Tidsskriftet
DEN NORSKE LEGEFORENING

Manglende data – sjelden helt tilfeldig

MEDISIN OG TALL

STIAN LYDERSEN

E-post: stian.lydersen@ntnu.no
Stian Lydersen er dr.ing. og professor i medisinsk statistikk ved Regionalt kunnskapssenter for barn og unge – psykisk helse og barnevern (RKBU Midt-Norge), Institutt for psykisk helse, Fakultet for medisin og helsevitenskap, Norges teknisk-naturvitenskapelige universitet (NTNU).
Forfatter har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

I de fleste medisinske studier vil det være «huller» i datasettet, slik at data mangler helt eller delvis for noen av personene. Dette medfører redusert statistisk styrke. Men det største problemet er at dette kan medføre skjevhet i resultatene. Det er sjelden helt tilfeldig hvilke personer man mangler data på.

Komplette data fra alle deltagerne i medisinske studier forekommer sjelden. Dette gjelder særlig i observasjonelle studier, men også i randomiserte kontrollerte studier (figur 1). Et eksempel er en randomisert studie av to behandlingsløp for hoftebrudd med 397 inkluderte pasienter (1). Pasientenes mobilitet ble målt ved inklusjon og etter 1, 4 og 12 måneder hos henholdsvis 344, 333, 325 og 284 pasienter. Bare 237 av pasientene ble evaluert på alle fire tidspunkt. Totalt 67 pasienter døde i løpet av studien. Det manglet altså data for mange av de pasientene som var i live.

		Variabel		
		1	*2*	*3*
Person	*1*	*1,4*	*1,9*	*2,5*
	2	*1,2*	*2,0*	*2,2*
	3	*0,7*	*?*	*1,8*
	4	*0,5*	*1,3*	*?*

Figur 1 *Eksempel på datamatrise med manglende data.*

Mekanismer

I hvilken grad mangler data tilfeldig? Dette kalles mekanisme for manglende data og er viktig ved valg av metode for å håndtere manglende data (2). Man skiller mellom tre typer antagelser (3): Mangler helt tilfeldig (missing completely at random, MCAR), mangler betinget tilfeldig (missing at random, MAR), og mangler ikke-tilfeldig (missing not at random, MNAR).

I det første tilfellet vil sannsynligheten for manglende data verken avhenge av observerte eller uobserverte data. Dette kan for eksempel være realistisk hvis blodtrykk ikke ble målt en av studiedagene fordi måleinstrumentet var defekt denne dagen.

Dersom sannsynligheten for manglende data er avhengig av observerte data, mangler ikke data helt tilfeldig, men kan mangle betinget tilfeldig. Dette kan være tilfelle dersom vi har registrert alder på alle pasientene, men ser at en større andel av de yngre enn de eldre møter opp for blodtrykksmåling. Den engelske betegnelsen *missing at random* kan lett misforstås, fordi sannsynligheten for manglende data avhenger av observerte data. Dersom det i tillegg er slik blant de eldre at de med dårligst helsetilstand møter opp i mindre grad enn friskere eldre, vil data mangle ikke-tilfeldig. I praksis vil dette til en viss grad gjelde i de fleste studier. Men da vil en metode som forutsetter betinget tilfeldighet, gi mindre skjevhet i resultatene enn en metode som forutsetter at data mangler helt tilfeldig (4).

Man kan skille mellom antagelsene om helt tilfeldig versus betinget tilfeldig ved å studere datasettet, som i eksemplet med alder ovenfor. Men man kan aldri skille mellom betinget tilfeldig og ikke-tilfeldig ved å analysere data. Graden av ikke-tilfeldighet vil alltid være en uverifiserbar antagelse, noe som mange ikke synes å være klar over.

Håndtering

Man skal rapportere omfanget av manglende data for alle variablene som inngår i analysene, og hvordan dette ble håndtert (5, 6).

Den enkleste metoden for å håndtere manglende data er å inkludere bare de observasjonene som har komplette data (complete case analysis). Dette kan være akseptabelt dersom andelen med manglende data er liten, for eksempel under 5 %. Men dette vil avhenge av grad av avvik fra antagelsen om at data mangler helt tilfeldig (2). Metoden vil være forventningsrett (unbiased) bare hvis data mangler helt tilfeldig.

Noen ganger kan man imputere manglende data, dvs. sette inn verdier som er estimert ved hjelp av de observerte data. Ved enkel imputering setter man inn et estimat i hvert hull i datamatrisen. Dette kan gi forventningsrette estimater hvis data mangler betinget tilfeldig. Men man underestimerer usikkerheten, dvs. får for smale konfidensintervaller og for små p-verdier, med mindre andelen manglende data er liten (2, s. 442–7). En bedre metode er multippel imputering, der man lager flere komplette datasett. I hvert datasett er de imputerte verdiene trukket fra sannsynlighetsfordelingen gitt de observerte verdiene. Det er vanligvis tilstrekkelig med 20–100 imputerte datasett (7, s. 58). Deretter kombineres resultatene fra de komplette datasettene, og man får forventningsrette estimater, konfidensintervaller og p-verdier hvis data mangler betinget tilfeldig (7).

I visse tilfeller kan man benytte full-informasjon-sannsynlighetsmaksimeringsmetoden (full information maximum likelihood), men denne er svært beregningsintensiv og ikke alltid mulig.

I longitudinelle studier, som i eksemplet nevnt innledningsvis (1), kan man bruke en blandet modell-regresjonsanalyse (mixed model regression analysis) uten å imputere data. Da bidrar alle pasienter som har data fra minst ett tidspunkt, og resultatene er forventningsrette når data mangler betinget tilfeldig (8, s. 285).

LITTERATUR:

1. Prestmo A, Hagen G, Sletvold O et al. Comprehensive geriatric care for patients with hip fractures: a prospective, randomised, controlled trial. Lancet 2015; 385: 1623–33. [PubMed][CrossRef]

2. Bjørnstad JF, Lydersen S. Missing data. I: Veierød M, Lydersen S, Laake P, red. Medical statistics in clinical and epidemiological research. Oslo: Gyldendal Akademisk, 2012: s. 429–61.

3. Lydersen S. Manglende data på norsk. Tidsskr Nor Legeforen 2019; 139. doi: 10.4045/tidsskr.18.0858. [CrossRef]

4. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7: 147–77. [PubMed][CrossRef]

5. Moher D, Hopewell S, Schulz KF et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869. [PubMed][CrossRef]

6. Statement STROBE. Strengthening the reporting of observational studies in epidemiology. https://www.strobe-statement.org/index.php?idstrobe-home (31.10.2018).

7. van Buuren S. Flexible imputation of missing data. 2. utg. Boca Raton, FL: CRC Press, 2018.

8. Thoresen M. Longitudinal analysis. I: Veierød M, Lydersen S, Laake P, red. Medical statistics in clinical and epidemiological research. Oslo: Gyldendal Akademisk, 2012. 259–87.

Publisert: 7. februar 2019. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0809