Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Introduction à la création et la gestion de bases de données

Stéphane Béchet

Biostatisticien / Data Manager

Institut Pasteur

Unité d'Epidémiologie des Maladies Emergentes

stephane.bechet@pasteur.fr

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Objectifs de la formation

- Comprendre la nécessité de l'utilisation d'un Système de Gestion de Bases de Données (SGBD) par rapport à de simples fichiers Excel

- Connaître les points essentiels dans la création d'une base de données

- Assimiler les étapes principales de gestion d'une base de données

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Objectifs de la formation

- Comprendre la nécessité de l'utilisation d'un Système de Gestion de Bases de Données (SGBD) par rapport à de simples fichiers Excel

- Connaître les points essentiels dans la création d'une base de données

- Assimiler les étapes principales de gestion d'une base de données

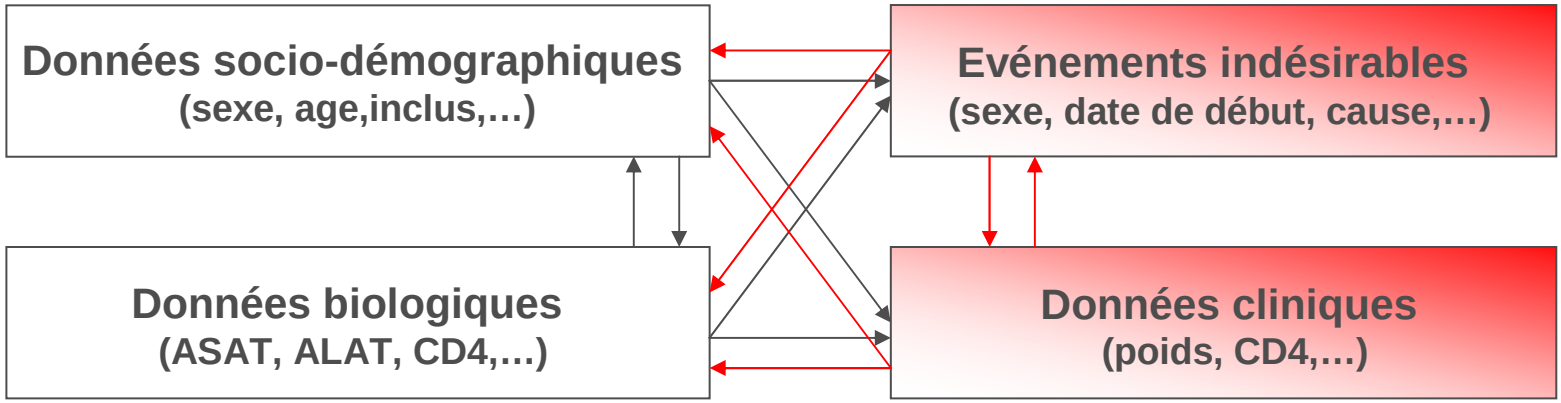
Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Définitions

- **Base de données (BD)** : Ensemble structuré de données informatisées, accessible de façon sélective par plusieurs utilisateurs.

- **Système de gestion de base de données (SGBD)**: Logiciel permettant d'interagir avec une base de données. C'est l'interface entre l'utilisateur et la base de données.

Limites à l'utilisation des fichiers type Excel



De telles applications sont :

- rigides
- contraignantes
- longues à mettre en oeuvre

Les données associées sont :

- redondantes
- peu fiables
- peu accessible de manière ponctuelle

Un grand effort de programmation est alors nécessaire pour exploiter et maintenir dans un état cohérent ces informations et leurs liens.

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Limites à l'utilisation des fichiers type Excel

- Doublons
- Contrôles des données (valeurs aberrantes, cohérence entre variables…)
- Format des données

