Atelier CNIL n°2 – Entrepôt de données de Santé

Visioconférence 12 octobre 2022

Groupement	Présents
HUGO	Christine RIOU, Marie LEBIGRE, Leslie GUILLON, Malvina DUTOT, Fanny GAUDIN, Emilie VAREY, Pascal VAN-HILLE, Morgan LE MAY, Jean-Marie CHRETIEN, Delphine TOUBLANT, Anne HESPEL, Emeric GUILLOU, Mathilde GRENTE, Claude PELLEN, Emeline LAURENT, Valentine GUITON
GGEST	Alban DUPOUX, Vincent VUIBELT, Sanae BOUALI, Claire BERTAUT, Bérenger MARTIN, Laurène ASSAILLY, Justine SELOSSE, Florent COLLOT, Melchior D'AGAY, Maxime SALAUN, Aurélie BANNAY, Vianney GUARDIOLLE, Jérôme FILIPONNE, Sébastien KRONZ, Nathalie THILLY, Sophie ZEVACO, Rémy BASSO, Igor MARCOUT
G4	Laetitia FIZET, Anthony BOUZIDI, Cécile CHEVASSUS-CLEMENT, Cédrick MORMANN, Leonardo BONILLA LOZANO, Hannah BOONE, Fabien CHAILLOT, Thierry GACHES, Charlotte GEAY FERRAN
CNIL	Marion JABOT, Erik BOUCHER DE CREVECOEUR, Manon DE FALLOIS

Ordre du jour
Introduction (Mme Fanny Gaudin, Déléguée Générale GCS HUGO) **1/** Présentation check list de conformité au référentiel Entrepôts de données dans le domaine de la santé (CNIL) **2/** Mise en œuvre du référentiel EDS, Questions du groupe de travail des interrégions Ouest, Nord Ouest et Grand Est et échanges sur la mise en œuvre par les établissements du référentiel EDS (Marie LEBIGRE CHU de Nantes, Dr Christine RIOU plateforme ODH) 3/ Position des patients sur la réutilisation de leurs données et les modalités d'information, présentation d'une enquête auprès des patients et échanges avec la CNIL (Pr Vincent VUIBLET CHU de Reims) **Report à une session ultérieure – en attente de date**

Préambule
Le présent CR sera mis à disposition des participants sur le site du GCS HUGO. La séance a été introduite par Mme GAUDIN.

1/ Présentation check list de conformité au référentiel Entrepôts de données dans le domaine de la santé

La check-list publiée par la CNIL est disponible sur internet : https://www.cnil.fr/fr/entrepots-de-donnees-de-sante-la-cnil-publie-une-check-list-de-conformite-son-referentiel

La check-list reprend tous les points du référentiel, elle se veut autoporteuse. L'outil a été développé afin d'aider les Responsables de Traitement (RT) à valider leur conformité au référentiel Entrepôt de Données de Santé (EDS). La mise à disposition d'un outil standardisé a été réalisée par la CNIL après avoir observé que plusieurs acteurs développaient leur propre outil. Il facilite ainsi la constitution et la revue des dossiers des demandes d'autorisation.

Chaque fois que l'on coche un faux, cela signifie que le traitement n'est pas conforme, dans ce cas, il convient de solliciter une autorisation spécifique auprès de la CNIL en justifiant chaque écart au référentiel.

Il est conseillé de joindre au dossier de demande d'autorisation cette check-list en justifiant les écarts et en présentant les mesures de compensation. Pour les dossiers de demande d'autorisation, les établissements peuvent continuer à utiliser leur propre check-list s'ils le souhaitent.

Série de Questions - Réponses	
A la lecture du référentiel EDS, il semble que certaines « petites » bases de données répondent désormais à la définition d'un EDS alors qu'elles sont gérées comme des bases de données de recherche. Le référentiel EDS, très complet et adapté pour les EDS, paraît disproportionné pour ce type de base – ces bases dépendent-elles désormais de ce référentiel ?	Le référentiel a été pensé pour couvrir un certain nombre de cas non problématiques, ce qui éviterait ainsi l'instruction de nombreux dossiers.
Quel est donc le périmètre que la CNIL envisage pour ce référentiel ? Est-il réellement applicable à des projets de cohortes ?	Il n'existe pas de réponse générale, il s'agit de raisonner au cas par cas. Si la cohorte est constituée dans le cadre d'un périmètre de recherche mais que cette dernière nécessite une collecte de données au fil de l'eau sur de longues durées, il est important de prévoir des mesures de sécurité adaptées. Toutes les bases recherche ne sont pas concernées par le référentiel. Ce qui caractérise un EDS c'est l'accès sur le long terme par plusieurs équipes et non un accès par une équipe avec un nombre limité de personnes pour une finalité de recherche précise. **Les anciennes bases ont effectivement une marche à monter importante si elles relèvent du référentiel.**

