mod otte i 2010 og ingen i 2008.

Formålet med denne artikel er at give et indblik i CS som metode og diskutere, hvilke muligheder der ligger i CS. Med den stigende interesse for fænomenet – også blandt patienter – er det væsentligt, at læger også kan forholde sig kritisk over for eventuelle begrænsninger ved CS.

Crowdsourcing (CS) er en ny studiemetode, hvor data fra mange brugere indsamles online. Læger bør kende til metoden, fordi der sandsynligvis kommer flere CS-baserede studier i fremtiden.

HVAD ER CROWDSOURCING?

Ordet *crowdsourcing* er en sammentrækning af ordene *crowd* og *outsourcing*. Begrebet blev populariseret af *James Surowiecki* i bogen »The Wisdom of Crowds« fra 2005 [3]. Det populære tv-program »Hvem vil være millionær?« bruges som et klassisk eksempel på CS. I quizzen kan deltageren vælge enten selv at svare på et quizspørgsmål, ringe til en ven for at bede om hjælp eller spørge publikum om deres mening. Det viser sig, at publikum (*the crowd*) har en højere succesrate end både vennen og deltageren selv [3]. Selv om »Hvem vil være millionær« er et illustrativt eksempel på styrken ved at indhente oplysninger fra mange, er selve CS-begrebet mere nuanceret.

CS kan defineres som brug af data, som er genereret af mange brugere eller interessenter, enten ved bevidst afgivelse af oplysninger eller via deres adfærd. Data vil typisk være indsamlede online og være mere usystematiske end egentlige registerdata.

Nogle af de nuværende trends inden for CS er dels at måle *gennemsnittet* af en stor gruppes bud på et faktuelt spørgsmål (jf. »Hvem vil være millionær«–eksemplet), dels at *indsamle adfærdsdata* fra en stor gruppe (såkaldt *crowd surveillance*) [8] samt at en gruppe i *fællesskab* løser et problem (såkaldt *citizen science* [9]).

CROWD SURVEILLANCE

Brugen af internettet er stigende på verdensplan. Alene i USA foretages der over otte millioner helbredsrelaterede søgninger dagligt [10], og omtrent 25% af Facebookprofilerne og 90% af Twitter-*feeds* er helt offentlige [11]. Den åbne brug af internettet giver muligheder for »*online crowd surveillance*«, dvs. en overvågning af aktiviteten online. I teorien vil man vha. onlineovervågning kunne følge spredningen af f.eks. smitsomme sygdomme ved at følge stigninger af *tweets* på Twitter e.l. vedrørende specifikke symptomer [12].

I 25 lande laver Google funktionen Google Flu Trends [2]; dette gøres i samarbejde med forskellige allerede eksisterende overvågningsmetoder, som European Influenza Surveillance Network og U.S. Centers for Disease Control. Ved hjælp af relevante søgeord mener Google at kunne opnå næsten øjeblik-

kelig status på influenzaniveauerne i disse lande. I et studie fra Nebraska, USA, var Google Flu Trends en god prædiktor for influenzaudbrud i perioden 2008-2012 [13]. Studier tyder på, at der også er høj korrelation mellem Googlesøgninger og andre sygdomme [14]. Google Flu Trends er endnu ikke nået til Danmark [2]. Sådanne overvågningsmetoder kan ikke erstatte, men snarere komplementere de nuværende metoder [15]; Google Flu Trends overestimerede for eksempel influenzaudbruddet i 2013.

I et studie fra april 2014 fremhæves muligheden for at *crowdsource* bivirkninger af lægemidler gennem Twitter og Facebook [11]. Dette er interessant, idet de nuværende metoder til identifikation af bivirkninger er mangelfulde og har høj grad af underrapportering [11].

CITIZEN SCIENCE

Fænomenet *citizen science* har påkaldt sig megen opmærksomhed. Studier, som er udgået fra den amerikanske hjemmeside PatientsLikeMe.com (PLM), nævnes ofte som et eksempel [5]. PLM er et virtuelt netværk for patienter med sjældne og kroniske sygdomme; over 250.000 medlemmer deler informationer om deres symptomer og oplevelser med andre brugere [5]. PLM har brugt informationer fra patienter med ALS til at analysere, om lithium har en effekt på progressionshastigheden af sygdommen [5]. Resultatet af dette CS-studie er senere blevet bekræftet i et regelret, randomiseret studie [5].

På en lignende hjemmeside, 23&Me, kan man for 99 USD få analyseret hele sit genom ved at indsende en simpel spytprøve [6]. Over 10.000 patienter med Parkinsons sygdom har allerede deltaget (disse spytprøver er dog sponsorerede), og to nye gener, som er associerede med Parkinsons sygdom, er blevet kortlagte [6].

