UNIVERSITE LILLE NORD DE FRANCE
ÉCOLE DOCTORALE BIOLOGIE SANTE
FACULTE DE MEDECINE HENRY WAREMBOURG

THESE D'UNIVERSITE
POUR L'OBTENTION DU GRADE DE
DOCTEUR DE L'UNIVERSITE DE LILLE 2

Soutenue le 25/09/2015 par Antoine Lamer

Contribution à la prévention des risques liés à l'anesthésie par la valorisation des informations hospitalières au sein d'un entrepôt de données

Jury :

Pr. Benoît Vallet	Directeur de thèse
Pr. Régis Logier	Directeur de thèse
Pr. Serge Molliex	Rapporteur
Pr. Stefan Darmoni	Rapporteur
Pr. Régis Beuscart	Examinateur
Pr. Marc Cuggia	Examinateur

Remerciements

J’exprime tout d’abord mes remerciements à mes directeurs de thèse, les Professeurs Régis Logier et Benoît Vallet, qui m’ont donné la chance de réaliser ce travail de doctorat dans un domaine de recherche passionnant. Leurs encouragements et leurs remarques m’ont permis de mener à bien cette thèse.

J’aimerais remercier les Professeurs Serge Molliex et Stefan Darmoni qui m’ont fait l’honneur de juger ce travail, ainsi que les Professeurs Régis Beuscart et Marc Cuggia pour avoir accepté de participer au jury.

Je voudrais remercier le Professeur Benoît Tavernier pour sa confiance ainsi que pour sa disponibilité tout au long de ce travail de recherche. Je remercie également le Professeur Alain Duhamel pour mon intégration au sein de l'équipe d'accueil EA2694.

Je remercie le Docteur Mathieu Jeanne pour son soutien, ses conseils, ses encouragements, son calme face à toute épreuve ainsi que ses remarques très pointues qui m’ont permis de mener à bien mon travail de thèse.

Je souhaite remercier le Docteur Romaric Marcilly pour ses nombreux conseils et encouragements lors de ces quatre dernières années.

Je remercie le Docteur Julien De jonckheere pour ses conseils lors de la rédaction des articles, des réponses aux appels à projet et de ce manuscrit de thèse.

Je tiens également à remercier tous mes collègues du CIC-IT, et en particulier Reza Jounwaz, Idir Ibarissene, Jessica Schiro et Juliette Thong pour leurs conseils avisés.

Je voudrais remercier tout particulièrement Gilles Ditilyeu, Ludovic Jacquinot et Elisabeth D'Alessandro pour m'avoir initié aux problématiques d'entrepôt de données et de traitements ETL. Je voudrais également remercie Christophe Dupire et Olivier Heulers pour m’avoir accompagné sur ce projet.

Je tiens à remercier le Docteur Hervé Menu et le Professeur Gilles Lebuffe du pôle d'Anesthésie-Réanimation pour leur soutien, ainsi que François Delaby et le docteur Amélie Bruandet du Département de l'Information Médicale pour leur collaboration.

J'aimerais remercier les étudiants Infirmiers-Anesthésistes pour l'aide qu'ils m'ont apporté au cours de leur stage de recherche au sein du CIC-IT. Je remercie également les internes d'Anesthésie-Réanimation (Jérôme Jaspart, Léa Sarte Buisson, Justine Mullie, Juliette Masse, Adrien Berthier et Baptiste Rosseel) avec qui j'ai pu concrétiser plusieurs projets.

Je remercie mes parents et ma famille, pour leur soutien inconditionnel.

Je remercie mes amis, Briac Laurenge, Vincent Lartiguet, Anthony Deraedt, Simon Boivinet, Jérémy Pestel et en particulier Elise Ténoglia pour m’avoir encouragé dans cette voie.

Résumé

Introduction

Le Système d'Information Hospitalier (SIH) exploite et enregistre chaque jours des millions d'informations liées à la prise en charge des patients : résultats d'analyses biologiques, mesures de paramètres physiologiques, administrations de médicaments, parcours dans les unités de soins, etc... Ces données sont traitées par des applications opérationnelles dont l'objectif est d'assurer un accès distant et une vision complète du dossier médical des patients au personnel médical. Ces données sont maintenant aussi utilisées pour répondre à d'autres objectifs comme la recherche clinique ou la santé publique, en particulier en les intégrant dans un entrepôt de données. La principale difficulté de ce type de projet est d'exploiter des données dans un autre but que celui pour lequel elles ont été enregistrées.

Plusieurs études ont mis en évidence un lien statistique entre le respect d'indicateurs de qualité de prise en charge de l'anesthésie et le devenir du patient au cours du séjour hospitalier. Au CHRU de Lille, ces indicateurs de qualité, ainsi que les comorbidités du patient lors de la période post-opératoire pourraient être calculés grâce aux données recueillies par plusieurs applications du SIH. L'objectif de ce travail est d'intégrer les données enregistrées par ces applications opérationnelles afin de pouvoir réaliser des études de recherche clinique. Ce travail est intégré au du projet d'établissement du CHRU de Lille et s'intitule DIAGnosTIC.

Méthode

Dans un premier temps, la qualité des données enregistrées dans les systèmes sources a été évaluée grâce aux méthodes présentées par la littérature ou développées dans le cadre ce projet. Puis, les problèmes de qualité mis en évidence ont été traités lors de la phase d'intégration dans l'entrepôt de données. De nouvelles données ont été calculées et agrégées afin de proposer des indicateurs de qualité de prise en charge. Enfin, deux études de cas ont permis de tester l'utilisation du système développé. Une gouvernance du projet DIAGnosTIC a été installée et s'intitulé CoPil DIAGnosTIC.

Résultats

Les données pertinentes des applications du SIH ont été intégrées au sein d'un entrepôt de données d'anesthésie. Celui-ci répertorie les informations liées aux séjours hospitaliers et aux interventions réalisées depuis 2010 (médicaments administrés, étapes de l'intervention, mesures, parcours dans les unités de soins, ...) enregistrées par les applications sources. Des données agrégées ont été calculées et ont permis de mener deux études de recherche clinique. La première étude a permis de mettre en évidence un lien statistique entre l'hypotension liée à l'induction de l'anesthésie et le devenir du patient. Des facteurs prédictifs de cette hypotension ont également étaient établis. La seconde étude a évalué le respect d'indicateurs de ventilation du patient et l'impact sur les comorbidités du système respiratoire.

Discussion

L'entrepôt de données développé dans le cadre de ce travail, et les méthodes d'intégration et de nettoyage de données mises en places permettent de conduire des analyses statistiques rétrospectives sur plus de 200 000 interventions. Le système pourra être étendu à d'autres systèmes sources au sein du CHRU de Lille mais également aux feuilles d'anesthésie utilisées par d'autres structures de soin. Le CoPil DIAGnosTIC manage l'utilisation de l'entrepôt de données à des fins de prévention des risques liés à l'anesthésie.

Abstract

Introduction

Every day, Hospital Information Systems (HIS) exploit and register millions of data related to patient care: biological analysis results, vital signs, drugs administrations, care process... These data are stored by operational applications and provide remote access and a comprehensive picture of Electronic Health Record. These data also serve other purposes as clinical research or public health, particularly when integrated in a data warehouse. The main difficulty is to exploit data in another way than the initial one.

Some studies highlighted a statistical link between the compliance of quality indicators related to anesthesia procedure and patient outcome during the hospital stay. In the University Hospital of Lille, the quality indicators, as well as the patient comorbidities during the post-operative period could be assessed with data collected by many applications of the HIS. The main objective of the work was to integrate data collected by operational applications in order to realize clinical research studies. This work was incorporated to the hospital project.

Methods

First, the quality of the recorded data by the operational applications has been assessed according to methods introduced in the literature or developed in the project framework. Then, data quality problems underscored by the evaluation were managed during the integration step of the ETL process. New data were computed and aggregated to propose indicators in terms of quality of care. Finally, two studies brought out the usability of the system. The governance du project was established and is entitled CoPil DIAGnosTIC.

Results

Pertinent data from the HIS have been integrated in an anesthesia data warehouse. This system stores data about the hospital stay and interventions (drug administrations, vital signs …) since 2010. Aggregated data have been developed and used in two clinical research studies. The first study points out statistical link between the induction and patient outcome. The second study also rated the compliance of quality indicators such as ventilation and the impact on comorbity.

Discussion

The data warehouse and the cleaning and integration methods developed as part of this work allow performing statistical analysis on more than 200 000 interventions. This system can be implemented with other applications used in the CHRU of Lille but also with Anesthesia Information Management Systems used by other hospitals. The CoPil DIAGnosTIC manage the use of the data warehouse to the prevention of risks related to the anesthesia procedure.

Table des matières

Introduction

Introduction

Depuis maintenant plusieurs dizaines d'années, les structures de soins tendent à informatiser leurs services (1,2). Cette informatisation se traduit par une augmentation du nombre d'ordinateurs et d'applications informatiques afin de remplacer progressivement les dossiers « papier » utilisés jusqu'alors.

Dans un premier temps, le travail d'informatisation s'est porté sur la gestion administrative et financière de la structure de soins afin d'optimiser son fonctionnement, puis s'est orienté progressivement vers la gestion des informations de soins en vue d'améliorer la prise en charge des patients. Si au départ, les applications informatiques étaient indépendantes les unes des autres, souvent propres à un service et développées localement, elles peuvent aujourd'hui être déployées sur plusieurs sites, être interconnectées entre elles et échanger des informations au sein d'une même structure de soins, voir entre plusieurs structures différentes.

Le Système d'Information Hospitalier (SIH) englobe l'ensemble des informations nécessaires au fonctionnement d'un établissement de santé, dont les différentes applications et bases de données permettent d'enregistrer, de stocker, d'interroger et d'exporter des informations médicales ou administratives. A l'origine, le SIH servait à répondre aux objectifs suivants :

- Proposer une vision complète du dossier médical du patient (constatations cliniques et paracliniques, imagerie, biologie, prescription des actes thérapeutiques, ...) ;
- Permettre un accès distant à l'information et un partage rapide des informations, en particulier dans les situations d'urgence ;
- Favoriser la traçabilité des soins : médicaments administrés, actes réalisés ;
- Proposer un suivi administratif et médico-économique de l'activité : gestion des places disponibles, génération de tableau de bord d'activité, facturation.

Dans un deuxième temps et en partie grâce aux progrès techniques (améliorations des réseaux informatiques, augmentation des capacités mémoires, baisse du coût du matériel informatique ...), il est maintenant possible :

- D'interconnecter des applications distinctes pour échanger des informations ;
- De développer des bases de données communes à plusieurs applications.

Ainsi, de plus en plus de données cliniques sont enregistrées quotidiennement (3,4) par des applications informatiques. Celles-ci augmentent la qualité des informations par rapport aux dossiers dits « papier » (5–7). La volumétrie et la qualité des données, ainsi que les différents types d'informations disponibles au sein du SIH offrent des possibilités de réutilisations des données (8–13) telles que la recherche clinique (14,15), le recrutement de patients pour des essais cliniques (16,17), la détection précoce d'épidémies (18,19), le remboursement des actes (20) ou encore la validation d'hypothèse et l'évaluation de la qualité (21).

Cependant, l'utilisation secondaire de ces enregistrements fait face à plusieurs difficultés liées par exemple à une variabilité de la documentation (termes utilisés, qualité, paramétrage de l'application...) au cours du temps ou en fonction des utilisateurs pour une application donnée, à des structures de données hétérogènes d'une application à l'autre, ou encore à une volumétrie de données importante (22–25).

Afin de maîtriser ces difficultés, les techniques d'entrepôts de données (26,27) déjà mises en places dans les secteurs bancaires, de la grande distribution ou du marketing, ont été adoptées dans le secteur médical (28–34).

Un entrepôt de données est une structure commune à plusieurs systèmes informatiques et permet de colliger des grandes quantités de données enregistrées par des systèmes sources différents et stockées initialement dans des bases de données distinctes et hétérogènes. Ces données peuvent ensuite être redistribuées dans plusieurs outils de restitution répondant à des problématiques définies (26,27). Les données transitent des systèmes sources vers l'entrepôt de données via un processus nommé ETL (*Extract, Transform, Load*) représenté figure 1. L'objectif consiste à extraire des systèmes sources les informations pertinentes (étape 1), puis à les transformer afin qu'elles puissent être interrogées facilement et rapidement (étape 2), pour enfin les charger dans l'entrepôt de données (étape 3). A ce stade, l'entrepôt de données représente des informations détaillées de l'ensemble du domaine étudié. Ces informations peuvent ensuite être filtrées et agrégées (étape 4) vers les magasins de données qui représentent chacun un sous-ensemble métier : dans le cas d'un entrepôt de données cliniques, un premier magasin de données peut proposer des indicateurs médico-économiques liés à l'activité des blocs opératoires par exemple, et un second magasin de données peut être orienté vers la qualité des soins afin de proposer des indicateurs de qualité des soins.

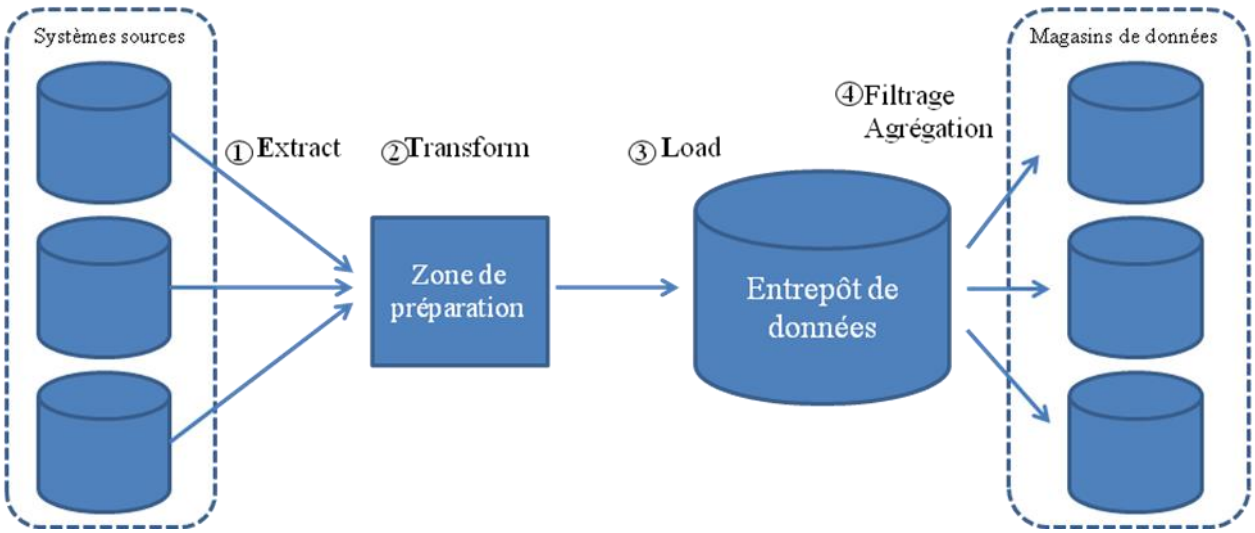


Figure 1 : Alimentation d'un entrepôt de données via le processus ETL

Les systèmes sources ont généralement des structures différentes. Il est donc difficile d'interroger conjointement les systèmes sources en raison de l'hétérogénéité de leurs structures. L'entrepôt de données, de par sa structure, permet de réaliser des analyses croisées des informations contenues à l'origine dans des systèmes sources distincts et réunies dans le modèle de données commun de l'entrepôt.

La figure 2 représente le modèle de données d'un entrepôt de données (schéma 3) développé à partir des modèles de données de deux systèmes sources différents (1er et 2ème schémas). Les systèmes sources A et B possèdent deux éléments en correspondance (qui décrivent le même élément du monde réel), respectivement les enregistrements contenus dans les tables "Données3" et "Données1". Cet élément commun fait office de liaison entre les deux modèles de données des systèmes sources, comme le montre le schéma 3.

Grâce à l'entrepôt de données, il est maintenant possible de faire des analyses croisées entre les données contenues dans les systèmes sources A et B, comme calculer la moyenne de la données 3 du système B par rapport aux enregistrements de la données 1 du système A.

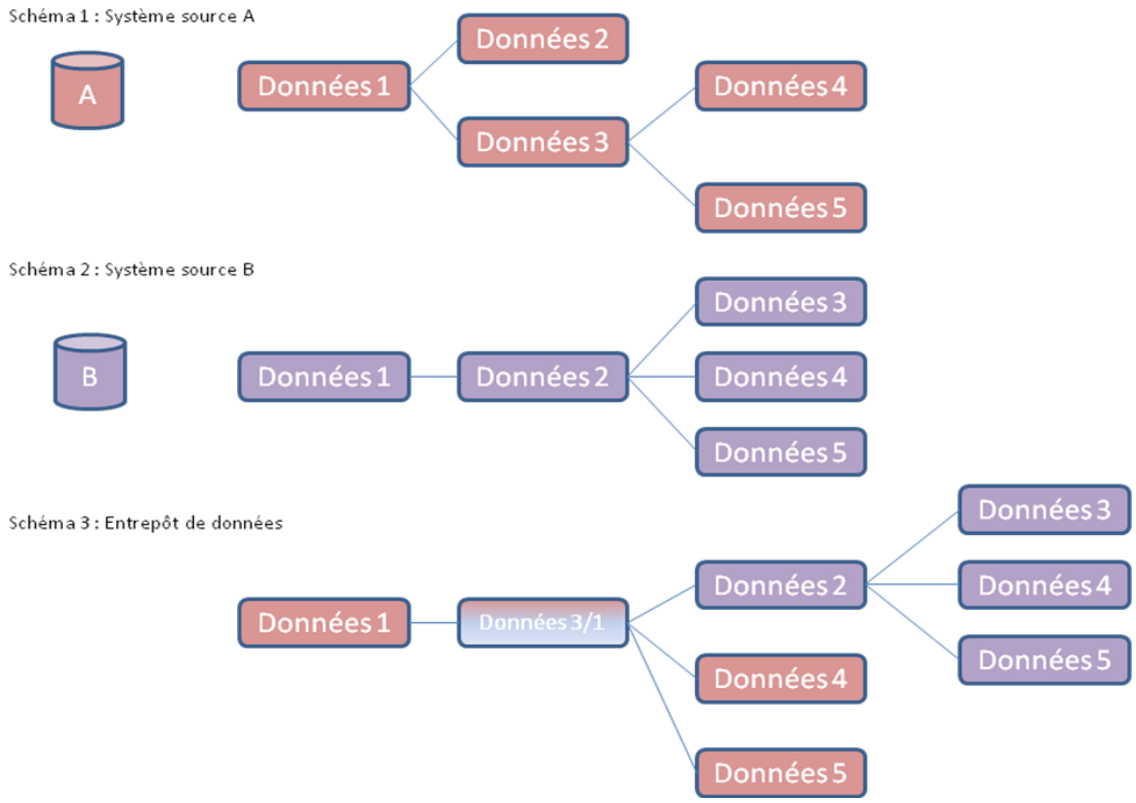


Figure 2 : Intégration des données au sein d'un entrepôt de données. Les données des systèmes sources A et B sont intégrées au sein d'un schéma de données commun dans l'entrepôt de données.

Grâce aux entrepôts de données, les informations disponibles dans différentes sources de données du SIH peuvent être croisées et analysées plus rapidement afin de répondre à de nouveaux objectifs. Il devient ainsi possible d'évaluer les pratiques et les innovations (28), de suivre l'incidence d'une maladie ou d'une infection (31), et de mettre en place des outils d'aide à la décision (tableau de bords, indicateurs) (35), de fouille de données (découverte de nouvelles connaissances) (36), de recherche clinique (37) et d'éducation (38).

