Reprezentarea datelor clinice în vederea prelucrării statistice.

A.D. Corlan, 2014.

Orice studiu clinic presupune extragerea unui set sintetic de date în vederea prelucrării statistice.

În cazul cel mai simplu, acest set constă într-un tabel, cu câte un rând pentru fiecare **caz** (fiecare pacient) şi câte o coloană pentru fiecare **variabilă** (cum ar fi vârsta sau prezenţa unui diagnostic). În cazurile mai complicate, pot fi mai multe tabele cu legături între ele, sau alte structuri de date.

Unele dintre coloane sunt destinate prelucrării statistice computerizate a datelor, în vreme ce altele sunt simple comentarii.

În cele de mai jos, vom numi 'variabile' doar acele coloane care sunt destinate prelucrării computerizate.

De obicei, realizarea unui studiu include următorii paşi:

A: Se stabilesc obiectivele studiului, se estimează pe baza unor rezultate anterioare din literatură de câte cazuri ar putea fi nevoie, se evaluează fezabilitatea.

B: Se stabileşte protocolul studiului, inclusiv "capul de tabel", adică se aleg şi se documentează variabilele corespunzătoare capului de tabel.

C: Se execută protocolul studiului, din care rezultă un set de documente primare (foi de observaţie, buletine de analiză, înregistrări, imagini, etc) pentru fiecare caz.

D: Se întocmeşte tabelul sintetic, completând valorile variabilelor în fiecare caz.

E: Se face o analiză computerizată a datelor din tabel, rezultând un număr de statistici. Prin **statistici** înţelegem aici numere calculate pe baza valorilor variabilelor, care caracterizează întreaga populaţie[1].

F: Se interpretează statisticile rezultate din E şi se întocmeşte un raport asupra acestora, care apoi se sintetizează într-o lucrare.

În faza E se dovedeşte însă, deseori, că statisticile necesare atingerii obiectivelor din faza A nu se pot determina din cauza unor probleme tehnice ivite în fazele B sau D, adică la stabilirea şi documentarea variabilelor şi la completarea valorilor.

Este atunci necesară corectarea capului de tabel şi reluarea fazei D, uneori şi C (ceea ce poate fi imposibil sau foarte costisitor), ceea ce poate duce la întârzieri substanţiale.

Scopul acestui document este să enumere cele mai frecvente erori şi probleme de acest tip, pur tehnic, în speranţa că astfel se vor preveni sau reduce întârzierile respective.

Cea mai utilă recomandare generală este: documentaţi în scris, în faza B, capul de tabel cu: (i) numele fiecărei variabile; (ii) unitatea de

[1] În termeni tehnici: parametrii estimaţi ai distribuţiei probabilităţii variabilei.

măsură; (iii) intervalul sau enumerarea valorilor pe care le poate lua; (iv) descrierea, narativă, a semnificaţiei variabilei.

Discutaţi apoi această documentaţie cu statisticianul, înainte de a începe colectarea datelor. Dacă nu aţi facut această documentare şi discutare, faceţi-o acum.

Toate valorile variabilelor prelucrabile sunt de fapt numere

Pentru a putea face prelucrarea statistică, calculatorul trebuie mai întâi să traducă automat fiecare valoare a unei variabile într-un număr.

Calculatorul nu poate interpreta, în acest scop, decât tipuri foarte limitate de expresii (formulări, date introduse), şi anume:

(1) Valorile unei variabile pentru fiecare caz sunt date chiar prin numere, întregi sau fracţionare (cu virgulă). În acest caz trebuie stabilit la documentarea capului de tabel ce fel de număr (întreg sau fracţionar), care e valoarea minimă sau maximă, şi ce reprezintă el. Unitatea de măsură trebuie să fie aceeaşi pentru toate valorile unei variabile.
(2) Valorile sunt simboluri—cuvinte sau acronime dintr-un set predefinit, de exemplu 'bărbat/femeie'—care sunt ulterior transformate în numere întregi în cursul prelucrării. Setul predefinit şi semnificaţia fiecărei valori trebuie stabilite la documentare. Este esenţial ca fiecare semnficaţie să fie

reprezentată prin exact acelaşi simbol. De exemplu, nu putem reprezenta sexul bărbătesc, în aceeaşi coloană, o dată prin ‘B’ şi altă dată prin ’M’ sau ’m’, ci trebuie sa alegem de la început un singur simbol.