	A	B	C	D	E	F	G	H	I
1	**NUM**	**INIT**	**DDN**	**SEXE**	**DVISIT**	**POIDS**	**TAILLE**	**PLVT**	**RSLT**
2	77001	mc	24/08/1998	Féminin	10/03/2003	55,4	160	Non	NA
3	77002	ca	13/11/1993	Féminin	17/03/2003	60,3	165	Non	NA
4	77003	pc	21/02/02	Féminin	24/03/2003	40.6	155	Non	NA
5	77004	br	11/12/1995	Masculin	31/03/2003	37,6	140	Oui	Pos
6	77005	mg	12/12/1991	Masculin	02/04/2003	35,4	141	Non	NA
7	77006	fp	09/10/1990	Féminin	08/04/2003	29,0	132	Non	Pos
8	77007	lt	4/19/2001	Masculin	10/04/2003	40,2	1455	Oui	NA
9	77008	na	29/10/1994	Masculin	25/04/2003	54.6	151	Non	NA
10	77009	se	27/08/1991	Féminin	26/05/2003	63,0	159	Oui	Neg
11	77010	vm	30/12/1999	Féminin	26/05/2003	60,1	17	Non	NA
12	77011	cn	09/09/1999	Masculin	17/06/2003	52,0	155	Oui	Pos
13	77012	dm	18/12/1993	Masculin	16/08/2003	38,8	141	Non	Neg
14	77003	ba	8/22/2000	Féminin	04/09/2003	39.2	143	Non	NA

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

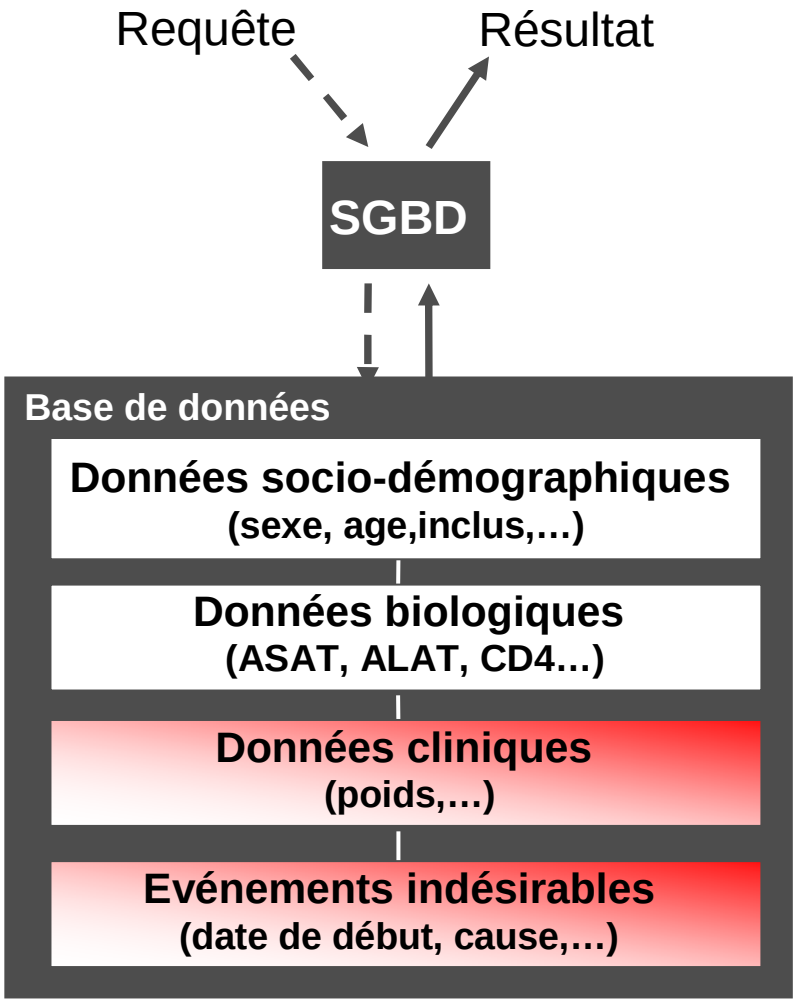
Limites à l'utilisation des fichiers type Excel

- Tri sur les données

	A	B	C	D	E
1	NUM	INIT	DDN	SEXE	POIDS
2	77006	fp	09/10/1990	Féminin	29,0
3	77005	mg	12/12/1991	Masculin	35,4
4	77004	br	11/12/1995	Masculin	37,6
5	77012	dm	18/12/1993	Masculin	38,8
6	77007	lt	4/19/2001	Masculin	40,2
7	77011	cn	09/09/1999	Masculin	52,0
8	77001	mc	24/08/1998	Féminin	55,4
9	77010	vm	30/12/1999	Féminin	60,1
10	77002	ca	13/11/1993	Féminin	60,3
11	77009	se	27/08/1991	Féminin	63,0
12	77003	ba	8/22/2000	Féminin	39.2
13	77003	pc	21/02/02	Féminin	40.6
14	77008	na	29/10/1994	Masculin	54.6

