Tidsskriftet Tidsskrift for Den norske legeforening

Husk å vaske dataene før bruk!

MEDISIN OG TALL

MARIANNE RIKSHEIM STAVSETH

m.r.stavseth@medisin.uio.no
Marianne Riksheim Stavseth er statistiker ved Seksjon for klinisk rus- og avhengighetsforskning (RusForsk) ved Oslo universitetssykehus og postdok i forskningsprosjektet PriSUD – Diagnostisering og behandling av ruslidelser i fengsel.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Data som er samlet inn, enten automatisk eller av mennesker, vil i de aller fleste tilfeller inneholde små og store feil eller mangler. Om de ikke håndteres, kan det skape problemer.

Feil i innsamlede data kan og vil oppstå i de aller fleste epidemiologiske og kliniske studier, til tross for godt design, nøye planlegging og god utførelse. Disse feilene må identifiseres, rettes opp i (hvis mulig) og rapporteres. Prosessen kaller vi gjerne *datavask* (1). Vasking bør foretas i forkant av analysering. En grundig vaskeprosess er tidkrevende, spesielt om datamaterialet er stort, og kan være teknisk utfordrende. Men om jobben gjøres på en god måte, er det vel verdt investeringen.

Det stilles i dag stadig høyere krav til rapportering av forskning gjennom forskjellige retningslinjer, som f.eks. CONSORT (*Consolidated Standards of Reporting Trials*) for randomiserte kliniske studier og STROBE (*Strengthening the Reporting of Observational Studies in Epidemiology*) for observasjonelle data. Her oppfordres det til økt rapportering av valg som blir tatt i arbeidet med dataene, og vaskingen er gjerne et første steg i denne prosessen.

Deskriptive analyser

I små datasett kan de aller fleste feil oppdages ved hjelp av visuell inspeksjon. Dette blir imidlertid fort uhåndterbart i større datasett. Da vil man ha behov for å ta i bruk forskjellige verktøy for å oppdage feil.

En god måte å starte en vaskeprosess på, er å gjøre *deskriptive analyser* av datamaterialet. Dette kan gjøres i din foretrukne statistikkprogramvare. Ved å sjekke minimums- og maksimumsverdier, gjennomsnitt og avviksmål kan man ofte enkelt oppdage om variabler inneholder ugyldige verdier, manglende verdier eller tall som er formatert feil. Dette kan også gjøres ved enkle grafiske fremstillinger eller krysstabeller.

La oss se på et tenkt eksempel. Personer som blir innlagt på sykehus, registreres i et eget system som skal brukes til å se på liggetid. Her registreres ID, kjønn, alder, fødselsår, start- og sluttdato for behandlingen samt informasjon om røyking (figur 1). En sjekk av variabelen «alder» vil i dette tilfellet raskt avsløre at alderen til person 6 er feil. Men det er ikke lett å se ved første øyekast at materialet inneholder flere feil.

ID	Kjønn	Alder	Fødselsår	Start	Slutt	Røyker	Sigaretter per dag	Feil
1	Mann	58	1956	05/12/2014		Ja	15	Mangler sluttdato
2	Mann	36	1978	19/04/2014	18/04/2014	Nei	-	Negativ behandlingstid
3	Mann	21	1993	15/08/2014	21/09/2014	Nei	-	
4	Kvinne	49	1965	02/06/2014	14/06/2014	Nei	-	Dato for ny innleggelse før dato for utskrivelse
4	Kvinne	49	1965	13/06/2014	30/06/2014	Nei	-	
5	Mann	34	1980	07/02/2014	19/02/2014	Nei	5	Nei til røyk, men antall sigaretter oppgitt
6	Mann	166	1948	03/04/2014	16/04/2014	Ja	10	Feil alder
7	Kvinne	52	1962	12/12/2014	23/12/2014	Ja	20	
8	Kvinne	55	1959	29/01/2014	31/01/2014	Nei	-	Ren duplikat
8	Kvinne	55	1959	29/01/2014	31/01/2014	Nei	-	
9	Mann	81	1933	13/11/2014	01/12/2014	Ja	20	Duplikat, men røyker vedkommende eller ikke?
9	Mann	81	1933	13/11/2014	01/12/2014	Nei	-	

Figur 1 Et tenkt eksempel på et datasett over personer innlagt på sykehus.

Økt kompleksitet

Jobber man med større og mer komplekse datafiler, trengs mer sofistikerte kontroller. Slike kontroller må settes opp som automatiserte søk i datafilen, da manuelle kontroller blir for omfattende. Søkene vil være spesifikke for hvert datasett og basert på bl.a. typen variabler og formatet til settet. Det er derfor umulig å gi helt generelle råd til vasking av større datafiler, og her ligger også den største utfordringen: Datavaskingen må tilpasses hvert enkelt tilfelle. Noen av de vanligste feilene som opptrer, er det dog nyttig å diskutere.

Manglende data er ofte en utfordring. I vårt eksempel mangler person 1 sluttdato for sitt sykehusopphold. Liggetiden er dermed umulig å regne ut. Her må man velge om man kun skal bruke komplette data, eller om man skal gjøre en form for imputering, det vil si estimere den manglende verdien basert på observerte data.

Duplikater forekommer ofte. I tilfellet hvor duplikatene er rene kopier av hverandre (to linjer er identiske, se person 8 i eksempelet), kan den ene linjen enkelt slettes, men i mange tilfeller vil en registrering være gjort to (eller flere)

ganger med små forskjeller, som f.eks. for person 9 i eksempelet – røyker vedkommende eller ikke?

I de fleste tilfeller er det også nødvendig å sjekke om dataene har *logiske brister*. F.eks. er det notert at person 5 i eksempelet ikke røyker, men også at vedkommende røyker fem sigaretter per dag. Slike feil kan være krevende å oppdage, men er viktige å behandle.

Om datamaterialet er samlet inn over tid, må man *sjekke sekvenser* – at hendelser er plassert logisk i tid. I eksempelet kan man finne en behandlingsslutt som ligger forut i tid for behandlingsstart (person 2) og to behandlingsepisoder som overlapper i tid (person 4).

Datavasking er å ta valg

I noen tilfeller vil man kunne rette opp feil i datamaterialet, men det vil ikke alltid la seg gjøre. Derfor vil datavasking, når det kommer til stykket, i stor grad handle om å ta valg. Dette kan f.eks. gjelde om ekstremverdier skal fjernes eller ikke, om lite utfylte spørreskjemaer skal utelates, eller om duplikater med små forskjeller skal fjernes. Valgene kan være avgjørende for videre analyser og må derfor vurderes nøye, rapporteres og i enkelte tilfeller sensitivitetstestes.

Akkurat hvordan et datamateriale kan og bør vaskes, vil variere fra studie til studie, men generelt er det viktig å ikke undervurdere innsatsen. Det settes sjelden av egen tid til å vaske data, og tidspress kan føre til at man tyr til enkle løsninger. Ofte har man brukt mye tid og ressurser på datainnsamlingen, og da er det er synd å ikke legge inn de ekstra arbeidsdagene som sikrer høykvalitetsdata. Et godt vasket datasett vil gjøre videre arbeid med materialet enklere, øke datakvaliteten og redusere sjansen for feilrapportering.

REFERENCES

1. Van den Broeck J, Cunningham SA, Eeckels R et al. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2005; 2: e267. [PubMed][CrossRef]

Publisert: 30. januar 2023. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.22.0825