77e congrès de la Société Nationale Française de Médecine Interne
Lyon 27 – 28 – 29 juin 2018

COMMENT REALISER SA BASE DE DONNEES?

Guide Méthodologique

Chaux Robin, Dr Oriol Mathieu, Tinquaut Fabien, Pr Trombert-Paviot Béatrice

SSPIM
Service de **S**anté **P**ublique et d'**I**nformation **M**édicale
Avenue Albert Raimond, 42270 Saint-Priest-en-Jarez
CHU Saint-Etienne

PLAN

1 – INTRODUCTION

2 – QUEL LOGICIEL UTILISER

3 – COMMENT CONSTRUIRE SON QUESTIONNAIRE

4 – COMMENT SAISIR LES DONNEES
- Les 12 Grandes Règles
- Les astuces

5 – QUELQUES SPECIFICITES

6 – RESUME

7 - SOURCES

1 - INTRODUCTION

- Thèse médecine
- Mémoire Master/DES
- Recherche

→ **Création préalable d'une Base de Donnée (BDD)**

Pourquoi a-t-on besoin de connaissances méthodologiques sur les BDD?

- Le fonctionnement particulier des logiciels de statistique conditionne entièrement la manière de récupérer les données et de les coder.

- Besoin d'un **FORMATAGE CORRECT** des données pour l'analyse des résultats.

etc...

BUTS:
- Surtout : **<u>Perte considérable de temps</u>**: la remise en forme, le nettoyage de BDD sont chronophages +++
- C'est à l'interne de nettoyer sa BDD, laborieux +++ si les règles ne sont pas respectées dès le début!
- *Ex: erreur de codage vue tardivement pour 1500 patients -> à récupérer manuellement, case par case…*
- Eviter les **erreurs dans la saisie**
- Eviter les **erreurs dans les résultats**
- *(Pour ne pas agacer son statisticien).*

2 - QUEL LOGICIEL UTILISER – LE RECUEIL des données

- **Sous forme de TABLEUR:** c'est la base.
- Sous Windows: **Microsoft Excel®**, **Open Office ™**

Avantages: - Simples d'accès, apprentissage rapide
- Navigation, corrections aisées

Inconvénients: - Pas de contrôle de la saisie de données
- Nombre de variables (colonnes) limité

- Autres tableurs:

- Numbers® sur Mac

- Open Office™ sur Linux

- Pour aller plus loin, vrai logiciel de gestion de bases de données:
- Access®. Eventuellement, moins utile pour thèses

Avantages: - Pas de limite au nombre de variables
- Contrôle de la saisie de données

Inconvénients: - Plus difficile à prendre en main
- Payant, moins facile d'accès.

- Dans tous les cas, il faudra **exporter vos résultats dans un format de type tableur** pour l'analyse (.xlsx, .xls, .odt).

2 - QUEL LOGICIEL UTILISER – RECUEIL EN LIGNE

Questionnaires

- **Google Forms**
  - Créer un questionnaire en ligne et le diffuser
  - La base de données se créé automatiquement
  - Utilisation simplissime

- Google doc
  - Gratuit, export sous Excel possible, pas de limite au nombre de questionnaires
  - Types de réponses limités, possibilité de répondre plusieurs fois

- Lime Survey
  - Gratuit, 20 types de questions, nombre de questionnaires illimités, export sous Excel
  - Nécessite de disposer d'un hébergeur

- Survey Monkey
  - 15 types de questions, outils statistiques intégrés, gestion des relances
  - Payant si > 10 questions et/ou > 100 réponses

3 – COMMENT CONSTRUIRE SON QUESTIONNAIRE?

- Réaliser sur un fichier Excel® un **DICTIONNAIRE DES VARIABLES.**

  - Contient l'ensemble des variables que vous souhaitez recueillir

  - Leur(s) modalité(s) de réponse

  - Leur type (quantitative, qualitative binaire/catégorielle)

  - Leur unité de mesure (Unités internationales +++)

3 – COMMENT CONSTRUIRE SON QUESTIONNAIRE?