2/ Mise en œuvre du référentiel EDS, Questions du groupe de travail des interrégions Ouest, Nord Ouest et Grand Est et échanges sur la mise en œuvre par les établissements du référentiel EDS

Article 3 : Objectif(s) poursuivi(s) par le traitement (finalités) et gouvernance		
Précisions de l'article	**Questions**	**Réponses**
3.1 Finalités **3.1.2 Finalités particulières** - fonctionnement d'outils d'aide au diagnostic ou à la prise en charge	Le libellé est très général. Doit-on comprendre qu'on parle là d'outils déjà existants en place et non du développement de nouveaux outils (auquel cas on serait dans la recherche) ?	Effectivement, **on parle des outils déjà existants**, tout ce qui est **développement de nouveaux outils, relève de la recherche**.
3.1 Finalités **3.1.2 Finalités particulières** - études de faisabilités (pré-screening) *« Lorsqu'ils sont mis en œuvre **exclusivement** à partir des données de l'entrepôt par les personnels habilités du responsable de traitement et pour son usage exclusif, les traitements répondant aux finalités suivantes peuvent être mis en œuvre dans le cadre de la déclaration de conformité au présent référentiel »*	Sur la première condition (=uniquement avec des données de l'entrepôt) : est-ce strict ? Est-ce que le fait de rajouter des données issues du dossier médical (pas encore dans l'entrepôt) que l'on croise avec les données de l'EDS nous fait sortir du cadre ?	L'objectif ici est de sécuriser les données de l'entrepôt en bordant les usages ; **à terme, on ne doit plus extraire de données de l'EDS**, ou alors **uniquement dans des espaces sécurisés** : modèle strict du SNDS. Il est donc primordial que les EDS soient enrichis et exhaustifs au plus vite. Dans le cadre du référentiel, on peut lancer une requête, à des fins de faisabilité, c'est à dire pour compter un nombre de patients dans l'EDS. Par contre, **les allers retours entre l'EDS et d'autres bases/supports**, ne sont pas prévus dans le référentiel qui n'a pas été écrit pour gérer ces situations, on rebascule sur une **demande d'autorisation**.

Article 5 : Données à caractère personnel pouvant être incluses dans l'entrepôt		
Précisions de l'article	**Questions**	**Réponses**
5.1 **On ne peut intégrer dans un EDS que les données qui figurent dans le dossier médical et administratif dont la collecte est justifiée par la prise en charge**	Ceci exclut l'appariement avec les données de décès sauf si on les rapatrie d'abord dans le DPI. Les données de décès de l'INSEE sont pourtant des données publiques et le statut vital est une catégorie de données permise au 5.2.1.2 Autres catégories de données à caractère personnel, comprenant des données sensibles. Cela est problématique car il est compliqué pour des raisons techniques et organisationnelles d'alimenter le DPI avec les données de l'INSEE, or le statut vital est une donnée cruciale en général en recherche, et particulièrement en épidémiologie ; s'en passer est une réelle moins-value des entrepôts Enrichir les entrepôts avec le statut vital ne nous parait pas augmenter le risque lié à la vie privée. Au contraire, avoir le statut vital exhaustif, permettrait de faciliter le respect des droits (d'information, d'opposition etc…) en ayant une meilleure visibilité sur les patients à ré-informer (ne pas compter les patients décédés pourrait aboutir à un retour plus facile des patients à ré-informer). Le référentiel n'interdit pas de faire une extraction d'un jeu de données dans un espace projet à partir de l'entrepôt et de l'apparier avec les données de décès ; dès lors, pourquoi ne pas autoriser cet appariement en amont, ce qui gagnerait du temps et des moyens ? Le statut vital est-il réellement juridiquement une donnée à caractère personnel dans la mesure où il s'agit d'une donnée publique ? Pourrait-on introduire une exception sur ce point ?	Après explications sur la difficulté d'introduire dans les DPI les informations du fichier INSEE (du fait de logiciel propriétaire, ou de manque de disponibilité des équipes soins et DSI notamment), **la CNIL n'est pas fermée à envisager une évolution du référentiel sur ce point**, il convient de bien en justifier la pertinence et de revoir la méthodologie de réalisation.
les données issues de recherches précédentes et dont la durée de conservation n'a pas expiré	Fait-on ici référence à la durée de conservation et donc à la durée de réalisation de l'étude ou bien à la durée d'archivage de l'étude ou les deux durées	Ceci indique ici que l'**on peut verser des données d'une étude dans l'EDS jusqu'à la date de fin d'archivage**, à l'issue de cette date, les données ne