(3) Valorile sunt momente în timp, cum ar fi data naşterii sau data unei intervenţii, care sunt introduse cu an/luna/zi/ora/minut, dar în vederea prelucrării vor fi transformate de calculator tot în numere simple—de pildă număr de secunde scurse de la o dată unică din trecutul mai îndepartat. Trebuie precizat, la documentarea capului de tabel, ce reprezintă fiecare dată şi cu ce acurateţe este măsurată (un an, o zi, ora şi minut, etc).

Lipsa unei valori poate fi înregistrată printr-un spaţiu sau, preferabil, printr-un simbol specific, cum este ’NA’.

Un exemplu de eroare frecventă în această privinţă ar fi o coloană cu titlul “diagnostic” sau “boli.asociate” care este completată cu un scurt text, eventual cu prescurtări, de genul “ICC, DZ.II”. O astfel de coloană va fi tratată ca un comentariu şi informaţia pe care o conţine nu va putea fi folosită în scop statistic, decât dacă, mai întâi, este recodificată sub forma unor variabile obişnuite—de exemplu: ce tip de diabet are pacientul (coloana cu titlul DZ şi cu valorile 0 reprezentând lipsa bolii, 1—diabet de tip I, 2—diabet de tip II).

Dacă datele disponibile în documentele primare—de exemplu, buletine de analiză—sunt cu diferite unităţi de măsură, cercetătorul care introduce datele trebuie să facă o conversie a acestor unităţi în unitatea stabilită pentru variabila respectivă.

În cazul că acest pas este dificil, se poate face o a doua variabilă, cu altă unitate de măsură, urmând să se realizeze conversia la prelucrarea datelor. De exemplu, glicemia *à jeun* în prima zi de internare, poate fi reprezentată în două coloane alăturate, GLIC0MGDL şi GLIC0MMDL. În cazul în care este disponibilă în mg/dl se scrie în prima coloană, iar în cea de a doua se notează NA, iar în cazul în care valoarea este disponibilă în mmol/dl, se scrie în a doua coloană, iar în prima se notează NA.

Unele variabile se pot calcula din altele

O dată ce datele introduse au fost traduse în numere, se pot calcula automat valori pentru bazate pe aceste numere, din care să rezulte noi variabile.

Un exemplu, în continuarea celui de mai sus cu glicemia, este calculul glicemiei în mg/dl, GLIC0 (care nu este completată sau măcar prezentă în tabelul iniţial) plecând de la GLIC0MGDL şi GLIC0MMDL. Calculatorul va folosi o formulă (introdusă de statistician) care va lua în seamă care dintre valori este prezentă şi va face conversia dacă este necesar.

Un alt exemplu este calculul vârstei când sunt prezente data naşterii şi o data curentă, a unei faze din protocol; sau calculul indicelui de masă corporală sau a suprafeţei corporale estimate plecând de la greutatea corporală şi înălţimea pacientului.

Aceste calcule nu trebuie făcute de către experimentator înainte de introducerea datelor deoarece consumă timp şi pot introduce erori. Este însă important să se stabilească de la început, în măsura posibilului, ce variabile se estimează a se calcula, pentru a verifica dacă toate datele necesare sunt prevăzute în tabel.

Este recomandat ca, pentru numele de variabile, se se folosească acronime scurte, care încep cu o literă şi conţin litere, cifre, şi în mod excepţional caractere de despărţire cum este '_' (liniuţa de subliniere).