DICTIONNAIRE des variables & Différents types de variables

	A	B	C	D	E
1	**Variable**	**Label**	**Type de variable**	**Modalités**	
2				**Codage**	**Valeur**
3	**id**	Numéro d'inclusion patient	Nombre		
4	**nom**	Trois 1ères lettres nom, deux 1ères lettres prénom	NNN-PP		
5	**cp**	Code postal	5 chiffres		
6	**sexe**	Sexe	Qualitative binaire	1	Homme
7				2	Femme
8	**ttt**	Efficacité du traitement	Qualitative binaire	0	Echec
9				1	Réussite
10	**ps**	Performans Status	Qualitative ordinale	1	PS= 0 -1
11				2	PS= 2 -3
12				3	PS= 4
13	**travail**	Catégories socioprofessionnelles	Qualitative catégorielle	1	Retraités
14				2	Employés
15				3	Cadres
16				4	Sans emploi
17	**age**	Age	Quantitative continue		
18	**taille**	Taille du patient	Quantitative continue		
19	**nbr_enfant**	Nombre d'enfants	Quantitative discrète		
20	**date_dc**	Date du décès	DD/MM/YY		
21	**heure_dc**	Heure du décès	HH:MM		

Ex: Identification

Variables QUALITATIVES

Variables QUANTITATIVES

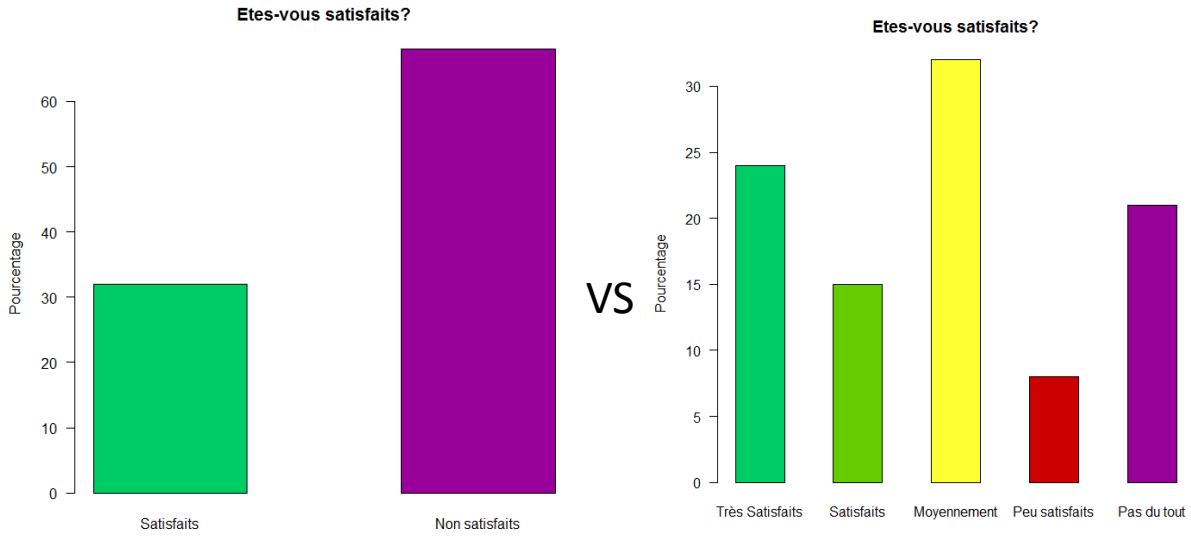
Variables particulières

3 – COMMENT CONSTRUIRE SON QUESTIONNAIRE?

- **Privilégier les questions fermées**. Eviter le texte libre dans la mesure du possible.
- *Le traitement des textes se fait manuellement, à utiliser avec parcimonie.*

- **Ne pas utiliser de réponse intermédiaire** « Moyennement satisfait », que l'on ne pourra rattacher ni à « Satisfait », ni à « Insatisfait ».

Etes-vous satisfaits de votre prise en charge?	
-Très satisfait	Satisfaits
-Satisfait	
-Peu satisfait	Insatisfaits
-Pas du tout satisfait	



- **Ne pas oublier la question portant sur votre Critère de Jugement Principal** et assurez-vous que le résultat que vous collectez soit exploitable pour répondre à votre Objectif Principal.