1. Problématique clinique

Sous l'impulsion de la Société Française d'Anesthésie et de Réanimation (SFAR) et de l'Institut National de la Santé et de la Recherche Médicale (Inserm), le décret du 5 décembre 1994 (39) a donné un cadre réglementaire à la pratique de l'anesthésie. Il prévoit que les services d'anesthésie garantissent aux patients les moyens nécessaires (notamment en monitorage minimal) à la réalisation de l'anesthésie ainsi qu'une organisation permettant de faire face à tout moment à une complication liée à l'intervention ou à l'anesthésie effectuée. De fait, la sécurité péri opératoire a considérablement progressé en France, comme en témoigne la réduction des décès par accident d'anesthésie d'un facteur 10 entre la première enquête publiée par l'Inserm en 1983 et celle publiée par Lienhart et collaborateurs en 2006 (40). Ces décès sont désormais de 1 pour 250 000 anesthésies chez les patients de classe ASA 1. Ces progrès sont indissociables de l'évolution de la pratique anesthésiologique, et cela dans trois domaines : évolution des techniques (en particulier la pharmacologie), évolution des moyens de surveillance, et évolution de l'organisation des soins.

- L'**évolution pharmacologique** se traduit par des médicaments plus souples d'usage et l'utilisation de modèles pharmacologiques embarqués sur les stations d'anesthésie.
- L'**évolution des matériels et du monitorage**, plus récente, enrichissant le monitorage minimal obligatoire, s'est traduite par l'apparition de stations d'anesthésie associant des respirateurs munis d'alarmes et de sécurités, ainsi que de systèmes de monitorage complets permettant le suivi personnalisé du patient anesthésie en particulier pour les trois composantes de l'anesthésie générale : la curarisation et les composantes hypnotique et antinociceptive.
- L'**organisation des soins péri opératoires**, quant à elle, s'est réellement structurée à partir des années 1990 avec la publication de diverses recommandations de la SFAR et de textes réglementaires imposant les salles de surveillance post-interventionnelles (SSPI), la consultation pré-anesthésique, la maintenance du matériel.

Deux sources de progrès supplémentaires peuvent donc être aujourd'hui envisagées : La première correspond à l'amélioration de la prise en charge des patients présentant des comorbidités (notamment cardiorespiratoires) ou bénéficiant d'une chirurgie « lourde » (acte très invasif et/ou prolongé) ; La seconde source de progrès concerne le respect des normes.

- Si les taux de décès ont été réduits d'un facteur 10, et se situent autour de 1/10 000 pour des patients de classe ASA 2 à 3, pour les patients présentant des comorbidités importantes et bénéficiant de chirurgies lourdes, cette mortalité reste comprise entre 1 et 10% (40,41). De plus, si en lieu et place de la mortalité, l'analyse chiffrée se porte vers la morbidité il faut considérer que les marges de progrès sont très importantes. En effet, une méta-analyse (42) a montré en « chirurgie à très haut risque » que lorsque la mortalité diminue de 10 à 5 % par l'application de protocoles d'optimisation personnalisés péri-opératoires, les complications postopératoires et la morbidité résiduelle concernaient encore de 20 à 10% des patients opérés. Au-delà du risque chirurgical, la responsabilité de la prise en charge anesthésique est engagée, puisque l'application de protocoles d'optimisation permet de réduire de moitié cette morbidité dont l'impact peut être mesuré en termes de défaillances d'organes, d'admissions en soins intensifs ou réanimation, ou de prolongations du séjour postopératoire. L'amélioration qualitative de la prise en charge passerait donc par une prise en charge « individualisée » du patient (on parle aujourd'hui de médecine de précision) et une meilleure mise en œuvre des technologies disponibles sur les plateformes d'anesthésie.

- Alors que les normes se sont fortement développées au travers de conférences de consensus ou d'avis d'experts, leur application pratique reste limitée. Plusieurs études rétrospectives ont mis en évidence un lien statistique entre le respect d'indicateurs de qualité d'anesthésie et le devenir du patient : la survenue per-anesthésique d'un indice BiSpectral abaissé, d'une hypo ou d'une hypertension artérielle, ou encore d'une tachycardie sont statistiquement liés à une durée de séjour plus longue et à une surmortalité (43–48). Ces événements pourraient être automatiquement détectées par les logiciels de feuilles informatisées d'anesthésie (49), qui s'avèrent même plus fiables que les enregistrements manuels (50).

Depuis une dizaine d'année, la feuille informatisée d'anesthésie est utilisée au sein des services d'anesthésie du CHRU de Lille : cette suite logicielle permet d'enregistrer les données du patient relatives à l'anesthésie (comprenant la consultation pré-anesthésique, la prise en charge du patient du bloc opératoire jusqu'à la sortie de la salle de soins post-interventionnels). D'autres applications recueillant les informations des séjours hospitaliers (durée de séjours, type d'unité de soins, diagnostics associés, …) permettent de suivre le parcours du patient au sein de l'établissement et de facturer les soins dispensés.

L'utilisation de cette base de données d'anesthésie qui s'enrichit des données de plus de 55 000 patients par an doit permettre de recenser l'impact de différentes campagnes d'amélioration de la qualité de prise en charge sur l'évolution de la mortalité et de la morbidité péri-opératoire en fonction des sites interventionnels, des risques propres au patient (classe ASA, comorbidités …) et des techniques d'anesthésie mises en œuvre. L'interconnexion entre cette base de données et la base de données PMSI (Programme de Médicalisation des Systèmes d'Information) du SIH peut également permettre d'apprécier l'impact économique des améliorations mises en œuvre : durée d'hospitalisation, typologie d'hospitalisation, complications, …

2. Problématiques et Objectifs

La feuille informatisée d'anesthésie DIANE (Bow Medical, Amiens, France) et les logiciels administratifs et de PMSI utilisés au CHRU de Lille fournissent potentiellement les données nécessaires pour évaluer la qualité de prise en charge anesthésique grâce à différents indicateurs de qualité d'anesthésie tel que l'ont fait Sessler *et al.* (46) ou Kertai *et al.* (48). Cependant, plusieurs interrogations doivent être levées :

- Les informations enregistrées par ces applications sont-elles exploitables ? En effet, les données sont renseignées par les utilisateurs ou enregistrées automatiquement par les appareils pour une utilisation précise déterminée dans le cahier des charges de l'éditeur du logiciel, c'est-à-dire essentiellement l'enregistrement continu de l'ensemble des éléments afférents à une anesthésie particulière chez un patient clairement identifié. L'utilisation de ces données dans un but purement statistique ne fait a priori pas partie du cahier des charges de ces logiciels.
- La qualité des données enregistrées est-elle affectée par le lieu où elles sont enregistrées ? Les différentes options de personnalisation du logiciel afin de s'adapter aux différents « métiers » de l'anesthésie, secteur par secteur, constituent-elles un frein à l'utilisation « statistique » de ces données à l'échelle d'un hôpital entier ?
- Est-il possible de relier les données enregistrées par deux applications opérationnelles différentes ?
- Les informations enregistrées par les différentes applications opérationnelles sont-elles cliniquement pertinentes et reflètent-elles la réalité ?

Ce travail de thèse comportait donc trois objectifs majeurs :

- Evaluer la pertinence des informations disponibles et la qualité des données enregistrées dans plusieurs applications déployées au CHRU de Lille en vue de leur utilisation secondaire ;
- Développer un entrepôt de données permettant de réaliser des analyses statistiques rétrospectives ;
- Tester l'exploitation de l'entrepôt de données et proposer de mesurer les indicateurs de qualité d'anesthésie en recherchant leur lien éventuel avec des critères objectifs tels que la morbidité et la mortalité postopératoire.

Ce travail de thèse intègre le projet DIAGnosTIC qui vise à utiliser les données enregistrées par la feuille informatisée d'anesthésie DIANE (DIA), pour produire de nouvelles connaissances (Gnos) grâce aux Technologies de l'Information et de la Communication (TIC). Il est inscrit dans le projet d'établissement du CHRU de Lille, dans le volet Qualité.

4. Plan

Ce document est organisé en trois parties.

Une première partie correspond à la phase d'analyse des bases de données disponibles :

- Le chapitre 1 liste les différents systèmes logiciels du SIH qui contiennent des données pertinentes pour notre problématique clinique ;
- Le chapitre 2 propose une évaluation de la qualité des données de ces systèmes en détaillant les différents problèmes de qualité de données ainsi que les méthodes d'évaluation définies dans la littérature ou développées dans le cadre de ce projet.

La seconde partie présente le développement de l'entrepôt de données :

- Le chapitre 3 détaille la chaîne d'alimentation de l'entrepôt de données à partir des bases de données sources ;
- Le chapitre 4 présente l'agrégation des données au sein des magasins de données afin de disposer d'indicateurs.

Enfin, la troisième partie deux études de cas illustratives de l'application de l'analyse des données de l'entrepôt à des questions cliniques :

- Le chapitre 5 présente une étude sur l'influence de l'induction sur la survenue d'hypotension ;
- Le chapitre 6 présente une étude sur le volume courant par unité de poids idéal au cours de l'anesthésie générale et de son impact sur la morbidité postopératoire.

Chapitre 1 : Description des systèmes sources

Chapitre 1 : Description des systèmes sources

1. Introduction

Le SIH désigne l'ensemble des informations nécessaires au fonctionnement d'un établissement de santé, ainsi que les différentes applications et bases de données permettant d'enregistrer, d'interroger ou d'exporter les informations médicales ou administratives. Si aujourd'hui les SIH tendent à interconnecter les différentes applications en usage sur leur parc informatique, celles-ci demeurent souvent indépendantes les unes des autres en raison de la variété des éditeurs, des usages et des techniques qu'ils mettent en œuvre.

Le CHRU de Lille met en œuvre plusieurs progiciels qui suivent le parcours administratif de chaque patient dans les unités de soins, tracent les administrations de produits sanguins labiles, permettent aux professionnels de visualiser les résultats d'examens complémentaires de radiologie ou de biologie, recensent les informations nécessaires à la pratique de l'anesthésie, etc. De plus, plusieurs entrepôts de données dans des domaines distincts sont également en service ou en cours de développement.

Dans le cadre de ce projet de thèse, nous nous sommes intéressés en particulier à trois applications déployées au CHRU de Lille, ainsi qu'à un entrepôt de données en cours de développement :

- La feuille informatisée d'anesthésie DIANE (Dossier Informatisé d'ANEsthésie) ;
- Le logiciel administratif GAM (Gestion Administrative des Malades) ;
- Le logiciel de PMSI CORA (Programme de Médicalisation des Systèmes d'Information) ;
- L'Infocentre d'Anesthésie.

La feuille informatisée d'anesthésie DIANE permet de regrouper toutes les informations concernant la procédure d'anesthésie depuis la consultation pré-anesthésique jusqu'à la sortie du patient de la salle de réveil après l'anesthésie. DIANE enregistre les paramètres physiologiques mesurés lors de l'intervention ainsi que les administrations de médicaments et les étapes de l'intervention. Les logiciels GAM et CORA gèrent les séjours administratifs des patients et permettent de facturer les séjours. L'Infocentre d'Anesthésie est un entrepôt de données développé au CHRU de Lille par la Délégation au Système d'Information (DSI) dans le but de mettre à disposition des praticiens des tableaux de bords et des indicateurs de suivi d'activité d'anesthésie, principalement d'ordre médico-économique.

DIANE, CORA et GAM contiennent des informations relatives à la prise en charge anesthésique du patient ainsi qu'au déroulement de son séjour hospitalier. L'exploitation de ces informations a permis de détecter des événements indésirables survenant au cours d'une anesthésie (hypotension, bradycardie, ...) et de suivre leurs conséquences sur la suite du séjour hospitalier (durée de séjour anormalement longue, décès, ...).

Les logiciels de traçabilité des administrations de produits sanguins labiles, d'analyses biologiques ou les autres entrepôts de données n'ont pas été étudiés lors de ce projet, mais des travaux collaboratifs pourront être envisagés à l'avenir.

Après avoir décrit brièvement le rôle de chacune des applications sélectionnées, nous nous sommes appliqués à détailler les données enregistrées, puis nous avons précisé comment les données de chaque

application s'intégraient dans notre projet. Enfin nous avons proposé un modèle de données hypothétique qui pourra être utilisé dans la suite du travail.

2. Système 1 : La feuille informatisée d'anesthésie (DIANE)

2.1 Description

La Société Française d'Anesthésie et de Réanimation recommande la tenue d'une feuille d'anesthésie (51) dans le cadre du dossier patient. En effet, ce document pourra être utilisé dans le cadre d'une procédure juridique et permettra d'attester de la bonne conduite de la procédure d'anesthésie. Au CHRU de Lille, la feuille d'anesthésie est informatisée via la suite logicielle DIANE (Dossier Informatisé d'ANEsthésie) depuis 2004. Elle est composée de plusieurs modules permettant de gérer :

- La consultation d'anesthésie ;
- La prise en charge per-opératoire ;
- La fusion a posteriori de dossiers patients ;
- La production de statistiques descriptives.

Développée par BOW Médical (Amiens, France) (52), DIANE est utilisée dans 160 établissements publics et privés. L'application est interfacée avec les moniteurs et respirateurs utilisés au bloc opératoire, ce qui permet de recueillir automatiquement les mesures collectées par ces appareils. Les informations enregistrées par DIANE sur les différents postes de travail d'un même établissement sont stockées dans une base de données commune. Au CHRU de Lille, cette application est déployée sur les 86 sites opératoires et enregistre les informations des 55 000 interventions annuelles. Deux bases de données miroirs liées à l'application sont hébergées sur deux serveurs différents et permettent d'assurer une continuité du service en cas de défaillance d'un des serveurs.

2.2 Données disponibles

Deux modules DIANE permettent de recueillir les données relatives à l'anesthésie : la feuille de consultation pré-anesthésique et la feuille d'anesthésie.

Feuille de consultation pré-anesthésique

Lors de la consultation pré-anesthésique, les caractéristiques et les antécédents du patient sont renseignés (Annexe 1). Les allergies du patient, les traitements en cours et les antécédents sont complétés à partir de menus préconfigurés ou en texte libre. Les résultats d'examens ou les paramètres mesurés le jour de la consultation sont également renseignés. Des consignes de prise en charge postopératoires peuvent être précisées. Un résumé de la consultation peut être généré et inclus dans le dossier patient.

Feuille d'anesthésie

Le module per-opératoire permet d'intégrer différents types d'informations. La figure 3 présente l'interface DIANE per-opératoire et les différents zones d'affichage ou de saisie d'informations : (1) informations du patient, (2) médicaments, (3) événements, (4) paramètres monitorés.

Figure 3 : Interface DIANE du module per-opératoire. Le cadre 1 présente les caractéristiques du patient le jour de l'intervention, le cadre 2 permet à l'utilisateur de renseigner les différents produits administrés lors de l'intervention, le cadre 3 concerne les étapes de l'intervention et le cadre 4 affiche les paramètres mesurés par le respirateur et le moniteur d'anesthésie ou tout autre appareil connecté à DIANE.

Informations du patient (cadre 1) : identité, âge, poids, taille, classe ASA, service, salle, type de chirurgie, type d'anesthésie, allergies.

Médicaments (cadre 2) : Pour chaque médicament administré au cours de l'intervention, plusieurs types d'informations peuvent être apportés : nom du médicament, heure d'administration, voie d'injection, posologie, concentration, heure de fin et quantité totale administrée dans le cas d'une perfusion.

Evénements (cadre 3) : Les événements enregistrés sont caractérisés par deux éléments : un nom et une heure d'occurrence. Les événements concernent plusieurs types d'informations : les étapes de l'intervention (début d'anesthésie, incision, fin de chirurgie, fin d'anesthésie, ...), le matériel utilisé (lames de laryngoscope, masques, ...), les techniques employées et les mouvements des personnels (entrée et sortie du bloc opératoire du personnel soignant).

Mesures (cadre 4) : Les paramètres recueillis automatiquement par le respirateur et le moniteur d'anesthésie, ou tout autre appareil disposant d'un driver communiquant avec DIANE (ex : BIS, ANI) sont enregistrés et affichés après paramétrage de l'interface.

Dans la partie haute du cadre 4 les courbes de saturation en oxygène (SpO2), fréquence cardiaque mesurée par l'ECG (FcECG) et fréquence cardiaque mesurées par le capteur de saturation en oxygène

(FcSpO2) sont affichées. Les mesures de pression artérielle diastolique, moyenne et systolique sont affichées dans la partie intermédiaire du cadre. Enfin, la partie inférieure du cadre 4 présente des paramètres mesurés par le respirateur d'anesthésie : la concentration en protoxyde d'azote (N2Oi et N2Oe), la concentration en halogéné (Sevi et Seve) et la concentration en dioxyde de carbone (CO2i et CO2e).

Tableau 1: Paramètres monitorés

Paramètre	Appareil de mesure
Concentration en gaz halogéné	Respirateur d'anesthésie
Fréquence respiratoire	Respirateur d'anesthésie
Fréquence cardiaque à partir de l'ECG	Moniteur d'anesthésie
Saturation en oxygène	Moniteur d'anesthésie
Pression artérielle non invasive	Moniteur d'anesthésie
Profondeur d'analgésie (ANI)	Moniteur ANI
Profondeur d'anesthésie (BIS)	Moniteur BIS

Les mesures sont enregistrées à intervalles irréguliers, dépendant de chaque paramètre : la pression artérielle non invasive est mesurée toutes les 2 à 5 minutes alors que les autres paramètres sont recueillis toutes les 30 secondes environ. Au total, plus de 200 paramètres différents peuvent être enregistrés pour chaque anesthésie. Le tableau 1 présente quelques paramètres disponibles dans DIANE ainsi que l'appareil de mesure associé.

Ces informations peuvent être renseignées de plusieurs manières : (1) éléments prédéfinis, menus et listes préconfigurés, (2) enregistrement automatique par des appareils de mesure, (3) commentaires libres.

1) Eléments prédéfinis, menus et listes préconfigurés

Le logiciel DIANE propose plusieurs listes d'éléments et de menus préconfigurés correspondant au processus d'anesthésie et de chirurgie afin de documenter l'intervention. Les listes sont organisées pour faciliter l'utilisation du logiciel et pour documenter de manière exhaustive le déroulement de la procédure d'anesthésie (Annexe 2). Les menus préconfigurés proposent pour chaque élément une liste d'éléments à compléter afin que l'exhaustivité de l'information soit optimale. Ainsi la figure 4 présente les informations à renseigner pour que l'administration de sufentanil soit complète. Plusieurs champs sont préconfigurés avec des valeurs ou données adaptées à l'administration de sufentanil : voie d'administration, posologie, concentration et unités.

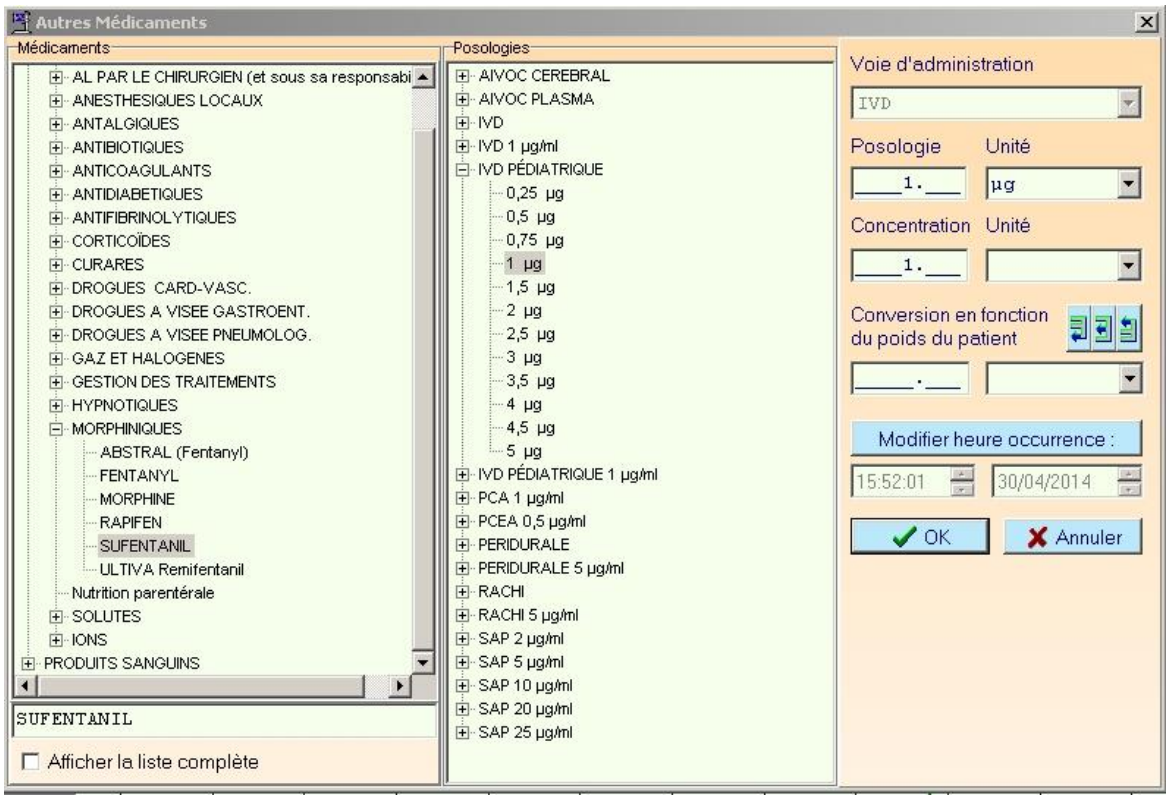


Figure 4: Menu préconfiguré pour le sufentanil

Des éléments prédéfinis permettent également de documenter rapidement une étape de l'intervention en proposant un ensemble d'événements associés à cette étape. La figure 5 présente la fenêtre associée à la procédure de rachianesthésie. Le matériel habituellement utilisé (aiguille de rachianesthésie), les médicaments employés pour cette procédure sont ainsi proposés pour permettre à l'utilisateur de gagner du temps lors de la saisie.