	A	B	C	D	E
1	NUM	INIT	DDN	SEXE	POIDS
2	77006	fp	09/10/1990	Féminin	29,0
3	77005	mg	27/08/1991	Masculin	35,4
4	77004	br	12/12/1991	Masculin	37,6
5	77012	dm	13/11/1993	Masculin	38,8
6	77007	lt	18/12/1993	Masculin	40,2
7	77011	cn	29/10/1994	Masculin	52,0
8	77001	mc	11/12/1995	Féminin	55,4
9	77010	vm	24/08/1998	Féminin	60,1
10	77002	ca	09/09/1999	Féminin	60,3
11	77009	se	30/12/1999	Féminin	63,0
12	77003	ba	8/22/2000	Féminin	39.2
13	77003	pc	4/19/2001	Féminin	40.6
14	77008	na	21/02/02	Masculin	54.6

Avantages de l'utilisation d'un SGBD



Utilisation d'un SGBD

- Définition des données (objets, attributs et liens),
- Centralisation des données,
- Manipulation simple des données (ajout, modification, tri, …),
- Obtention de renseignements spécifiques à partir des données,
- Contraintes d'intégrités,
- Accès multiples aux données,
- Accès sécurisé aux données

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Objectifs de la formation

- Comprendre la nécessité de l'utilisation d'un Système de Gestion de Bases de Données (SGBD) par rapport à de simples fichiers Excel

- Connaître les points essentiels dans la création d'une base de données

- Assimiler les étapes principales de gestion d'une base de données

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Développement d'une application

- La partie la plus importante du développement d'une application se fait avant même d'allumer un ordinateur !!!

- Travail à réaliser en collaboration avec les membres de l'équipe (Investigateurs, Attachés de Recherche Clinique, Médecins d'Etude Clinique, Biostatisticiens, …)

- Etapes dans le développement d'une application :
  - spécification des besoins (principaux intervenants et principales activités)
  - l'analyse
  - la conception (élaboration de la structure de la base)
  - l'implémentation (écriture des programmes informatiques)

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Type d'utilisateurs d'une base de données

- L'administrateur
  - maintenance de la base de données,
  - Sauvegardes et restauration du système en cas de pannes,
  - définition des droits d'accès pour chaque utilisateur
- Le programmeur
  - construction de la base et des interfaces utilisateurs,
  - mise en place des droits d'accès
- L'utilisateurs
  - Selon les droits : ajout, modification, suppression, consultation des données
  - L'utilisateur ne peut en aucun cas modifier la structure de la base

Élaboration d'une base de données (1/2)

1) Annotation du questionnaire

2) Dictionnaire des données
   - nom des variables
   - libellés
   - unité des mesures
   - page du CRF où se trouve la variable

3) Structure de la base
   - tables
   - variables (nom, type de données, format, clé primaire)
   - schéma relationnel

4) Cahier des charges (incohérences)
   - numéro du contrôle
   - nom des variables impliquées
   - description

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Élaboration d'une base de données (2/2)

5) Masques de saisie
   - Mise en page
   - Liste déroulante, cases à cocher
   - Boutons de commandes…

6) Contrôle de cohérences
   - bornes
   - saut conditionnel…

7) Rapports automatiques

8) Test de la base

9) Mode d'emploi de la base de données

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Tracking / Audit Trail

Suivi des modifications apportées à une base de données :

- ✓ Qui
- ✓ Quand
- ✓ Quoi
- ✓ Pourquoi
- ✓ Ancienne valeur
- ✓ Nouvelle valeur

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Sécurité

- ✓ Un contrôle d'accès à votre réseau / ordinateur : firewall, antivirus, login, mot de passe, maintenance régulière (matériels, logiciels…)
- ✓ un contrôle d'accès à la base de données : login, mot de passe
- ✓ un contrôle d'accès aux données par des droits différents selon les utilisateurs :
  - Consultation
  - Ajout
  - Modification
  - Suppression
  - Modification de la structure de la base
  - …

Mot de passe

FAIRE

- Longueur d'au moins sept caractères
- Lettres majuscules et minuscules
- Chiffres
- Symbole
- Au moins quatre caractères différents (sans répétitions)
- Ressemble à une séquence aléatoire de lettres et de nombres