4 – COMMENT SAISIR LES DONNEES

LES 12 REGLES:

N°1: La plus importante, ELABORATION du PLAN D'ANALYSE

- **La règle d'or**
- **Avant de commencer à récolter des données**
- Prendre contact avec un spécialiste/méthodologiste
- Evite les erreurs méthodologiques: **IRRATTRAPABLES** au moment de l'analyse, et très fréquentes.
- Prévoir les informations indispensables à récupérer.
- Facilite le processus: protocole, questionnaire, analyses prévues
- Au final, gain de temps +++

PLAN D'ANALYSE:

A l'avance! Pas une fois la BDD terminée!

A quelles questions voulez-vous répondre?

- Critère de jugement principal
- Critère (s) de jugement secondaire(s)

Formaliser le plan d'analyses comparatives:

- Quelles variables voulez-vous comparer?
- Cibler les demandes sur les analyses utiles pour l'interprétation, la compréhension, la comparaison (avec les données de la littérature, les données cliniques consensuelles par exemple).
- Savoir à l'avance ce que l'on veut essayer de démontrer, ce que l'on veut faire de ses variables et de ses comparaisons.

4 – COMMENT SAISIR LES DONNEES

LES 12 REGLES:

N°2: Il est obligatoire de déclarer votre base de données à la CNIL

- Avant de débuter le recueil de données.
- Consulter le site:
- Démarches préliminaires, conseils légaux selon le type de BDD (changent souvent; RGPD)
- Déclaration Normale ou inscription au registre ou autre.
- Conseils sur l'information des patients et l'utilisation de leurs données.
- Conseils sur l'anonymisation.
- La réponse peut prendre du temps (plusieurs mois), prévoir à l'avance.
- **Attention= N° d'agrément CNIL obligatoire pour publication!**

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

https://www.cnil.fr/fr

4 – COMMENT SAISIR LES DONNEES

LES 12 REGLES:

N°3: Une ligne par patient. Une colonne par variable.

- Repérer l'unité élémentaire de l'étude. Le plus souvent le patient.

N°4: La 1ère ligne contient le nom des variables

- Donner des noms simples aux variables, préférer un seul mot ou code
- Noms explicites
- Noms uniques

	A	B	C
1	Abcès (cm)	Délai entre chirurgie première & fistule	Catégories autre (1 : Ennui, 2 : Divertisseme
2	2	12	2
3	5	6	4

NON

	H	I	J
	abces	delai_chir	cat_autre
	2	12	2
	5	6	4

OUI

Une colonne par variable

1ère ligne= Nom des colonnes

Une ligne par patient

	A	B	C	D	E	F
1	ID	sexe	age	prof	maladie	grp_ttt
2	1	1	52	2	1	1
3	2	1	52	3	0	1
4	3	2	28	4	0	1
5	4	2	47	1	1	0
6	5	1	65	1	0	0
7	6	2	58	1	1	0
8	7	1	63	2	0	1
9	8	2	33	2	1	1
10	9	1	69	4	0	1
11	10	2	48	5	0	1
12	11	1	51	2	1	0
13	12	1	28	2	1	0
14	13	1	30	1	1	0
15	14	1	88	1	1	1
16	15	2	57	2	0	0
17	16	2	63	2	0	1
18	17	2	57	4	0	0
19	18	2	56	4	1	1
20	19	1	49	5	1	0
21	20	1	55	6	1	1
22	21	1	18	4	1	0
23	22	1	48	4	0	1
24	23	2	44	1	0	0
25	24	2	88	1	0	0

4 – COMMENT SAISIR LES DONNEES

Numéro de patient en 1ère colonne

A	B	C	D
id	**nom**	**prenom**	**ddn**
1	Rodriguez	Bender	01/02/1954
2	Dimagio	John	15/12/1994
3	Vercherion	Paul	28/10/1987
4	Tinquetau	Fabrice	09/09/1942
5	Baden	Baden	24/07/1963
6	Lyu	Lucie	06/06/1914

INTERDIT !

N°5: La 1ère colonne contient un numéro patient unique.

- Chaque patient doit avoir un numéro identifiant.

- [Avec les correspondances sur un fichier à part par exemple, pour pouvoir retourner au dossier patient en cas de besoin].

- L' ANONYMISATION est OBLIGATOIRE.

- Attention aux données identifiantes.

- Faire appel à la CNIL pour connaître les modalités d'anonymisation selon son type d'étude +++.