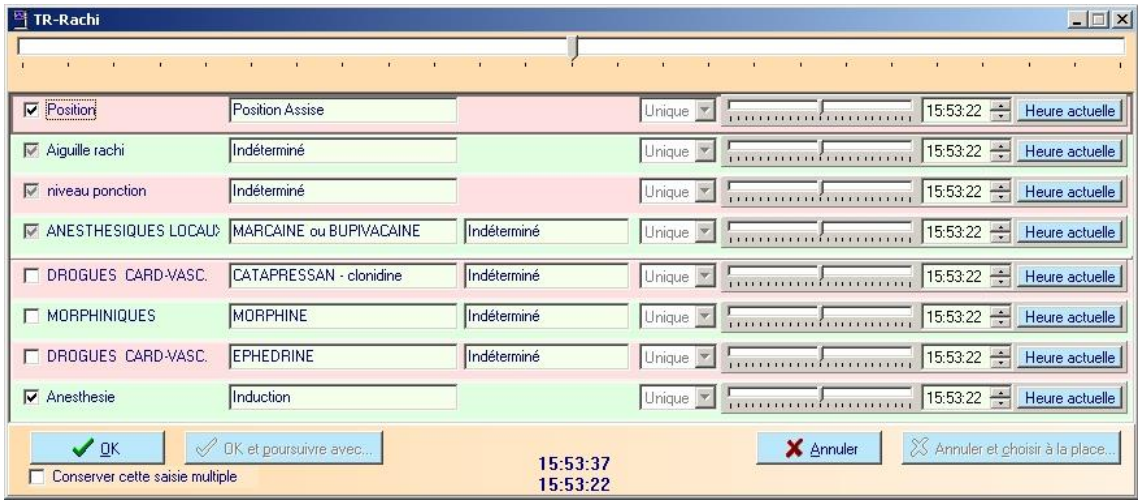


Figure 5: Eléments prédéfinis associés à la procédure de rachianesthésie

2) Enregistrement automatique

Les paramètres sont enregistrés automatiquement par les appareils de mesures. La qualité de l'information dépend de l'appareil, de son paramétrage et des connexions avec le poste de travail.

3) Commentaires libres

L'utilisateur peut renseigner manuellement un événement ou un médicament (figure 6). Dans ce cas, la qualité de l'information est dépendante de l'utilisateur et peut comporter des fautes d'orthographe, des abréviations. Dans le cas de l'administration d'un médicament, rien n'oblige l'utilisateur à fournir toutes les informations nécessaires (posologie, unité ...) et celui-ci peut préciser uniquement le nom du médicament.

Figure 6 : Saisie manuelle d'une administration de sufentanil

2.3 Synthèse

Le logiciel DIANE est le point de départ de notre projet puisqu'il enregistre les données relatives à chaque anesthésie réalisée au CHRU de Lille. Il contient des données caractérisant le patient au moment de l'intervention (âge, classe ASA, poids, taille, IMC). Les mesures enregistrées permettent de qualifier l'état du patient et de ses différents systèmes (cardio-pulmonaire, hémodynamique...) au cours de l'intervention. Elles offrent la possibilité de détecter les événements indésirables tels que l'hypotension, l'hypertension, la tachycardie, la bradycardie, etc. Les événements apportent des informations sur la prise en charge de l'anesthésie (administrations de médicaments, procédures [induction, intubation, ...) et caractérisent l'environnement du patient. Ces événements ont une importance par rapport à l'état du patient puisqu'ils peuvent justifier l'évolution de l'état de ce dernier (administration de médicaments déclenchant une hypotension, gestes chirurgicaux douloureux provoquant une hypertension, etc.). Enfin, les informations de la consultation préopératoire permettent de filtrer les patients en fonction de leurs antécédents et de caractériser leur environnement avant l'intervention.

3. Système 2 : Le dossier administratif patient (GAM)

3.1 Description

Le logiciel GAM (Gestion Administrative des Malades), édité par McKesson, est utilisé au CHRU de Lille depuis 1999. Il gère la partie administrative des dossiers patients (entrée et sortie des unités de soins, coûts des soins) quelque soit le type de prise en charge (soins externes, Médecine-Chirurgie-Obstétrique, ...) et permet de calculer la prise en charge des coûts du séjour hospitalier par la sécurité sociale, la mutuelle de santé et le patient pour les soins externes. Le calcul des coûts des séjours relevant de la Médecine-Chirurgie-Obstétrique se fait en lien avec le logiciel CORA.

3.2 Données disponibles

Patient : L'identité de l'ensemble des patients entrant au CHRU de Lille sont disponibles dans GAM. Les informations des patients enregistrées dans GAM sont très complètes puisqu'en plus des informations d'identités habituelles, les adresses et identités des médecins traitants sont disponibles.

Séjour : Les séjours enregistrés dans GAM concernent tous les passages des patients dans la structure de soins : consultation, séjour externe, MCO...

3.3 Synthèse

Les données de séjours enregistrées par GAM sont plus exhaustives que celle de CORA. En revanche, elles sont moins détaillées et ne renseignent que sur la durée de séjour dans la structure de soins. Néanmoins, cette source de données peut être utilisée au cas où le lien entre DIANE et CORA ne pourrait être établi, mais également dans l'optique de futurs travaux visant à intégrer les données issues d'autres systèmes sources.

4. **Système 3 : Le logiciel PMSI (CORA)**

4.1 Description

Le logiciel CORA est une suite logicielle éditée par Maincare (France), utilisée dans le cadre du Programme de Médicalisation des Systèmes d'Information (53). Ce programme vise à réduire les inégalités de ressources entre les établissements de santé et permet de suivre l'activité de chaque établissement. Il ne couvre que les séjours hospitaliers de type Médecine-Chirurgie-Obstétrique (MCO).

4.2 Données disponibles

Les informations enregistrées grâce au logiciel CORA permettent de retracer le parcours du patient dans les différentes unités, ainsi que les actes qui réalisés et les diagnostics associés à son séjour hospitalier. Ce parcours est détaillé sur plusieurs niveaux :

Le séjour hospitalier : définit le séjour hospitalier et est caractérisé par une date d'entrée et de sortie de l'établissement de soins.

Le RUM (Résumé d'Unité Médicale) : caractérise un passage dans une unité de soins. Pour chaque RUM est précisé le mode d'entrée et de sortie du patient dans l'unité de soins, ce qui permet de détecter deux « événements » importants : l'entrée du patient par les urgences et le mode de sortie correspondant au décès. D'autres parts, l'unité de soins hébergeant chaque patient comporte une autorisation précisant si cette unité est une unité de soins lourds (réanimation ou soins intensifs) ou non. Le RUM répertorie l'ensemble des actes médicaux administrés au patient ainsi que les diagnostics justifiant leur réalisation. Les actes médicaux sont codés en utilisant la Classification Commune des Actes Médicaux (CCAM) (54). Cette nomenclature regroupe les actes diagnostics ou thérapeutiques dans 19 chapitres correspondant à des structures anatomiques ou fonctionnelles (figure 7). Chaque acte est identifié par un code composé de chiffres et de lettres. Exemple : La figure 8 détaille les actes du chapitre 14 correspond aux actes en lien avec l'appareil ostéoarticulaire et musculaire du membre inférieur et l'acte NBEP001 identifie une "Réduction orthopédique progressive de fracture du fémur, par traction continue collée".

Figure 7 : Classification commune des actes médicaux

Les diagnostics sont référencés par la Classification Internationale des Maladies, 10ème version (CIM10) (55). Ils sont qualifiés de diagnostics principaux, associés ou secondaires. La CIM10 est organisée en 21 chapitres, correspondant à un appareil fonctionnel. Les chapitres sont ensuite divisés en sous-chapitres, catégories et sous-catégories. Le diagnostic est identifié par un code en 7 caractères. Ainsi le chapitre XI correspond aux maladies de l'appareil circulatoire et le diagnostic I10 identifie une "Hypertension essentielle (primitive)" (figure 8).

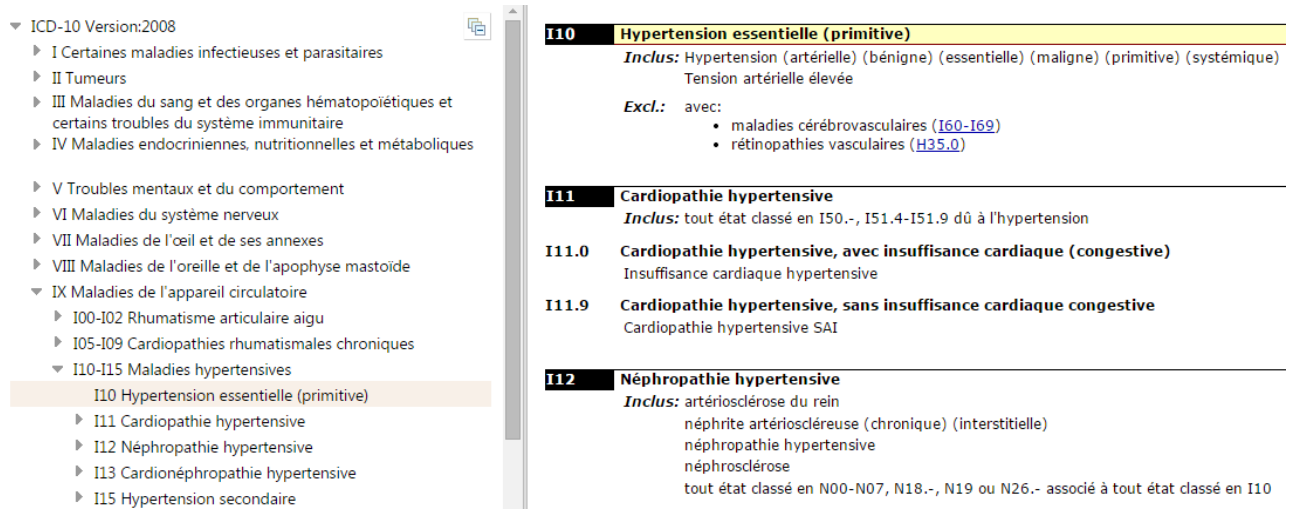


Figure 8 : Classification Internationale des Maladies

Le RSS (Résumé de Sortie Standardisé) est défini à partir des RUM du séjour hospitalier en fonction des diagnostics et des actes médicaux réalisés lors de la prise en charge du patient. Ces informations sont traitées par un algorithme de groupage qui associe chaque séjour à un groupe homogène de malades (GHM), caractérisé par des coûts similaires de prise en charge. Le RSS peut ensuite être utilisé lors d'analyses médicaux-économiques.

4.3 Synthèse

Les données fournies par CORA sont consolidées car elles sont utilisées pour la facturation des séjours. Les diagnostics et les actes médicaux sont codés en utilisant deux terminologies différentes. Ils peuvent être utilisés pour segmenter les patients par types d'intervention ou pour détecter certaines complications post-opératoires (infarctus, embolie pulmonaire, infections, ...). Cependant, la saisie des diagnostics n'est pas toujours exhaustive, les comorbidités des patients pouvant ainsi être sous-estimées. Deux informations sont robustes : la durée de séjour et le mode d'entrée et de sortie des unités de soins, permettant ainsi de caractériser chaque séjour en fonction de sa durée, du décès éventuel du patient et de son passage éventuel dans une unité de soins lourds.

5. Système 4 : L'Infocentre d'anesthésie

5.1 Description

Au sein du CHRU de Lille, plusieurs entrepôts de données ont été réalisés par l'équipe « Pilotage » de la Délégation du Système d'Information (Infoservice des Dépenses, Infoservice d'Imagerie et Explorations Fonctionnelles, Infoservice de Biologie, Infoservice Patient, Infoservice Bloc opératoire, etc...). La construction d'un entrepôt de données d'anesthésie a d'ors et déjà été initiée par la Délégation du Système d'Information (DSI) sur la base de données DIANE.

L'objectif initial de l'Infocentre d'anesthésie était avant tout médico-économique et consistait à mettre à disposition des équipes de soins des tableaux d'activités : nombres d'interventions ou durées d'interventions par semaine, par mois, par années, par service, par chirurgien, par anesthésiste, par tranche de poids, ou d'âges des patients, par classe ASA, par types d'anesthésie, par types d'interventions.

Les restitutions fournissent également les durées des étapes des interventions définies en fonction des événements saisis dans Diane. Ces durées sont détaillées en fonction de l'année et du trimestre, des types d'anesthésie et types d'intervention primaires, des personnels anesthésistes et chirurgiens, des services du CHRU réalisant ou demandant l'intervention.

L'entrepôt de données est stocké sur une base de données Oracle (version 10g puis 11g) et alimenté par un processus ETL (Extract, Transform, Load) développé avec le logiciel Infofusion Integration Center (anciennement appelé Genio ou O.T.I.C. - Editeur Open Text). Le processus ETL propose une opération majeure de nettoyage des données : le dédoublonnage des interventions et des patients. En effet, après un arrêt intempestif du logiciel, lors de son utilisation au bloc opératoire, les utilisateurs sont souvent amenés à créer une nouvelle intervention plutôt que de reprendre celle abandonnée lors de l'arrêt du logiciel. De même, lors de la consultation d'anesthésie, l'utilisateur ne vérifie pas toujours si le patient existe déjà dans la base de données (si il a déjà été opéré au CHRU) et peut être amené à recréer à tort un « nouveau patient ». Il en résulte des doublons de patients et d'interventions.

5.2 Données disponibles

Au début de ce travail, l'Infocentre d'Anesthésie enregistre les informations liées aux patients et aux interventions :

<u>Données relatives aux interventions</u>

- Dates, heures, durées des principales étapes des interventions (dates prévues, prise en charge du patient, installation, induction, chirurgie, fin de l'anesthésie, passage en salle de soin post-intervention) ;
- Services du CHRU en relation avec l'intervention ;
- Informations relatives au patient au moment de l'intervention : âge, taille, poids, classe ASA ;
- Evénements de l'intervention : administration de médicaments, événements opératoires. Les événements restitués sont les événements directement saisis dans Diane sans retraitement de l'information. Ils sont par ailleurs organisés selon la hiérarchie paramétrée dans Diane ;

- Personnels, chirurgiens et anesthésistes réalisant l'intervention, dates de début et fin de leur présence lors de l'intervention ;
- Types d'intervention et d'anesthésie réalisés.

Données relatives aux patients pris en charge

- Identité, date de naissance.

Indicateurs d'activité

- Nombre d'anesthésies réalisées, nombre de patients pris en charge, nombres de consultations pré et post-opératoires, âges moyens des patients lors des interventions, durées des interventions.

L'alimentation de ce bloc fonctionnel initial est déclenchée chaque lundi matin et couvre généralement deux ans glissants. Elle met à jour intégralement les informations des interventions réalisées au cours de ces deux années dans le cas où des données auraient été mises à jour a posteriori et dure environ 4 heures.

5.3 Synthèse

L'Infocentre d'Anesthésie présente une structure adaptée pour ce travail puisque le système est un entrepôt de données. Les informations disponibles sont davantage consolidées que celles qui sont présentes dans DIANE sans recoupement avec d'autres données de DIANE. Cette structure pourrait être complétée en implémentant d'autres blocs fonctionnels avec les données cliniques de DIANE, mais également avec d'autres sources, comme GAM ou CORA.

6. Discussion

Cette première étape du projet avait pour but de déterminer quelles applications proposaient des informations potentiellement pertinentes pour le développement d'un entrepôt de données et nécessitait une bonne compréhension des flux d'informations au sein du SIH (56).

Si ce travail a permis de sélectionner quatre systèmes sources, plusieurs interrogations doivent encore être levées, en particulier pour les trois systèmes opérationnels DIANE, GAM et CORA (22). Tout d'abord, les bases de données associées à ces trois applications doivent être accessibles, en effet, les bases de données de systèmes opérationnels ne sont pas développées pour être interrogées par une application tierce. La variabilité associée aux habitudes d'utilisation et aux paramétrages de chaque application constitue également un frein à l'utilisation automatique de ces données.

L'évaluation de la qualité des informations enregistrées dans les systèmes sources sélectionnés dans ce chapitre constitue le thème du prochain chapitre.

7. Conclusion

Ce chapitre a permis de faire l'inventaire des systèmes sources en lien avec notre projet. Les logiciels sélectionnés et les données dont ils disposent ont été présentés. Nous avons choisis d'intégrer les données enregistrées par DIANE (mesures, événements, médicaments et données de la consultation), CORA et GAM (patients et séjours). Les informations de CORA et GAM spécifiques aux recettes des séjours n'ont pas été

utilisées puisque le but de ce travail était avant tout de pouvoir répondre à des problématiques de recherche clinique.

Pour répondre à notre problématique sur l'incidence des événements indésirables, DIANE permet de fournir différentes variables sur l'état du patient, son évolution ou la présence d'événements indésirables (hypotension, tachycardie, ...).

Les logiciels CORA et GAM présentent des informations sur le devenir à court terme des patients, c'est à dire après l'intervention, dans le cadre de leur séjour hospitalier. Les deux variables explicatives les plus importantes disponibles grâce à ces logiciels sont la durée de séjour et le décès (disponible seulement dans CORA). D'autres variables secondaires comme des diagnostics liés à des complications postopératoires ou des admissions non programmées dans des unités de soins lourds peuvent également être utilisées (disponible seulement dans CORA).

L'Infocentre d'Anesthésie a été le point de départ du travail de développement puisqu'il collige les informations des patients ayant bénéficié d'une procédure d'anesthésie. Ce sont les patients pour lesquels le lien avec les séjours hospitaliers disponibles dans GAM et CORA devra être réalisé. Les informations relatives aux mesures et événements enregistrées dans DIANE, ainsi que les données des séjours hospitaliers de CORA et GAM seront greffées au schéma existant de l'Infocentre d'Anesthésie.

Le tableau 2 présente un récapitulatif des données sélectionnées dans chaque système source, ainsi que l'intérêt de ces données pour notre projet. A l'avenir, les données d'autres systèmes pourront être intégrées ou des passerelles pourront être développées entre les différents entrepôts de données dont dispose le CHRU de Lille.

La qualité des données est étudiée dans le prochain chapitre. Ce chapitre démontre dans quelle mesure les données sélectionnées sont réellement utilisables, et quels sont les modules de nettoyage à développer pour le processus ETL d'alimentation de l'entrepôt de données.