	cumulées ?	sont plus censées exister et ne peuvent donc plus être versées dans l'EDS. **Une fois versée dans l'EDS, les données ont une nouvelle durée d'archivage, celle prévue dans le cadre de l'EDS**.
5.2.1 Liste des données patient permises **5.2.1.1 données directement identifiantes et administratives, devant être conservées dans un espace distinct des autres données** **5.2.1.2 autres**	Quelle est la position pour les données de santé des détenus pris en charge à l'hôpital, en particulier les compte-rendus ? Peuvent-ils être versés dans un EDS ? Les comptes rendus peuvent mentionner le statut de détenu ou le statut de détenu peut être déduit du fait de l'unité de prise en charge si il y a une unité dédiée. Ils peuvent bénéficier de l'information individuelle concernant la constitution de l'EDS. Il n'y a pas de risque sur la personne. En ce sens une RNIPH se distingue d'une RIPH. Les données pourraient être utilisées dans une étude RNIPH. Ne pas inclure leurs données de santé pourrait être une perte de chance en cas d'alerte sanitaire.	**Les données d'infraction ne sont pas citées dans le référentiel, elles sont donc interdites dans l'EDS**. Leur collecte doit être justifiée et minimisée.
5.5 les données directement identifiantes ne peuvent être réunies dans l'entrepôt que pour les finalités suivantes : *– recontacter les patients pour leur proposer de participer à des études ou pour les informer régulièrement des projets de recherche n'impliquant pas la personne humaine, réutilisant les données de l'entrepôt les concernant;* *– recontacter les patients à la suite de découvertes de caractéristiques génétiques pouvant être responsables d'une affection justifiant des mesures de prévention ou de soins à leur bénéfice ou au bénéfice de leur famille, à l'exception des cas dans lesquels le patient s'y*	Il manque selon nous 2 situations, peut être les plus importantes (et fréquentes) : - la mise à jour de l'entrepôt au fil de l'eau (retrouver les patients déjà dans l'entrepôt – cela se fait par un programme mis en œuvre de façon systématique) - screening : identifier les patients pouvant rentrer dans une étude RIPH ou pour une étude RNIPH, donner aux demandeurs la liste de patients résultante (identifiante) afin qu'ils puissent ensuite aller consulter les dossiers médicaux ou d'autres sources lorsque toutes les données nécessaires à l'étude ne sont pas dans l'entrepôt – on est bien là dans une démarche nécessitant d'identifier les patients, au contraire des faisabilités pour lesquelles	**La mécanique d'alimentation de l'EDS se fait par un SAS d'entrée qui transforme l'IPP**. Ce dernier n'est pas importé dans l'EDS mais **conservé dans le SAS d'entrée** sur un espace à part. **L'appariement doit être réalisé dans ce SAS** sur un dérivé de l'IPP. Pour les données identifiantes, il doit y avoir une table, à part, indexée avec le même identifiant technique dans un espace cloisonné de la base de données de santé de l'EDS. L'accès y est réduit au strict minimum. **L'utilisation de l'EDS pour du screening avec export de données et appariement à d'autres sources n'est pas prévue dans le référentiel.**