4 – COMMENT SAISIR LES DONNEES

N°6: 1 variable = 1 colonne = 1 INFORMATION

- Par exemple, stade TNM = 3 informations = 3 variables = 3 colonnes (T | N |M)
- Le décès = 2 informations = 2 variables = 2 colonnes (une colonne décès (Oui/Non); une colonne date de décès).
- Ne pas mélanger les types dans une même colonne.

L
Temps de suivi
décès à 13,5 ans
en vie à 16,8 ans
décès à 8
décès à 10,3 ans
PDV date inconnue
en vie à 2,9 ans

2 informations dans la même colonne = INUTILISABLE

N	O
tps_suivi	etat
13,5	1
16,8	0
8	1
10,3	1
NA	0
2,9	0

2 informations = 2 colonnes
(Codage: décès: 1 En vie/PDV: 0)

15	Catégories socioprofessionnelles	csp
16	Agriculteurs exploitants	1
17	Artisants, commerçants, chefs d'entreprises	2
18	Cadres et professions intellectuelles supérieures	3
19	Professions intermédiaires	4
20	Employés	5
21	Ouvriers	6
22	Retraités	7
23	Autre sans activité	8

Préférer un codage de ce type *(ici façon INSEE)*

N°7: Codage des variables:

- **Privilégier les chiffres.**
- Oui/Présence= 1 Non/Absence= 0
- Homme= 1 Femme= 2
- Saint-Etienne= 1 Lyon= 2 Paris= 3
- Allemagne= 7 [...] Brésil= 1
- **Dates et heures**, toujours de la même façon
- ex: JJ/MM/AAAA et HH:MM

4 – COMMENT SAISIR LES DONNEES

LES 12 REGLES:

N°8: Ecriture des données et des noms de variables:

- Supprimer les accents, la ponctuation
- Pas de caractères spéciaux (&, %, >, <)
- Pas d'espaces: utiliser l'underscore « _ »

- **Eviter le texte libre,** sinon standardiser la saisie:
tout en minuscules, «_»

- Attention au séparateur décimal, soit virgule, soit point.

- **PAS DE COULEURS!** Les logiciels ne les reconnaissent pas

- En bref: être constant dans le remplissage des valeurs.

ATTENTION
Caractères spéciaux
Non standardisé
Accents
Espaces
Couleur
Majuscule/minuscule

Q	R
Abcès (cm)	Cicatrisation complète
<5	Oui
< 5	oui
> 10	non
entre 5 et 10	Non
entre 5et 10	OUI
<5	OUI

OUI
Standardisé
Codage taille abcès
Aucun caractère spécial
tout en minuscules

T	U
abces	cic_compl
1	oui
1	oui
3	non
2	non
2	oui
1	oui

4 – COMMENT SAISIR LES DONNEES

LES 12 REGLES:

X
Délais chirugie
22 mois et 3 jours
5 semaines
11 jours
7 jours
20 mois

NON!
Pas les mêmes unités de temps
Inutilisable
(et ne pas préciser l'unité sur le tableur)

OUI
Choisir une unité
exemple: en jours,
Toujours la même

Z
delai_chir
674
35
11
7
610

N°9: Valeurs manquantes.

- Notez « NA » (Not Assigned)
- Afin de les différencier d'une donnée non recueillie (laisser la case vide)

N°10: Utiliser toujours la même unité (cf dictionnaire des variables)

- Et ne pas l'écrire dans la variable (ex: pas de « 125 µmol/l » -> écrire seulement « 125 », et toute la colonne doit être en µmol/l).

4 – COMMENT SAISIR LES DONNEES

LES 12 REGLES:

N°11: in fine, donner une BDD DEFINITIVE

- Plus d'ajout de patient.
- Plus d'ajout de variable.
- Pas de changement de Critères de Jugement.
- Vérifier et revérifier l'intégrité des données, les erreurs de saisies avant de définir votre base finale parfaite.
- **Plus question de la modifier ensuite.**

N°12: Valorisation du travail du statisticien

- Pour tout travail d'analyse substantiel, le nom du statisticien, de l'interne ou du praticien ayant réalisé une aide méthodologique et/ou des analyses statistiques doit être mentionné dans les documents finaux.

