Document de référence – Bureau qualité recherche PMU DUMSC

Élaboration du codebook et de la base de données

Destinataire : CRD

1 Création du codebook

Pour chaque étude, une liste exhaustive des données récoltées et des valeurs calculées doit être établie. Ce codebook doit permettre de comprendre la signification des données récoltées, ce qui est non seulement utile pour l'investigateur mais également pour les autres personnes chargées de travailler sur ces données. La présentation sous forme de tableau est la plus pratique à utiliser. Le format informatique recommandé est Excel. Les éléments qui doivent nécessairement figurer dans ce tableau sont, pour chaque variable :

- sa dénomination (=label, chaîne de lettres et de chiffres sans espace, limité à 10 caractères, commençant toujours par une lettre, sans caractères spéciaux) ;
- sa définition en texte avec les références le cas échéant ;
- sa catégorie (sociodémographique, clinique, etc.) ;
- son but (« description du sujet à l'inclusion », « mesure d'outcome primaire », « mesure d'outcome secondaire », « mesure explicative », etc.) ;
- son type (continue, discrète, catégorielle, etc.) ;
- son format (texte, numérique, etc.) ;
- pour les variable numériques : l'unité, le domaine de validité des réponses (valeurs minimale et maximale);
- pour les variables catégorielles : les modalités les plus exhaustives possibles i.e. le codage univoque s'y rapportant avec description et mention de l'unité s'il a lieu ;
- le code établi en cas de donnée manquante.

Il convient évidemment de ne pas utiliser la même dénomination pour 2 variables différentes. Le code d'identification des sujets d'étude (ID) doit être univoque. Il convient en outre d'utiliser si possible les codes standards pour les variables courantes (cf. plus bas) et favoriser les codes numériques aux chaînes de caractères, qui seront elles-mêmes réservées aux variables pouvant présenter des modalités nombreuses, non définissables de façon exhaustive a priori. Pour limiter les réponses "texte libre" prenant beaucoup de temps à saisir et à interpréter, il est souvent possible de prévoir une variable catégorielle (d'après les réponses obtenues lors d'une pré-étude, par exemple) avec un item "autre" et de restreindre le texte libre à cette dernière catégorie.

Il est également à noter que dans certains logiciels statistiques, le nombre de caractères pour une variable est limité.

Parallèlement au codebook, il peut être utile de rédiger un protocole de saisie qui sera utile aux chargés de saisie, en précisant certaines règles décidées a priori, par exemple les réponses entre 2 cases ne sont pas prises en compte et doivent être marquées comme manquantes, etc.

Un codebook générique a été établi. Il peut être repris pour chaque étude. L'idée de cet outil est de faciliter le travail préparatoire du chercheur mais également de rendre compatibles les différentes bases de données entre elles. Il convient donc de ne modifier les variables mentionnées et leurs modalités qu'en cas d'absolue nécessité. Des variables jugées potentiellement utiles à de futures études doivent être ajoutées à ce codebook générique.

F000_codebook_PMU (à venir).

2 Choix du format et codage de la variable

Pour le traitement statistique des données, il est préférable d'avoir des variables numériques à des variables "texte".

Règles à respecter concernant les variables catégorielles (*vcat*) :

- Variables dichotomiques ou booléennes → 0/1
- Non/Oui → 0/1
- Aucun/ jamais/en aucun cas → 0
- Femme/Homme → 0/1
- Commencer le codage des variables catégorielles à "0" et non pas à "1".
- La réponse potentiellement la plus fréquente devrait avoir le code le plus petit *(vcat)*
- Autre → 6 ou 66 ou 666

Autres règles à respecter :

- Les chiffres ne doivent pas contenir d'unité (5 et non 5mg ou 5%) → l'unité sera référencée dans le codebook
- Age → Année de naissance
- Ville → code postal
- Date → yyyy-mm-dd (= format international)
- Pour chaque variable numérique, le nombre de chiffres avant et après la virgule doit être précisé.
- La standardisation du codage pour les valeurs manquantes, les données non cohérentes et les questions non applicables, est difficile car il dépend des possibilités du logiciel utilisé pour la récolte de données. Dans le cas d'une variable numérique, il est plus simple de mettre un code au format texte si le logiciel statistique utilisé permet de mélanger les formats pour une même variable. Sinon, il faut éviter que le code ne soit confondu avec une valeur effectivement prise par la variable. Cela peut néanmoins poser des problèmes lors du contrôle de la saisie. Voici quelques suggestions :
  - Valeur manquante :
    - sous Epidata (calcul direct du score) et questionnaires scannés → « » (vide)
    - Stata et autre logiciel permettant un format texte → « .m »
    - « -1 » si le logiciel utilisé permet de définir une valeur séparée en plus du domaine de validité
  - Pour les données non cohérentes par rapport à la variable :
    - Stata et autre logiciel permettant un format texte → « .inc »
    - « -2 » si le logiciel utilisé permet de définir une valeur séparée en plus du domaine de validité
  - Question ou point non applicable pour ce sujet :
    - Stata et autre logiciel permettant un format texte → « .na »
    - « -3 » si le logiciel utilisé permet de définir une valeur séparée en plus du domaine de validité