Tableau 2: Récapitulatif des données sélectionnées

Système source	Information	Intérêt
DIANE	Mesures	Etat d'un patient et de ses différents systèmes (hémodynamique, respiration, système nerveux central ...).
DIANE	Administration de médicaments	Environnement du patient
DIANE	Etapes de l'intervention	Environnement du patient
DIANE	Gestes chirurgicaux	Segmenter les patients
CORA	Décès	Complications post-opératoires
CORA et GAM	Durée de séjour	Complications post-opératoires
CORA	Actes thérapeutiques et diagnostics	Sélection des patients, environnement du patient.
CORA	Diagnostics	Complications post-opératoires

Chapitre 2 : Evaluation de la qualité des données

Chapitre 2 : Evaluation de la qualité des données

1. Introduction

Dans le premier chapitre, nous avons identifié quatre systèmes enregistrant des informations pertinentes pour notre problématique. Trois systèmes sont des *logiciels opérationnels* (la feuille informatisée d'anesthésie DIANE, le logiciel de gestion administrative des malades GAM et le logiciel de facturation CORA). Le quatrième est un *entrepôt de données d'anesthésie* développé par la DSI du CHRU de Lille. Nous souhaitons maintenant montrer que les données enregistrées et/ou stockées par ces systèmes sont (i) réellement disponibles, (ii) correctes, (iii) et dans quelle mesure elles peuvent être intégrées au sein d'un entrepôt de données. La qualité des données peut être définie de nombreuses manières. Dans notre étude, nous avons abordé une évaluation des dimensions extrinsèques, c'est à dire dépendantes de l'usage envisagé que nous devons faire de ces données.

(i) Nous abordons ici la notion de *data completeness* (57–61). Dans le chapitre précédent, nous avons relevé les différents types d'informations théoriquement enregistrées par les systèmes abordés ici. Nous tentons maintenant d'évaluer dans quelle mesure ces informations sont *disponibles*.

(ii) La présence d'une information dans un enregistrement n'implique pas qu'elle soit *correcte*. Ainsi, les artefacts dus aux gestes chirurgicaux ou à une mauvaise mise en place d'un capteur provoquent fréquemment des erreurs de mesures, qui sont enregistrées sans discernement par la feuille informatisée d'anesthésie DIANE. De plus, les informations saisies manuellement par le professionnel de santé comportent fréquemment des fautes de frappes, des abréviations, et peuvent de plus être incomplètes. En effet, les différentes catégories d'utilisateurs ont des formations différentes, tant de part des cursus médico-universitaires différents que par des différences de niveau dans leurs connaissances des logiciels qu'ils ont à utiliser dans leur pratique professionnelle. De plus, dans le cas de l'utilisation de DIANE, le temps dont ils disposent pour saisir ou compléter différentes informations sur le déroulement de l'anesthésie ou les administrations thérapeutiques réalisées dépend de chaque travail au bloc opératoire. C'est pourquoi la qualité des données enregistrées dans DIANE doit être analysée selon plusieurs dimensions complémentaires : *correctness*, *concordance* ou *plausibility* (61).

(iii) L'objectif du projet DIAGnosTIC est d'intégrer les informations stockées par les trois systèmes opérationnels distincts que sont DIANE, GAM et CORA. Ces trois logiciels émanent de trois éditeurs de logiciels différents, ont été élaborés selon des architectures différentes et sont utilisés par opérateurs différents : les modèles de données mis en œuvre et les informations stockées par ces systèmes sont donc hétérogènes. Le travail d'évaluation de la qualité des données proposé dans ce chapitre doit donc garantir que malgré les différences entre les systèmes qui les ont produites, les données pourront être intégrées dans une structure unique dont l'exploitation devrait conduire aux indicateurs de qualité de prise en charge recherchés.

Différents travaux proposent des taxonomies permettant de répondre aux questions de qualité de données (62–65). Kim *et al.* proposent de classer les problèmes de qualité de données selon qu'elle sont *manquantes* (*missing data*), *présentes* mais *erronées* (*not missing but wrong data*), ou *présentes* mais *inutilisables* (*not missing and not wrong but unusable*) (62). Cette classification ne tient pas compte des problèmes liés aux schémas de données et le premier niveau correspond plutôt à l'incidence des problèmes de qualité sur l'utilisation. Une autre taxonomie (63) se limite aux problèmes de qualité liés à une seule source de données, selon que la *syntaxe* est *incorrecte* (*syntactical anomalies*), comme par exemple le format des champs, ou que la *représentation des données* (*semantic anomalies*), comme par exemple la présence de doublons, ou encore qu'il y ait des *valeurs manquantes* (*coverage anomalies*). Cette terminologie ne tient pas

compte des problèmes de qualité liés à l'intégration de données provenant de systèmes aux architectures hétérogènes. Rahm et Do distinguent les problèmes de qualité liés à une ou plusieurs sources de données hétérogènes, puis ceux relatifs au schéma de données ou aux enregistrements (64). Enfin, Oliveira *et al.* présentent une taxonomie des problèmes de qualité de données basée sur leur granularité, des problèmes de qualité lié à un seul champ d'un seul enregistrement (par exemple une valeur manquante) jusqu'aux problèmes de qualité liés à plusieurs systèmes sources de données (par exemple les doublons inconsistants) (65). Cependant, cette taxonomie ne tient pas compte des problèmes de qualité liés à la structure des données. Dans ce travail, nous proposons de suivre cette taxonomie parce qu'elle se rapproche le plus des besoins d'analyse qui concernent le projet DIAGnosTIC, en y ajoutant l'analyse des problèmes de qualité issus d'autres travaux ou de notre expérience, en particulier pour l'analyse de la qualité liée à des structures de bases de données.

Les méthodes d'évaluation de la qualité des données ont été présentées dans plusieurs articles (60,61,66,67). Dans la mesure où les méthodes définies par Woodall *et al.* (67) sont organisées en suivant la granularité des éléments utilisés par la taxonomie d'Oliveira *et al.* (65), nous nous sommes basés sur ce travail pour le choix des méthodes d'évaluation. Nous nous sommes également inspirés des travaux de Wieskpof et Weng (61), ainsi que de notre propre expérience lorsque des problèmes de qualité n'ont pas été traités par Woodall *et al*. Afin de faciliter la présentation des résultats, nous avons abordé d'abord les problèmes de qualité liés au *schéma de données*, puis les problèmes de qualité liés aux *enregistrements*.

L'ensemble des informations présentes dans les systèmes sources n'a pas été évaluée dans ce travail parce que seulement une partie des tables présentes dans ces systèmes intègrent l'entrepôt de données. Les tables que nous avons sélectionnées pour l'analyse des données et qui ont été intégrées à l'entrepôt de données sont présentées annexe 3 selon un schéma simplifié qui ne contient pas les tables de dimensions. Le tableau 3 récapitule les informations qui ont été évaluées en précisant pour chaque table les tables de dimensions associées, et s'il y a lieu, les enregistrements sur lesquels ont porté les analyses. Pour chaque système source, les tables PATIENT ont été évaluées puisqu'elles permettent de faire le lien entre les données issues des différents systèmes. Pour les bases de données GAM et CORA, les séjours hospitaliers, ainsi que les mouvements, les diagnostics et les actes médicaux ont été évalués. Pour DIANE, l'évaluation a porté sur les mesures, les événements et les médicaments. Dans le cas des *mesures*, les enregistrements de FcECG, FcSpO2, SPO2, PNIm, PART-m, O2i, O2e, CO2e, BIS, Entropie, Température, Vce, Dese et de Seve ont été étudiés. Concernant les *médicaments*, l'accent a été porté sur les administrations d'Alfentanil, Sufentanil, Rémifentanil, Clonidine, Kétamine, Lidocaïne, Propofol, et Penthotal dans la mesure où les classes thérapeutiques de ces médicaments (antinociceptifs et hypnotiques) définissent en soi l'anesthésie générale. Enfin, quatre *événements* permettant de définir des périodes clés de toute procédure réalisée sous anesthésie ont été recherchés systématiquement dans chaque enregistrement : le début et la fin de l'anesthésie et le début et la fin de la chirurgie. Pour l'ensemble de ces données, nous nous sommes intéressés aux interventions ayant eu lieu entre le 01/01/2010 et le 31/12/2012.

Tableau 3 : Informations des systèmes sources évaluées dans ce travail

Tables	Enregistrements	Tables de dimensions associées
Patient Infocentre - Patient GAM - Patient CORA	2010 - 2012	
Intervention Infocentre	2010 - 2012	Classe ASA
Séjour GAM - Séjour CORA	2010 - 2012	
RSS CORA		Groupe Homogène de Malades

RUM CORA	-	Unité de soins, Autorisation de l'unité, Mode d'entrée, Mode de sortie
Acte Médical CORA		Acte médical (CCAM)
Diagnostic CORA		Diagnostic (CIM10)
Mesure DIANE	FcECG, FcSpO2, SPO2, PNIm, PART-m, O2i, O2e, CO2e, BIS, Entropie, Température, Vce, Dese, Seve	Paramètre, Unité
Evénement DIANE	Début d'anesthésie, Fin d'anesthésie, début de chirurgie, Fin de chirurgie	Evénement
Médicament DIANE	Alfentanil, Clonidine, Kétamine, Lidocaïne, Rémifentanil, Propofol, Penthotal, Sufentanil	Médicament

Nous avons étendu l'évaluation de la qualité aux tables de dimensions référencées par les clés étrangères dans les tables déjà présentées : Classe ASA pour l'Infocentre, Groupe Homogène de Malades, Unité de soins, Autorisation de l'unité, Mode d'entrée, Mode de sortie, Acte médical (CCAM), Diagnostic (CIM10) pour CORA et Paramètre, Unité, Evénement et Médicament pour DIANE.

Le chapitre est organisé comme suit :

Dans un premier temps, nous présentons les problèmes de qualité que nous avons évalué et nous les illustrons avec des exemples en lien avec notre projet, puis nous passons en revue les méthodes d'évaluation qui ont été employées, enfin nous présentons les résultats de l'évaluation.

2. Méthode

La mesure de la qualité des données peut être différente en fonction du besoin de l'utilisateur (60) : c'est pourquoi nous ne nous sommes intéressés qu'aux patients ayant bénéficié d'une anesthésie sans chercher à évaluer les problèmes de qualité intrinsèques aux bases de données GAM et CORA. En effet, l'objectif de ce travail était d'intégrer au sein de l'Infocentre d'anesthésie les séjours des patients ayant bénéficié d'une procédure d'anesthésie et non pas tous les séjours de tous les patients.

2.1 Problèmes de qualité

La taxonomie définie par Oliveira *et al.* (65) permet de hiérarchiser les problèmes de qualité de l'évaluation d'un champ d'un enregistrement jusqu'aux problèmes de qualité posés par l'association de plusieurs bases de données. Dans notre travail, nous avons utilisé cette taxonomie et les problèmes de qualité sont abordés en suivant le même ordre (tableau 4).

Nous avons intégré les problèmes de qualité liés aux structures de données qui n'étaient pas abordés dans la taxonomie développée par Oliveira *et al.* (65). Dans les sections suivantes, nous présentons les différents problèmes de qualité qui ont été évalués sur les données des systèmes sources de ce travail.

Tableau 4 : Problèmes de qualité sélectionnés pour l'évaluation de la qualité dans les systèmes sources

Problèmes de qualité	Source
Un champ d'un enregistrement	
Valeur manquante	Oliveira et al.
Valeur incorrecte	Oliveira et al.
Violation du domaine de valeurs	Oliveira et al.
Erreur de saisie	Oliveira et al.
Valeur imprécise	Oliveira et al.
Un champ, plusieurs enregistrements	
Violation de contrainte d'unicité	Oliveira et al.
Synonymes	Oliveira et al.
Format inapproprié	Rahm et al.
Plusieurs champs, un enregistrement	
Violation de dépendance fonctionnelle	Oliveira et al.
Violation d'une règle métier	Oliveira et al.
Doublons similaires	
Doublons similaires	Oliveira et al.
Doublons incohérents	Oliveira et al.
Enregistrements manquants	
Violation de contrainte d'unicité globale	Oliveira et al.
Violation d'une règle métier	Oliveira et al.
Plusieurs tables	
Violation d'intégrité référentielle	Oliveira et al.
Différence de structure	Rahm et al.
Différence de représentation	Rahm et al.

Différence de syntaxe	Rahm et al.
Plusieurs sources de données	
Différence de structure	Rahm et al.
Différence de représentation	Rahm et al.
Différence de syntaxe	Rahm et al.
Absence de lien entre deux systèmes sources	
Doublons similaires entre deux systèmes sources	Oliveira et al.
Doublons incohérents entre deux systèmes sources	Oliveira et al.

Un champ d'un enregistrement

Valeur manquante : La valeur d'un champ est nulle pour un enregistrement donné. Dans l'exemple de la figure 9, la date de naissance du patient 125876 n'est pas renseignée et est considérée comme manquante. Dans notre étude, nous considérons également comme nulle les valeurs par défaut telle que 'Inconnu' pour un identifiant ou '31/12/9999' pour une date. En effet, dans ces exemples, l'information associée à la colonne est manquante et nous souhaitons dissocier ce type d'erreur d'une valeur renseignée mais incorrecte. En suivant cette définition, l'IPP du patient 125896 est également considéré comme manquant.

PATIENT				
ID_PATIENT	IPP	NOM	PRENOM	DATE_NAISSANCE
56248	157896540	Lartiguet	Vincent	12/12/1991
125876	157896550	Pestel	Jérémy	
125896	Inconnu	Hammoudi	Leïla	04/06/1984

Figure 9 : Illustration du problème de qualité "Valeur manquante" avec des enregistrements de la table PATIENT[1].

Valeur incorrecte : Un champ possède une valeur incorrecte lorsqu'elle est différente de la valeur réelle. Ainsi quand un patient est né le 12/01/2001 et que la date de naissance renseignée est le 12/10/2001, la valeur est considérée comme incorrecte.

Violation du domaine de valeurs : Certains champs d'une table ne peuvent prendre qu'une plage de valeurs déterminées. Par exemple, la date de naissance des patients ne peut pas être postérieure à la date où à lieu l'enregistrement, ni à la date où à lieu l'analyse. Dans le cas de la figure 10, le patient 125876 possède une date de naissance en dehors du domaine de valeurs possibles pour ce type de champ.

[1] Les noms de patients utilisés dans les différentes figures sont fictifs.

PATIENT				
ID_PATIENT	IPP	NOM	PRENOM	DATE_NAISSANCE
56248	157896540	Lartiguet	Vincent	12/12/1991
125876	157896550	Pestel	Jérémy	10/08/2023
125896	Inconnu	Hammoudi	Leïla	04/06/1984

Figure 10 : Illustration du problème de qualité " Violation du domaine de valeurs" avec des enregistrements de la table PATIENT.

Erreur de saisie : Un champ textuel peut être mal orthographié. Le deuxième enregistrement de la table EVENEMENT de la figure 11 présente une faute d'orthographe pour la valeur du paramètre INTITULE.

EVENEMENT			
ID_INTERVENTION	ID_EVENEMENT	INTITULE	HEURE_OCCURENCE
157896540	25	Induction	12:18:21
157896540	252	Incsion	14:45:23
157896540	301	Fin de chirurgie	15:32:20
157896540	652	Fin d'anesthésie	15:45:36

Figure 11 : Illustration du problème de qualité "erreur de saisie" avec des enregistrements la table EVENEMENT.

Valeur imprécise : Quand la valeur renseignée n'apporte aucune précision, elle est qualifiée d'imprécise. Dans la figure 12, le libellé de l'événement 524 ("Chirurgien") est imprécis et ne permet pas de savoir si l'occurrence de l'événement correspond à l'arrivée ou la sortie du chirurgien du bloc opératoire.

LISTE_EVENEMENT	
ID_EVENEMENT	LIB_EVENEMENT
25	Début d'anesthésie
301	Mise en place du garrot
450	Fin de chirurgie
524	Chirurgien

Figure 12 : Illustration du problème de qualité "valeur imprécise" avec des enregistrements de la table LISTE_EVENEMENT.

Un champ, plusieurs enregistrements

Violation de contrainte d'unicité : Un champ autre que le champ de clé primaire peut obéir à une contrainte d'unicité, c'est à dire que les valeurs contenues dans ce champ sont uniques au sein de la table considérées. Dans la table PATIENT présentée figure 13, le champ ID_PATIENT est le champ de clé primaire. Le champ IPP doit respecter une contrainte d'unicité. Les patients 125876 et 125896 possèdent des IPP identiques, ce qui viole la contrainte d'unicité.

PATIENT				
ID_PATIENT	IPP	NOM	PRENOM	DATE_NAISSANCE
56248	157896540	Lartiguet	Vincent	12/12/1991
125876	157896550	Pestel	Jérémy	27/02/1982
125896	157896550	Hammoudi	Leïla	04/06/1984

Figure 13 : Illustration du problème de qualité "violation de contrainte d'unicité" avec des enregistrements de la table PATIENT.

Synonymes : Une information est présente sous différentes orthographes dans le même champ. Ainsi les enregistrements d'événements "Début d'anesthésie" et "Induction" sont des synonymes.

Format inapproprié : Le format d'une colonne est inapproprié lors qu'il est inadapté au type de valeurs qu'il contient. Par exemple, si une colonne est typée comme un champ texte (par exemple VARCHAR2(50)) et qu'elle contient des dates, le format est considéré comme inapproprié, alors qu'un format DATE aurait été approprié. Ce problème de qualité est illustré par la figure 14, qui représente une table PATIENT; le type de données attendues dans chaque champ est précisé. Le champ DATE_NAISSANCE est "typé" comme un champ texte (VARCHAR2) mais ne contient que des dates : le format est inapproprié.

PATIENT				
ID_PATIENT NUMBER(10)	IPP CHAR(10)	NOM VARCHAR2(50)	PRENOM VARCHAR2(50)	DATE_NAISSANCE VARCHAR2(50)
56248	157896540	Jounwaz	Reza	07/04/1969
125876	157896550	De jonckheere	Julien	13/07/1977
125896	157896550	Pestel	Jérémy	27/02/1982

Figure 14 : Illustration du problème de qualité "format inapproprié" avec des enregistrements de la table PATIENT.

Plusieurs champs, un seul enregistrement

Violation de dépendance fonctionnelle : Il y a une dépendance fonctionnelle entre plusieurs champs d'un même enregistrement lorsqu'un ou plusieurs champs déterminent la valeur d'un ou plusieurs autres champs. Ainsi, dans l'exemple de la figure 15, le contenu du champ UNITE est déterminé par le contenu du champ PARAMETRE. Dans le cas de la fréquence cardiaque ECG, la mesure doit être exprimée en battement par minute (bpm), ce qui n'est pas le cas de la première mesure, exprimée en millimètre de mercure (mmHg). Il y a une violation de dépendance fonctionnelle.

MESURE				
ID_INTERVENTION	PARAMETRE	HEURE_OCCURENCE	VALEUR	UNITE
157896540	FcECG	12:18:21	80	mmHg
157896550	FcECG	14:45:23	62	bpm
157896570	PNIm	08:25:45	73	mmHg

Figure 15 : Illustration du problème de qualité "violation de dépendance fonctionnelle" entre plusieurs attributs de la table MESURE.

Violation d'une règle métier : Le domaine de valeurs d'un champ peut être défini par la combinaison de plusieurs autres champs au sein d'une même table. Par exemple, le domaine de valeurs de la posologie d'un médicament dépend du médicament et de l'unité dans laquelle sa posologie est exprimée. La figure 16 illustre ce problème avec des administrations de sufentanil. Celles-ci sont exprimées en microgramme (µg), leur domaine de valeurs ayant été défini *a priori* dans l'intervalle [1 - 350 µg]. La posologie de la troisième administration est donc en dehors du domaine de valeurs.

MEDICAMENT				
ID_INTERVENTION	MEDICAMENT	HEURE_OCCURENCE	POSOLOGIE	UNITE
157896540	Sufentanil	12:18:21	20	µg
157896550	Sufentanil	14:45:23	100	µg
157896570	Sufentanil	08:25:45	1000	µg

Figure 16 : Illustration du problème de qualité "violation du domaine de valeurs" entre les champs MEDICAMENT et POSOLOGIE de la table MEDICAMENT.

Une seule table

Doublons similaires : Si deux enregistrements d'une même table présentent une majorité de champs identiques, ils sont considérés comme doublons similaires. Dans l'exemple présenté figure 17, les patients 15789 et 65874 présentent des prénoms et dates de naissance identiques, et ne diffèrent que par une lettre du nom, ce qui peut être une erreur de saisie. Ils sont donc potentiellement des doublons similaires.