Este dezirabilă evitarea numelor lungi, descriptive, mai ales conţinând spaţii, care crează tot felul de dificultăţi şi confuzii ulterior.

Variabilele reprezintă măsurabile

Rezultatul publicat al cercetării va fi o metodă predictivă prin care, pe baza unor măsurători, să se poată estima probabilitatea de apariţie a unor evenimente clinice. De exemplu, pe baza măsurătorilor hemoglobinei glicozilate se prezice incidenţa neuropatiei diabetice.

Prin măsurătoare se înţelege o determinare obiectivă a unei valori. Esenţială pentru aplicabilitatea rezultatelor este reproductibilitatea independentă a măsurătorii.

De exemplu, părerea subiectivă a clinicianului privind tipul de germen care produce o infecţie, chiar dacă—să presupunem—bazată pe o vastă experienţă şi frecventă confirmată prin investigaţii independente, nu este o măsurabilă, pentru că nu este reproductibilă

obiectiv în mod independent. Cititorul rezultatelor studiului poate avea şi el păreri personale privind tipul de germen dintr-o infecţie, dar nu ne putem aştepta ca aceste păreri să aibă aceeaşi valoare predictivă—pentru că depind de experienţa lui diferită—ceea ce face rezultatul obţinut inaplicabil.

Studiile trebuie să fie cât mai uşor de comparat cu alte studii

După publicare, rezultatele vor fi valorificate în continuare, în alte studii, de exemplu în meta-analize. Oricum, ele vor trebui confruntate cu rezultatele altor studii. Din acest motiv, este important ca, dacă există deja în literatură nişte practici în alegerea variabilelor într-un anumit tip de studiu, să se urmeze aceste practici dacă se poate.

Evitarea pierderii de informaţie

Variabilele alese reprezintă întotdeauna nişte reprezentări simplificate ale realităţii. Această simplificare înseamnă că o parte din informaţia prezentă în documentele primare se pierde.

De exemplu, în documentul primar (buletin de analiză) se găseşte valoarea glicemiei *à jeun*. Aceasta este deja o reprezentare simplificată a fenomenului mai complex al dinamicii glicemiei în cursul zilei, funcţie de alimentaţie, efort şi alţi parametri.

Experimentatorul poate face eroarea de a reprezenta această valoare sub forma unei variabile simbolice: 1: hipoglicemic, 2: normoglicemic, 3: hiperglicemic alegând mai mult sau mai puţin arbitrar nişte praguri.

Această reprezentare este o măsurabilă (presupunând că pragurile au fost bine documentate şi respectate) şi se poate prelucra statistic, dar o mare parte din informaţia prezentă în valoare exactă a glicemiei se pierde.

Este mai simplu în acest caz să se reprezinte informaţia primară (valoarea glicemiei) urmând ca, daca din diverse motive—cum ar fi comparaţia cu datele altui studiu—dorim ulterior să folosim nişte praguri să facem asta automat, printr-o variabilă calculată din valoarea primară.

Putem ilustra superioritatea reprezentării valorii primare prin faptul că ne permite comparaţia cu studii diferite care folosesc seturi diferite de praguri. Dacă alegem noi nişte praguri de la început nu mai putem face această comparaţie decât cu studii care folosesc acelaşi set.

În general, este bine să introducem în tabel datele aşa cum sunt disponibile în documentele primare, urmând ca orice reprelucrare să se facă ulterior, automat.

Independenţa variabilelor

Uneori se face eroarea de a se introduce variabile care se pot calcula unele din altele, ceea ce reprezintă o irosire de efort.

De exemplu, un tabel ar putea avea două câmpuri: externat viu (EXTV) şi externat decedat (EXTD), fiecare luând valorile 'da' sau 'nu'. Însă de fiecare dată când EXTV=da, EXTD=nu (evident) şi invers. În astfel de situaţii trebuie păstrate numai minimum de variabile, celelalte putând fi, la nevoie, calculate.