Raccolta dei dati in Excel

Ufficio Studi Clinici e Statistica

15 giugno 2022

Indice

1 Alcune indicazioni preliminari

Premettendo che non costituisce sicuramente lo strumento migliore per la raccolta dati[1], rimane il fatto che Excel sia spesso utilizzato soprattutto in ambito retrospettivo e nel contesto di studi piccoli. Pertanto, di seguito, alcune indicazioni di minima per garantire qualità lungo tutto il processo; sono rivolte ad utenti che abbiano già un minimo di padronanza con lo strumento.
Le sezioni cui prestare particolare attenzione in questo documento dipendono dalla provenienza dei dati nello studio:

- qualora i dati provengano da archivi informatici già presenti, ci si può limitare a seguire le sezioni 2 e 5 (su pianificazione della raccolta ed invio dei dati);
- qualora viceversa lo studio richieda una raccolta dati vera e propria (es revisione cartelle), si consiglia caldamente di leggere e applicare tutto quanto segue.

Qualora l'analisi statistica venga svolta dall'ufficio scrivente le procedure descritte sono mandatorie, rimanendo a disposizione per supporto/implementazione.
Senza pretesa di completezza (per la quale esistono manuali), questo documento è stato scritto con lo scopo di massimizzare il rapporto utilità/pagine da leggere; commenti, suggerimenti e dubbi, tutti ben accetti, possono essere indirizzati all'autore.

2 Pianificazione della raccolta dati

Occorre innanzitutto pianificare le informazioni da raccogliere (prima della sottoposizione dello studio al Comitato Etico) compilando un template come quello riportato in tabella 1 (e allegato: file `excel_struttura_dataset.xlsx`).
Una volta completata la predisposizione del file l'ufficio scrivente rimane a disposizione per un *controllo* dello stesso (prima della sottomissione al Comitato Etico).

2.1 Variabili da raccogliere e tipi da impiegare

Alcune indicazioni:

- utilizzando le denominazioni di Excel, i *tipi di variabili che vanno utilizzati* sono: Numero intero (es età in anni compiuti), Decimale (es altezza in metri), Data/Ora (es della chirurgia), Elenco (variabili categoriche con o senza ordinamento, es Sì/No, Lic. elementare/Lic. media/Lic. superiore/Laurea);
- ai fini dell'analisi vanno evitate *colonne testuali a libero inserimento*: queste vanno ricondotte a Elenchi (variabili categoriche) individuandone le relative modalità;

[1] Tra le criticità vanno sicuramente annoverate: possibilità che il file si corrompa, difficoltà ad effettuare check di inserimento un minimo elaborati, assenza di autenticazione degli utenti, assenza di backup (che deve essere gestito autonomamente dall'utente).

- eventuali *item a scelta multipla* vanno ricondotti a molteplici Sì/No, uno per ciascuna scelta possibile;
- non è necessario, salvo casi particolari raccogliere *variabili derivate* (es BMI, gruppi determinati da cutoff etc): è sufficiente e meglio, da un punto di vista di riproducibilità, fornire le variabili da cui derivarle (es peso e altezza, score quantitativi) e fornire in sede di analisi le regole di calcolo/determinazione delle variabili derivate (laddove non banali, es scoring questionari);
- i *colori/evidenziazione* delle celle (cosiccome i commenti) non sono importati/letti dai programmi di statistica; pertanto, invece di codificare l'informazione in questo modo in sede di raccolta dati (es gruppo1 evidenziato in verde, gruppo2 in rosso), in sede di pianificazione/raccolta dati introdurre una variabile binaria che includa tale informazione.

