Crap in? Crap out!

Tankar om utbildning för forskarstuderande om datakvalitet

Vad är datakvalitet?

Korrekt redovisade data, från korrekt inmatade data, från korrekt insamlade data, insamlade med korrekt metodik, som mäter det fenomen man avser att studera i ett urval som korrekt avspeglar den population fenomenet berör.

Datakvalitet berör alltså hela den vetenskapliga processen, från hypotesformulering via studiedesign, selektionskriterier, mätmetodik, insamlingsmetodik, bearbetning av data, analys av data fram till och med redovisning av resultat. Om ett eller flera av dessa steg fallerar kan resultaten bli missvisande eller omöjliga att tolka.

Syftet med en utbildning om datakvalitet bör vara att ge handfasta råd om vad man bör tänka på i olika stadier av en studie. Tyngdpunkten för mitt intresse ligger vid den datorunderstödda organisationen, bearbetningen och analysen av insamlade data.

Förståelsen av hur olika dataprogram fungerar är lika betydelsefull som att förstå andra verktyg, t ex röntgenapparater och kemiska analysmetoder, som används i arbetet. Deltagarna ska förstå för- och nackdelar med olika typer av dataprogram och kunna utnyttja dem för olika steg i studien. De ska utifrån en given frågeställning och metodik kunna planera enkla databaser för hantering av insamlade data, med hjälp av enkla metoder förebygga inmatningsfel och hantera saknade värden redan på inmatningsstadiet.

De ska kunna skapa och använda enkla inmatningsformulär, som ger stöd för inmatning och förebygger felvärden.

Deltagarna ska från ett inmatat datamaterial kunna identifiera troliga felvärden grafiskt och med deskriptiv statistik, skapa nya variabler dynamiskt eller statiskt, studera betydelsen av bortfall. De ska kunna förstå och åtgärda de vanligaste problemen vid överföring av data mellan program.

De ska kunna skapa pivottabeller, enkla diagram och basal statistik för hela materialet och för olika grupper eller urval av data.

Utbildningen förutsätter att deltagaren är insatt i grunderna för studiedesign och elementär statistik. Deltagaren måste ha grundläggande datorkunskap, och bör förstå grunderna i kalkylbladsprogram som Excel. Programvara som används är Windows-baserad.

Studiedesign och metodik

Typ av och syfte med studie. Sambandet mellan påvisbara effekter och studiestorlek i olika studiedesigner. Hur stor population kan man teoretiskt välja studieobjekt från? Hur stor skillnad mellan grupper förväntas? Hur stor studie kan man hantera (tid, pengar, personal, lokaler, obehag för deltagare osv). Vilken typ av data vill man studera? Vilken typ av data vill man få ut av studien? Vilken typ av data kan man samla in? Kan dessa data besvara frågan?

Statistisk signifikans kontra klinisk signifikans!

Använda EpiInfo för beräkning av studiestorlek.

Metodik

Adekvat, verifierad metod? Metodjämförelser? Detektionsgränser? Upplösning?

Samma och stabil mätmetod i hela studien?

Bortfall

Hur stort bortfall förväntas? När i studien? Vilken typ och storlek av bias uppstår? Hur ska bortfallet analyseras?

CKG
Carin Kullberg
Beräkning och Bild

Mätvärdenas natur

En eller flera end-points, transienta stadier/klasser?

Dummy-variabler, kategoridata med eller utan inbördes ordning, kontinuerliga värden.

Normalfördelning och andra fördelningar, log-transformering, klassning av kontinuerliga variabler, summerande index

Autokorrelation i mätserier?

Confounders?

Data är svåra att dela, men lätta att slå ihop. Registrera ”minsta tänkta analysenhet” för sig redan från början!.

Håll rätt på enheterna. Ange tydligt i protokoll, inmatningsformulär och all output!

Datahantering

Val av programvara

Jämförelse och beskrivning av olika programtyper

- Studiedesign-stöd
- Kalkylbladsprogram
- Databasprogram
- Statistik/analysprogram
- ”Plottningsprogram”
- Grafikprogram

Hur avancerat behöver det vara?

Vilken erfarenhet av olika program finns i omgivningen? Varför just det programmet? Fungerar det med befintliga standardprogram? Prova programmet för att se hur hög ingångströskeln är.

Säkerhetsrutiner

Server-backup kontra manuell backup. Datummärkta kopior!

Säkerhet: vem kommer åt data för inmatning och output? Skydd för virus, Internetaccess, spel, uppdateringar av operativsystem och program, trådlösa nätverk osv.

Kontroll av inmatningsfel

Metoder för felkontroll, manuella och programstyrda

Förstår den som matar in värdena vad de står för? Viktigt för att minska fel.

Hantering av saknade data

Exkludering av deltagare före/efter studie. Exkludering av deltagare ur del av studie, t ex avhopp.

Enstaka saknade data: Null, 0, specifikt ”Ej svarat”-värde, ersättning med medelvärde för individ eller population

Kontroll av rådata

Nyckelvariabler: kontroll av dubbletter, omatchade värden, rimliga datum etc

Min- och maxvärden för alla numeriska variabler, jämför med detektionsgräns och medelvärden.