est opposé, conformément à l'article L. 1130-5 du code de la santé publique; *– recontacter les patients à la suite de découvertes annexes liées à l'identification de facteurs de risques et/ou d'identification syndromiques à même de modifier leur prise en charge (thérapeutique ou de suivi);* *- avertir une personne d'un risque sanitaire auquel elle est exposée.*	un simple dénombrement (anonyme) suffit. Le NIR et l'INS figurent parmi les données pouvant être incluses dans l'entrepôt. Initialement les établissements n'ont pas tous déclaré le NIR comme donnée traitée dans l'entrepôt. Cela n'était pas autorisé. Peuvent-ils aujourd'hui l'intégrer ? L'INS est aujourd'hui l'identifiant patient dans le dossier médical informatisé. Il est accessible pour toute personne consultant le dossier médical. Sous quelles conditions le NIR ou l'INS peuvent-ils être utilisés localement pour apparier plusieurs sources de données ? Certaines informations peuvent ne pas être présentes dans l'entrepôt mais présentes dans le système d'information de l'établissement avec un lien possible avec l'INS.	Depuis 2018, **la CNIL autorise le traitement du NIR et de l'INS hors cadre de recherches, il est donc possible de l'introduire dans l'EDS**. En terme de formalité, si l'EDS dispose d'une autorisation et que l'on souhaite l'ajouter comme données dans l'EDS, il est possible de l'intégrer dans l'EDS, ceci ne nécessite pas une demande d'autorisation auprès de la CNIL.
5.8 Dans le cas où des données directement identifiantes, des tables de correspondance, des données génétiques ou des données de suivi de localisation sont versées dans l'entrepôt, celles-ci doivent être stockées séparément des données pseudonymisées, en utilisant les procédés décrits dans les exigences de sécurité SEC-LOG-4 à SEC-LOG- 6.	Qu'est-il entendu par données génétiques ? Un compte-rendu (CR) de génétique devrait-il être isolé du reste ? Il n'est pourtant pas identifiant une fois les identifiants directs enlevés ? Il arrive par ailleurs que des comptes-rendus de consultation d'autres disciplines mentionnent également les mutations, comment doivent être traités ces CR ?	**Les CR de génétique relèvent des données génétiques, il ne peut être versé dans l'EDS que des données génétiques interprétées**. Les CR de génétique au même titre que les données doivent être traités comme des **données hautement sensibles**. Elles doivent faire l'objet de mesures de sécurité particulières. Afin de clarifier ces interprétations, **la CNIL va mener une analyse pour mettre en place une définition plus précise**. Il y a une différence entre la description d'une partie du génome et la présence d'une mutation qui existe chez un grand nombre de personnes. Pour le moment ce **point doit être traité au cas par cas (maladies rares notamment) et relève, sauf données interprétées d'une demande d'autorisation si il y a un besoin urgent.**

8. Information des personnes		
Précisions de l'article	**Question**	**Réponse**
8.4. Les personnes concernées doivent en outre être informées de chacune des réutilisations des données les concernant à des fins de recherche, d'étude ou d'évaluation, sauf lorsque les responsables de traitement se trouvent dans l'impossibilité de réaliser l'information ou qu'elle exigerait des efforts disproportionnés.	Dans le cas où le responsable de traitement (RT) est dans l'impossibilité de réaliser l'information ou qu'elle exigerait des efforts disproportionnés, la dérogation est-elle acquise de fait avec inscription de la justification de la dérogation au registre des traitements de l'établissement, une information individuelle initiale sur la réutilisation des données de soins ayant été délivrée ? Lorsqu'une étude est réalisée à partir de données déjà collectées avec un suivi prospectif et une réutilisation des données nouvellement recueillies dans le dossier patient au cours du suivi habituel du patient, l'information individuelle se révèle difficile surtout si l'étude est multiservices ou multi-établissements. Le responsable de l'étude n'est pas en contact avec le patient. Peut-on considérer que l'information collective avec publication de la note d'information sur le site internet fait foi ?	**Une dérogation à l'information des patients doit toujours faire l'objet d'une demande d'autorisation à la CNIL.** **Si le patient a bénéficié d'une information individuelle sur la réutilisation de ses données de soins qui renvoie vers un portail de transparence pour les études qui seront réalisées sur l'EDS : on est en conformité avec la MR004**.