Inden for astronomi og zoologi bruger man også CS til at indsamling og bearbejdning af data [16]. Hjemmesiden Zooniverse.org har netop rundet en million tilmeldte frivillige, hvis indsats har resulteret i mere end 50 peer reviewede artikler [16]. Zooniverse står bl.a. bag hjemmesiden Galaxy Zoo, hvor folk inviteres til at assistere i klassifikationen af galakser [17]. Et andet eksempel er eBird, som er en online database, der forsyner ornitologer med data om fugleobservationer verden over [18].

Der er således utallige eksempler på CS. På én hjemmeside kan brugerne sågar hjælpe forskere med at folde proteinstrukturer [19]. Brugerfladen er udformet som et spil; men i virkeligheden hjælper brugerne forskere med at finde den korrekte, og hidtil ukendte, tredimensionelle struktur på proteiner.

! FAKTABOKS

Crowdsourcing (CS) er en ny metode til indsamling af data.

Ved CS indhentes information fra mange mennesker, oftest via internettet.

Eksempler på CS i medicinsk sammenhæng er f.eks.:

Crowd surveillance; f.eks. overvågning af spredning af sygdomme (Google Flu Trends) via analyse af internetaktivitet

Citizen science; ikkeforskere deltager i videnskabelige projekter via internettet

CS af diagnoser; vanskelige patientcases bliver lagt online mhp. diskussion af mulige diagnoser.

Studier, hvor man benytter CS-teknikker, bliver formentlig hyppigere i fremtiden; CS er et supplement til de mere klassiske studiedesign.

BRUGEN AF CROWDSOURCING TIL DIAGNOSER

Det er blevet almindeligt, at patienter søger information om sygdom på internettet, og at personlige helbredsproblemer bliver diskuteret på sociale netværkssider. 8,5% af de engelsksprogede Twitter-*tweets* er således relaterede til sygdom, og op mod 21,5% er relaterede til helbred [8]. Der findes tilmed en Twitterkonto ved navn »Radiopedia Twitter«, hvor røntgenbilleder lægges op, og diagnoser diskuteres [20].

Hjemmesiden CrowdMed er et forsøg på at systematisere CS til diagnoser [21]. På CrowdMed opretter syge en profil og giver læger fri adgang til al data vedrørende deres sygdom. Den »gennemsnitlige sygdomsramte« på CrowdMed har været syg i seks år, set otte læger og haft medicinske udgifter på omkring 50.000 USD, før vedkommende forsøger sig med CrowdMed. Indtil videre har CrowdMed haft over 200 sager, og ca. 80% af de involverede patienter har meldt tilbage, at de bedste gæt på deres diagnoser faktisk viste sig at være rigtige [21].

I et dansk studie har man peget på, at brugere af Facebook kan benytte deres virtuelle netværk til at få en fornemmelse af alvorligheden af symptomer [22]. For nylig har medier fortalt historien om en treårig pige, der blev diagnosticeret med Coats syndrom på grund af billeder, som moderen havde lagt ud på hjemmesiden [23].

Måske kan CS-metoden også bruges til »diagnosticering« af fejl og forbedring af sjældent brugt udstyr og udførte procedurer. En CS-lignende teknik er eksempelvis blevet brugt til »diagnosticering« af muligheden for at forbedre design af udstyr til intraossøs adgang [24].

DISKUSSION

CS har mange fordele til visse typer af studier; ikke mindst er den høje hastighed, hvorved man kan opnå materiale, usammenlignelig med andre metoder. Derudover er muligheden for nemt at inkludere en

stor, bred geografisk population unik. CS-studier er desuden relativt billige, og deltagernes frivillige og selvmotiverede tilgang betyder måske, at risikoen for økonomisk interessekonflikter er mindre.

Desværre fører CS ikke altid til det rette svar eller den smarteste løsning på et problem. For det første er den aldersgruppe, som internettets brugere udgør, ikke repræsentativ for befolkningen som helhed, eftersom de helt unge og den ældre population i mindre grad er repræsenteret på internettet [20]. For det andet kan brugere fra lande og socialgrupper uden adgang til stabil og billig internetforbindelse ikke forventes at deltage.

Som ved andre metoder er der også mulighed for snyd inden for CS og risiko for forsøg på at manipulere resultatet. Dette har været særligt diskuteret for Wikipedia, der jo i høj grad også er en form for CS-projekt. På Wikipedia kan alle brugere oprette eller redigere artikler. Konsekvensen af denne åbne struktur er, at der ingen garanti er for indholdets rigtighed. Wikipedia er derfor ofte blevet kritiseret for at være unøjagtig sammenlignet med f.eks. Encyclopedia Britannica, hvor kun eksperter på områder får lov til at oprette artikler. En gennemgang, som blev lavet i 2005 af Nature og var baseret på 42 artikler på Wikipedia og Encyclopedia Britannica, viste dog få forskelle i nøjagtighed på de to. Nature konkluderede, at den gennemsnitlige artikel på Wikipedia indeholder omkring fire unøjagtigheder, hvor Encyclopedia Britannica indeholder omkring tre [25].