NE PAS FAIRE

- N'utilisez AUCUNE PARTIE de votre nom de connexion .
- N'utilisez ni votre nom, ni votre prénom
- N'utilisez aucune lettre ou nombre consécutif tel que "abcdefg" ou "23456"
- N'utilisez aucune touche adjacente sur votre clavier telle que "azerty"

Ne communiquez pas votre mot de passe, ne l'écrivez pas

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Sauvegarde

La **sauvegarde** est un paramètre fondamental pour toute étude et reste l'unique garantie de ne pas perdre votre travail

- ✓ sauvegardes fréquentes
- ✓ différents supports (Serveur, bandes magnétiques, DVD…)
- ✓ gestion des anciennes versions

Synthèse

- L'élaboration de la base de données exige :
  - du temps
  - de la rigueur
  - Une collaboration avec toute l'équipe projet
- La base de données doit être :
  - évolutive
  - conviviale
  - cohérente
- La sécurité des données des sujets doit être garantie :
  - droits différent selon les utilisateurs
  - mots de passe efficaces
- Sauvegarde régulière

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Objectifs de la formation

- Comprendre la nécessité de l'utilisation d'un Système de Gestion de Bases de Données (SGBD) par rapport à de simples fichiers Excel
- Connaître les points essentiels dans la création d'une base de données
- Assimiler les étapes principales de gestion d'une base de données

Définition

Data Management : Processus de gestion de données permettant d'obtenir une base de données cohérente et exploitable par les statisticiens afin de répondre aux objectifs de l'étude.

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Plan

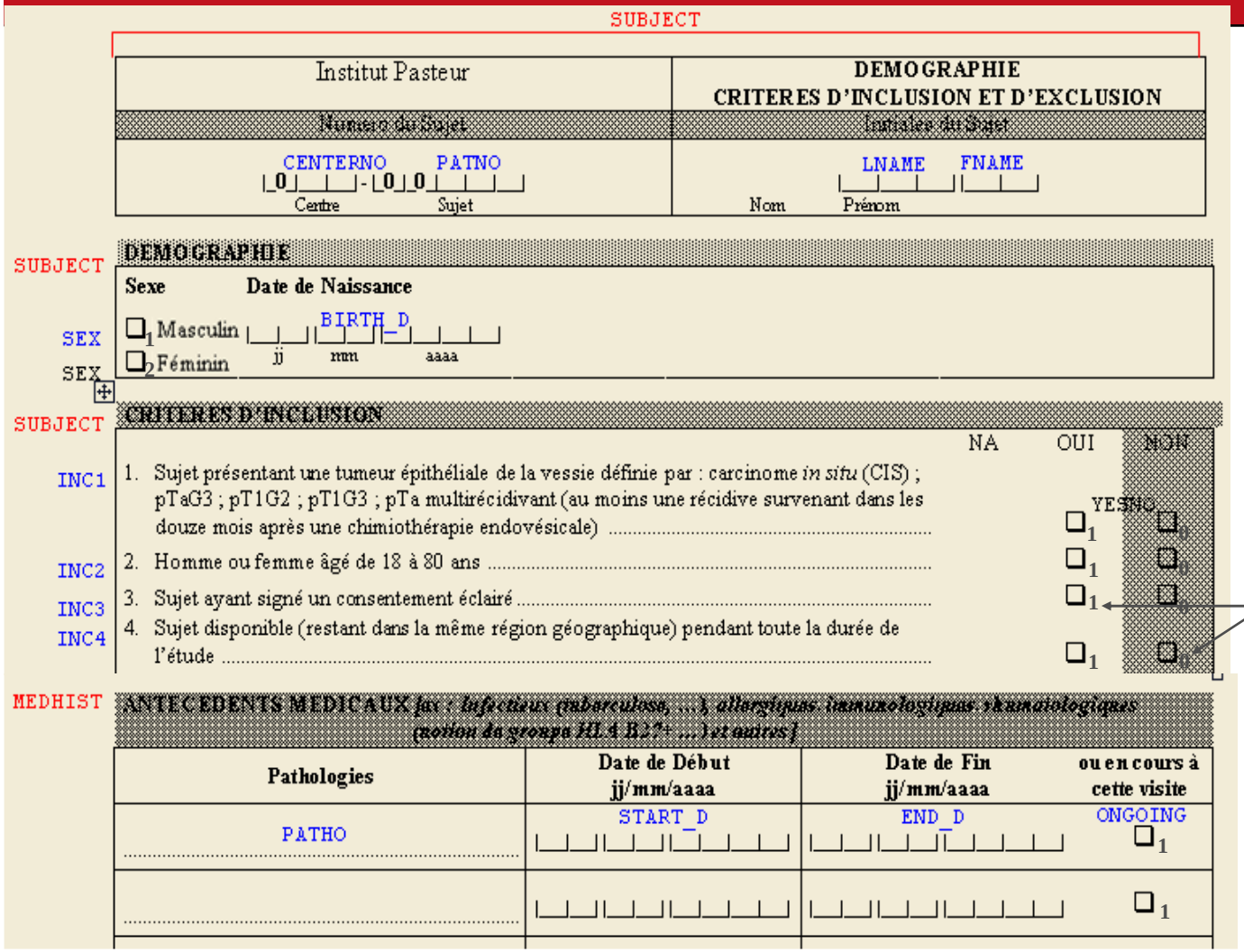
- CRF (Case Report Form) annoté
- Cahier des charges
- Création de la base de données
- Saisie des données
- Programmation des contrôles de cohérence
- Edition des demandes de clarification
- Correction des données
- Codage des données à l'aide de dictionnaires spécifiques
- Revue de données en aveugle (essais cliniques)
- Gel de la base de données
- Traçabilité
- Règlementation / Guidelines