10. Sécurité		
Précisions de l'article	**Question**	**Réponse**
SEC-LOG-4 Dans le cas où des données directement identifiantes ou des tables de correspondance sont stockées dans l'entrepôt, celles-ci doivent être séparées logiquement des données pseudonymisées par des moyens cryptographiques. Par exemple, les données administratives des patients et les tables de correspondance doivent être chiffrées avec des clés différentes de celles utilisées pour chiffrer les données de santé de l'entrepôt.	La séparation entre données identifiantes et pseudonymisées était la demande de la CNIL au moment de l'autorisation Ehop Rennes. Pourquoi avoir ajouté dans le référentiel une exigence de cryptage, des moyens cryptographiques ?	**Le simple cloisonnement physique ne suffit plus**, les cyber-attaques sont plus sophistiquées qu'il y a quelques années ; une attaque à distance permet maintenant de reconstituer l'architecture. **Le chiffrement est une garantie suffisante en matière de risque**, qui fait que la CNIL n'a pas besoin d'étudier une proposition alternative. Si le chiffrement n'est pas possible selon les modalités décrites dans le référentiel **d'autres mesures de sécurité peuvent être suffisantes**, mais dans ce cas la CNIL doit vérifier et juger, d'où l'exclusion au

		référentiel et le **passage par une demande autorisation spécifique**.
SEC-LOG-5 L'accès aux deux catégories de données séparées définies à l'exigence SEC-LOG-4 doit être effectué via des comptes utilisateur différents, ou via un seul compte utilisateur devant choisir à la connexion un des profils d'habilitation différents qui lui sont attribués.	Les profils applicatifs concernant Ehop sont conformes, mais il peut exister par ailleurs, dans certains centres, aussi des profils "administrateurs", qui de fait, techniquement, ont accès à tout dans le cadre de leurs missions, notamment via des outils de requêtage en direct sur la base, sans passer par une application comme Ehop (par exemple R). Ils sont peu nombreux, formés, habilités, dans un service spécifique. Cela est-il conforme ? Pour les accès à partir de Rstudio on n'est plus sur des droits ehop mais sur des droits ORACLE donnés à un utilisateur identifié et habilité, à Rennes les datascientists, qui peuvent pour une requête complexe par l'interface eHop interroger l'entrepôt global. Un datamart est ensuite constitué dans eHop pour mise à disposition de l'utilisateur ou accessible par Rstudio pour analyses pour les utilisateurs disposant de droits ORACLE sur ce datamart, en pratique les utilisateurs habilités sont les datascientists.	Il convient **d'éviter des comptes avec trop de privilèges**. L'idée ici est **qu'une personne,** même administrateur, **ne doit pas se connecter tout le temps avec un profil qui donne accès à tout**. *Exemple : un data scientist qui peut être amené à faire du contrôle qualité sur toute la base n'a pas besoin d'accéder à toutes les données tous les jours à tout moment. Il doit avoir **un profil de tous les jours avec uniquement les droits suffisant ; et un profil avec plus de privilèges** dont il se sert pour les opérations le nécessitant.* On peut faire le parallèle avec les administrateurs système : un compte en lecture pour tous les jours et un compte en écriture dont on se sert juste ponctuellement pour installer quelque chose. **On se sert du profil avec le moins de privilège par défaut, et on augmente ses droits quand on en a besoin uniquement.** Pour Ehop, on ne va pas tous les jours requêter sur la partie identifiante, les profils de tous les jours ne doivent donc pas avoir ces droits. Il en est de même sur R. Ainsi si un pirate usurpe les droits, il usurpe les droits d'un profil avec moins de privilèges. Les pirates font de l'escalade de privilèges, il convient donc d'être vigilant à cet égard pour stopper les pirates.