PATIENT				
ID_PATIENT	IPP	NOM	PRENOM	DATE_NAISSANCE
15789	157896540	Dupond	Gérard	05/09/1962
65874	Inconnu	Dupont	Gérard	05/09/1962
89527	157896560	Dureil	Françoise	04/06/1954

Figure 17 : Illustration du problème de qualité "doublons similaires" entre les enregistrements de la table PATIENT.

Doublons incohérents : Des doublons sont considérés comme incohérents au sein d'une même table lorsqu'ils partagent une information, souvent un identifiant, mais que les autres informations sont différentes. Dans la figure 18, Les patients 65874 et 89527 possèdent un IPP identique mais des noms, prénoms et dates de naissance différents. L'IPP étant un numéro identifiant le patient de manière unique, il y a dans cette situation deux erreurs possibles : soit un des IPP est erroné, soit la saisie des informations d'un des deux patients (nom, prénom, date de naissance) est fausse. Dans ce cas, les doublons sont considérés comme incohérents.

PATIENT				
ID_PATIENT	IPP	NOM	PRENOM	DATE_NAISSANCE
15789	157896540	Dupond	Gérard	05/09/1962
65874	157896550	Agez	François	22/10/1988
89527	157896550	Dureil	Françoise	04/06/1954

Figure 18 : Illustration du problème de qualité "doublons incohérents" entre les enregistrements de la table PATIENT.

Enregistrements manquants : Des enregistrements sont manquants lorsqu'ils sont absents d'une table. Ainsi, si un patient bénéficie d'un monitorage de la fréquence cardiaque par électrocardiogramme et que la table stockant les paramètres vitaux ne répertorie aucune mesure de fréquence cardiaque, les enregistrements sont considérés comme manquants.

Violation de contrainte d'unicité globale : Lorsqu'au sein d'une même table, la combinaison des valeurs de plusieurs champs est unique, ces champs respectent une contrainte d'unicité globale. Dans le cas de la figure 19, la table MEDICAMENT possède une contrainte d'unicité globale sur les champs ID_INTERVENTION, ID_MEDICAMENT, HEURE_OCCURRENCE et POSOLOGIE. Pour les deux premiers enregistrements, cette contrainte n'est pas respectée.

MEDICAMENT				
ID_INTERVENTION	ID_MEDICAMENT	HEURE_OCCURRENCE	POSOLOGIE	UNITE
157896540	252	12:41:50	200	mg
157896540	252	12:41:50	200	mg
157896570	62	14:36:47	60	µg

Figure 19 : Illustration du problème de qualité "violation de contrainte d'unicité globale" entre les enregistrements de la table MEDICAMENT.

Violation d'une règle métier : Les enregistrements d'une table représentant des événements temporels doivent être cohérents vis à vis de la réalité (règle métier). Ainsi, la figure 20 présente les enregistrements d'événements d'une intervention. L'événement représentant le début de la chirurgie a été enregistré avant

l'événement représentant le début de l'anesthésie (12:18:21/14:45:23) alors qu'en pratique, le début de l'anesthésie a lieu avant le début de la chirurgie : il y a violation d'une règle métier.

EVENEMENT			
ID_INTERVENTION	ID_EVENEMENT	INTITULE	HEURE_OCCURENCE
157896540	25	Début de chirurgie	12:18:21
157896540	252	Début d'anesthésie	14:45:23
157896540	301	Fin de chirurgie	15:32:20
157896540	652	Fin d'anesthésie	15:45:36

Figure 20 : Illustration du problème de qualité "violation d'une règle métier" entre les enregistrements de la table EVENEMENT.

Plusieurs tables

Violation d'intégrité référentielle : Lors d'une relation clé primaire/clé étrangère (68) entre deux tables d'une même base de données, l'identifiant de clé étrangère de la première table ne trouve pas de correspondance dans la colonne de clé primaire de la seconde table. Ce problème de qualité est illustré par la figure 21, la colonne ID_EVENEMENT de la table EVENEMENT référence la colonne ID_EVENEMENT de la table LISTE_EVENEMENT. L'identifiant 252 n'existe pas dans cette seconde table, il y a violation d'intégrité référentielle.

EVENEMENT		
ID_INTERVENTION	ID_EVENEMENT	HEURE_ OCCURENCE
157896540	25	12:18:21
157896550	252	14:45:23
157896570	301	08:25:45

LISTE_EVENEMENT	
ID_EVENEMENT	LIB_EVENEMENT
25	Début d'anesthésie
301	Mise en place du garrot
450	Fin de chirurgie

Figure 21 : Illustration du problème de qualité "violation d'intégrité référentielle" entre les champs ID_EVENEMENT de la table EVENEMENT et LISTE_EVENEMENT.

Différence de structure/représentation/syntaxe : Deux tables différentes peuvent enregistrer la même information. Elles peuvent cependant employer pour cela des structures, des représentations ou des syntaxes différentes.

Il y a une différence de structure quand une information est représentée par un enregistrement dans une table et par un attribut dans une autre table. La différence de structure peut aussi se matérialiser par des types de colonnes différents pour la même information. Ainsi, sur la figure 22, l'information concernant l'unité de la mesure est représentée par le champ de texte UNITE dans la table MEDICAMENT du premier système source alors qu'elle est représentée par des enregistrements de la table UNITE dans le deuxième système source.

Les différences de représentation et de syntaxe sont présentées dans la section suivante.

MEDICAMENT				
ID_INTERVENTION NUMBER(10)	ID_MEDICAMENT NUMBER(6)	HEURE_OCCURENCE DATE	POSOLOGIE NUMBER(6)	UNITE VARCHAR2(50)
157896540	252	12:41:50	200	mg
157896550	124	09:12:52	50	mg
157896570	62	14:36:47	60	µg

MESURE				
ID_INTERVENTION NUMBER(10)	ID_PARAMETRE NUMBER(6)	VALEUR NUMBER(12,6)	ID_UNITE NUMBER(3)	HEURE DATE
157896540	5	55	1	11:55:21
157896550	15	45	2	09:15:35
157896570	22	78	32	14:45:47

UNITE	
ID_VILLE NUMBER(3)	CODE_POSTAL CHAR(5)
1	bpm
25	%
133	µg/ml

Figure 22 : Illustration du problème de qualité "différence de structure" entre les tables MEDICAMENT et MESURE concernant la représentation de l'unité.

Plusieurs sources de données

Nous avons ici plusieurs systèmes sources avec des informations similaires. Trois tables répertorient les patients dans les trois systèmes, et deux tables répertorient les séjours dans les systèmes GAM et CORA. L'objectif étant de retrouver pour les patients enregistrés dans DIANE, les séjours correspondants dans GAM et CORA, nous aborderons les problèmes de qualité associés (i) à la cohérence des informations enregistrées dans les différents systèmes, mais aussi (ii) à la liaison des données entre DIANE d'un côté et GAM et CORA de l'autre.

(i) L'information commune aux trois systèmes est l'identifiant unique patient au sein de l'hôpital, l'IPP. Ainsi, tous les patients bénéficiant d'une anesthésie, et enregistrés dans DIANE, doivent correspondre à un patient dans GAM et CORA par le biais de l'IPP.

(ii) Les patients identifiés par un même IPP dans deux systèmes différents doivent correspondre d'un point de vue identité (nom, prénom, date de naissance).

Différence de structure/représentation/syntaxe : Deux systèmes sources différents peuvent enregistrer la même information. Ils peuvent cependant employer pour cela des structures, des représentations ou des syntaxes différentes. La figure 23 présente plusieurs de ces problèmes de qualité.

Il y a une différence de structure quand une information est représentée par un enregistrement dans une table pour un système source et par un attribut d'une table pour l'autre système. La différence de structure peut aussi se matérialiser par des types de colonnes différents pour la même information. L'information concernant la ville du patient est représentée par les attributs CODE_POSTAL et VILLE dans la table PATIENT du premier système source alors qu'elle est représentée par des enregistrements de la table

VILLE dans le deuxième système source. Les champs IPP, SEXE et CODE_POSTAL ont des types différents dans les deux systèmes, respectivement CHAR(10)/CHAR(12), NUMBER(1)/CHAR(1) et NUMBER(5)/CHAR(5).

Une différence de représentation correspond à des valeurs d'attributs différents dans chaque système pour exprimer la même information. La figure 23 présente deux tables représentant des patients dans deux systèmes sources différents. Les modalités du sexe sont représentées par la valeur 0 et 1 dans le premier système contre H et F dans le second. Les noms des tables et de certains attributs sont également différents alors qu'ils représentent les mêmes objets. Les champs DATE_NAISSANCE et DATE_NAISSANCE_PAT sont exprimés selon la syntaxe JJ/MM/AA dans le premier système contre la syntaxe JJ/MM/AAAA dans le second système. Cela illustre le problème de différence de syntaxe.

SYSTEME SOURCE 1 : PATIENT						
IPP CHAR(10)	NOM VARCHAR2(50)	PRENOM VARCHAR2(50)	DATE_NAISSANCE DATE	SEXE NUMBER(1)	CODE_POSTAL NUMBER(5)	VILLE VARCHAR2(50)
157896540	Dupond	Jean-François	05/09/62	0	59000	Lille
157896550	Agez	François	22/10/88	0	59100	Roubaix
157896570	Dureil	Françoise	04/06/54	1	59140	Dunkerque

SYSTEME_SOURCE_2 : PAT_HOSP					
IPP CHAR(12)	NOM_PAT VARCHAR2(50)	PRENOM_PAT VARCHAR2(50)	DATE_NAISSANCE_PAT DATE	SEXE_PAT CHAR(1)	ID_VILLE NUMBER(5)
157896540	Dupond	J-François	05/09/1962	H	1
157896550	Agez	François	22/10/1988	H	25
157896570	Dureil	Françoise	04/06/1954	F	133

SYSTEME_SOURCE_2 : VILLE		
ID_VILLE NUMBER(5)	CODE_POSTAL CHAR(5)	VILLE VARCHAR2(50)
1	59000	Lille
25	59100	Roubaix
133	59140	Dunkerque

Figure 23 : Illustration du problème de qualité "différence de structure" entre les tables PATIENT et PAT_HOSP de deux systèmes sources différents.

Absence de lien entre deux systèmes sources différents : Deux systèmes sources possèdent deux entités similaires. Ces entités possèdent un identifiant unique commun aux deux systèmes. Il y a une absence de lien entre les deux systèmes quand les entités d'un système ne trouvent pas de correspondance dans le deuxième système. C'est le cas dans l'exemple défini sur la figure 24, où le patient 15789670 du système source 1 n'a pas de correspondance dans le système source 2.

SYSTÈME 1			
IPP	NOM	PRENOM	DATE_ NAISSANCE
157896540	Dupond	Gérard	05/09/1962
157896550	Agez	François	22/10/1988
157896570	Dureil	Françoise	04/06/1954

SYSTÈME 2			
IPP	NOM	PRENOM	DATE_ NAISSANCE
157896540	Dupond	Gérard	05/09/1962
157896550	Agez	François	22/10/1988
157896580	Silva	Toni	12/06/1977

Figure 24 : Illustration du problème de qualité "absence de lien entre deux systèmes sources différents".

Doublons incohérents entre deux systèmes sources différents : Deux entités sont identifiées par le même numéro unique (commun aux deux systèmes), mais possèdent des attributs de valeurs différentes. Ici (figure 25), les patients possèdent les mêmes identifiants uniques dans les deux systèmes sources, mais leur identité est différentes.

SYSTÈME 1			
IPP	NOM	PRENOM	DATE_ NAISSANCE
157896540	Dupond	Gérard	05/09/1962
157896550	Agez	François	22/10/1988
157896570	Dureil	Françoise	04/06/1954

SYSTÈME 2			
IPP	NOM	PRENOM	DATE_ NAISSANCE
157896540	Schiro	Caroline	24/08/1979
157896550	Leveau	Florian	30/05/1999
157896570	Deraedt	Anthony	10/04/1980

Figure 25 : Illustration du problème de qualité "doublons incohérents" entre les enregistrements de patients de deux bases de données.

Doublons similaires entre deux systèmes sources différents : Deux entités sont identifiées par le même numéro unique, et possèdent un ou plusieurs attributs de valeurs différentes mais similaires. Dans l'exemple présenté en figure 26, le lien a pu être établi entre les deux systèmes source grâce à l'identifiant unique IPP. En revanche, chaque enregistrement présente un champ différent.

SYSTÈME 1			
IPP	NOM	PRENOM	DATE_ NAISSANCE
157896540	Dupond	Gérard	05/09/1962
157896550	Agez	François	22/10/1988
157896570	Dureil	Françoise	04/06/1954

SYSTÈME 2			
IPP	NOM	PRENOM	DATE_ NAISSANCE
157896540	Dupont	Gérard	05/09/1962
157896550	Agez	Françoise	22/10/1988
157896570	Dureil	Françoise	04/06/1956

Figure 26 : Illustration du problème de qualité "doublons similaires" entre les enregistrements de patients de deux bases de données.

Synthèse

L'annexe 4 présente pour chaque problème de qualité les informations des systèmes sources sur lesquelles portera l'évaluation.

2.2 Méthodes d'évaluation

Plusieurs auteurs proposent des méthodes d'évaluation de la qualité des données (60,61,64,67). Une même méthode peut être utilisée pour évaluer plusieurs problèmes de qualité, mais certains problèmes de qualité ne peuvent pas être résolus par une méthode automatique (67). Néanmoins, des méthodes manuelles permettent de détecter si ces problèmes sont présents, mais ne sont pas capables de mesurer avec exhaustivité la fréquence de ces problèmes de qualité (voir Analyse lexicale pour la recherche de synonymes ci-dessous).

Le tableau 5 présente les associations entre les problèmes de qualité décrits dans la première partie de ce chapitre et les méthodes d'évaluation présentées dans la littérature pour évaluer chaque problème.

Tableau 5 : Sélection des méthodes pour évaluer les problèmes de qualité

Problème de qualité de données	Méthodes d'évaluation sélectionnées	Méthodes de la littérature non utilisées
Un champ, un enregistrement		
Valeur manquante	Analyse de colonne/Analyse de domaine/Element presence	Gold standard
Valeur incorrecte	Analyse lexicale/Analyse de colonne/Vérification des données	Gold standard/Data Element Agreement
Violation du domaine de valeurs	Analyse de domaine	
Erreur de saisie	Analyse lexicale, Vérification/validation des données	Gold standard/Data element agreement
Valeur imprécise	Analyse de colonne/Analyse lexicale	
Un champ, plusieurs enregistrements		
Violation de contrainte d'unicité	Analyse de colonne	
Synonymes	Analyse lexicale	
Format inapproprié	Analyse de colonne	
Violation d'une règle métier	Semantic profiling/Validity check/Data Element Agreement	
Plusieurs champs, un enregistrement		
Violation de dépendance fonctionnelle	Analyse de colonne	
Violation du domaine de valeurs	Analyse de domaine/ Data Element Agreement	
Violation d'une règle métier	Semantic profiling/Validity check/ Data Element	

	Agreement	
Une table		
Doublons similaires	Algorithmes de comparaison	Schema matching
Doublons incohérents	Algorithmes de comparaison	Schema matching
Enregistrements manquants		Gold standard/ Data Element Agreement/Data source agreement
Violation de contrainte d'unicité globale	Analyse de colonne	
Plusieurs tables		
Violation d'intégrité référentielle	Analyse clé primaire/étrangère	
Différence de structure	Analyse de colonne	
Plusieurs bases de données		
Différence de structure	Analyse de colonne	Cross-Domain Analysis, Schema matching
Différence de syntaxe	Analyse de colonne	Cross-Domain Analysis, Schema matching
Différence de représentation	Analyse de colonne	Cross-Domain Analysis, Schema matching
Absence de lien entre deux systèmes sources différents	Algorithmes de comparaison	
Doublons incohérents entre deux systèmes sources différents	Algorithmes de comparaison	Cross-Domain Analysis, Schema matching
Doublons similaires entre deux systèmes sources différents	Algorithmes de comparaison	Cross-Domain Analysis, Schema matching

Après une brève description de chaque méthode, nous détaillons les problèmes de qualité qu'elles sont capables de traiter ainsi que le score d'évaluation qui leur est associé : celui-ci est généralement calculé comme le ratio entre le nombre d'éléments présentant un problème de qualité ou une erreur, et le nombre total d'éléments évalués, ces éléments pouvant être des colonnes de tables ou des enregistrements.

Pour les problèmes de qualité liés aux enregistrements, plusieurs dénominateurs peuvent être utilisés pour le calcul du score en fonction de la table étudiée : patient, intervention, événement, mesure, médicament. Ainsi, dans les cas des enregistrements de mesures, on peut rapporter le nombre de mesures incorrectes au nombre de mesures total ou le nombre d'interventions avec des mesures incorrectes sur le nombre total d'interventions. Le choix dépend du but dans lequel l'analyse des problèmes de qualité est mise en œuvre.

Analyse de colonne

L'analyse de colonne (67) est sans doute l'une des premières méthodes à appliquer pour tous les champs des tables d'une base de données. Elle peut être utilisée pour détecter plusieurs problèmes de qualité. Dans la majorité des cas, le langage SQL suffit à extraire les informations nécessaires au calcul des scores.

Les informations qui peuvent être exploitées grâce à une analyse de colonne sont les suivantes :

Type des colonnes (DATE, NUMBER, VARCHAR...) (requête 1)

(1)

```
DESC INTERVENTION;
```

Longueur maximales des champs ou de la partie entière d'une valeur numérique (requête 2)

(2)

```
SELECT MAX(LENGTH(VALEUR)),
MAX(LENGTH(VALEUR) - LENGTH(ROUND(VALEUR)) - 1)
FROM MESURE;
```

Nombre de valeurs distinctes (requête 3)

(3)

```
SELECT COUNT(DISTINCT(LIBELLE_ASA))
FROM INTERVENTION;
```

Fréquence des valeurs (requête 4)

(4)

```
SELECT LIBELLE_ASA, COUNT(*)
FROM INTERVENTION
GROUP BY LIBELLE_ASA
ORDER BY COUNT(*) DESC;
```

Valeur manquante (requête 5)

(5)

```
SELECT COUNT(*)
FROM PATIENT
WHERE IPP IS NULL;
```

Valeur minimale, maximale, 1er quartile, médiane, 3ème quartile, moyenne, écart-type (requête 6)

(6)

```
SELECT MIN(VALEUR), MAX(VALEUR),
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY VALEUR),
PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY VALEUR),
PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY VALEUR),
AVG(VALEUR), STDDEV(VALEUR)
FROM MESURE;
```

Valeur unique (requête 7)

(7)

```
SELECT IPP, COUNT(*)
FROM INFOCENTRE.PATIENT
GROUP BY IPP
HAVING COUNT(*) > 1;
```

Combinaisons uniques (requête 8)

(8)

```
SELECT ID_INTERVENTION, ID_EVENEMENT, HEURE_OCCURENCE, COUNT(*)
FROM EVENEMENT
GROUP BY ID_INTERVENTION, ID_EVENEMENT, HEURE_OCCURENCE
HAVING COUNT(*) = 1;
```

La méthode d'analyse de données est comprise dans la méthode d'évaluation plus vaste de profilage de base de données (data base profiling) (66). Aux requêtes précédentes peut s'ajouter la requête (9), qui retourne le nombre de lignes d'une table. Cette requête est utile pour estimer la volumétrie d'une table, lorsqu'elle est associée aux métadonnées des colonnes (volume utilisé pour une colonne).

(9)

```
SELECT COUNT(*)
FROM MESURE;
```

Les différentes requêtes présentées dans cette section peuvent être employées pour détecter les problèmes de qualité présentés ci-dessous.

Valeur manquante

La requête 5 permet de comptabiliser le nombre d'enregistrements de la table PATIENT où l'attribut IPP n'est pas renseigné.

Score = nombre d'enregistrements dont le champ n'est pas renseigné / Nombre total d'enregistrements analysés.

Format inapproprié

Plusieurs requêtes sont nécessaires pour déterminer si le format d'une colonne est approprié aux valeurs qu'elle contient. Tout d'abord, la requête 1 permet de déterminer le type des colonnes des tables étudiées. La requête 2 permet de connaitre la longueur de champ réellement utilisée ainsi la longueur de la partie entière si le champ est numérique. Cette requête permet de déterminer si l'espace alloué par le type de colonne est réellement utilisé par les enregistrements. Enfin la requête 4 liste les valeurs enregistrées dans le colonne étudiée dans le but de détecter si des champs numériques, binaires ou temporels sont stockés dans des champs textuels.