Endvidere viser eksempler som Google Flu Trends også, at CS ikke kan levere præcise resultater hver gang [15], hvilket tyder på, at der er lang vej igen, før CS vil kunne erstatte de nuværende overvågningssystemer. Ifølge *Surowiecki* er det heller ikke alle grupper, der er egnede til CS [3]. Han mener, at en CS-gruppe bør opfylde følgende fire kriterier: 1) Mangfoldighed i opfattelserne: Hver enkelt skal råde over egne oplysninger, også selvom de bare er en utraditionel fortolkning af kendte fakta. 2) Uafhængighed: Folks opfattelser skal være upåvirkede af omgivelsernes meninger. 3) Decentralisering: Folk skal have mulighed for at specialisere sig og trække på lokal viden. 4) Aggregering: Der skal findes en mekanisme, der kan føre de individuelle standpunkter sammen til en kollektiv beslutning. Kun når disse kriterier er opfyldt, mener *Surowiecki*, at CS kan udføres med succes.

I bogen »You are not a gadget« [26] argumenterer den amerikanske forfatter *Jaron Lanier* for, at *the crowd* med større sandsynlighed træffer en klog beslutning, når den ikke selv definerer spørgsmålene, og når evalueringen af svarets korrekthed kan udtrykkes enkelt (som f.eks. en enkelt numerisk værdi). Endvidere skal informationerne filtreres af en kvalitetskontrolmekanisme, der i vidt omfang er underlagt enkeltpersoner. De bør altså sorteres af folk med mere traditionel faglig indsigt, enten egentlig professionelle forskere eller amatører med særlig indsigt og træning. Under sådanne omstændigheder kan mængden træffe klogere beslutninger end en enkelt person. Men hvis en af disse betingelser ikke overholdes, bliver mængdens resultat usikkert eller dårligere, skriver *Lanier* [26].

FREMTIDSPERSPEKTIVER

CS giver nye muligheder for at indsamle data. Et endnu uudnyttet aspekt kunne vedrøre patienters muligheder for aktiv deltagelse i sundhedsvæsenet [27]. I øjeblikket inkorporerer sundhedsvæsenet ikke mange af de data, som patienter indsamler. Det er oplysninger om f.eks. hjemmeblodsukkermålinger, blodtryksmålinger, fitnessaktiviteter samt utilsigtede hændelser. Som led i den amerikanske regerings »EHR Incentive Programs«, hvor man forsøger at øge forekomsten af elektronisk sundhedsdata, er det forslået, at man også prøver at implementere patientgenereret og -registreret data. Man vil herved kunne spare sygehuse for arbejde i form af f.eks. indregistrering af spørgeskemaer og undersøgelser samt komme op på et højere niveau af detaljeringsgrad [28]. Ved at gøre bedre brug af patientopnåede data [27] kunne sundhedsvæsenet både behandle den enkelte patient bedre og nemmere opdage systematiske fejl.

Læger bør kende til metoden, fordi der sandsynligvis kommer flere CS-baserede studier i fremtiden. Resultaterne af CS-studier bør dog tolkes ud fra viden om usikkerhed i metoden. Inden for medicin vil CS ikke kunne erstatte allerede eksisterende studiemetoder som f.eks. case-kontrol-studier eller randomise-

! FAKTABOKS

Tips til vurdering af *crowdsourcing*-studier

Er problemstillingen egnet til *crowdsourcing* (CS) (er den f.eks. et sjældent fænomen, eller er den en repetitiv rutineopgave)?

Er der bias i forhold til adgangen til internet og målgruppen for undersøgelsen?

Hvem står bag studiet? Fagfolk? Superviseres/vejledes der?

Er der taget højde for tilfældige/ukorrekte data? Kan der være en økonomisk interesse bag studiet?
Ville problemstillingen være bedre egnet til en anden form for studie, f.eks. et randomiseret studie?

CS bør bruges som guide og inspiration, ikke som facit.

rede studier. Men CS kan være hypotesegenererende og danne grundlag for ny viden.