4 – COMMENT SAISIR LES DONNEES

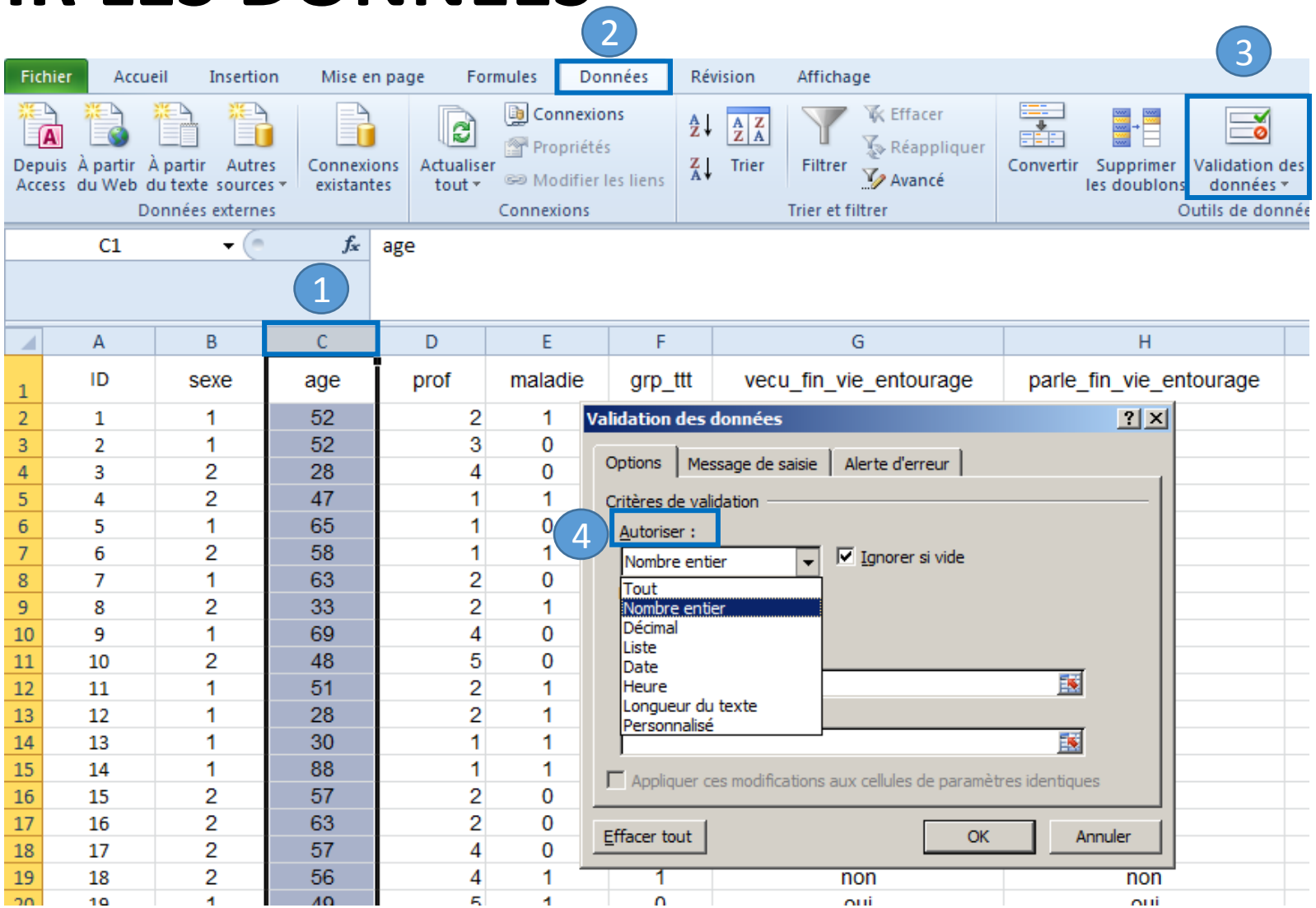
LES ASTUCES:

N°1: UNIFORMISER les réponses et éviter les erreurs de saisies et d'orthographe.

- Sélectionner une colonne -> « Données » -> « Validation des données ». Dans le menu « Autoriser », choisir le type de données correspondant à votre variable.

- On ne peut saisir ainsi que le type de données choisi.

Car attention, une différence de majuscule ou un espace oublié change la valeur de la donnée pour le logiciel!



