VIDENSKAB OG PRAKSIS | STATUSARTIKEL

R – en programpakke til statistisk databehandling og grafik

Overlæge Ram Benny Dessau &
post-doc. Christian Bressen Pipper

Næstved Sygehus, Klinisk Mikrobiologisk Afdeling, og
Københavns Universitet, Det Sundhedvidenskabelige Fakultet,
Biostatistisk Afdeling

R er et software for statistisk databehandling og grafik. R er et internetbaseret redskab, som udvikler sig med bidrag fra statistikere over hele verden. Det er en software, som frit kan hentes på internettet [1] og bruges. R indeholder gængse, grundlæggende og avancerede statistiske metoder. Herudover findes der *packages* til epidemiologi, overlevelsesanalyse etc. Grafikken er fleksibel og kan anvendes til publikationer. R betegnes som et miljø, idet funktionerne fra datahåndtering, grafik og statistisk analyse fungerer sammenhængende. R er ikke helt nemt, og man må regne med at skulle investere noget tid for at lære de grundlæggende principper. I det følgende bringes en kort gennemgang af R, tip til at komme i gang samt visioner om at udbrede R i det sundhedsfaglige miljø som et værktøj til forskning og kvalitet. I **Tabel 1** er der givet en oversigt over, hvilke fordele brugen af R rummer.

Hvordan kommer jeg i gang med R?

1. Hent programmet på internettet (www.R-project.org) og installer det
   a. Kommentarer begynder med # og opfattes ikke som en kommando af programmet
   b. Når du har installeret og startet R, går du op i *file* og vælger *new script*
   c. Forfatterne har konstrueret et skripteksempel med grundlæggende statistiske test, foruden koden til Figur 1, som vi gerne sender på e-mail.
   d. Skriptet køres linje for linje med ctrl-r, og resultatet ses i konsolvinduet eller et grafikvindue. Obs. Cursoren skal stå på den linje, du vil køre
   e. Et vigtigt koncept er tildeling af objektnavne med »<-«. Bemærk, hvordan disse objektnavne genbruges i de følgende linjer til grafik og statistiske test. Dermed bliver koden effektiv
   f. Bemærk, hvordan der genereres kunstige datasæt
2. Hvis du vil gå videre så find en *tutorial* f.eks. på www.R-project.org under *manuals*, hvor der også er links til *contributed documentation. Crawleys* bog [2] kan anbefales som en detaljeret trin for trin-lærebog i statistik, hvor man samtidig lærer at bruge R, inklusive grafikken. Datasæt og kode kan hentes på bogens hjemmeside. Men der er flere lignende bøger at finde, formentlig (også) udsprunget af undervisningsnoter fra universitetskurser. De grafiske muligheder og teknikker er beskrevet i *Murrells* bog [8]
3. Forudsætninger er almindeligt brugerkendskab til computere. Man skal kunne forstå fil- og stinavne, så man ved, hvor på harddisken man har gemt sit arbejde, og man skal kunne gemme data fra f.eks. et regneark som kommasepareret fil etc.

Formålet med denne artikel er dels at orientere om, at der til de kommercielle statistikpakker findes en seriøs, licensfri konkurrent, som specielt har nogle fremragende grafiske muligheder, og dels at inspirere den læser, som har behov for en statistikpakke med de grafiske muligheder og nogle af de mere specielle statistiske metoder, hvor R er førende.

Baggrund og udvikling

Programmeringssproget R [1, 2] er udviklet på basis af S, som oprindelig blev udviklet til professionelle statistikere ved AT&T's Bell Laboratorier. S blev senere kommercialiseret som S-plus. Men på grund af licensomkostninger, især for universitetskurser med mange studerende, udviklede to statistikere på New Zealand, *Ross Ihaka* & *Robert Gentleman*, en simpel version af S til undervisning. På grund af deres fornavne, og fordi R kommer før S i alfabetet, kom programmet til at hedde R. Programmet blev frigivet i 1995 under *general public license*. Siden er programmet vokset med bidrag fra statistikere over hele verdenen og bliver brugt på universiteter til både forskning og uddannelse. Der har været en frygt for, at R på grund af de mange forskellige bidragydere kunne være temmelig fejlbehæftet. Det har vist sig ikke at være tilfældet [3]. Alle nye programpakker godkendes og frigives af R Core Team inden de bliver lagt på R-projects egen hjemmeside, og R-interesserede statistikere følger udviklingen. Dermed afprøves nye programpakker, og i praksis bliver eventuelle fejl hurtigt rettet. Den åbne programkode fremmer både udvikling og fejlretning.