2.2 Struttura del dataset

Alcune indicazioni:

1. il dataset deve avere una variabile iniziale con l'*id paziente* (es un progressivo numerico), che permetta di riferirsi in maniera anonima al caso. Al fine dei check precedenti l'analisi statistica, lo sperimentatore deve essere in grado di risalire al caso una volta fornitogli l'id;
2. il *dataset* deve essere il più possibile *anonimizzato*: non servono e vanno evitati nomi, cognomi, iniziali, numeri di telefono, email, codici fiscali etc. Salvo necessità particolari, alla data di nascita è preferita l'età del paziente (in un dato momento comune di interesse, es alla diagnosi);
3. il dataset deve avere nomi di variabile facilmente importabili dalla maggior parte dei programmi di statistica, ovvero:
   - iniziare con una lettera, non con un numero;
   - niente spazi (usare "_": es "id_paziente" invece di "id paziente");
   - al massimo 32 caratteri.
4. nel caso di molteplici rilevazioni di alcune variabili entro paziente e/o nel tempo (e ciascun paziente può averne un numero differente[2]) sono opportune *più tabelle di dati* e quindi più fogli. Ciascun foglio deve presentare un id paziente nella prima colonna (es `id_pz`) per poter esser collegabile e un id specifico della misurazione entro paziente (es lesione, tempo). Si forniscono due esempi:
   (a) nel primo caso, tabella 2, ad ogni paziente sono associate più lesioni (e il numero di lesioni può essere variabile a seconda del paziente);
   (b) nel secondo caso, tabella 3, ad ogni paziente sono associati gli item dei questionari misurati lungo il follow-up.

[2] es in ragione alla gravità della situazione o della perdita al follow up

Variabile	**Descrizione** (unità misura)	**Tipo**	**Modalità**
`id_pz`	Identificativo paziente	Intero	
`sex`	Sesso	Elenco	Maschio, Femmina
`age`	Età alla diagnosi (anni compiuti)	Intero	
`peso`	Peso (kg)	Decimale	
`altezza`	Altezza (m)	Decimale	
`diabete`	Paziente diabetico	Elenco	No, Sì
`data_arr`	Data arruolamento	Data	

Tabella 1: Struttura: tabella pazienti

Variabile	**Descrizione** (unità misura)	**Tipo**	**Modalità**
`id_pz`	Identificativo paziente	Intero	
`id_les`	Identificativo lesione	Intero	
`diam_radio`	Diametro (mm) alla radiografia	Decimale	
`diam_eco`	Diametro (mm) all'ecografia	Decimale	
`diam_tac`	Diametro (mm) alla TAC	Decimale	

Tabella 2: Struttura: tabella lesioni

Variabile	**Descrizione** (unità misura)	**Tipo**	**Modalità**
`id_pz`	Identificativo paziente	Intero	
`time`	Tempo	Elenco	T0, T1, T2
`sf12_1`	SF12: item 1 (salute)	Elenco	Eccellente, Molto buona, Buona, Passabile, Scadente
`sf12_2`	SF12: item 2 (salute limita att. fisica)	Elenco	Sì parecchio, Sì parzialmente, No per nulla

Tabella 3: Struttura: tabella valutazioni

3 Creazione del file per la raccolta dati

Una volta pianificata la struttura occorre creare il file che verrà utilizzato per la raccolta dati effettiva. Fornendone in allegato un esempio (file `excel_esempio_file_dati.xlsx`), nel seguito si illustreranno le fasi per realizzarne uno analogo.

3.1 Struttura del file nel suo complesso

In questa fase creiamo il layout del file. Partendo dal file di struttura creato in sezione 2:

- aggiungere una scheda per ciascuna tabella dati (nell'esempio `pazienti`, `lesioni` e `valutazioni`) e una scheda `modalità` se si vogliono implementare elenchi (menù a scelta) per le variabili categoriche;
- copiare i nomi delle variabili nella prima riga di ciascuna tabella dati, uno per colonna[3];
- aggiungere eventuali variabili di servizio (ad esempio `note`) funzionali all'inserimento più che all'analisi dati;
- può essere utile evidenziare in grassetto e bloccare la prima riga (in maniera tale che i titoli non spariscano se si scorre in basso): posizionarsi nella cella `A2` e selezionare:

  `Visualizza > Blocca riquadri > Blocca riquadri`

3.2 Strumenti di validazione dell'inserimento

A questo punto al fine di garantire una raccolta dati il più possibilmente rapida, facilitante e priva di errore, vediamo come introdurre suggerimenti per chi inserisce dati; come effettuare check sul valore inserito; come implementare *menù a scelta* per le variabili categoriche (es genere, oppure stadio di malattia).