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

CRF annoté (1)

- L'annotation du CRF permet de structurer toute l'information contenu dans le questionnaire
- On indique directement sur un CRF vierge :
  - le nom des tables
  - le nom et type des variables
  - les listes de codes utilisés

CRF annoté (2)



En rouge, le nom de la table

En bleu, le nom de la variable

En indice, le code utilisé

Cahier des charges (1)

- Objectif : lister les contrôles de cohérences qui seront effectués sur les données
- Il est réalisé à partir du protocole et du CRF
- Il doit comprendre :
  - le numéro du contrôle
  - la ou les variable(s) impliquée(s) dans le contrôle
  - le message destiné aux cliniciens

↳Le cahier des charges doit être discuté/validé lors d'une réunion avec tous les participants à l'étude (chef de projet, ARC, DM, statisticien...)

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Cahier des charges (2)

N°	Variable	Contrôle
1	INIT	• Les initiales du patient doivent être renseignées.
2	DTCONS	• La date de signature du consentement doit être renseignée, complète et calendaire.
3	DTNAIS	• La date de naissance doit être renseignée, complète et calendaire.
4	DTNAIS / DTCONS	• La date de naissance doit être antérieure à la date de consentement.
5	SEXE	• Le sexe du patient doit être renseigné et être indiqué à 1 (Masculin) ou 2 (Féminin).
6	DTV1	• La date de visite doit être renseignée, complète et calendaire.
7	DTV1	• La date de la visite 1 doit être postérieure au 01/01/2002 et antérieure à la date de saisie du CRF.
8	DTCONS / DTV1	• La date de signature du consentement doit être antérieure ou égale à la date de la visite 1.

Création de la base de données (1)

Choix du logiciel en fonction :

- du type d'étude (essai clinique, enquête…)
- du nombre de sites impliqués
- du type de saisie (simple ou double)
- du nombre d'accès simultanés à la base

EpiData	Access	EasyPHP / Voozanoo
Petites enquêtes et saisies en local	Etudes plus complexes et accès multiples à la base	Etudes multicentriques et saisie via le web

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Création de la base de données (2)

- Elaboration de la structure de la base de données sur papier à partir du CRF
- Création de la base de données sous le SGBD :
  - Tables et variables
  - Schéma relationnel
  - Masques de saisie
  - Premiers contrôles (à partir du CRF, du protocole et du cahier des charges)
  - etc…

Saisie des données

- Saisie : simple ou double ?
  - Double saisie : permet d'éviter les erreurs en cours de saisie (lecture, frappe…)
  - Choix avant l'élaboration de la base de données car répercution sur la structure
- Si double saisie :
  1) Saisies indépendantes
  2) Comparaison des deux bases pour identification des différences
  3) Retour au CRF pour vérification et correction des données de chaque base

Contrôle des données

CRF et protocole

Cahier des charges

Programmation des contrôles

Le cahier des charges suit l'ordre du CRF et la programmation est le reflet du cahier des charges

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Contrôle des données après saisie

Le programmes de contrôles doivent détecter :

- Les déviations de protocole

  Ex : Respect des critères d'inclusion

- Les erreurs de remplissage du cahier (incohérence sur une donnée ou entre plusieurs données)