Statistik

På mange områder er R også førende med at implementere nye statistiske analyser. Eksempler er metoder til analyse af ikkelineære dosis-respons-modeller [4], multivariatanalyse [5] og analysemodeller inden for overlevelsesanalyse [6, 7] . R har stærke funktioner til simuleringsmetoder som f.eks. *bootstrap*, der kan bruges, når data ikke er normalfordelte, eller når forudsætningerne for de gængse regressionsmodeller ikke er opfyldt. Multipel regressionsanalyse med brug af såkaldte *generalized additive models* [8], hvor man i samme model kan kombinere kategoriske variable og nonlineær regression, er velegnet til lægevidenskabelige data, hvor man ofte har en blanding af kategoriske og kontinuerte data. Disse komplekse regressionsanalyser er praktisk tilgængelige at bruge, når man først har forstået de grundlæggende principper for opstilling af modellen i en »formel«. Man skal have hjælp af en kyndig person for at komme i gang, men herefter kan man selv arbejde videre med at finjustere modellerne.

Tabel 1. Egenskaber ved programmet R.

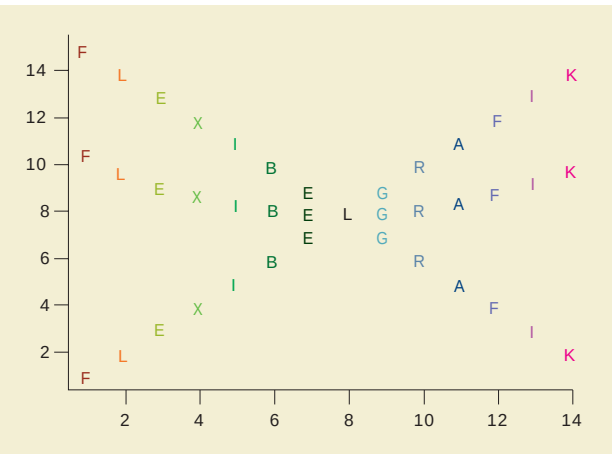
Ingen licens*	Programmet kan frit bruges også på hjemmecomputeren, hvor en del forskningsarbejde typisk foregår. Yngre læger skifter ofte ansættelsessted. I værste fald har man ikke noget statistikprogram det nye sted, eller man skal starte forfra med at lære et nyt program at kende
Forsøgsplanlægning*	R har stærke værktøjer til generering af kunstige datasæt og simuleringer. Man kan dermed f.eks. simulere forsøgsudfald med statistisk usikkerhed, hvilket kan bruges til powerberegning
Grafik*	Den største fordel ved R er de fleksible grafiske funktioner, f.eks. multiple plot. Det kræver forholdsvis lidt at fremstille en pæn graf
Rettelser og ændringer	Det kan tage en del tid at få opstillet den statistiske analyse af nogle resultater eller at tilpasse en god grafikfigur. Men når man efterfølgende skal indføre rettelser eller gentage lignende analyser eller grafiske plot, er det nemt at lave ændringer
Dokumentere databehandling	R håndteres ved hjælp af programstumper kaldet *scripts*. Disse filer kan gemmes, tages frem til senere brug og sendes med mail til medforfattere
Hjælpefunktion*	R har en særdeles brugbar hjælpefunktion med *fuzzy matching*, så man ikke behøver at stave helt korrekt. Hjælpen er illustreret med skript-eksempler, så man konkret kan se, hvordan en bestemt funktion fungerer. Man kan kopiere disse skripter og tilpasse dem til eget brug. R har et righoldigt bibliotek af datasæt, hvor man kan se eksempler på datastrukturer og brugen af statistiske modeller. Herudover er der en del hjælp at hente både på R-projects hjemmeside og på internettet, typisk undervisningsmateriale fra universitetskurser. Man kan søge med hjælp af søgemaskiner som f.eks. Google
Kvalitetsformål*	Resultater, som analyseres med jævne mellemrum, vil kunne automatiseres. Dette gælder også grafiske fremstillinger
Undervisning og kurser*	I undervisning, hvor der indgår statistiske metoder, kan R bruges som trin for trin-øvelser med grafiske figurer. Kursisterne kan få kursusmaterialet med hjem i form af skripter
Installation*	Dette er nemt, og der er ingen påvirkning af computerens drift i øvrigt ud over lidt forbrug af plads på harddisken

*) Punkter, hvor R kan have fordele frem for andre statistikprogrammer.