Score = Nombre de colonnes avec un format inapproprié / Nombre total d'enregistrements analysés.

Violation de contrainte d'unicité

L'analyse de colonne peut être utilisée pour déterminer si les valeurs d'une colonne sont uniques au sein de la table. La requête 5 retourne les valeurs de la colonne IPP utilisée par plusieurs enregistrements de la table PATIENT de l'Infocentre. Il y a violation de la contrainte d'unicité sur la colonne IPP dans le cas où le résultat de la requête serait non nul.

Score = Nombre d'enregistrements avec violation de la contrainte d'unicité / Nombre total d'enregistrements analysés.

Violation de contrainte d'unicité globale

Dans l'exemple suivant, les champs dont la combinaison est unique sont les champs ID_INTERVENTION, ID_EVENEMENT, HEURE_OCCURENCE. La requête 8 permet de retourner les enregistrements pour lesquels la combinaison des trois champs n'est pas unique.

Score = nombre d'enregistrement pour lesquels la clé composite n'est pas unique / nombre d'enregistrements analysés.

Violation de dépendance fonctionnelle

On définit les dépendances fonctionnelles entre chaque colonne. La figure 8 présente les dépendances fonctionnelles entre les paramètres et les unités dans lesquelles les valeurs sont exprimées. Ainsi, les mesures du paramètre FcECG doivent être exprimées en battement par minute (bpm). Lors de l'évaluation, les enregistrements pour lesquels la dépendance fonctionnelle entre les champs PARAMETRE et UNITE n'est pas respectée sont comptabilisés.

DEPENDANCE FONCTIONNELLE	
PARAMETRE	UNITE
FcECG	bpm
PNIm	mmHg
SpO2	%

MESURE				
ID_INTERVENTION	PARAMETRE	HEURE_OCCURENCE	VALEUR	UNITE
157896540	FcECG	12:18:21	80	mmHg
157896550	FcECG	14:45:23	62	bpm
157896570	PNIm	08:25:45	73	mmHg

Figure 27 : Dépendance fonctionnelle entre les paramètres et les unités.

Score = nombre d'enregistrements avec dépendance fonctionnelle non respectée / Nombre total d'enregistrements analysés.

Différence de structure

L'objectif étant de relier les tables et les champs d'un schéma à ceux d'un autre schéma, la différence de structure entre plusieurs systèmes est habituellement évaluée en comparant la structure des tables. Cependant, les métadonnées relatives aux schémas (contraintes, clés, type de champs, tables) ne sont pas toujours disponibles ni suffisantes pour évaluer la qualité des données. De plus, les applications opérationnelles ne sont pas toujours documentées. C'est pourquoi il faut généralement étudier les données enregistrées pour obtenir des informations sur les contraintes qui les régissent (ex : plage de données) et les problèmes de qualité qui peuvent survenir.

Il existe des méthodes automatiques permettant d'évaluer les différences de structure entre plusieurs schémas (69). Ces méthodes ont été développées pour apparier des schémas avec un nombre important de tables. Elles peuvent également être utilisées pour évaluer des structures d'une même source de données. Dans le cas présent, seulement sept tables (PATIENT pour GAM, CORA et l'Infocentre, SEJOUR pour GAM et CORA, MESURE et MEDICAMENT pour DIANE) vont être étudiées du point de vue de leur structure, ce qui peut donc être réalisé manuellement.

Score = Nombre d'éléments avec différence / Nombre total d'éléments comparés.

Différence de syntaxe

La différence de syntaxe entre deux champs est difficile à évaluer automatiquement. Il est nécessaire de comparer les valeurs des deux champs, pour déterminer si les syntaxes sont différentes. Ainsi, la figure 28 illustre cette méthode. La comparaison des champs NOM, PRENOM et DATE_NAISSANCE des deux tables permet de détecter que les champs textuels du premier système sont en majuscules et que la date de naissance est au format anglo-saxon.

SYSTÈME 1			
IPP	NOM	PRENOM	DATE NAISSANCE
157896540	DUPONT	GERARD	1962/05/09
157896550	AGEZ	FRANCOIS	1988/10/22
157896570	DUREIL	FRANCOISE	1954/06/12

SYSTÈME 2			
IPP	NOM	PRENOM	DATE NAISSANCE
157896540	Dupond	Gérard	05/09/1962
157896550	Agez	François	22/10/1988
157896580	Silva	Toni	12/06/1977

Figure 28 : Différence de syntaxe entre deux tables de deux systèmes sources différents

Score = Nombre de champs avec des syntaxes différentes / Nombre total de champs analysés.

Différence de représentation

Pour évaluer la différence de représentation entre plusieurs champs, la requête (4) permet de trouver des correspondances entre les modalités des deux colonnes comparées, comme illustré figure 29. Les modalités de la variable sexe et le nombre d'enregistrements associés sont clairement distincts entre les deux systèmes.

Système source 1	
Modalités	Nombre d'enregistrements
0	32547
1	27665

Système source 2	
Modalités	Nombre d'enregistrements
H	32601
F	28412

Figure 29 : Comparaison des modalités de la variable "Sexe" entre deux systèmes sources

Score = Nombre de champs avec des modalités différentes / Nombre total de champs étudiés.

Analyse de domaine

Violation du domaine de valeurs

L'analyse de domaine (67) permet de vérifier que les valeurs d'un champ se trouvent dans une plage autorisée ; en cela, elle est comparable à la validation des données (validity check (61)). Ainsi, une table peut contenir les domaines de valeurs de référence ; les enregistrements de la table sont parcourus pour vérifier que chaque valeur enregistrée se situe bien à l'intérieur de la plage de valeurs autorisées. Cette méthode peut être utilisée pour évaluer le nombre de violations de domaines de valeurs aux niveaux "un seul champ, un seul enregistrement" et "plusieurs champs, un seul enregistrement". La requête 10 permet de calculer le nombre d'enregistrements de la table MESURE pour lesquels les valeurs se trouvent en dehors des bornes définies dans la table BORNE pour le paramètre associé à la mesure.

(10)

```
SELECT COUNT(*)
FROM MESURE, BORNE
WHERE MESURE.ID_PARAMETRE = BORNE.ID_PARAMETRE
AND MESURE.VALEUR NOT BETWEEN
BORNE.BORNE_INFERIEURE AND BORNE.BORNE_SUPERIEURE
```

Score = nombre d'enregistrements avec une valeur en dehors du domaine de valeurs / nombre d'enregistrements évalués.

Analyse lexicale

L'analyse lexicale (67) s'applique aux champs textuels en vue de découvrir le sens des mots saisies par l'utilisateur. Dans notre cas, ce type de méthode sera utilisé pour détecter les synonymes et les erreurs de saisie.

Recherche de synonymes / erreurs de saisie

Comme précisé dans (67), aucune méthode automatique n'est disponible pour détecter automatiquement les synonymes. En revanche, Woodall et al. préconisent d'utiliser un dictionnaire pour détecter les erreurs de saisies. Dans notre cas, aucun dictionnaire en rapport avec la procédure d'anesthésie n'est disponible. Nous nous contenterons de définir et de rechercher, pour les éléments étudiés (tableau 1), une liste de mots clés comportant des synonymes et des fautes de frappe. Ces mots clés sont ensuite recherchés à l'aide de la requête 11. La clause LIKE ('%MOT_CLE%') recherche une chaîne de caractères (ici MOT_CLE) au sein d'une autre chaîne de caractères (INTITULE).

Tableau 6 : Exemples de synonymes et de fautes de frappe pour le sufentanil

Sufentanil	
Synonymes	Suf, Sufenta
Fautes de frappe	Sfuenta, Sfenta

(11)

```
SELECT *
FROM DIANE.MEDICAMENT
WHERE INTITULE LIKE ('%MOT_CLE%');
```

Cette méthode ne pouvant pas être exhaustive dans la mesure où les utilisateurs peuvent créer de nouvelles fautes de frappe, le score ne peut être calculé à partir du nombre d'enregistrements présentant des synonymes ou des fautes d'orthographes. Nous déterminerons plutôt le nombre d'informations pour lesquelles des synonymes ou des fautes d'orthographes ont été détectées pour au moins un enregistrement.

Par exemple, dans le cas des médicaments, nous évaluerons combien de médicaments présentent des synonymes sur les huit médicaments étudiés dans ce travail.

Score = nombre d'informations avec des synonymes / nombre d'informations étudiées.

Score = nombre d'informations avec des fautes d'orthographe / nombre d'informations étudiées.

Algorithmes de comparaison

Les algorithmes de comparaison (ou d'appariement, de correspondance) (67) peuvent être utilisés pour évaluer plusieurs problèmes de qualité : leur principe consiste à rechercher des liens entre plusieurs sources de données, les sources de données pouvant être des tables de bases de données distinctes ou d'une même base de données. Ces algorithmes peuvent ainsi être utilisés pour rechercher des correspondances entre deux relations (*data linking*) ou pour détecter des doublons (*deduplication*) (70,71).

Les liens entre les données comparées sont réalisés à partir d'opérateurs (*égal à, différent de, existe dans, n'existe pas dans*) ou de fonctions (*nombre de différences entre deux chaînes de caractères, similarité entre deux chaînes de caractères*).

Le système de gestion de bases de données Oracle (72) propose plusieurs fonctions de calcul de distance entre deux chaînes de caractères (73).

La fonction EDIT_DISTANCE calcule le nombre de changements nécessaires pour transformer une première chaîne de caractère en lune seconde. Ces deux chaînes de caractères étant passées en argument à la fonction. La requête 12 illustre cette fonction en comparant les deux noms DUPOND et DUPONT et retourne 1 comme résultat, les deux chaînes n'étant différentes que d'un caractère.

La fonction EDIT_DISTANCE_SIMILARITY propose un score de 0 à 100 en fonction du nombre de changements calculé avec la fonction EDIT_DISTANCE, 100 correspond à deux chaînes identiques. La requête 13 illustre cette fonction. Elle retourne une similarité de 25 pour les deux chaines DUPOND et DELCROIX.

(12)

```
SELECT UTL_MATCH.EDIT_DISTANCE('DUPOND', 'DUPONT')
FROM DUAL;
> 1
```

(13)

```
SELECT UTL_MATCH.EDIT_DISTANCE_SIMILARITY('DUPOND','DELCROIX') FROM
DUAL;
> 25
```

L'une des sources de données est employée comme référence. Ainsi, le nombre d'enregistrements présents dans le premier système est le dénominateur, le nombre d'enregistrements du premier système pour lesquels la condition de comparaison est établie étant le numérateur.

Absence de lien entre deux systèmes sources différents

La requête 14 compte le nombre d'enregistrements de la table PATIENT de l'Infocentre pour lesquels aucune correspondance (sur la base d'un IPP identique) n'est trouvée dans la table PATIENT de GAM.

(14)

```
SELECT COUNT(*)
FROM INFOCENTRE.PATIENT
WHERE NOT EXISTS (
      SELECT * FROM GAM.PATIENT
      WHERE GAM.PATIENT.IPP = INFOCENTRE.PATIENT.IPP
);
```

Score = Nombre d'enregistrements du premier système sans correspondance sur l'identifiant unique dans le deuxième système / Nombre total d'enregistrements du premier système.

Doublons incohérents entre deux systèmes sources différents

Dans le cas où un lien entre deux enregistrements est trouvé, par exemple un identifiant unique, il est possible de comparer les autres champs des enregistrements afin de vérifier s'ils présentent des différences importantes. Ainsi, dans l'exemple présenté dans la figure 28, les liens entre les trois enregistrements sont possibles grâce à l'IPP, mais les autres informations (NOM, PRENOM, DATE_NAISSANCE) différent dans plusieurs cas. La fonction EDIT_DISTANCE_SIMILARITY permet de mettre ces différences en évidence : les enregistrements dont au moins un des champs obtient un score de similarité inférieur à 80 sont considérés comme "doublons incohérents". La requête 15 illustre ce cas de figure : elle retourne le nombre d'enregistrements de la table PATIENT de l'Infocentre dont le score de similarité pour le champ NOM de GAM est inférieur à 80.

(15)

```
SELECT COUNT(DISTINCT T1.ID_PATIENT)
FROM INFOCENTRE.PATIENT T1, GAM.PATIENT T2
WHERE T1.IPP = T2.IPP
AND UTL_MATCH.EDIT_DISTANCE_SIMILARITY(T1.NOM, T2.NOM) <= 80;
```

Score = nombre d'enregistrements du premier système avec lien sur l'identifiant unique dans le deuxième système et au moins un champ différent / Nombre total d'enregistrements du premier système avec lien sur l'identifiant unique dans le deuxième système.

Doublons similaires entre deux systèmes sources différents

Selon le même principe, il est possible de détecter si les champs présentent une légère différence, de l'ordre d'un ou deux caractères différents par exemple. Dans le cas présenté figure 26, le lien entre les

patients est bien établi grâce à l'identifiant unique, mais chaque patient ainsi identifié présente une différence d'un caractère dans un champ des deux sources. Ce cas de figure peut être détecté grâce à la requête 16 qui retourne le nombre de patient de l'Infocentre pour lesquels l'identifiant unique IPP est commun avec un patient de GAM mais dont le nom présente une différence d'un ou de deux caractères entre les deux sources.

(16)

```
SELECT COUNT(DISTINCT T1.ID_PATIENT)
FROM INFOCENTRE.PATIENT T1, GAM.PATIENT T2
WHERE T1.IPP = T2.IPP
AND UTL_MATCH.EDIT_DISTANCE(T1.NOM, T2.NOM) IN(1, 2);
```

Analyse clé primaire/clé étrangère

Violation d'intégrité référentielle

L'analyse de clé primaire/clé étrangère a pour objectif de déterminer si les colonnes de deux tables différentes présentent une relation de clé primaire/clé étrangère même si cette contrainte n'est pas spécifiée par le modèle de données. Cette technique consiste à compter les enregistrements pour lesquels la colonne candidate de clé étrangère ne trouve pas de correspondance dans la colonne candidate de clé primaire. Le nombre d'enregistrements ainsi détectés est rapporté au nombre d'enregistrements total.

La requête 17 comptabilise le nombre d'enregistrements de la table EVENEMENT pour lesquels le champ ID_EVENEMENT ne trouve pas de correspondance dans la colonne ID_EVENEMENT de la table LISTE_EVENEMENT. Ce nombre est rapporté au nombre total d'enregistrements la table EVENEMENT.

(17)

```
SELECT COUNT(*)
FROM EVENEMENT
WHERE EVENEMENT.ID_EVENEMENT NOT IN (
      SELECT ID_EVENEMENT FROM LISTE_EVENEMENT
);
```

Score = Nombre d'enregistrements avec contrainte non respectée / Nombre total d'enregistrements étudiés.

Semantic profiling

Erreur de saisie/Valeur incorrecte/Violation de règle métier

La méthode du *semantic profiling* (67), comparable à la méthode *validity check* (61), consiste à définir des règles métier entre plusieurs colonnes d'une ou plusieurs tables pour vérifier la cohérence des données. Dans le cas des variables poids, taille et âge, il est possible de définir plusieurs règles du type : SI Age > 15 ans alors Taille > 100 cm et poids > 25 kg. Les trois colonnes sont ensuite parcourues et comparées pour détecter d'éventuelles incohérences. La figure 30 illustre ce cas de figure. Le patient 157896540 est âgé de 52 ans, mais le poids renseigné est de 9 kg. Il y a une erreur de saisie.

PATIENT			
ID_PATIENT	POIDS	TAILLE	AGE
157896540	9	170	52
157896550	50	162	41
157896570	87	178	35

Figure 30 : Semantic profiling - Vérification de règles métier entre les colonnes Poids, Taille et Age.

Score = Nombre d'enregistrements ne respectant pas la règle métier / Nombre total d'enregistrements étudiés.

Data Element Agreement

La méthode *Data Element Agreement* compare deux éléments d'une même source de données pour vérifier qu'elles sont identiques ou compatibles. Cette méthode permet ainsi de vérifier la présence d'enregistrements dans une table particulière, si une autre table du même système source permet d'affirmer leur occurrence.

Score = Nombre d'enregistrements identiques ou cohérents entre les deux tables / Nombre total d'enregistrements étudiés.

Element Presence

La méthode *element presence* (61) a pour objectif de déterminer les champs ou enregistrements manquants. Elle est comparable à l'analyse de colonne détaillée plus haut dans le cas des champs. La méthode du *Gold Standard* peut également être employée quand une deuxième source est disponible. En revanche, comme précisé dans (67) (Gap 2), il n'existe pas de méthode automatique pour détecter les enregistrements manquants si aucune autre source ne peut être utilisée comme gold standard. En effet, dans le cas de la procédure d'anesthésie, aucune autre source n'enregistre les signaux physiologiques, les administrations de médicaments et les heures d'occurrence des différentes étapes de l'intervention. Ainsi, si la feuille informatisée d'anesthésie ne recueille pas ces informations, il est impossible de savoir si les signaux physiologiques ont été mesurés ou non par les appareils de monitorage ou si un médicament a été administré au patient.

En revanche, certaines informations doivent obligatoirement être enregistrées au cours de la procédure d'anesthésie ou certains événements comme l'induction ou le début de chirurgie ont obligatoirement lieu au cours de l'intervention. Il est donc possible de détecter si ces événements ont été enregistrés dans DIANE. Ici, nous cherchons à détecter les interventions pour lesquelles des enregistrements d'événements, de médicaments ou de mesures sont manquants.

Nous nous intéressons aux quatre événements "Début de l'anesthésie", "Début de la chirurgie", "Fin de la chirurgie", "Fin de l'anesthésie". Nous comptabilisons les interventions pour lesquelles au moins l'une des occurrences des quatre événements n'est pas renseignée.

Pour les médicaments, lors d'une intervention chirurgicale nécessitant une anesthésie, au moins l'un des médicaments suivants est administré au patient : Propofol, Penthotal, Alfentanil, Remifentanil,

Sufentanil, Ketamine, Lidocaine, Clonidine. Nous relevons le nombre d'interventions pour lesquelles aucun de ces médicaments n'est renseigné.

Pour les mesures, nous nous intéressons aux 14 paramètres sélectionnés dans l'introduction de ce chapitre. Pour chaque paramètre, nous comptabilisons le nombre d'interventions pour lesquels aucun enregistrement du paramètre n'est retrouvé.

La requête 18 permet de détecter les interventions pour lesquelles les quatre événements "Début de l'anesthésie", "Début de la chirurgie", "Fin de la chirurgie", "Fin de l'anesthésie" sont enregistrés dans la table d'événements.

(18)

```
SELECT COUNT(DISTINCT I.ID_INTERVENTION
FROM INTERVENTION I
WHERE EXISTS (
      SELECT *
      FROM EVENEMENT E1
      WHERE E1.ID_EVENEMENT = 1
      AND I.ID_INTERVENTION = E1.ID_INTERVENTION)
AND EXISTS (
      SELECT *
      FROM EVENEMENT E2
      WHERE E2.ID_EVENEMENT = 2
      AND I.ID_INTERVENTION = E2.ID_INTERVENTION)
AND EXISTS (
      SELECT *
      FROM EVENEMENT E2
      WHERE E2.ID_EVENEMENT = 3
      AND I.ID_INTERVENTION = E2.ID_INTERVENTION)
AND EXISTS (
      SELECT *
      FROM EVENEMENT E2
      WHERE E2.ID_EVENEMENT = 4
      AND I.ID_INTERVENTION = E2.ID_INTERVENTION)
```

Score = nombre d'interventions sans l'événement / nombre total d'interventions étudiées.

Cross-Domain Analysis

La méthode Cross-Domain Analysis (67) compare plusieurs colonnes de tables différentes (parfois de systèmes sources différents) pour détecter si ces colonnes sont redondantes et proposent les mêmes valeurs. Cette méthode est utile lors de l'intégration de bases de données comportant un nombre important de tables pour lesquelles une revue manuelle est impossible ou lorsque la documentation n'est pas disponible.

Vérification des données

La méthode de vérification des données (*data verification*) (67), aussi appelée validation des données (*data validation*) consiste à rechercher si les données enregistrées se trouvent dans un set de valeurs de référence, celui-ci pouvant être un dictionnaire d'adresse, de médicaments, d'événements ...