3 Création de la base de données brute

La base de données doit être créée avant que la récolte des données ne commence. Lors de ce processus, il convient d'adopter certaines règles :

- Mettre les variables en tête de colonnes, le code d'identification des sujets d'étude (ID) (ou autre unité d'analyse) en ligne et les observations dans les cases d'intersection (=format longitudinal).
- Si un sujet est vu à plusieurs reprises, créer une variable « Date de visite» et/ou « Type de visite» (V0, V1, etc.).
- Ne pas faire de sections en fonction des types de données, par exemple patients traités et contrôles, mais mettre toutes les caractéristiques en variables.
- Les lignes doivent être numérotées si elles ne le sont pas automatiquement.
- N'utiliser dans la mesure du possible qu'une feuille de données.
- Ne pas effectuer de calcul sur la base de données brute, ni cacher de colonnes, mais n'enregistrer que des valeurs. L'investigateur peut réaliser ce genre d'opération pour son intérêt personnel dans un autre fichier ou sur une autre feuille. Cependant, il transmettra au statisticien la base de données brute.
- Le choix du logiciel pour la création de la base de données et la saisie de celles-ci se fait selon les critères propres de l'étude. Un tableau comparatif des logiciels les plus courants a été établi pour faciliter le choix de l'investigateur (cf. ci-dessous).
- Appliquer les contrôles de saisie (restriction du format, nombre de caractères, etc.) en fonction des propriétés du logiciel.

3.1 Comparaison des logiciels de création et de gestion de bases de données

Cette liste contient les logiciels les plus souvent utilisés dans l'institution et n'est évidemment pas exhaustive ; d'autres logiciels sont disponibles à l'instar d'Acrobat Pro ou designer. Pour certaines études, il est à noter que la FDA (Food and Drug Administration) exige, de même que Swissmedic à l'avenir, un suivi complet des modifications réalisées dans la base de données. Ceci n'est pas réalisable avec les logiciels présentés ci-dessous. Il convient alors de contacter le CRC qui met à disposition des chercheurs un logiciel de création et gestion de bases de données (Sécutrial).

Caractéristiques	**Excel (Microsoft Office 2007)**	**Access (Microsoft Office 2007)**	**Epidata**
Accessibilité	Installé d'office	Demander l'installation à l'OIH (coût : 39 CHF)	Gratuit ; installation par l'OIH
Capacité	1 048 576 lignes par 16 384 colonnes	32 768 objets dans une base de données	En théorie, pas de limite du nombre d'observations saisies, pratiquement testé avec 250'000 enregistrements
Avantages	Utilisation instinctive Tri/transformation/calculs réalisables de façon systématique Nettoyage facile Exportation en .txt facile Format des cellules pouvant être prédéfini Validation des données possible Réalisation de tableaux croisés dynamiques possible Présentation des données sous forme de graphiques Tutoriaux disponibles en ligne	Erreurs de saisie limitées par l'utilisation de masques de saisie Pas de suppression de cellules possible Exportation en .txt facile Très utile pour mettre des données en relation et faire des requêtes de données Création de plusieurs tables de données par étude (différents questionnaires, points de mesure et de suivis dans le temps, etc) avec possibilité de faire des bases de données relationnelles. Tutoriaux disponibles en ligne	Fonction de contrôle par double saisie disponible Saisie facile, peu de risque d'erreur Création du masque de saisie très simple Possibilité d'ajouter les contrôles de base sans notion de programmation Exportation très facile des données dans divers applications ou en format texte (la signification du codage des données catégorielles est conservée lors de l'exportation dans stata, ex.: 0=non; 1=oui) Avantage par rapport à Excel: base de données possible. Collaboratrice de recherche bien formée et disponible pour réaliser les masques de saisie et les fichiers de contrôle → Mme S. Payot Sylvie.Payot@hospvd.ch) Quelques masques de saisie déjà disponibles (données socio-démographiques générales, certains scores et questionnaires courants ; cf. Mme S. Payot Sylvie.Payot@hospvd.ch) Tutoriaux disponibles en ligne

Caractéristiques	Excel (Microsoft Office 2007)	Access (Microsoft Office 2007)	Epidata
Inconvénients	Risque important de mouvement de cellules non volontaire entraînant un décalage des données Erreurs de saisie probables A ne réserver qu'aux études de faible puissance	Utilisation complexe Prise en main trop laborieuse pour une étude de faible puissance	Un seul utilisateur simultané. Aspect graphique des formulaires très pauvre (police, taille et style de caractère uniformes, difficultés à aligner les champs). Caractère pour valeur manquante dépendante du format de la variable (date, chiffre ou texte)