  Ex: Valeurs aberrantes de taille, poids…

  Test de grossesse = NA pour les hommes

- Les données manquantes qui seront potentiellement récupérables

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Contrôle des données après saisie

420 - Queries listing by volunteer

num=1028 init=SAA

Item	Query	Correction	Variable
CRF - Lab report V1	CRP result in CRF equal to 'Normal' but at least one result of lab report is equal to CS/NCS. Please verify.		crp
CRF : Systolic BP at V5	Value '153' out of limits [90-139]. Please verify.		bpsys
CRF : Heart rate at V5	Value '111' out of limits [50-100]. Please verify.		heart
Diary : Temperature (morning at D14)	Value '34.9' out of limits [35-37.4]. Please verify.		temp_am
Lab report : CRP at V1	Value '3.2' out of range [0-2.5] but NCS/CS not specify. Please verify.		crp

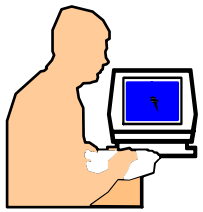
NB : 7=NA, '.' and 8=Missing data, ''=Missing date, 9=ND/NK *Edité sous SAS le 10/02/06 à 16:08* *Date : _ _ / _ _ / _ _* *Signature :_____*

Les demandes de clarification (DCF) doivent être datées, signées et archivées dans le dossier du sujet concerné

Correction des données

Opérateur

Donnée initiale :
Poids = 180 kg



Opérateur

Réponse à une query :
Poids = 80 kg

AUDIT-TRAIL

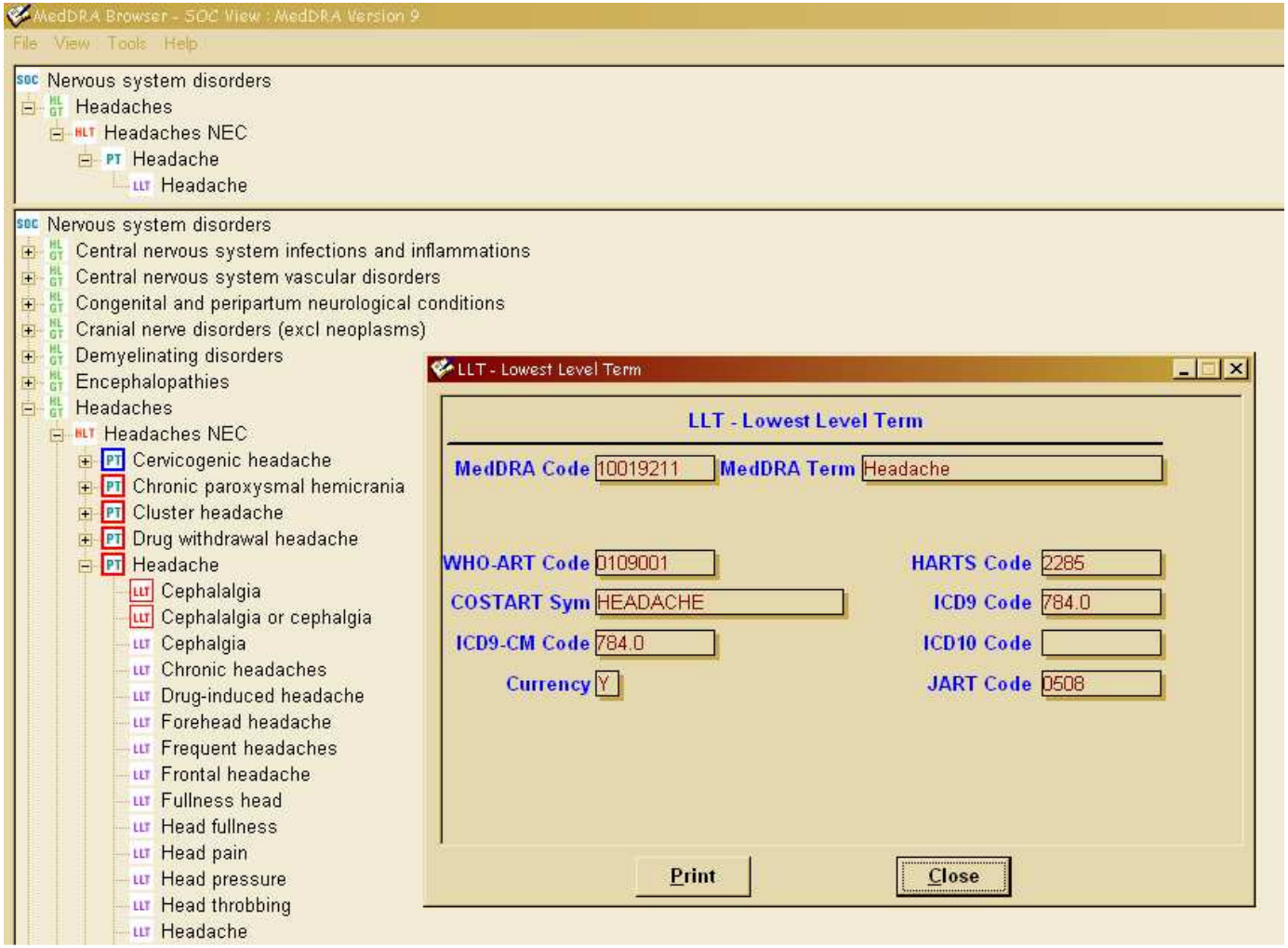
Date	Utilisateur	Variable	Ancienne valeur	Nouvelle valeur	Query
14/03/07	Opérateur	Poids	180	80	Veuillez vérifier le poids du patient (180 Kg).