Comparaison de schéma

La méthode de comparaison de schéma (schema matching) (67) a pour objectif de retrouver des similitudes entre plusieurs schémas en comparant les structures (tables, colonnes), les contraintes et les enregistrements.

Gold standard

La méthode du Gold Standard (61) consiste à utiliser des enregistrements d'une autre source comme référence pour l'évaluation. Cette méthode permet de détecter les problèmes de valeurs manquantes et de valeurs incorrectes. L'inconvénient de cette méthode est que dans la majorité des cas, la source de référence n'existe pas.

Enregistrements manquants

La requête 19 permet de comptabiliser les enregistrements présents dans la source de référence (Table EVENEMENT_GOLD_STANDARD) absents de la source à évaluer (Table EVENEMENT).

(19)

```
SELECT COUNT(*)
FROM EVENEMENT_GOLD_STANDARD T1
WHERE NOT EXISTS (
            SELECT *
            FROM EVENEMENT T2
            WHERE T1.ID_INTERVENTION = T2.ID_INTERVENTION
            AND T1.ID_EVENEMENT = T2.ID_EVENEMENT
            AND T1.DATE_OCCURENCE = T2.DATE_OCCURENCE);
```

Score = Nombre d'enregistrements absents / Nombre total d'enregistrements évalués.

Valeurs incorrectes

La requête 20 permet de vérifier que les valeurs des enregistrements à évaluer (champ POSLOGIE de la table MEDICAMENT) sont identiques aux valeurs des enregistrements de la source de référence (champ POSOLOGIE de la table MEDICAMENT_GOLD_STANDARD).

(20)

```
SELECT COUNT(*)
FROM MEDICAMENT_GOLD_STANDARD T1, MEDICAMENT T2
WHERE T1.ID_INTERVENTION = T2.ID_INTERVENTION
AND T1.ID_MEDICAMENT = T2.ID_MEDICAMENT
AND T1.POSOLOGIE = T2.POSOLOGIE;
```

Score = Nombre d'enregistrements avec valeurs différentes / Nombre total d'enregistrements étudiés.

Data source agreement

La méthode *Data source agreement* (61) consiste à comparer les données à évaluer avec les données d'une autre source.

Comparaison de distribution

La méthode de comparaison de distribution (61) compare une la distribution d'une variable étudiée avec sa distribution théorique, définie par une référence. Cette méthode peut être comprise dans la méthode d'analyse de colonne.

Log Review

La revue des fichiers journaux (61) consiste à étudier les pratiques d'utilisation du logiciel.

2.3 Synthèse

Les méthodes *Gold Standard*, *Data Element Agreement* et *Data Source Agreement* ne peuvent pas être utilisées pour détecter les valeurs manquantes, les valeurs incorrectes, les erreurs de saisies, ou les enregistrements manquants parce qu'il n'existe pas de base de données de référence ou d'autres sources de données. De même, les méthodes *Cross-Domain Analysis* et *Schema matching* ne seront pas utilisées car nous savons déjà quelles données vont être extraites de chaque système.

3. Résultats

Les résultats sont présentés en deux sections : dans la première, les éléments analysés sont les structures des bases de données (tables, colonnes), dans la seconde, ce sont les problèmes de qualité propres aux enregistrements.

3.1 Schéma

Les résultats complets de l'évaluation des problèmes de qualité de données liés aux schémas des bases de données présentés en annexe 7. Parmi les quatre systèmes étudiés, l'Infocentre d'Anesthésie est

celui-ci pour lequel les formats des champs sont les plus appropriés aux données qu'ils contiennent. Les trois autres systèmes ont généralement des champs plus grands que nécessaires, en particulier DIANE (exemple : un champ typé comme un NUMBER(63) pour stocker une valeur de mesure, un champ typé comme un VARCHAR(255) pour tous les champs texte comme le prénom, le nom et le nom marital).

D'un point de vue sémantique, les tables PATIENT et SEJOUR possèdent des nom de tables et de colonnes différents d'un système à l'autre, ainsi que des types de colonnes différents.

Concernant les structures de données, les tables MEDICAMENT et MESURE de DIANE possèdent toutes les deux un champ référant à une unité. Cependant, dans la première table l'unité est stockée comme un champ textuel alors que dans le seconde la colonne contient un identifiant de clé étrangère référençant un enregistrement de la table de UNITE.

La différence la plus importante concerne les colonnes NOM/NOMMARITAL et NOM_USUEL/NOM_NAISSANCE des tables PATIENT des systèmes DIANE et CORA. Les informations sont inversées dans les deux systèmes. La figure 31 illustre ce problème avec le patient 157896550.

DIANE : PATIENT				
IPP	NOM	PRENOM	NOMMARITAL	DATE_NAISSANCE
157896540	Dupond	Jean-François		05/09/62
157896550	Schiro	Séverine	Wargnier	22/10/88
157896570	Dureil	Françoise		04/06/54

CORA : PATIENT				
IPP	NOM_USUEL_PATIENT	PRENOM	NOM_NAISSANCE	DATE_NAISSANCE
157896540	Dupond	Jean-François		05/09/62
157896550	Wargnier	Séverine	Schiro	22/10/88
157896570	Dureil	Françoise		04/06/54

Figure 31 : Hétérogénéité de représentation entre les tables PATIENT des bases de données DIANE et CORA

3.2 Enregistrements

Les résultats complets de l'évaluation des problèmes de qualité de données liés aux enregistrements des bases de données sont présentés en annexe 8. Nous détaillons ci-dessous les principaux résultats.

Un champ, un seul enregistrement

Concernant les problèmes de qualité liés à **un champ d'un enregistrement**, trois champs présentent un nombre important de **valeurs manquantes**, les champs POIDS, TAILLE et IMC de la table PATIENT de l'INFOCENTRE. Cinq champs de la table MEDICAMENT sont également peu renseignés mais l'information qu'ils contiennent s'avère être facultative et dépend du mode d'administration du médicament (CONCENTRATION, UNITE_CONCENTRATION, TOTAL, UNITE_TOTAL, HEURE_FIN).

Les tables de références MEDICAMENT, PARAMETRE et EVENEMENT contiennent des **erreurs de saisie** et des **valeurs imprécises** pour la plupart des enregistrements étudiés. Le tableau 7 illustre ces deux problèmes.

Tableau 7: Exemple d'erreurs de saisie, valeurs imprécises et synonymes pour les médicaments, paramètres et événements.

Champ	Erreurs de saisie	Valeurs imprécises	Synonymes
MEDICAMENT		-	Sufenta PCA 1 µg/ml, Sufenta PCA 1µg/ml, SUFENTANIL, SUFENTANIL IV (**Sufentanil**)
PARAMETRE		Groupe O2, Groupe CO2, Fréquence cardiaque	Pression ART moyenne, Pression invasive moyenne 1, Pression arterielle sanguine moyenne. (**Pression invasive moyenne**)
EVENEMENT	Inductino, Incsion, fin dechir	-	Début de l'acte, Début de l'acte ou incision (**Début de la chirurgie**)

Les champs POIDS, TAILLE et IMC de la table PATIENT de l'Infocentre contiennent peu d'enregistrements avec une **violation du domaine de valeurs**, respectivement 0,43%, 0,68% et 0,72%.

Un champ, plusieurs enregistrements

Au niveau de granularité "**un champ, plusieurs enregistrements**", le problème de qualité le plus important concerne les **synonymes** sur les tables MEDICAMENT, PARAMETRE et EVENEMENT de DIANE (voir tableau 7). Seulement six cas présentent une **violation de contrainte d'unicité**. Dans le cadre de la documentation d'une procédure d'anesthésie, les événements "début d'anesthésie" et "début de chirurgie" doivent être renseigné avant respectivement les événements "fin d'anesthésie" et "fin de chirurgie". Pour un peu plus de 1% des interventions, les événements de fin de période sont renseignés avant les événements de début de période. Pour ces interventions il y a une **violation de règle métier**.

Plusieurs champs d'un seul enregistrement

Une ou plusieurs unités de référence ont été définies pour chaque paramètre et chaque médicament étudiés (voir tableau 8). Pour certains paramètres et tous les médicaments, plusieurs unités sont utilisées. Certaines sont correctes et dépendent du paramétrage de l'appareil de monitorage ou du mode d'administration du médicament. Ainsi, les paramètres liés à la ventilation peuvent être exprimés en millimètres de mercure (mmHg) ou en pourcentage, suivant que le paramétrage de l'appareil suit les normes françaises ou anglo-saxones. Le tableau 8 illustre les différentes unités utilisées pour exprimées les administrations du propofol en fonction du mode d'administrations. La proportion d'enregistrements de médicaments ou de mesures présentant une **violation de dépendance fonctionnelle** entre les champs PARAMETRE et MEDICAMENT et le champ UNITE n'excède pas 0,01%.

Tableau 8 : Mode d'administration du propofol et unités associées

Mode d'administration	Unités
AIVOC	µg/ml plasma
Bolus	mg
Seringue auto-pulsée	ml/h
AIVOC	µg/ml

La proportion d'enregistrements de mesures ou de médicaments présentant une **violation du domaine de valeurs** reste inférieure à 0,75%. Cinquante pourcent des erreurs concernant les médicaments proviennent des administrations de Kétamine.

Aucun enregistrement de la table INTERVENTION de l'Infocentre ne présente de **violation de la règle métier** "Date début intervention < date de fin intervention".

Une table

Il y a très peu de **doublons similaires** (0,7%) pour les enregistrements de la table PATIENT de DIANE et aucun doublon incohérent n'ont été détectés (avec le même identifiant unique IPP). En revanche, la proportion d'interventions avec des enregistrements manquants est importante. En effet, au moins un événement est manquant dans environ 30% des interventions (voir le détail Tableau 9) et aucun médicament n'est renseigné dans 13% des interventions.

Tableau 9 : Nombre d'interventions avec événement manquant

Evénement manquant	Nombre d'enregistrements (%)
Début d'anesthésie	21974 (13,6%)
Début de chirurgie	34823 (21,6%)
Fin de chirurgie	37938 (23,6%)
Fin d'anesthésie	30571 (19,0%)
Au moins un événement manquant	54048 (33,6%)

Une minorité des enregistrements de mesures, d'événements et de médicaments (<1%) présente une **violation de contrainte d'unicité globale** sur les tables MESURE, EVENEMENT et MEDICAMENT. Seule la table SEJOUR de GAM présente un nombre important d'enregistrements pour lesquels il y a violation de la contrainte d'unicité globale (avec comme contrainte d'unicité les champs NOM, PRENOM et DATE_NAISSANCE).

Plusieurs tables

Deux tables seulement présentent une proportion significative d'enregistrements avec **violation d'intégrité référentielle**, les tables EVENEMENT et MEDICAMENT de DIANE. Ce problème de qualité est du à une autorisation de saisie en texte libre (intitulé d'événement et de médicament). Pour ces enregistrements, il n'y a donc aucune correspondance avec les tables de référence d'événements et de médicaments.

Plusieurs bases de données

Pour 3,64% des patients de DIANE il n'a pas été possible de retrouver le patient correspondant dans GAM et CORA avec l'identifiant unique utilisé dans les trois systèmes (IPP).

Pour moins de 1% des patients pour lesquels le lien a pu être établi par l'IPP, il y a une incohérence dans l'identité (nom, prénom, date de naissance) et pour 1,86% il existe une légère différence dans le contenu de ces champs entre les différentes bases de données.

4. Discussion

4.1 Discussion générale

Dans ce chapitre, nous avons passé en revue les problèmes de qualité de données recensés par la littérature et présenté les principales méthodes d'évaluation de la qualité des données. Nous avons présenté les résultats de l'évaluation de la qualité des données des bases de données étudiées dans ce travail : la feuille informatisée d'anesthésie DIANE, le logiciel de gestion administrative des malades GAM, le logiciel de facturation CORA et l'Infocentre d'Anesthésie.

L'évaluation de la qualité des données permet dans un premier temps de valider les hypothèses émises lors de la sélection des systèmes sources pour la constitution d'un entrepôt de données. Elle permet également de déterminer les limites de l'utilisation secondaire des données, de fournir les informations nécessaires aux développements des modules ETL, et enfin de proposer des indications concernant les systèmes sources ou les pratiques des utilisateurs afin d'améliorer la qualité des données.

Dans notre travail, nous avons choisi de parcourir les problèmes de qualité par niveau de granularité des entités évaluées, en commençant par l'analyse de la qualité de chaque champ individuel, en élargissant progressivement jusqu'à l'analyse de la qualité de la réunion de plusieurs bases de données. Une technique alternative aurait consisté en une analyse de qualité variable par variable, mais nous avons trouvé utile pour ce projet de présenter et de discuter les méthodes et résultats question par question, dans l'optique *in fine* d'être en mesure d'étendre le travail effectué à d'autres systèmes sources et donc à d'autres variables, qui ne pouvaient intégrer l'analyse variable par variable puisqu'elles n'étaient pas encore définies.

Contrairement à certaines publications, les résultats de notre analyse n'ont pas conduit à une présentation selon des dimensions de qualité de données (61,74,75), ni selon un score global de qualité. L'intérêt d'une telle présentation aurait été de pouvoir comparer l'évolution de la qualité des données année par année et entre différents établissements de soins, mais les scores que nous avons présentés pour chaque méthode d'évaluation fournissent une vue globale suffisante de la qualité des données et des éléments importants pour la phase ultérieure d'intégration des données.

4.2 Discussion sur les méthodes d'évaluation

Les méthodes d'évaluation présentées dans la littérature et utilisées dans ce travail sont relativement hétérogènes : d'une part parce que les scores proposés par chaque méthode ne peuvent pas être discutés conjointement, et d'autre part parce que la difficulté de mise en œuvre de chaque méthode dépend du problème de qualité à traiter.

Scores

Un résultat satisfaisant pour une méthode d'évaluation n'induit pas automatiquement que la donnée est de bonne qualité : en effet, plusieurs méthodes peuvent être appliquées à une même donnée selon la dimension à évaluer, de sorte que le résultat obtenu avec une méthode renseigne plutôt sur le type de défaut (de qualité) qu'il faut résoudre pour exploiter la donnée.

Les scores obtenus avec les différentes méthodes ne conduisent pas toujours à des résultats similaires, car **chaque défaut de qualité constaté n'a pas les mêmes effets quant à l'utilisation secondaire des données étudiées.** Par exemple les résultats des scores de qualité obtenus avec les items "différence de structure/syntaxe/représentation" sont généralement élevés car les systèmes évalués ont été développés indépendamment et sont donc hétérogènes. Ces défauts de qualité peuvent cependant être facilement corrigés lors de la phase d'intégration des données vers l'entrepôt de données, alors que des enregistrements manquants ou incorrects sont définitivement perdus si aucune autre source permettant un recoupement des données recherchées n'est disponible.

Mise en application

La facilité d'application des méthodes d'évaluation varie d'une méthode à l'autre, dépendant des données évaluées, de leur volumétrie et de la disponibilité de la documentation. Certaines méthodes sont sans appel et fournissent directement un résultat :

- Analyse de colonne pour la détection de valeur manquante, les différences de structure/ syntaxe/représentation ou les violations de contrainte d'unicité ;
- Analyse de clé primaire/clé étrangère pour l'évaluation de la violation d'intégrité référentielle.

D'autres méthodes présentées sont plus difficiles à appliquer :

- Analyse de domaine et *semantic profiling* pour l'évaluation de la violation du domaine de valeurs et de la violation des règles métier : pour certaines variables, il est difficile de définir le domaine de valeurs et les règles métier.

Comme précisé dans (67), certains problèmes de qualité ne peuvent qu'être identifiés et pas corrigés: les méthodes semi-automatiques ne sont pas exhaustives, elles ne fournissent qu'une indication sur l'occurrence d'un problème de qualité :

- Analyse de colonne, analyse lexicale et *element presence* pour détecter et évaluer les problèmes d'enregistrements manquants, de synonymes, d'erreurs de saisie, de valeurs imprécises et de valeurs incorrectes ;
- Algorithmes de comparaison pour détecter les doublons similaires et les doublons inconsistants. Cette méthode automatique nécessiterait une vérification manuelle pour valider les doublons.

Enfin, d'autres méthodes ne peuvent pas toujours être appliquées, en particulier dans notre travail où nous ne bénéficions pas de données de comparaison ou de référence :

- *Gold standard*/*Data element agreement*/*Data source agreement* pour les valeurs incorrectes/erreurs de saisie/enregistrements manquants.

4.3 Discussion sur les résultats de l'évaluation

Le résultat principal de ce travail d'évaluation est que les informations décrites dans le chapitre 1 sont réellement disponibles et utilisables pour une analyse rétrospective.

L'identifiant unique utilisé pour chaque patient pris en charge au CHRU de Lille, l'IPP, peut être utilisé comme lien entre les différents systèmes. Des difficultés dans le traitement des données apparaissent cependant du fait de différences de structures/représentation/syntaxe entre les différents systèmes, mais celles-ci peuvent être résolues.

Certaines données sont cependant définitivement perdues lorsqu'elles sont :

- Non renseignées par l'utilisateur (par exemple: poids, taille du patient, certaines administrations de médicaments, certaines étapes de l'intervention) ;
- Non enregistrées ou non reliées à la feuille informatisée d'anesthésie (par exemple lorsqu'un moniteur externe n'est pas relié à l'ordinateur de la feuille d'anesthésie ; ceci concerne par exemple l'entropie, le BIS et parfois la température) ;
- Imprécises ou incorrectes (valeurs hors domaine, par exemple en cas de défaut de mise à zéro d'une constante (violation du domaine de valeurs)).

Certaines données présentant des problèmes de qualité peuvent être nettoyées par plusieurs méthodes de nettoyage de données lors de la phase d'ETL :

- Doublons, violation d'intégrité référentielle, violation de contrainte d'unicité globale, violation de contrainte d'unicité, format inapproprié : intégration des données ;
- Synonymes, fautes d'orthographe : standardisation des données ;
- Enregistrements manquants d'événements : imputation des données manquantes.

Les erreurs proviennent de plusieurs causes (annexe 7), et certains problèmes de qualité de données peuvent être corrigés a posteriori, d'autres nécessitent d'informer les utilisateurs afin de modifier leur pratique et améliorer ainsi l'exhaustivité des bases de données (enregistrements manquants), d'autres enfin nécessitent d'améliorer le paramétrage du logiciel (par ex. synonymes, doublons). Enfin, certains problèmes de qualité sont plus compliqués voire impossibles à corriger, car leur correction nécessiterait une action directe de l'éditeur du de logiciel.

Les problèmes de qualité peuvent se répercuter de différentes manières sur l'utilisation secondaire des données (annexe 8): certains problèmes de qualité comme les violations du domaine de valeurs ou les valeurs incorrectes peuvent fausser les résultats dans la mesure où les valeurs représentées existent mais sont fausses à cause d'un problème technique lors de leur recueil.

Enfin, il faut souligner que de nouveaux problèmes de qualité pourraient survenir dans l'avenir suite à l'utilisation de nouveaux appareils, de nouvelles interfaces ou de nouveaux logiciels.

5. Conclusion

Les systèmes sources évalués dans ce travail présentent de nombreux problèmes de qualité de données. Cependant, les données principales nécessaires pour répondre à notre problématique sont disponibles et le lien entre les différents systèmes sources pourra être réalisé lors de l'étape d'ETL. Le principal problème de qualité concerne les enregistrements pour lesquels les données sont imprécises ou incorrectes, ces données étant définitivement perdues, parce qu'aucune autre source de données n'est disponible pour corriger ces problèmes de qualités.

Les problèmes de qualité détectés lors de cette évaluation sont communs pour ce type de données (12,76), ce sont les principaux freins quant à la réutilisation secondaire des EHRs.

Les méthodes employées dans ce chapitre pourront être réutilisées à l'avenir pour évaluer d'autres systèmes sources. De plus, après mise en place d'actions correctives, la qualité des données pourra être évaluée de nouveau.

L'état des lieux réalisé lors de cette évaluation permettra de définir les différents modules d'intégration et de nettoyage des données présentés dans les chapitres suivants.