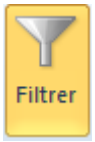
4 – COMMENT SAISIR LES DONNEES

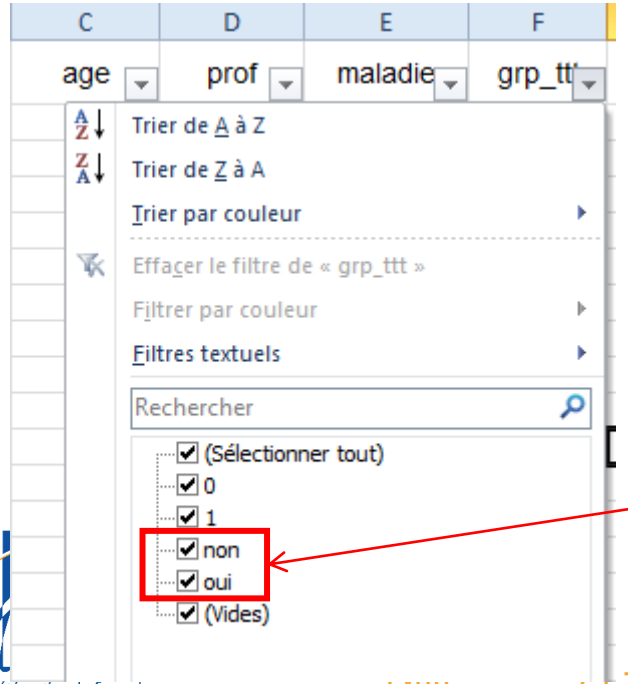
LES ASTUCES:

- Copier/Coller les réponses.

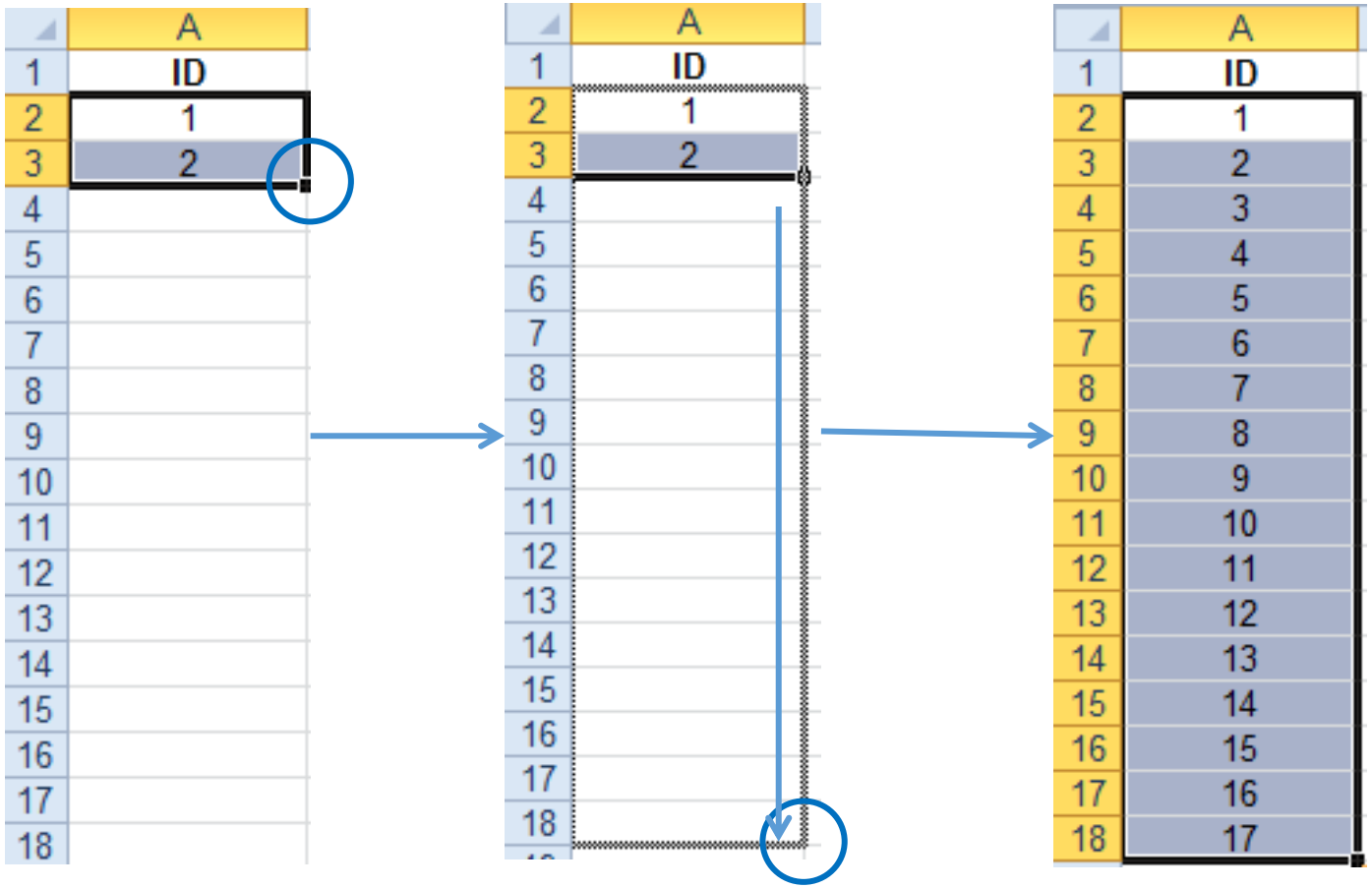
- Recopie incrémentée.