Fördelning av värden: Frekvens, jmf med normalfördelning, inom hela populationen och för resp grupp av förväntat lika värden, resp förväntat olika värden

Grafisk kontroll

Plottning av data för att identifiera extremvärden och möjliga felvärden

Bortfallsanalys

Selektionsbias i hela pop och för subgrupper, t ex åldersgrupper

Jämför bortfall med resultat efter analys, för att se hur mycket det kan tänkas påverkat resultatet.
Kontrollera confounders i bortfallsgrupp

Beräknade värden

Logg över när och hur beräknade variabler skapats, t ex ålder, BMI, index av enkätfrågor.

Kontrollera att strataspecifika beräkningar endast påverkat rätt strata!

Dynamiska kontra statiska beräkningar.

Organisation av insamlade data

Databas vs kalkylblad

Registertillstånd. Personuppgifter

Kommentera och dokumentera allt! Använd logg för varje studie!

Enkla tabeller:

Kolumner= variabler, rader =poster

Hellre många rader än många kolumner. Använd grupperingsvariabler.

Datatyper: booleanskt, kategori, numeriskt, text, datum/tid (olika system), (lång) fritext; bild

Standardvärden

Verifiering: värdegränser, listor

Obligatoriskt värde

Unikt kontra upprepat värde

Löpnummer

Beräknade värden

Enkla databaser

Enkla tabeller

Inmatningsformulär

Rapportfunktion, ev med viss analysmöjlighet

Exempel: (Excel)

Relationsdatabaser

Flera tabeller hopkopplade

Minskar mängd data att mata in

Minskar risk för fel vid upprepad inmatning

Standardisering.

Större möjlighet att kontrollera inmatning.

Relationstyper; 1-1, 1-n, n-n

Exempel: Access, FileMaker, 4D, EpiInfo

Inmatningsformulär

Så likt originalprotokoll som möjligt

Utveckla gärna originalprotokoll och inmatningsformulär samtidigt under studiedesign. Klarlägger ofta möjliga problem med utformning av protokoll!

Undvik scrollning

Enhetlig layout

Omfång (om för jobbigt att mata in lär inte deltagarna vara särskilt noga …)

Stöd för upprepade värden; listor, kopiering av föregående värde mm

Alltid klartext, aldrig kodvärden!

Flervalsfrågor: val av endast ett eller flera alternativ ?.Hur hantera om deltagare angivit fler alternativ när bara ett ska anges?

Kommentarfält!

Ej svarat!

Beräkningar görs av datorn, aldrig för hand!

Huvud- och underformulär

Kontinuerliga vs enstaka formulär

Alla data eller urval

Rapporter

Sammanfattar resultat

Redovisar resultat från flera källor tillsammans

Pivottabeller

Diagram

Enkel deskriptiv statistik, men kolla alltid med specialiserat statistikprogram

Frågor

SQL-frågor vs grafisk representation

Urvalsfrågor: undergrupper, söka dubbletter/omatchade värden etc

Tillägg/borttagning

Uppdatering

Beräkningar

Korsfrågor

Koppling av tabeller och/eller frågor: matchning mha nyckelvariabler

Sammanfattning av grupper av poster: aggreggering

Organisation i praktiken

Skapa en tabell

Lista variabler

Namnge variabler: Korta, klara namn, helst utan diakritiska tecken, mellanslag eller interpunktion, samt inled med bokstav, ej siffra

Ange variabelbeskrivning, om programmet tillåter

Välj datatyp och storlek

Standardvärde: lägger in ett valt värde som standard i ny post

Verifiering: värdegränser, begränsa till val från lista

Obligatoriskt värde?

Unikt värde? Nyckelvariabel? Löpnummer?

Beräknade värden: sparade eller beräknas vid behov?

Dolda fält, t ex regdatum.

Pröva inmatning av realistiska data, inkl förutsägbara fel

Undersök tabellens innehåll

Var förekommer upprepade datavärden?

Behövs inmatningshjälp i form av lista men giltiga värden? Val enbart från lista?

Många uppgifter gemensamma för grupper av poster: Dela upp på två eller flera tabeller?

Verifieringsgränser för snäva/vida?

Skapa en ny tabell med ”gruppdata”

Välja fält

Unika värden

Relationer mellan tabeller

Typer av relationer

Relationsintegritet

Standardisering av data

Jämförelse av olika nivåer

Balans mellan standardisering och hanterbarhet vid analys

Inmatningsformulär

Excel!

Access snabbformulär

Konventioner för utformning

Ledtext och inmatning

Grupp-fält för koder

Listor och kombinationsrutor

Datablad/kontinuerliga formulär/enstaka poster

Visningsformat för datatyper

Inmatningsmasker

Beräknade fält

Underformulär

Länka data mellan formulär

Rapporter

Listor

Sammanfatta alla poster

Sammanfatta gruppvis

Pivottabeller

Diagram

Frågor

Skapa frågor i grafiskt gränssnitt

Olika frågetyper

Aktionsfrågor vs urvalsfrågor

Använda som datakälla till formulär och rapporter

Exportera och importera data

Filformat

Vanliga felkällor

Statistik- och analysprogram

SPSS , EpiInfo, Excel?, Statistica, Systat etc

Import/exportformat för data och resultat

Analysmöjligheter

Programmerbarhet, Loggfiler

Hantering av ”missing values”

Urval och gruppvis analys

”Merge” och ”Aggreggate”-funktioner

Basal statistik: min, max, frekvens, fördelning, medel, median, spridning, gruppering, extremer

Plotta data; enskilda variabler och parvis