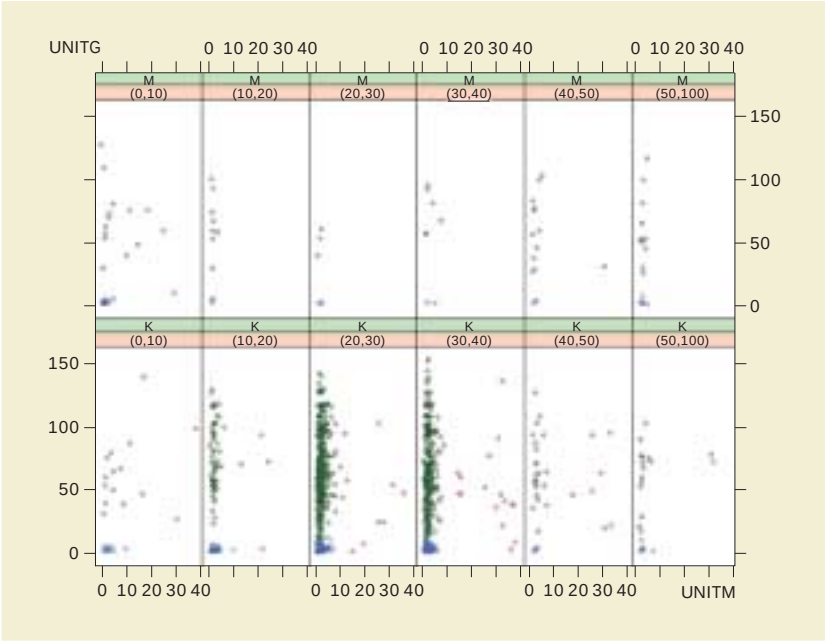
Ved overlevelsesanalyse bruges *frailty*-modeller til korrelation mellem overlevelsestider, og de er derfor populære inden for familiestudier og ved andre design med klyngestruktur. Hvis man vil regne på sådanne meget beregningstunge modeller, er R i praksis den eneste mulighed [6], dette gælder også, hvis man ønsker at inkludere tidsafhængige effekter af sine risikofaktorer i f.eks. en Cox-regression [7]. Der synes ikke at mangle statistiske metoder. For eksempel er de nyere metoder vedrørende konfidensintervaller for binomiale proportioner [9] implementeret i R i pakker, som hedder Epi, Hmisc og binom. *Receiver operating curves* bruges til evaluering af diagnostiske test. Denne metode findes også som menuvalg i SPSS, men R er nemmere at gå til og indeholder et større udvalg af statistiske test til dette formål, foruden at grafikken fungerer bedre.

Grafik

En af de helt store fordele ved R er den overlegne grafik [10]. Det er relativt nemt at producere flotte, informative og fleksible plot (**Figur 1**) og gemme dem i standard Windows-billedformater. Der er grafiske funktioner til analyse af komplekse datasæt (**Figur 2**). Opsætningen af Figur 2 er defineret ud fra data ved hjælp af navnene på de relevante datavariable som er skrevet op i en kortfattet formelsyntaks. Farven på punkterne er defineret ved de fire mulige positive og negative udfald af immunglobulin (Ig)M- og IgG-analysen (»groups = resultat«).



Figur 1. Eksempel på et plot produceret med fire kommandolinjer og lagt ud i en *.jpeg-fil.Grafen illusterer det grundlæggende princip i R (*painters model*): hvordan man linje for linje kan tilføje nye elementer til samme graf. Som eksempel på fleksibilitet vises, at hvert datapunkt kan defineres forskelligt.