Codage des données à l'aide d'un dictionnaire

- Pourquoi ?
  - Facilité de manipulation
  - Facilité d'analyse
  - Présentation standard des informations
- Les dictionnaires les plus utilisés :
  - MedDRA ou ICD9CM pour coder les antécédents, pathologies, actes chirurgicaux
  - MedDRA ou WHO-ART ou COSTART ou HARTS pour coder les événements indésirables
  - WHO-DRUG pour les médicaments

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Codage des données : Le dictionnaire MedDRA



Revue des données en aveugle
(nouveaux produits)

Objectifs :

- Définir les déviations par rapport au protocole et les classer en « mineures » ou « majeures »
- Rechercher les incohérences non détectables par des contrôles automatiques

➔ Edition et revue (en aveugle des traitements) des listing de données par patient

Gel d'une base de données

- But : Empêcher toute modification de données par un verrouillage physique de la base ou une copie des données

- Quand : Lorsque les données sont « propres » et la revue des données en aveugle terminée

- Pourquoi : - Analyse statistique intermédiaire
  - Analyse statistique finale

Dégel d'une base de données

Dégel de la base possible :

- Obtenir les accords
- Documenter et motiver
- Re-verrouiller la base

Traçabilité

Il est indispensable de conserver toutes traces d'opération sur les données :

- Journal de saisie
- Comparaison des saisies
- Demandes de clarification
- Corrections évidentes de la base
- …

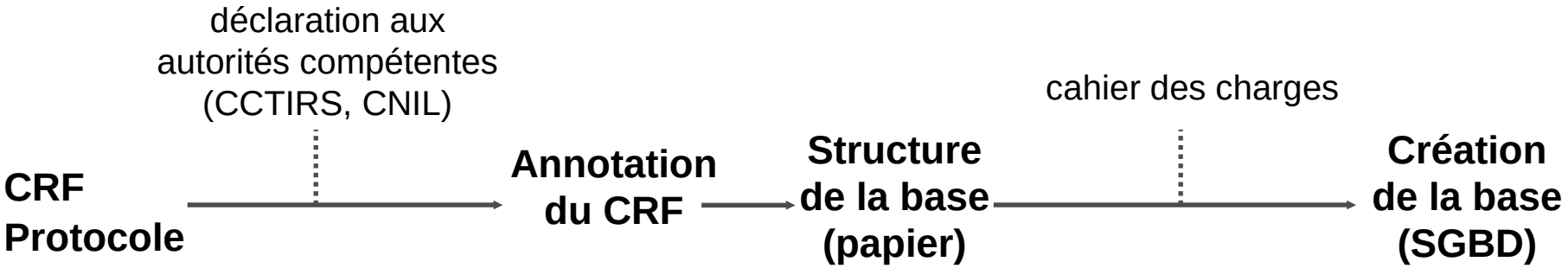
et de les joindre soit au questionnaire de chaque sujet, soit à la documentation relative à la base de données

