VIDENSKAB OG PRAKSIS | ORIGINALARTIKEL

Personhenførbare og fortrolige data og resultater på internettet

Adjunkt Mads Ronald Dahl & seniorforsker Peter Vedsted

Aarhus Universitet, Institut for Folkesundhed,
Sektion for Sundhedsinformatik, Afdeling for Biostatistik, og
Forskningsenheden for Almen Praksis i Århus

Resume

Introduktion: Sundhedsvidenskab er baseret på indsamling, opbevaring og analyse af persondata. For at sikre fortrolighed og anonymitet bygger omgangen med disse data på en række love, vejledninger og regler. Når forskere mv. kommunikerer om deres resultater, sker det ofte ved brug af computer og internet. I den forbindelse kan der ske en utilsigtet udbredelse af følsomme persondata. Forskningsdata analyseres og fremstilles ofte grafisk med MS Excel eller MS Word. Efterfølgende indsættes en figur eller tabel i MS PowerPoint eller MS Word for distribution. De kan f.eks. publiceres på internettet som: PowerPoint-præsentation, PowerPoint-show eller som et Word-dokument. Desværre vil man ofte finde hele Excel-dataarket indsat også, og dataene bliver derved tilgængelige. I andre tilfælde benyttes der billeder, såsom røntgenbilleder eller *screen dumps* i præsentationer, der indeholder persondata.

Materiale og metoder: Vi lavede en simpel søgning på den danske Google-søgemaskine og afgrænsede søgningen til sundhedsrelaterede emner og filer af typen PowerPoint-præsentation eller Word-dokument. Vi downloadede og testede disse filer for skjulte persondata.

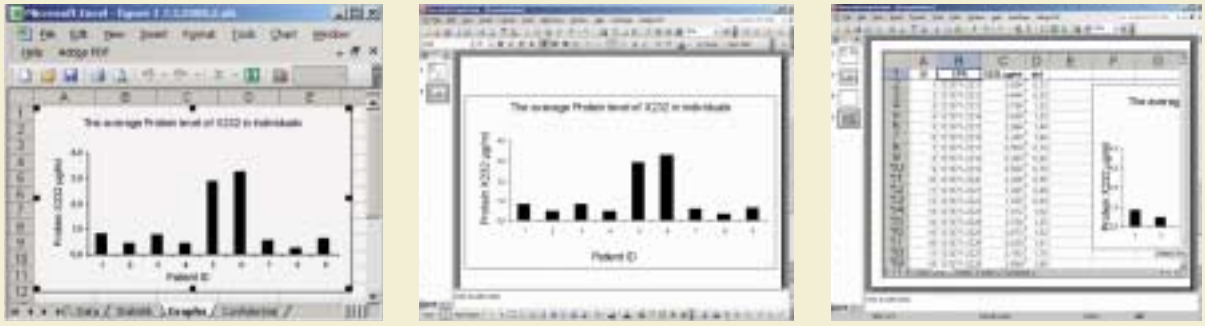
Resultater: Vi fandt skjulte data i 37% (spændvidde: 26-49%) af de downloadede PowerPoint-filer, men ikke i Word-filerne. Nogle af filerne indeholdt særligt personfølsomme oplysninger.

Konklusion: Det var nemt for os at finde persondata på internettet via utilsigtet publicering af skjulte data. Vi giver nogle simple råd mht. at undgå dette.

Forskningsbaseret viden inden for sundhedsvidenskaben har som grundlæggende fundament, at der kan indsamles og opbevares valide data om f.eks. sundhed og sygdom. Dataene stammer ofte fra identificerbare personer, der med eller uden samtykke direkte eller indirekte leverer oplysninger. Praktisk talt al klinisk forskning og sundhedstjenesteforskning er baseret på en eller anden form for personhenførbare data. Dette gælder også for den basale biomedicinske grundforskning på en lang række delområder. Arbejdet med data stiller store krav til forskernes evner og vilje til at sikre, at følsomme oplysninger udelukkende bliver udnyttet på en korrekt forskningsmæssig måde. Det gælder opbevaring, bearbejdning og udveksling i henhold til love og regler mellem forskere og mellem forskere og samfund.

