

Peut-il y avoir des données identifiantes dans une base de données de recherche ?

Rédaction CRC et CCER

Vous êtes investigateur et vous prévoyez de récolter les données de votre projet de recherche via un formulaire (CRF) électronique, créé par exemple avec REDCap. En plus de vos données de recherche, vous avez prévu d'enregistrer le nom du patient et son numéro de téléphone. Le CRC et la CCER vous demandent pourtant de n'enregistrer aucune donnée identifiante (nom, prénom, date de naissance etc.) dans votre fichier de données de recherche.

Pourquoi ne peut-on pas noter le nom du patient dans le CRF ?

Quand on a le patient en face de soi, c'est vite fait de confondre le No. 43567 avec le No. 42567 et ainsi de mettre les valeurs de Mme Dupont à la place de M. Duval. S'il y avait nom, prénom et date de naissance, de telles erreurs ne seraient-elles pas plus facilement détectées ?

Certains systèmes de récolte de données (CRF) permettent l'intégration de données identifiantes dans des formulaires spécifiques qui ne seront pas exportés lors de l'analyse des données. C'est notamment le cas de REDCap. Cette solution a priori attrayante, laisse ces systèmes vulnérables en cas de piratage informatique.

Comment faire alors ?

Il est nécessaire d'avoir deux fichiers séparés, dont un avec nom, prénom, date de naissance, et l'autre avec les données de recherche. Le lien se fait par un identifiant unique, ainsi défini dans les Bonnes Pratiques cliniques (ICH E6):

- **Article 1.58** : Code d'identification du sujet : Identificateur unique attribué par l'investigateur à chacun des sujets participant à l'essai pour protéger son identité et que l'investigateur utilise au lieu du nom du sujet lorsqu'il consigne des effets indésirables ou d'autres données se rapportant à l'essai.

- **Article 5.5.5** : Le promoteur doit utiliser des codes d'identification des sujets non ambigus (voir 1.58) permettant l'identification de toutes les données consignées pour chaque sujet.

Reste le problème de confusion entre les patients identifiés seulement par un code : une solution pragmatique est celle d'ajouter, dans le fichier qui contient les données de l'étude, quelques indications d'identité, mais pas l'identité complète. L'étude suisse de cohorte VIH par exemple, avec plus de 20'000 sujets, utilise le sexe, la taille en cm, et l'année de naissance.

Cette solution pragmatique a fait ses preuves, mais n'est pas toujours applicable. Dans une étude en pédiatrie par exemple, la taille n'est pas un paramètre invarié et dans une étude aux soins intensifs, elle n'est pas connue. Il faudra donc individualiser selon l'étude. Beaucoup d'études étrangères demandent la date de naissance complète combinée aux initiales, mais en Suisse on considère que cela pourrait permettre d'identifier la personne*.

Alors quelles données utiliser ?

Voici deux idées pour éviter les confusions:

- dans une étude pédiatrique, cela pourrait être le mois de naissance combiné aux 4 derniers chiffres du portable de la maman;
- la troisième lettre du prénom, le nombre de lettres du prénom, la troisième lettre du nom, le nombre de lettres du nom, suivi de l'année de naissance à l'envers (Bernard Hirschel: **r7r86491**). Un tel système fut utilisé pour ne pas compter en double les personnes VIH-positives.

Par ailleurs, noter le numéro de patient de l'étude dans son dossier médical permettra de ne pas confondre les patients lors de la récolte de données.

D'autres solutions plus sophistiquées sont possibles avec des logiciels (coûteux) qui font un lien temporaire entre 2 bases de données (exemple du Bus Santé à Genève).

Au niveau suisse, une solution est à l'étude. Elle repose sur la possibilité de créer un code à partir de l'identité, l'inverse étant impossible. Il n'est cependant pas clair si les Hôpitaux Universitaires de Genève vont adopter cette solution, et si oui, quand. Si vous êtes intéressé : ideal-project.ch.

Conclusion

1. Il est obligatoire de séparer les données identifiantes des données de l'étude quel que soit le logiciel utilisé.
2. Les procédés qu'offrent certains systèmes de récolte des données (champs non exportables) ne doivent pas être utilisés (moindre protection contre le piratage).
3. Pour éviter des confusions entre patients, des données identifiantes partielles peuvent être ajoutées au fichier qui contient les données de l'étude. Ces solutions doivent être adaptées à chaque projet en fonction de la population étudiée.
4. Au niveau suisse, le projet IDEAL sera peut-être une solution à l'avenir.

A noter

- Journée de formation commune des membres de la CCER et CER-VD : le 24 novembre 2021 à Lausanne.
- Colloque du CRC du 13 décembre : **Nouveautés à la CCER** : visites de suivi, consentement général, ...
- Formation aux bonnes pratiques (GCP niveau II - promoteur) : **les 19 et 20 janvier 2022**
Lien pour l'inscription : <https://crc.hug.ch/enseignement/formations-bonnes-pratiques-des-essais-cliniques>

*Voir la recommandation de swissethics sur le codage : https://swissethics.ch/assets/Themen/akzeptierte_verschlueselung_f.pdf

Rédacteur principal - Pr Bernard Hirschel