3.2.1 Convalida dati

La funzione di Convalida dati serve innanzitutto per fornire suggerimenti di inserimento ed effettuare controlli sullo stesso. Per accedervi:

- selezione la colonna interessata;
- selezionare "Dati" > sezione "Strumenti dati" > "Convalida Dati".

Nella finestra che si apre abbiamo tre schede:

- in "Messaggio di input" vengono impostati eventuali suggerimenti dati nell'inserimento;

[3] Ad esempio copiando la prima colonna della struttura e trasponendo nell'incolla speciale, per chi lo conosce.

- in “Impostazioni” viene scelto il tipo di inserimento accettato (tipo di dato e criteri di validità); di default è consentito qualsiasi valore immesso, che non è good practice al fine di assicurare qualità nella raccolta dati;
- in “Messaggio di errore” viene impostato il messaggio custom da visualizzare nel caso in cui un inserimento non risponda ai criteri dettati in “Impostazioni”.

Nel seguito si illustrano alcuni esempi tra quelli ritenuti più utili per il nostro contesto; si specificheranno “Messaggio di input” e “Messaggio di errore” solamente nel primo caso, sottendendo che nei rimanenti possano essere compilati analogamente con gli aggiustamenti del caso.

Altezza Per validare l’altezza in metri come un numero decimale positivo e non superiore a 3:

- in “Impostazioni” selezionare Consenti: Decimale, tra 0 e 3;
- in “Messaggio di input”: inserire ad esempio Titolo: “Altezza in metri”; Messaggio di input: “Inserire altezza in metri (es 1.8 = 1 metro e 80 centimetri)”;
- “Messaggio di errore”: inserire ad esempio Titolo: “Altezza erronea”, Messaggio di errore: “Altezza in metri deve essere compresa tra 0 e 3”.

Se tutto ha funzionato non sarà possibile inserire valori illegali (es stringhe o numerici) nella colonna così impostata.

Età alla diagnosi Per gli anni compiuti (quindi ignorando i decimali) in “Impostazioni” selezionare Consenti: Numero intero, maggiore o uguale a 0.

Data di arruolamento In “Impostazioni” selezionare Consenti: Data; se si vogliono impostare range, specificare le date di inizio e fine (es quelli previsti l’arruolamento nel protocollo).

3.2.2 Creazione di menù a scelta

Molto utili per le variabili categoriche (es maschio/femmina, sì/no, basso/medio/alto etc) si implementano sempre attraverso la “Convalida dati”.

Genere

- prima di accedere a “Convalida Dati”, nella scheda `modalità` del foglio Excel che stiamo modificando inseriamo le modalità (`Maschio` e `Femmina`, appunto) in celle separate e vicine;
- nella tabella dati selezionare la colonna del genere e accedere alla convalida selezionando “Dati” > sezione “Strumenti dati” > “Convalida Dati”;
- in “Impostazioni” selezionare Consenti: Elenco;

- cliccare su "Origine" e spostandosi nella scheda `modalità` selezionare l'intervallo di celle che contiene `Maschio` e `Femmina`. Dopodiché confermare chiudendo la finestra[4].

Se tutto ha funzionato si materializzerà un menu di scelta appena cerchiamo di inserire informazioni nella colonna del genere.

3.3 Utilizzo di formule

L'adozione delle regole descritte in 3.2 ed esemplificate nel file allegato permette già di ottenere un buon file per la raccolta dati. In questa sezione si descrivono alcune funzionalità più avanzate, talvolta utili, facenti uso delle formule.

3.3.1 Dati derivati

In alcuni casi è opportuno, per esigenze funzionali alla raccolta dati, raccogliere variabili derivate. Ad esempio sapere se il BMI è superiore a 25 perché in quel caso il protocollo prescrive la raccolta di altre informazioni. Il modo migliore per procedere è creare una formula che crei il dato derivato (vediamo in seguito come fare) e proteggerla da modifiche accidentali (in sezione 3.4).