For at sikre at disse regler overholdes, findes der en række fornuftige retningslinjer, der er baseret på f.eks. persondataloven [1], der sikrer borgerens ret med hensyn til, hvilke oplysninger der må indsamles, hvordan de må distribueres, samt forhold vedrørende aktindsigt og datatilsyn. I forbindelse med forskningsmæssig udnyttelse registreres det via Datatilsynet [2], hvilke forskningsregistre der oprettes, hvem der har adgang til data, og hvem der databehandler. Datatilsynets regler for opbevaring og behandling af data er klare og omfatter bl.a. en adskillelse af personidentificerbare variable og variable, der indeholder følsomme oplysninger. I andre sammenhænge er der endvidere love og regler for Det videnskabsetiske komitésystem [3] og sundhedsloven via Sundhedsstyrelsens Enhed for Tilsyn [4], der er involveret i forbindelse med præcisering af f.eks., hvilke data om patienter man må indsamle, og hvordan disse data må håndteres.

Mellem forskere er der grundlæggende en stor respekt og tillid omkring det ofte meget store og omfattende arbejde, som den enkelte forsker eller forskningsgruppe udfører med dataindsamling, analyser og formidling. Inden for lægevidenskaben har man en udbredt professionel tradition og norm,



Figur 1. A. Data bliver brugt til at generere en graf i Excel. **B.** Grafen markeres og kopieres med genvejstasten Ctrl + C hvorefter den indsættes i PowerPoint med genvejstasten Ctrl + V. **C.** Hele det oprindelige Excel-regneark åbnes i PowerPoint ved et dobbeltklik på grafen.

der skal beskytte den enkelte forskers arbejde, indtil det er publiceret. Samtidig er der ofte en stor tillid til, at analyser foretages efter bedste praksis. Som for personhenførbare data gælder det derfor også for forskeres rådata og analyser, at disse beskyttes mod almindelig spredning og kun udnyttes i den rette forskningsmæssige sammenhæng. Det beskytter forskerens integritet i forhold til det bagvedliggende arbejde og beskytter mod misbrug af data og mod at andre fristes til at benytte de opnåede resultater uden tilladelse. I forbindelse med uretmæssig brug af forskningsdata, -resultater og -analyser er der mulighed for at indklage misbrug eller plagiat for Udvalgene vedrørende Videnskabelig Uredelighed [5].

Ved formidlingen af forskningsresultater benytter man ofte det elektroniske medie, og det gør man selvfølgelig også i forbindelse formidling via internettet. Det teknologiske grundlag for, at data kan spredes utilsigtet via elektroniske medier, er brugen af dokumenter fra især MS Word og MS PowerPoint. I forbindelse med en grafisk præsentation baseret på en række observationer lagret i et databaseprogram som MS Excel vil de fleste forskere benytte *copy-paste*-funktionen, hvor man enten med musen eller ved at taste (Ctrl + C/Ctrl + V) kopierer et diagram ind i MS PowerPoint eller MS Word (**Figur 1**).

Herved indsættes ud over figuren også det bagvedliggende datamateriale. Det gælder både data, der blev benyttet til at fremstille figuren og alle andre data, man evt. også havde gemt i det regneark (inklusive andre faner i regnearket). Via diagrammet kan man umiddelbart hente data frem igen ved at dobbeltklikke på figuren.

I andre sammenhænge indsættes der billeder, hvor der grafisk er anonymiseret vha. »bjælker« over personidentificerende steder. Disse anonymiseringer kan ofte let fjernes i PowerPoint og Word.

I denne artikel vil vi med en gennemgang af nogle søgninger påvise utilsigtede brister i den ovenfornævnte beskyttelse af personhenførbare data og forskningsdata samt komme med forslag til, hvordan dette undgås i forbindelse med brugen af elektronisk formidling.