SEC-LOG-6 Dans le cas où des données génétiques ou de suivi de localisation sont collectées, celles-ci doivent faire l'objet d'un chiffrement distinct avec une clé spécifique par rapport aux autres données de l'entrepôt. La clé de déchiffrement des données génétiques ou de suivi de localisation ne doit être mobilisable que par les profils d'habilitation responsables de l'alimentation de l'entrepôt et de l'exportation de données vers un espace de travail.	Qu'appelle-t-on ici le suivi de localisation ? Le géocodage IRIS peut-il être collecté comme le reste sans être ni séparé ni chiffré ?	La réponse dépend du niveau de granularité de ce qui est collecté et en quoi cette collecte est nécessaire : si c'est très précis alors cela peut être très ré-identifiant, on est donc dans **la catégorie à séparer/chiffrer**. **On n'autorise jamais de géolocalisation exacte.** La CNIL étudie si le code IRIS peut être utilisé pour les zones rurales et revient vers nous.
PARTIE Espace de travail SEC-ESP-1, 2, 3	Qu'est-il entendu par espace de travail ? Des datamarts (= sous entrepôt) ? Ou bien des espaces de travail sur lesquels on dépose les extractions faites à partir de l'entrepôt pour analyse et traitement ?	C'est **un accès donné à un chercheur** (membre de l'équipe de recherche dument autorisée : médecin, statisticien, …) **pour travailler sur un jeu de données minimisées** par rapport à l'étude et par rapport à la justification de l'étude **avec les outils nécessaires pour faire les analyses**. Techniquement cela peut être différentes choses : une vue, une vue matérialisée, une table temporaire, un espace physiquement séparé qui sert uniquement au projet de recherche. Le plus sûr est un espace séparé (une simple vue offre toujours un risque en cas de bug logiciel qui ferait que, avec une mauvaise requête, on accèderait à d'autres données) ; d'un autre côté, l'espace séparé oblige à dupliquer les données, ce qui créé un autre risque. **Remarque de Pascal Van Hille : sur Ehop, les datamarts sont bien cloisonnés, les données sont dupliquées. Sur l'ODH, les espaces de travail seront bien cloisonnés comme demandé dans le référentiel.**
PARTIES Espace de travail SEC-ESP-1, 2, 3 ET Exportation de données SEC-EXP-1, 2, 3, 4, 5	Le niveau de sécurité demandé dans les 2 parties « Espaces de travail » et «extraction» est impactant sur	**L'objectif à terme est que le travail de recherche sur des données se fasse**

	les infrastructures classiques d'un SIH et sur l'organisation générale. C'est un niveau de sécurité analogue à celui du SNDS, qui sont de très grands ensembles de données, provenant de multiples sources, et pour lesquels une sécurité accrue se justifie (exporter que des données anonymes ; pour les données non anonymes ne travailler que sur des espaces cloisonnées avec audit trail etc…). Or, concernant l'entrepôt d'un établissement de santé : *(1) on est sur des extractions de données réalisée pour une étude donnée : ce type de jeu de données, hors entrepôt et études sur données massives, est analogue, en type de données, volume et risque, à ceux manipulés tous les jours à des fins de recherche sous le régime des MR001, 003 et 004 selon le type d'étude, ou pour des cohortes ; les MR acceptent que ces jeux de données soient codés/pseudonymisés et non anonymes, et que les analyses et extractions soient faites sur des serveurs classiques du SIH ; (2) les données sont uniquement celles des patients de l'établissement, et ce sont uniquement des données provenant des bases de soin ; l'entrepôt n'est qu'un intermédiaire entre les bases de soin et les chercheurs.* ➢ Qu'est ce qui justifie un tel écart d'exigence de sécurité, alors que les données sont les mêmes (par exemple dans le cadre d'une MR004, seule la source diffère : entrepôt versus les bases de soin) ? ➢ L'entrepôt peut aussi être utilisé pour compléter une base de données déjà constituée, registre par exemple, notamment des données structurées (biologie, médicaments administrés…). Dans ce cas un export est nécessaire.	**exclusivement sur des espaces de travail sécurisés (modèle SNDS).** Extraire, c'est toujours un risque (que les données se retrouvent sur des supports non sécurisés USB ou autre), c'est toujours une perte de contrôle. Le principe de l'EDS est d'éviter les bases séparées. Cela fait en effet un gap avec d'autres circuits et d'autres études, et la CNIL a conscience que les deux modes de travail vont coexister. Néanmoins, l'idée est **qu'à terme les EDS soient la seule source d'accès à des données pour la recherche avec des espaces de travail sécurisés**, et que cela soit pratique pour les chercheurs. On offre un espace projet sécurisé au chercheur qui n'a plus de questions à se poser sur la sécurité des données. Il faut tendre vers ce modèle. Juridiquement, les MR continuent à exister. **Ce point va nécessiter un engagement fort des directions pour se mettre en conformité.** **La CNIL est à notre disposition pour organiser un séminaire ou des réunions techniques avec des interlocuteurs recherches, informatiques afin de convaincre et expliquer ou rechercher des façons de faire, pour passer de nos serveurs actuels aux espaces de travail souhaités.**