Figur 2. Eksplorativ grafisk dataanalyse af rutineprøver for parvovirus-immunglobulin (Ig)M- og IgG-antistoffer. Analysen viser klinikernes valg af patienter fordelt på køn og aldersgrupper med de fundne resultater af IgM og IgG i arbitrære enheder. Grafikpakken kaldet *lattice* er anvendt med følgende syntaks: xyplot(UnitIgG~UnitIgM|ALDERSGRUPPE+SEX,groups=Resultat).

Programmering i R

R betjenes ved at skrive kommandoer i ren tekst, og det giver en række fordele samt fleksibilitet i forhold til et menustyret program (Tabel 1). Har man først investeret noget tid i at lære at programmere i R, kan man strømligne og effektivisere sine statistiske analyser. Dertil kommer, at man i høj grad er selvhjulpen på grund af den gode hjælpefunktion og de velfungerende internetsider for R-brugere, hvor man kan få besvaret sine spørgsmål og finde relevante programkoder.

Det kræver noget tålmodighed i begyndelsen, hvor man ofte får fejlmeddelelser på grund af småfejl i syntaksen. Bemærk at stinavne i R skrives med »/« og ikke »\« som i Windows. Sådanne »småting« kan drille frygteligt i begyndelsen, før man får lidt erfaring. Det er efter forfatternes opfattelse en investering, der kan betale sig, og den objektorienterede syntaks er forholdsvis let at læse og forstå.

Datahåndtering

Håndtering af datasæt (*data frames*) er ikke let, og her virker hjælpen i R ikke godt, men *Crawleys* bog [2] har et indledende kapitel om *data frames*. R er ikke velegnet som egentlig database eller til indtastning af rådata. Der er ikke faciliteter til håndtering af relationelle tabeller. Derimod er der funktioner til at opdele data i intervaller (f.eks. aldersklasser), omkode kategoriske data, krydstabellere etc.

Brug af R i sundhedsvæsenet og til undervisning

R kan være et værktøj til at udvikle og modernisere den statistiske kompetence i det lægefaglige miljø. Det foreslås at spare licenspengene til de kommercielle statistikpakker og i stedet investere i R-kurser for de ansatte som en faglig kompetenceudvikling. Man kunne tænke sig en model med nøglepersoner, ligesom det bruges på andre områder inden for kvalitetsudvikling, hygiejne, utilsigtede hændelser m.m. Disse nøglepersoner kan så hjælpe kolleger med konkrete problemstillinger. En anden grund til at bruge R på institutioner er, at man kan programmere grafik (f.eks. til benchmarking), som kan opdateres automatisk og dermed betjenes af f.eks. kontorpersonale. Grafikken kunne være i form af søjlediagrammer over ventetider, komplikationsrater m.m. til brug på sygehusets hjemmeside, som skal opdateres regelmæssigt. Hvis R bruges formelt i institutionens regi, kan det overvejes at give støtte til R-foundation. Listen over støttende organisationer omfatter universiteter, medicinalfirmaer og Folkhälsoinstituttet i Finland. R er velegnet til undervisning, idet man til undervisningsbrug kan udvikle programmer inklusive grafik, som kursisterne vil kunne benytte på deres egne computere bagefter.

Korrespondance: *Ram Dessau*, Klinisk Mikrobiologisk Afdeling, Næstved Sygehus, DK-4700 Næstved. E-mail: rde@cn.stam.dk

Antaget: 6. august 2007
Interessekonflikter: Ingen

Litteratur

1. The R Project for Statistical Computing. 2007. www.r-project.org /okt. 2007.
2. Crawley JC. Statistics. An introduction using R. Chichester: John Wiley & Sons, 2005.
3. Dalgaard, P. Statistical software development: selected war stories. Research report 05/7. Københavns Universitet, Institut for Biostatistik, 2005.
4. Ritz C. Bioassay analysis using R. J Stat Software 2005;12:1-22.
5. Dalgaard, P. New R functions for multivariate analysis. Research Report 06/11. Københavns Universitet, Institut for Biostatistik, 2006.
6. Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. J Computational Graphical Stat 2003;12:156-75.
7. Martinussen T, Scheike TH. Dynamic regression models for survival data. New York: Springer-Verlag, 2006.
8. Wood SN. Generalized additive models. An introduction with R. Boca Raton: Chapman & Hall/CRC, 2006.
9. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat.Med 1998;17:873-90.
10. Murrell P. R Graphics. Boca Raton: Chapmann & Hall/CRC, 2006.