SUMMARY

Josefine Stokholm Bækgaard & Peter Hallas:

Crowdsourcing is a new method for generating data for scientific research

Ugeskr Læger 2015;177:V07140415

Crowdsourcing (CS) is a rapidly emerging method in scientific research. In CS, large groups of people generate new data or try to find solutions to specific research questions, mainly by online collaboration. Examples of the current use of CS in medicine include disease surveillance as well as diagnosis of rare conditions. CS techniques are rapid, low cost and geographically independent – traits lacking in traditional types of study design. However, CS as a method has not yet found its place in the evidence rating scale and a standard for conducting and reporting CS studies does not yet exist.

KORRESPONDANCE: *Josefine Stokholm Bækgaard*, Øster Voldgade 20, 1350 København K. E-mail: kqs486@alumni.ku.dk

ANTAGET: 4. november 2014

PUBLICERET PÅ UGESKRIFTET.DK: 5. januar 2015

INTERESSEKONFLIKTER: Forfatternes ICMJE-formularer er tilgængelige sammen med artiklen på Ugeskriftet.dk

LITTERATUR

1. Cheng CKY, Lau EHY, Ip DKM et al. A profile of the online dissemination of national influenza surveillance data. BMC Public Health 2009;9:339.
2. GoogleFlutrends, 2014. Mountain View, Californien: Google Inc. www.google.org/flutrends/ (21. jul 2014).
3. Surowiecki J. The wisdom of crowds. New York: Random House, 2005.
4. Stoové MA, Pedrana AE. Making the most of a brave new world: opportunities and considerations for using Twitter as a public health monitoring tool. Prev Med 2014;63:109-11.
5. Wicks P, Vaughan TE, Massagli MP et al. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. Nat Biotechnol 2011;29:411-4.
6. 23andme. New Parkinson's Findings Published , 2014. www.blog.23andme.com/23andme-research/new-parkinsons-findings-published/ (20. maj 2014).
7. PubMed Health. Bethesda, MD: National Library of Medicine (US), 1996. www.ncbi.nlm.nih.gov/pubmedhealth/crowdsourcing (6. jun 2014).
8. Hill S, Merchant RM, Ungar L. Lessons learned about public health from online crowd surveillance. Big Data 2013;1:160-7.
9. Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. J Med Internet Res 2012;14:46.
10. Susannah Fox. Online health search 2006. Pew Internet & American Life Project , 2006. www.pewinternet.org/PPF/r/190/report_display.asp (21. jul 2014).
11. Freifeld CC, Brownstein JS, Menone CM et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Saf 2014;37:343-50.
12. st Louis C, Zorlu C. Can Twitter predict disease outbreaks? BMJ 2012;344:e2353.
13. Araz OM, Bentley D, Muelleman RL. Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. Am J Emerg Med 2014;32:1016-23.
14. Pelat C, Turbelin C, Bar-Hen A et al. More diseases tracked by using Google trends. Emerg Infect Dis 2009;15:1327-8.
15. Butler D. When Google got flu wrong. Nature 2013;494:155-6.
16. Bonney R, Shirk JL, Phillips TB et al. Citizen science Science 2014;343:1436-7.
17. Galaxy Zoo , 2007. www.galaxyzoo.org (21. jul 2014).
18. Sullivan BL, Wood CL, Iliff MJ et al. eBird: a citizen-based bird observation network in the biological sciences. Biol Conserv 2009;142:2282-92.
19. Foldit. Center for Game Science at University of Washington in collaboration with UW Department of Biochemistry, 2008, (opdateret 2014). http://fold.it/portal/ (7. jul 2014).
20. Gaillard F. Radiopaedia, 2005. http://radiopaedia.org (15. jul 2014).
21. Heyman J. CrowdMed. San Francisco, 2013 (opdateret 2014). www.crowdmed.com/ (21. jun 2014).
22. Folkestad L, Brodersen JB, Hallas P et al. Lægfolk kan bruge deres Facebook-venner til at få hjælp vedrørende medicinske diagnoser. Ugeskr Læger 2011;173:3174-7.
23. Billington J. Facebook users spot rare Coat's disease in child's photo. News.com.au, 2014. www.news.com.au/technology/online/facebook-users-spot-rare-coats-disease-in-childs-photo/story-fnjwnhzf-1226873463826 (17. jul 2014).
24. Hallas P, Brabrand M, Folkestad L. Complication with intraosseous access: Scandinavian users' experience. West J Emerg Med 2013;14:440-3.
25. Giles J. Internet encyclopaedias go head to head. Nature 2005;438:900-1.
26. Lanier J. You are not a gadget: a manifesto. New York: Vintage Books, 2011.
27. Israel Green-Hopkins. Patient-generated health data: is health care ready to absorb it? Vector, 2014. www.vectorblog.org/2014/02/patient-generated-health-data-is-health-care-ready-to-absorb-it/ (20. jun 2014).
28. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. J Med Internet Res 2005;7:e3. doi:10.2196/jmir.7.1.e3.