Metode

For at undersøge forekomsten af utilsigtet udbredelse af følsomme persondata foretog vi en søgning på internettet via søgemaskinen Google. Vi lavede søgninger på hhv. *Clinical trial, Double blind* og *Health informatics* og indskrænkede disse til at omfatte PowerPoint-præsentationer eller Word-dokumenter. For hver søgning gjorde vi op, hvor mange hit, der fremkom, og vi tog de første 110-140 PowerPoint og Word-dokumenter afhængigt af adgangen til de pågældende sites. Hver fil blev gennemgået med henblik på, om der var bag-

Tabel 1. Stikprøve af henholdsvis PowerPoint (.PPT)- og Word (.DOC)-filer fundet ved en Google-søgning, hvor der blev søgt på hhv. *Clinical trial, Double blind* og *Health informatics* samt filtype (PPT eller DOC).

Filtype/søgeord	Hit på Google, n	Stik-prøver, n	Indeholdt graf eller tabel, n (%)	Baggrunds-data i graf/-tabel, n (%)	Personidentifi-cerbare data i graf/tabel, n (%)
.PPT					
Clinical trial	94.200	133	35 (26)	5 (14)	5 (14)
Double blind	29.200	112	75 (67)	28 (37)	10 (13)
Health informatics	37.800	121	26 (21)	7 (27)	6 (23)
.DOC					
Clinical trial	249.000	103	5 (5)	0 (0)	0 (0)
Double blind	158.000	106	21 (20)	2 (10)	0 (0)
Health informatics	107.000	101	14 (14)	0 (0)	0 (0)

vedliggende data i grafer eller tabeller. I bekræftende fald undersøgte vi, om det var muligt at fremdrage de bagvedliggende data, og om de var personhenførbare.

Endvidere lavede vi en Google-søgning, der blev indskrænket til hjemmesider fra Danmark. Vi gennemsøgte de første ca. 1.500 filer, der fremkom på søgeordene personlig, data, cpr og/eller patient afgrænset til PowerPoint-filer. I disse undersøgte vi, om der var bagvedliggende data, der indeholdt personhenførbare data (navn og cpr-nummer), og om der var tilknyttet særligt følsomme helbredsdata. Vi præsenterer i denne artikel kun nogle eksempler på disse.

Resultater

I forbindelse med den generelle søgning på Google fandt vi, at det i stikprøverne var muligt at fremdrage skjulte data fra de filer, der indeholdt grafer eller tabeller (**Tabel 1**). I op mod 37% (95% sikkerhedsinterval: 26-49%) af PowerPoint-præsentationerne fandt vi skjulte data. Kun i få tilfælde gjaldt det for Word-dokumenterne. I op til 23% (9-44%) af PowerPoint-præsentationer med grafer eller tabeller var der personidentificerbare data. Det fandt vi ikke i Word-dokumenter.

Ved en gennemgang af PowerPoint-præsentationer fra danske sites var det relativt nemt at finde navne og cpr-numre, der var sammenhængende med endog meget følsomme sundhedsoplysninger (**Tabel 2**). Det gjaldt oplysninger om f.eks. helbred, operative indgreb og resultater fra afprøvning af behandlinger.

Tabel 2. Antal cases hvor vi fandt personhenførbare data som navne og cpr-numre på patienter, og hvor data var kombineret med følsomme oplysninger om helbred.

Eksempler	Navne	Cpr-numre
En PowerPoint fra et universitetshospital	161	161
Fire PowerPoint fra f.eks. en offentlig sundhedsmyndighed	1	1
En PowerPoint fra en region	0	25.763
En PowerPoint fra en offentlig institution	29	29

Tabel 3. Information om at undgå utilsigtet spredning af personhenførbare data og forskningsresultater.

Generelle retningslinjer

- Indsæt altid tabeller og figurer som billeder (ikke med Ctrl + C/Ctrl + V) ved at benytte:
  Rediger → Indsæt speciel → Billede
- Adskil altid identifikation og følsomme data jf. Datatilsynets retningslinjer
- Publicer ikke Word, WordPerfect, Excel, PowerPoint eller Access på internettet
- Udlever ikke filer fra Word, WordPerfect, Excel, PowerPoint eller Access til andre, medmindre det er i en arbejdsgruppe med aftaler om fortrolighed
- Publicer altid som PDF-dokumenter
- Slet filer, der er lagt på andre pc'er i forbindelse med foredrag mv. og tøm pc'ens papirkurv

Datatilsynet – sådan undgår du at få personoplysninger indlejret i PowerPoint-præsentationer mv. [6].
Take good care of your data – Svend Juul om datahygiejne [7].
Sektion for Sundhedsinformatik – Aarhus Universitet [8].