BMI Supponendo di avere il peso in colonna `B`, l'altezza (in metri) in colonna `C` e vogliamo calcolare automaticamente il BMI per i pazienti inseriti in colonna `D`:

- ci posizioniamo in `D2` (prima osservazione di BMI) e inseriamo nella cella:

  ```
  = B2 / (C2^2)
  ```

- estendiamo a tutta la colonna la formula mediante copia/incolla oppure trascinando l'angolo in basso a destra di `D2` verso il basso (fino a dove necessario, meglio abbondare che deficere).

In questo modo appena si avranno dati in colonna `B` e `C` verrà calcolato il corretto BMI del paziente in colonna `D`.

3.3.2 Controlli incrociati

Excel, se pur con alcune limitazioni, permette di effettuare controlli incrociati tra variabili del dataset. Questo si implementa attraverso l'uso di formule: se la formula restituisce `VERO` il check è passato, alternativamente con `FALSO` verrà segnalato errore.

[4] Per casi semplici come questo in "Origine" si possono specificare le modalità intermezzate da separatore di campo (questo dipende dalle impostazioni internazionali settate, ma dovrebbe essere punto e virgola; ad esempio `Maschio; Femmina`). Si consiglia tuttavia di procedere come mostrato per uniformità (es nel caso di molteplici Sì/No, item SF12 che hanno le stesse modalità di risposta), minor possibilità di errori e maggior trasparenza/intelleggibilità.

Sesso e tumore alla prostata Supponendo che il protocollo preveda la raccolta della variabile tumore alla prostata (Sì/No) in colonna `H` e che il sesso del paziente sia raccolto in colonna `G` vogliamo che tumore alla prostata sia compilato solo per i maschi, per logica. Come evitare che le donne abbiano qualsiasi valore inserito? Per inserire questo controllo incrociato:

- posizionarsi su `H2` ed accedere alla Convalida dati avendo attiva solo questa cella selezionando "Dati" > sezione "Strumenti dati" > "Convalida Dati";
- in "Impostazioni", Consenti: Personalizzato e inserire la formula

  ```
  = SE(G2 = "Femmina"; VAL.VUOTO(H2); VERO)
  ```

  la quale restituirà `FALSO` (sollevando il messaggio di errore) solo qualora `G2` sia compilato `Femmina` e `H2` abbia *qualsiasi* valore inserito;
- "Messaggio di errore": inserire ad esempio Titolo: "Inserimento erroneo", Messaggio di errore: "Tumore alla prostata deve esser compilato solo per pazienti maschi";
- confermare, uscendo dalla finestra di Convalida dati;
- estendere il check da `H2` a tutta la colonna `H` mediante copia/incolla della cella, oppure trascinando l'angolo in basso a destra di `H2` verso il basso.

Alcune criticità evidenziabili:

- un check semplice come questo presuppone che la colonna `G` sia compilata interamente prima di procedere a `H`, cosa che è tutto sommato ragionevole per l'esempio proposto;
- di fatto se si imposta questo check in `H` non è possibile l'utilizzare un menù/elenco Sì/No sempre in `H`; questo è limite dello strumento.
  Una soluzione alternativa potrebbe essere, banalmente, impostare normalmente l'elenco Sì/No su `H` e introdurre una colonna strumentale di check in `I`; questa colonna dovrebbe contenere la formula restituente `VERO` se il check è passato o `FALSO` in caso contrario (ossia si tratterebbe di inserire in `I2` la formula di sopra ed estenderla a tutta la colonna `I`).
  Una ispezione visiva di queste colonne di check a dataset completato permetterebbe di correggere eventuali sviste, beneficiando comunque della comodità dell'elenco in `H` in fase di inserimento dati.

3.4 Protezione del file da modifiche erronee/involontarie

Una volta ottenuto il file con cui si effettuerà la rilevazione dati, può essere opportuno proteggere tutto il lavoro fatto sinora da modifiche involontarie. Di default *Excel protegge tutto*: dovremo pertanto andare a specificare cosa non proteggere, al fine di permettere l'effettivo inserimento dati.