N°2: La fonction filtre sur Excel® permet de visualiser et de sélectionner une donnée aberrante.





Des erreurs se sont glissées dans notre variable



- La recopie incrémentée -

4 – COMMENT SAISIR LES DONNEES

LES ASTUCES:

N°3: Ne pas oublier une variable « groupe » si besoin.

- Exemple, Groupe placebo= 0, Groupe traitement actif= 1 dans la même colonne
- A nouveau, les couleurs ne sont pas détectées.

NON

AC
Date De Décès
11/05/2015
04/07/2015
27/08/2015
20/10/2015
13/12/2015

OUI

AE	AF
dte_deces	grp_ttt
11/05/2015	0
04/07/2015	1
27/08/2015	1
20/10/2015	0
13/12/2015	0

N°4: Le tableau de données doit se trouver sur une seule page (dans un format rectangulaire)

4 – COMMENT SAISIR LES DONNEES

LES ASTUCES:

N°5: Simplifier au maximum les variables pour avoir des analyses pertinentes.

- En dehors d'un intérêt purement descriptif

Modalités de réponses:
1: Phrase Positive,
2 : Phrase Négative SANS exemple,
3 : Phrase Négative AVEC exemple (conséquence Santé),
4 : Phrase Négative AVEC exemple (conséquence Socio/Scolaire),
5 : Phrase Négative AVEC exemple (TOUT),
6 : Phrase avec notion de nuance/limite SANS exemple,
7 : Phrase avec notion de nuance/limite AVEC exemple (conséquence Santé),
8 : Phrase avec notion de nuance/limite AVEC exemple (conséquence Socio/Scolaire),
9 : Phrase avec notion de nuance/limité AVEC exemple (TOUT),

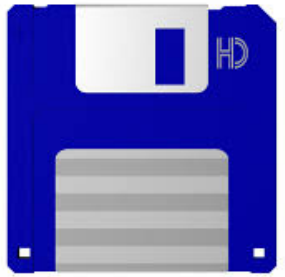
- Difficilement interprétable.
- Peu intuitif.

Modalités de réponses:
Positives
Négatives
Nuancé

- Modalités regroupées/ simplifiées
- Compréhension aisée

N°6: Important, sauvegarder les données.

- Créer régulièrement un nouveau fichier et nommez-le avec la date du jour.
- Stockez vos données dans le Cloud (Google Drive, Dropbox) et/ou sur un disque dur externe.
- (*Personne ne veut faire partie des 2 internes sur 100 qui perdent leur BDD).*



4 – COMMENT SAISIR LES DONNEES

LES ASTUCES:

N°7: Ne pas sauter de lignes. Maintenir la BDD « compacte »

5 – QUELQUES SPECIFICITES

> Variables qualitatives à plusieurs modalités de réponses, dont une modalité « autre » avec réponse ouverte. Comment faire?

Exemple: Quel est votre fruit préféré?