Diskussion

Det var med en simpel søgning på internettet muligt at finde relativt mange præsentationer og dokumenter, der indeholdt data om identificerbare personer, grunddata til forskningsprojekter og forskeres analyser. I relativt mange tilfælde var der tale om ret store mængder af personidentificerbare data med meget følsomme oplysninger om helbred og sundhed.

Mange af de data, vi fandt på internettet, overholdt ikke persondataloven i forhold til beskyttelse af og omgang med patienters data, idet identifikationsdata ikke var adskilte fra følsomme oplysninger. Det er således muligt for uvedkommende at få indsigt i disse oplysninger og evt. misbruge dem.

Ingen af de fundne fremstillinger havde karakter af, at uvedkommende tilsigtet skulle kunne have adgang til dem, og ingen data blev fremstillet på en måde, så det så ud til, at andre skulle kunne have adgang til dataene. Vi må derfor konkludere, at fundet af disse mere eller mindre følsomme data skyldes en utilsigtet indarbejdning af data i fremstillingen.

Vi har med denne simple søgning på internettet påvist, at trods et fornuftigt og velfungerende regelsæt med hensyn til at beskytte persondata var det muligt at tilegne sig dem i forbindelse med, at forskerne formidlede deres resultater. Det er således nødvendigt at indarbejde regler og rutiner for, hvorledes man indsætter figurer og lignende i elektronisk fremstillet materiale, og hvordan man distribuerer det elektronisk, samt understrege vigtigheden af gældende regler for, hvordan man adskiller de følsomme oplysninger fra identifikatorer.

Der findes en række simple procedurer, der, hvis de overholdes, vil hindre, at uvedkommende får adgang til følsomme oplysninger (**Tabel 3**). Samtidig har Datatilsynet udarbejdet retningslinjer på baggrund af disse oplysninger.

Vi er bevidste om, at dette problem også findes inden for andre forskningsdiscipliner end den sundhedsvidenskabelige og inden for den finansielle verden.

Korrespondance: *Mads Ronald Dahl,* Institut for Folkesundhed, Afdeling for Biostatistik, Sektion for Sundhedsinformatik, Aarhus Universitet, DK-8000 Århus C.
E-mail: mrd@folkesundhed.au.dk

Antaget: 30. maj 2008
Interessekonflikter: Ingen

Litteratur

1. Persondataloven. Lov nr. 429 af 31/05/2000. Justitsministeriet. www.retsinformation.dk/Forms/R0710.aspx?id=828 (15. maj 2008).
2. Datatilsynet. www.datatilsynet.dk (15. maj 2008).
3. Komiteloven. Lov nr. 402 af 28/05/2003. Indenrigs- og Sundhedsministeriet. www.retsinformation.dk/Forms/R0710.aspx?id=29142&exp=1 (15. maj 2008).
4. Sundhedsloven. Lov nr. 546 af 24/06/2005. Indenrigs- og Sundhedsministeriet. www.retsinformation.dk/Forms/R0710.aspx?id=10074 (15. maj 2008).
5. Udvalgene vedrørende Videnskabelig Uredelighed. Forretningsorden. http://fi.dk/site/forside/raad-komiteer-udvalg/udvalgene-vedroerende-videnskabelig-uredelighed/forretningsorden-uvvu (15. maj 2008).
6. www.datatilsynet.dk/offentlig/internettet/oplysninger-indlejret-i-PowerPoint (15. maj 2008).
7. www.folkesundhed.au.dk/uddannelse/software/takecare.pdf (15. maj 2008).
8. http://hilist.au.dk/index.php/it-security (15. maj 2008).