3.4.1 Protezione dei fogli `struttura` e `modalità`

Vogliamo proteggere interamente i fogli `struttura` e `modalità` dove sono conservate la descrizione del dataset e le modalità delle variabili categoriali (per la creazione di Elenchi/menu). Per farlo spostarsi nella scheda `struttura` (rispettivamente `modalità`) e:

- cliccare su "Revisione", sezione "Proteggi", "Proteggi Foglio";
- nella finestra che si apre possiamo lasciare invariati i valori di default, inserire e confermare eventuale password (da ricordare poi) e cliccare OK.

Se tutto ha funzionato non si potrà modificare nessuna cella del foglio.

3.4.2 Protezione di nomi variabili e formule

Può essere utile proteggere la prima riga di ciascuna tabella (quella con il nome delle variabili) ed eventuali formule:

- selezionare tutte le celle che si vorranno poter modificare durante l'inserimento dati (quindi tutto ad eccezione di nomi variabili e formule);
- cliccare col tasto destro; Formato celle; Protezione; De-spuntare Bloccata; OK;
- nella scheda Struttura, cliccare su Revisione, Proteggi Foglio, inserire e confermare eventuale password (da ricordare sempre), OK.

Se tutto ha funzionato, si potranno effettuare inserimenti/modifiche solo nelle celle che non includono nomi variabili o formule.

3.4.3 Preservare la struttura del file nel suo complesso

Al fine di preservare la struttura (ossia evitare che interi fogli vengano cancellati o ne vengano aggiunti nuovi): selezionare "Revisioni" > sezione "Proteggi" > "Proteggi cartella di lavoro" > inserire e confermare eventuale password; OK.
Se tutto ha funzionato, ad esempio, non sarà possibile eliminare la tabella pazienti involontariamente.

4 Raccolta dati

Una volta completata la predisposizione del file con cui si effettuerà la raccolta dati, qualora lo studio sia:

- *monocentrico*: si può procedere alla raccolta effettiva mediante lo stesso;
- *multicentrico*: idealmente si procederà a test, utilizzando una copia del file vuoto ed alcuni casi del centro locale. Se l'esito sarà positivo si provvederà a inviare la copia bianca a ciascun centro. Una volta completata la rilevazione, sarà cura di chi effettua l'analisi ottenere il dataset complessivo a partire da quelli dei singoli centri.

Durante la raccolta dati sono importanti alcuni accorgimenti:

- per non creare problemi di importazione dati, ciascuna colonna essere compilata con dati dello stesso tipo: es dati numerici oppure caratteri/stringhe, non un mix dei due. Questo è garantito se si sono adottati criteri di convalida dati;
- in caso di eventuali dati mancanti, lasciare semplicemente la cella vuota; *non* indicare il dato mancante come "nd" o simili. Ancora, questo è garantito se si sono adottati criteri di convalida dati;
- evitare di porre sintesi "di comodità" dei dati (es media sd), tipicamente come ultima riga nel file.

5 Invio del dataset per l'analisi

In seguito alla raccolta dati e dopo un *attento check/consolidamento del dataset* assieme ai co-investigator, rinominare il file aggiungendo la data odierna (es `acronimoprogetto_data.xlsx`) ed inviarlo a chi si occuperà dell'analisi.
Nel caso sia l'ufficio scrivente:

- aprire una richiesta di collaborazione mediante l'applicativo aziedale, disponibile qui, allegando il protocollo nella sua versione finale approvata dal Comitato Etico;
- quando si verrà ricontattati per l'effettiva presa in carico sarà necessario inviare i dati via mail.

Infine, in sede di analisi:

- eventuali/desiderati *check* di coerenza da verificare sul dataset andranno comunicati/specificati; sulla base degli stessi verranno impostate opportune *query* e ed eventuali violazioni verranno comunicate, sempre anonimamente (facendo riferimento all'id paziente);
- qualora sia necessario apportare eventuali modifiche (es in seguito alle *query*), farlo prestando attenzione a non modificare la struttura del dataset (numero colonne, ordinamento delle stesse, nomi delle variabili); dopodiché aggiornare la data nel nome del file e re-inoltrarlo con le medesime modalità.