1 - Créer une variable avec les modalités de réponse prévues:

- 1= banane
- 2= pomme
- 3= poire
- 4= autre, précisez:

2 - Si « autre »: Créer une deuxième variable avec du texte libre: POUR AMENER LA PRECISION

1	2
fruit_pref	autre_fruit_pref
1	
1	
2	
4	kiwano
3	
4	akebia

5 – QUELQUES SPECIFICITES

> Question à réponses multiples (plus d'une seule réponse possible)

- Exemple: plusieurs complications sont possibles en même temps**: <u>créer une variable par complication</u>**

- « cpl_abces »; Oui= 1, Non= 0, Donnée manquante= NA
- « cpl_adenite »; Oui= 1, Non= 0, Donnée manquante= NA
- « cpl_fistule », « cpl_allergie »: idem
- « cpl_autre » *(pour celles qui ne tombent pas dans nos cases prévues)*
  - Si oui: nouvelle variable avec texte libre cpl_autre_precis: « texte libre »

AH	AI
ID	**complications**
1	abces, adenite
2	fistule
3	allergie, hemorragie
4	encore une suspicion de Lyme

Deux ou plus cpl° par colonne
Non analysable par les logiciels

AJ	AK	AL	AM	AN	AO	AP
ID	**cpl_abces**	**cpl_adenite**	**cpl_fistule**	**cpl_allergie**	**cpl_autre**	**cpl_autre_precis**
1	1	1	0	0	0	
2	0	0	1	0	0	
3	0	1	0	1	0	
4	0	0	0	0	1	encore une suspicion de Lyme

Parfait!

5 – QUELQUES SPECIFICITES

> Chaque ligne correspond à un sujet.

> Mais parfois plusieurs lignes pour un seul sujet, données répétées, plusieurs consultations etc.

- Une colonne = numéro d'observation
- Une colonne = numéro de patient.

AR	AS
id_visite	**id_patient**
1	1
2	1
3	1
4	2
5	2
6	3
7	4
8	5
9	6
10	6

6 - RESUME

> RESUME DES REGLES D'OR DE LA CREATION DE BASE DE DONNEES

N°1: ELABORATION d'un PLAN D'ANALYSE.

N°2: Déclarer obligatoirement la BDD à la CNIL.

N°3: Une ligne par patient. Une colonne par variable.

N°4: La 1ère ligne contient le nom des variables.

N°5: La 1ère colonne contient un numéro identifiant unique.

N°6: 1 variable = 1 colonne = 1 INFORMATION.

N°7: Coder les variables.

N°8: Règles d'écriture des données, des noms de variables.

N°9: Traitement des valeurs manquantes.

N°10: Utiliser toujours la même unité pour une même VA.

N°11: Donner une version des données DEFINITIVE.

N°12: Valorisation du travail statistique.

6 - RESUME

> RESUME DES TIPS & TRICKS

N°1: UNIFORMISER les réponses +++.

N°2: Contrôle des valeurs aberrantes, fonction filtre.

N°3: Variable « groupe » si besoin.

N°4: Tableau de donnée sur une seule page.

N°5: Simplifier les variables au maximum.

N°6: Sauvegarder ses données.

N°7: Ne pas sauter de ligne.

7 - SOURCES

> « Comment créer sa Base de Données exploitable? »
Dr. Mathieu ORIOL, Médecin de Santé Publique, Fabien TINQUAUT, Statisticien
Centre Hygée, ICLN, Saint-Etienne, www.centrehygee.fr

> « Présentation des données pour une analyse statistique »
Dr. Caroline TOURNOUX-Facon et Alexandre ROLLET
http://medphar.univ-poitiers.fr/santepub/images/staff_2009/091028_base_donnees.pdf

> « Outils méthodologiques et astuces pour la thèse de médecine, Les statistiques, comment faire ? »
Cyril Ferdynus, USM, CHU La Réunion
http://docplayer.fr/10102284-Outils-methodologiques-et-astuces-pour-la-these-de-medecine-les-statistiques-comment-faire.html

> « Aide pour la création d'une base de données »
Elsa Parot-Schinkel, interne de santé publique
CHU d'Angers, Département Universitaire de Santé Publique
http://dusp.angers.free.fr/fichiers/aide_creation_bdd_plana.pdf

> « Elaboration d'un guide pratique à lire avant la mise en place d'une étude »
Auteurs non retrouvés. Université de Poitiers.
http://medphar.univ-poitiers.fr/santepub/images/staff_2009/091028_EXCEL.pdf

Merci de votre attention

robinvchaux@gmail.com
Interne de Santé Publique