SEC-JOU-3 Un contrôle des traces doit être réalisé régulièrement et a minima bimestriellement, ainsi qu'à la fin de chaque période d'habilitation liée à un projet de recherche. Ce contrôle doit être réalisé par: *– une solution réalisant une surveillance automatique avec une remontée d'alertes traitées manuellement par un opérateur habilité;* *– ou par un contrôle semi-automatique via exécution de programmes permettant une sélection des traces anormales, suivi d'une relecture manuelle par un opérateur habilité*	Les volumes de log sont énormes à l'échelle d'un entrepôt. Comment en pratique cela peut-il être exploité ?	Si **le système remonte des alertes, il n'est pas nécessaire de garder les traces aussi longtemps, elles doivent être traitées au fur et à mesure**. Il existe une seconde stratégie : **une remontée ponctuelle des traces avec la réalisation d'une analyse plus approfondie sur un gros volume à échéances régulières**. Il convient d'en rediscuter avec le RSSI et la DSI, a priori, ce sont les mêmes outils que pour le SIH, ce serait juste des règles métiers spécifiques à mettre en place pour détecter les incidents (accès de l'étranger, de nuit…). Il conviendrait également de mettre des règles fonctionnelles sur le risque de réidentification : par exemple un chercheur qui lance des requêtes régulièrement avec des résultats sur un ou 2 patients, etc… **Ce point est à réfléchir. Une réunion technique pourrait être organisée pour l'explorer.** Il existe déjà des outils de surveillance de cyber sécurité classiques : contrôle des connexions, fréquence de connexion, volume de données traité, exporté ; il faut prendre en compte des cas d'usage métier variés et la technique. **Une analyse de risque est à réaliser.**
Partie Procédures de ré-identification SEC-REI-1, 2, 3, 4, 5	Nous avons interprété toute cette partie comme n'étant pas applicable dans le cas d'Ehop car les identifiants directs sont présents dans l'entrepôt dans un schéma séparé. La réidentification se fait directement (sans besoin de recourir à une procédure) par une personne habilitée à accéder aux tables de correspondance, et l'application des droits des patients ne pose pas de problème de ce point de vue.	**L'exigence du référentiel ne porte pas sur la façon de gérer ce process mais sur l'existence d'une procédure.** Automatiser une réidentification n'est pas conforme, il faut faire intervenir un contrôle humain. Il convient de : ➢ mettre en place une procédure technique et humaine qui doit expliquer qui a le droit de

	Notre interprétation est-elle correcte ?	réidentifier, où sont stockées les données réidentifiantes, etc… ➢ conduire une analyse de risques pour ces process.
SEC-PSE-1 : Aucun numéro interne, tel qu'un numéro de dossier patient ne peut être directement réutilisé comme identifiant au sein de l'entrepôt. Seul un identifiant pseudonyme unique peut être utilisé, permettant le cas échéant la correspondance entre les données pseudonymisées stockées dans l'entrepôt et des données directement identifiantes. Cet identifiant doit être dédié à un seul entrepôt. Il doit être généré par une fonction de hachage cryptographique résistante aux attaques par force brute ou un générateur de nombres pseudo-aléatoires cryptographiquement sûr. Les données doivent être pseudonymisées préalablement à leur intégration dans l'entrepôt. *SEC-ESP-2 : Les jeux de données importées dans un espace de travail spécifique à un projet de recherche doivent être minimisés et limités aux seules données nécessaires au projet. Un numéro pseudonyme unique spécifique à chaque espace de travail devra être généré dans les mêmes conditions qu'à l'exigence SEC-PSE-1*	Le pseudonyme entrepôt et les pseudonymes étude sont générés par un algorithme de façon aléatoire sans risque de collision. La longueur de l'identifiant est de 40 caractères, seuls 8 caractères seront affichés. Malgré cela la lecture est difficile pour l'utilisateur final et un second numéro séquentiel sera créé, identifiant_protocole, pour une meilleure visibilité pour l'utilisateur. L'identifiant protocole sera un rang calculé à partir d'une date système et l'identifiant de l'étude Le numéro pseudonyme unique spécifique à l'étude peut-il être généré selon la méthode décrite ci-dessus ? Ce qui éviterait de générer systématiquement un troisième identifiant et la table de correspondance avec l'identifiant étude ?	L'attribution des deux numéros peut être réalisée en une seule étape. Il convient de revoir techniquement comment cela est faisable. Les pseudonymes issus du hachage pourraient être renumérotés de façon aléatoire.