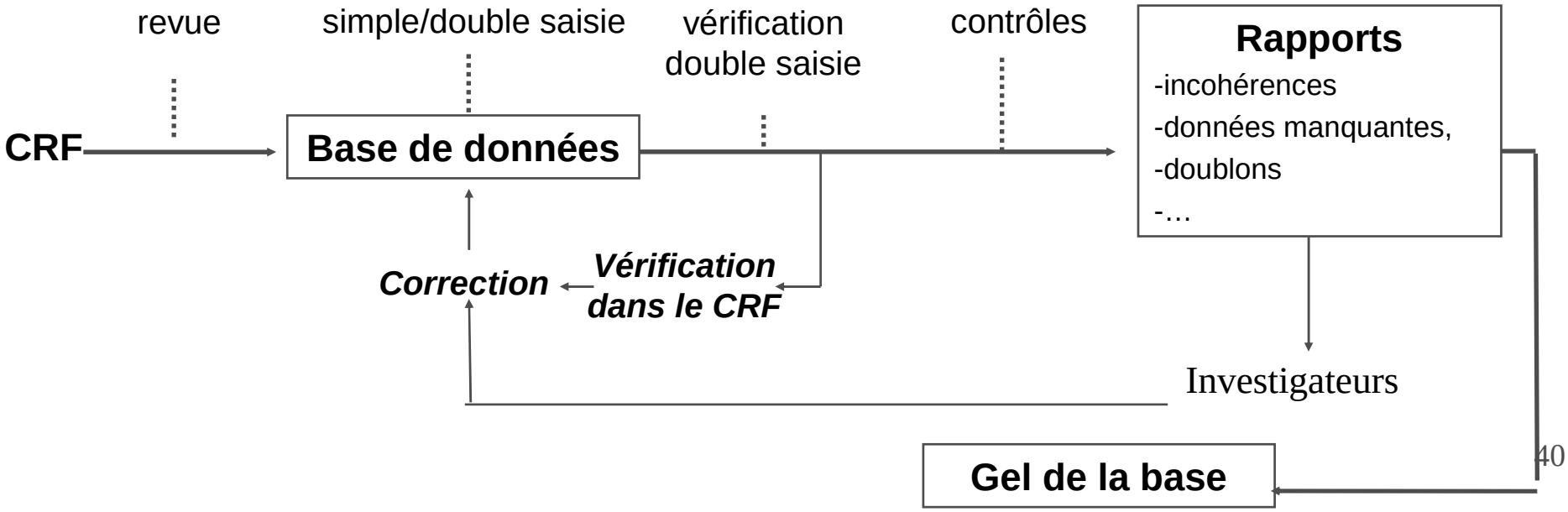
Réglementation / Guidelines

- Réglementation (données traitées en France) :
  - Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé (CCTIRS) www.recherche.gouv.fr/cid20537/cctirs.html
  - Comission Nationale de l'Informatique et des Libertés (CNIL) www.cnil.fr

- Guidelines :
  - ICH www.ich.org
    - ➔ Bonnes Pratiques Cliniques (ICH E6)
    - ➔ General Consideration of Clinical Trials (ICH E8)
  - Bonnes Pratiques de Data Management (www.scdm.org)
  - …

Conclusion


déclaration aux autorités compétentes (CCTIRS, CNIL)
cahier des charges
CRF Protocole
Annotation du CRF
Structure de la base (papier)
Création de la base (SGBD)


revue
simple/double saisie
vérification double saisie
contrôles
Rapports
-incohérences
-données manquantes,
-doublons
-…
CRF
Base de données
Correction
Vérification dans le CRF
Investigateurs
Gel de la base

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Conclusion

- Le Data Management est une étape très importante dans une étude car il conditionne la qualité des données et par conséquent celle des résultats statistiques
- C'est aussi l'étape la plus longue, exigeant rigueur et organisation
- L'efficacité du DM dépendra pour beaucoup de votre connaissance du protocole et de la rédaction du cahier des charges

Solthis
SOLIDARITÉ THÉRAPEUTIQUE & INITIATIVES CONTRE LE SIDA

Cours de l'Institut Pasteur

www.pasteur.fr rubrique Enseignement

- Création et Gestion de Bases de données sous Access et Epidata
- Analyse de données avec Stata (Tests statistiques, régression logistique, survie)
- Introduction à l'Epidémiologie et aux Biostatistiques et Validation des Tests Diagnostiques
- Essais Cliniques et Maladies Infectieuses et Tropicales
- …