Chapitre 3 : Intégration des données

Chapitre 3 : Intégration des données

1. Introduction

Dans le chapitre 2, l'évaluation de la qualité des données a mis en évidence plusieurs problèmes de qualité de données. Cependant, les quatre systèmes étudiés peuvent être intégrés au sein d'une structure commune grâce à l'identifiant unique des patients (IPP) utilisé au CHRU de Lille.

Les données issues de systèmes sources opérationnels hétérogènes sont intégrées au sein d'un entrepôt de données par le biais du processus ETL (Extract, Transform, Load) (27,64) qui doit permettre de disposer de données de qualité optimale afin de répondre aux problématiques pour lesquelles l'entrepôt de données est développé. Ce processus ETL est composé de trois étapes majeures qui sont développées dans la première partie de ce chapitre. Chacune de ces étapes met en application des méthodes d'intégration et de nettoyage de données afin de corriger ou tout du moins maîtriser les problèmes de qualité détectés précédemment.

Les méthodes définies dans la première partie de ce chapitre ont été appliquées à l'intégration et au nettoyage des données de DIANE (mesures, médicaments, événements), de CORA (séjours, RSS, RUM, actes médicaux, diagnostics) et de GAM (séjours). L'agrégation et le calcul de nouvelles données à partir de ces données sont abordés dans le chapitre 4.

2. Méthode

Le processus ETL comporte plusieurs étapes (27,64). A chacune de ces étapes, des opérations de nettoyage des données peuvent être mises en place. Nous détaillons ci-dessous les trois étapes du processus ETL ainsi que les méthodes d'intégration et de nettoyage de données qui peuvent être employées.

2.1 Etapes

Le processus ETL est composé de trois étapes majeures (27,64) représentées figure 32 : *Extract*, *Transform* et *Load*. L'alimentation des magasins de données sera abordée dans le chapitre 4.

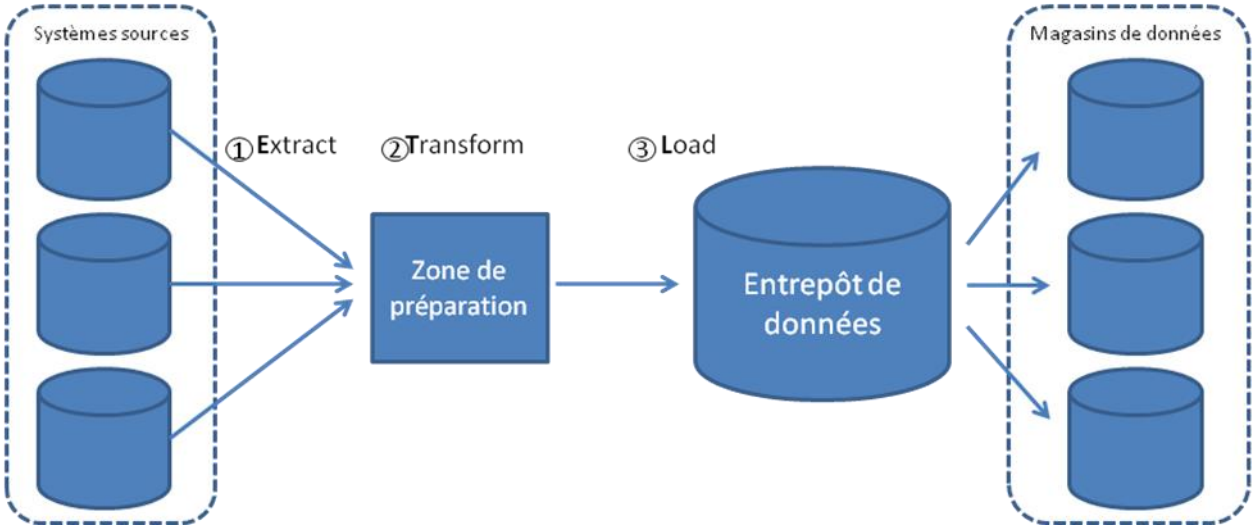


Figure 32 : Etapes du processus ETL

Extraction des données sources

La première étape du processus d'ETL consiste à sélectionner et importer des données des différents systèmes sources vers la zone de préparation (*Extract*) où elles sont enregistrées selon un schéma de base de données commun, rendant ainsi leur exploitation possible.

- Sélection des données : les systèmes opérationnels comptent généralement un grand nombre de tables : seules les tables stockant les informations pertinentes définies dans la première phase du projet (chapitre 1) sont extraites afin de ne pas surcharger inutilement le serveur.
- Sélection des champs: de même, les tables des systèmes opérationnels peuvent comporter un grand nombre de champs. Cependant, certains d'entre eux n'ont un intérêt que pour les besoins opérationnels du système.
- Sélection des enregistrements : les bases de données sources enregistrent de nouvelles données en continu. Afin de ne pas travailler sur l'historique complet de ces systèmes, les enregistrements peuvent être extraits en fonction d'une plage temporelle. Un premier filtre peut également être appliqué sur les valeurs incorrectes ou nulles dans les systèmes sources afin de ne pas importer d'enregistrements erronés.

Lors de cette première étape, les noms des entités de chaque système peuvent être renommés afin d'obtenir une homogénéité des libellés exploités ultérieurement.

Transformation et intégration

La phase de transformation et d'intégration des données est sans doute la plus compliquée du processus ETL (*Transform*) : elle consiste à transformer et nettoyer les données, puis à calculer les agrégations des données sources. Ces opérations sont généralement réalisées dans l'ordre suivant (63) et seront détaillées dans la section 2.2 :

1) Standardisation / Consolidation des données
2) Mise en application des contraintes d'intégrité
3) Remplacement des valeurs manquantes à partir des valeurs disponibles
4) Suppression des erreurs entre enregistrements ou au sein d'un même enregistrement
5) Fusion et suppression des doublons.

6) Détection des valeurs extrêmes et des enregistrements potentiellement incorrects

Ces opérations de nettoyage peuvent également être suivies par une mise au format finale et une suppression des données qui n'étaient nécessaires qu'au travail de nettoyage, habituellement des champs de calcul.

Chargement des données

La dernière étape consiste à insérer ou mettre à jour les données dans l'entrepôt de données (*Load*). Pour cela, il faut tenir compte des données déjà présentes dans l'entrepôt de données et déterminer si elles doivent être mises à jour. Puis le flux d'alimentation des nouvelles données est inséré dans l'entrepôt.

Les tables de l'entrepôt peuvent être partitionnées ou indexées (77,78) en fonction des champs utilisés par les requêtes des utilisateurs afin d'optimiser le temps de réponse. Le partitionnement et l'indexation des tables doivent également tenir compte de cette étape de chargement de données, la mise à jour des indexes pouvant être couteux en ressources en fonction des opérations réalisées (insertions, suppressions ou modification).

2.2 Méthodes de nettoyage des données

Plusieurs méthodes de nettoyage de données sont définies dans la littérature (63,67,79). Nous les détaillons dans le tableau 10 ci-dessous.

Tableau 10 : Méthodes d'intégration et de nettoyage de données pour maîtriser les problèmes de qualité de données

Méthode de nettoyage	Problèmes de qualité de données
Standardisation/consolidation des données	Différences de représentation, différences de syntaxe, synonymes, erreurs de saisies, valeurs imprécises
Enrichissement des données	Données manquantes
Intégration des données	Différences de structure, différence de représentation, différences de syntaxe
Mise en application des contraintes d'intégrité	Violation des contraintes d'intégrité
Génération d'identifiant unique	Violation de contrainte d'unicité
Détection des valeurs incorrectes	Violation du domaine de valeurs, violation des règles métier
Remplacement des valeurs manquantes	Valeurs manquantes

Standardisation/consolidation des données

La standardisation/consolidation des données (64) vise à représenter de manière homogène et cohérente une information, aussi bien du point de vue du schéma que des enregistrements.

Ainsi, comme nous l'avons montré dans le chapitre 2, une même information peut être représentée différemment dans plusieurs sources de données hétérogènes : un premier système peut représenter l'information sous forme de colonne alors que le deuxième système peut la représenter sous forme d'enregistrements. De même, les enregistrements peuvent présenter des synonymes ou des erreurs de saisies.

La standardisation/consolidation des données aura pour but d'homogénéiser les représentations en définissant une représentation homogène unique de l'information : création de terminologies, utilisation de dimensions communes (liste d'unités pour les mesures et les médicaments par exemple), format commun pour les représentations de valeurs (mesure, posologie, concentration etc...).

Enrichissement des données

La méthode d'enrichissement des données (*data enrichment*) (67) consiste à intégrer des données d'une source de référence pour compléter les données des systèmes sources principaux. Cette méthode est souvent appliquée pour intégrer des données concernant le personnel, les services ou des terminologies de référence. Ces nouvelles données permettront de proposer des axes d'analyse supplémentaires.

Intégration des données

L'intégration des données consiste à rassembler au sein d'un même schéma, des données enregistrées par des systèmes différents et stockées sous des formats hétérogènes. Pour cela, les formats de données, ainsi que les codes utilisés pour les variables (comme le 0/1 et H/F pour le sexe) sont harmonisés.

Mise en application des contraintes d'intégrité

Les techniques de mise en application des contraintes d'intégrité (*integrity constraint enforcement*) (63) ont pour objectif d'éviter les violations de contraintes d'intégrité. Ce problème de qualité défini dans le chapitre précédent survient lorsqu'un champ d'un enregistrement fait référence à un enregistrement absent d'une autre table; dans un système opérationnel, des vérifications du respect des contraintes d'intégrité peuvent être mises en place à chaque insertion, modification ou surpression d'un enregistrement. Dans le cas d'un système tel qu'un entrepôt de données, ces opérations ont déjà été effectués dans les systèmes sources. Il est cependant possible d'insérer un enregistrement par défaut à la place de l'enregistrement manquant. Ainsi, même si le champ de clé étrangère pointe vers un enregistrement fictif, la contrainte d'intégrité est respectée et l'enregistrement de la première table peut être conservé.

Dans l'exemple présenté par la figure 33, le code de mode d'entrée de l'enregistrement 157896540 n'est pas renseigné. Afin de ne pas perdre l'enregistrement et afin que la contrainte d'intégrité soit respectée, l'identifiant de mode d'entrée 0 pourra être associé à l'enregistrement lors de la phase d'ETL.

MOUVEMENT			
ID_MOUVEMENT	DATE_DEBUT	DATE_FIN	CODE_MODE_ENTREE
157896540	22/10/2012	25/10/2012	-
157896550	01/03/2013	10/03/2013	2
157896570	04/05/2013	06/05/2013	3

LISTE_MODE_ENTREE		
ID_MODE_ENTREE	CODE_MODE_ENTREE	LIB_MODE_ENTREE
0	-	Non défini
1	8	Domicile
2	85	Domicile Urgences

Figure 33 : Utilisation d'un enregistrement par défaut pour respecter une contrainte d'intégrité dans le cas d'un code de mode d'entrée manquant.

Génération d'identifiant unique

Lorsque les enregistrements des systèmes sources ne possèdent pas de clés candidates, ou que les formats de celles-ci les rendent inutilisables, il est nécessaire de définir une clé primaire artificielle (référence). La clé primaire artificielle est définie par la génération d'un identifiant unique. La plupart des SGBD permettent la génération d'identifiant unique (SEQUENCE pour Oracle, et IDENTITY pour Sql Server par exemple).

Détection des valeurs incorrectes

Certaines informations peuvent être incorrectes pour une application donnée mais utiles pour une autre application : par exemple dans Diane, certaines administrations de médicaments renseignées manuellement sont incomplètes et ne comportent que le nom du médicament sans indication de posologie; cette donnée ne pourra pas être utilisée pour mesurer la dose totale de médicament administrée, mais elle peut être utilisée par une requête qui ne nécessiterait que l'heure d'administration du médicament.

Ainsi, certaines informations pourront être identifiées et caractérisées par le biais d'un champ "qualité", et ne seront utilisées que lorsque le niveau de qualité sera adapté au niveau requis par la requête.

Remplacement des valeurs manquantes

Les données manquantes sont un problème récurrent lors de l'utilisation secondaire des bases de données, même si des méthodes automatiques d'imputation des valeurs existent (80). Nous avons aussi développé une méthode permettant de remplacer des événements temporels non documentés en utilisant d'autres événements temporels de substitution (81).

Synthèse

La décomposition du processus ETL en plusieurs étapes, et l'application de différentes méthodes d'intégration et de nettoyage des données est nécessaire pour maîtriser les problèmes de qualité identifiés dans les systèmes sources, et faciliter ainsi l'intégration des données au sein de l'entrepôt.

3. Application

Le processus d'alimentation d'un entrepôt de données doit prendre en compte les spécificités des systèmes sources telles que volumétrie et fréquence de mise à jour des données. Ainsi, la chaîne d'alimentation utilisée dans l'Infocentre d'anesthésie du CHRU de Lille est exécutée chaque semaine et intègre ou mets à jour les données des interventions sur une année glissante, ce qui permet de prendre en compte les éventuelles mises à jour des données (dédoublonnage d'interventions ou de patients).

Le processus d'alimentation des modules développés dans ce travail sera intégré à la chaîne d'alimentation déjà en place pour l'Infocentre d'Anesthésie. Comme illustré par la figure 34, le processus d'alimentation couvre deux périodes, en fonction des données extraites des systèmes sources:

- les données relatives aux mesures, aux médicaments et aux événements de l'anesthésie présentent une volumétrie importante (environ 7 000 000 mesures / semaine ainsi que 150 000 événements et médicaments/semaine) et ne sont plus mis à jour après l'enregistrement. Ces données sont intégrées chaque semaine avec un flux de deux semaines glissantes.
- les informations liées aux séjours peuvent être mises à jour plusieurs mois après la fin d'un séjour pour le codage des actes médicaux par exemple, mais représentent une volumétrie moindre. Le flux d'alimentation correspondant aux séjours peut ainsi être identique à celui utilisé pour identifier les interventions et les patients.

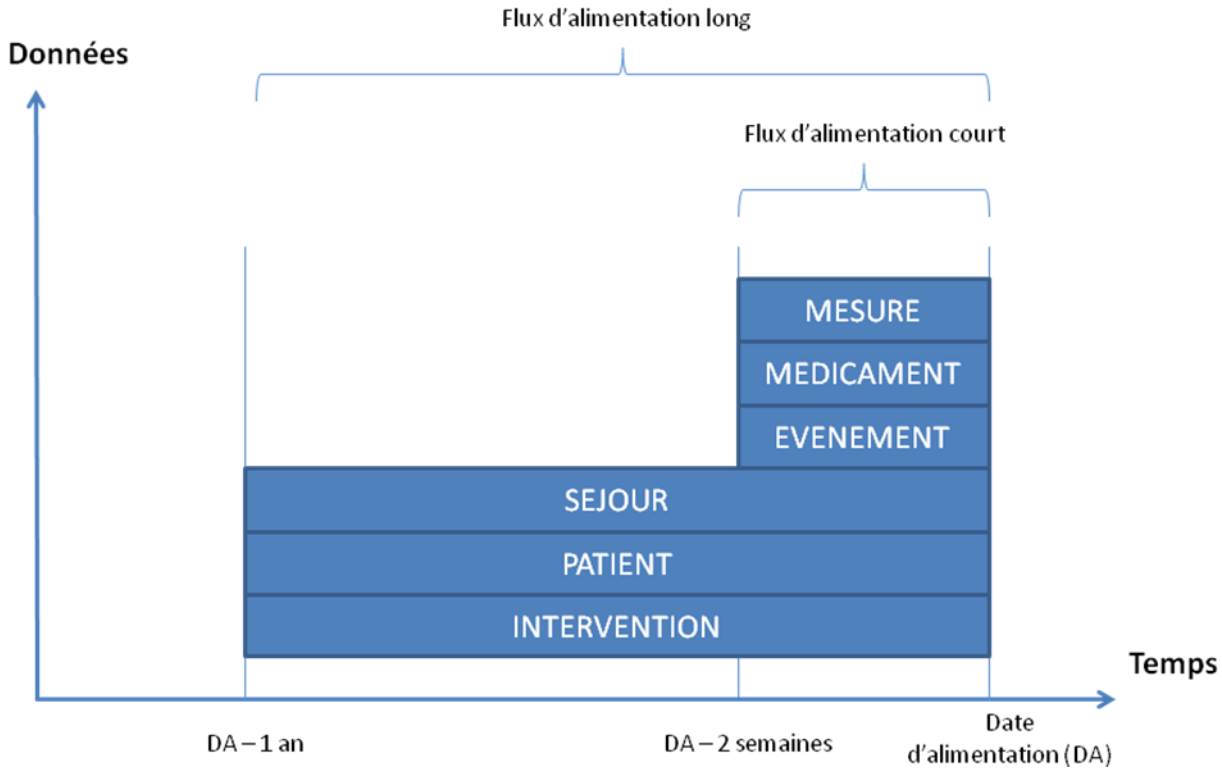


Figure 34 : Flux d'alimentation de l'entrepôt de données

3.1 Intégration des séjours

Les informations relatives aux séjours des patients sont enregistrées à partir de deux systèmes sources différents, GAM et CORA (voir chapitre 1). Le lien entre les patients enregistrées dans l'Infocentre

d'anesthésie et les séjours stockés par GAM et CORA est réalisé grâce à l'identifiant patient unique (IPP). Les différences de représentations, de syntaxe et de structure de ces deux structures avec l'Infocentre d'Anesthésie (voir chapitre 2) nécessitent de plus d'intégrer les informations stockées par ces deux systèmes en les homogénéisant. Le tableau 11 récapitule les différentes méthodes utilisées pour cela.

Tableau 11 : Méthodes utilisées pour le nettoyage et l'intégration utilisée des données de séjours

Etapes	Méthode de nettoyage de données	Problèmes de qualité
Extraction	Sélection des enregistrements et des champs - Mise au format	Format inapproprié Différence de structure / représentation / syntaxe
Transformation et intégration	Mise en application des contraintes d'intégrité	Violation d'intégrité référentielle
	Identification des valeurs incohérentes	Violation du domaine de valeurs Violation des règles métiers
	Dédoublonnage	Violation de contrainte d'unicité globale
Chargement	Alimentation des tables de l'entrepôt de données	-

Extraction des données sources

Les données de GAM et CORA relatives aux séjours sont extraites en fonctions des dates de début et de fin de séjour et des dates du flux d'alimentation long présenté dans la figure 34.

Transformation et intégration

L'Infocentre d'anesthésie, GAM et CORA possèdent chacun une table PATIENT, qui représente l'information commune utilisée dans ce travail d'intégration de données. Pour chaque patient dont les données sont stockées dans l'infocentre, le ou les séjours qui lui sont attachés dans GAM et CORA, les RSS, RUM, actes médicaux et diagnostics associés doivent être correctement extraits. Dans un second temps, chaque intervention doit être reliée au séjour pendant lequel elle a été réalisée.

Lien Patient - Séjour

Le travail d'évaluation de la qualité des données a permis de mettre en évidence la présence de doublons similaires au sein des bases de données sources. Dans le cas de GAM et CORA, ces doublons sont produits par des créations répétées d'enregistrements d'un même patient et d'un même séjour. De plus, les patients et séjours enregistrés dans CORA sont théoriquement également enregistrés dans GAM, mais CORA propose davantage d'informations sur le séjour de chaque patient que GAM (voir chapitre 1).

Le travail de liaison des séjours de CORA et GAM avec les patients de DIANE est réalisé en 3 étapes (figure 35). Les patients de l'Infocentre sont d'abord reliés aux patients de CORA et GAM grâce à l'IPP (étape 1), puis les séjours des patients de CORA et GAM sont rattachés aux patients correspondants dans DIANE (étape 2). A ce stade, un patient peut être relié à un ou plusieurs séjours, de même qu'un séjour peut être relié à un ou plusieurs patients. Il s'agit ensuite de sélectionner pour chaque séjour, le patient le plus probable. Pour cela, quand plusieurs patients sont concurrents pour un même séjour, le séjour est relié au patient dont l'IPP est le plus petit, parce qu'il correspond à l'IPP créé en premier. Selon ce processus, le séjour

sera relié au premier enregistrement du patient (étape 3). Au total, 1 902 021 de séjours de GAM et CORA ont ainsi été reliés à des patients (tableau 12).

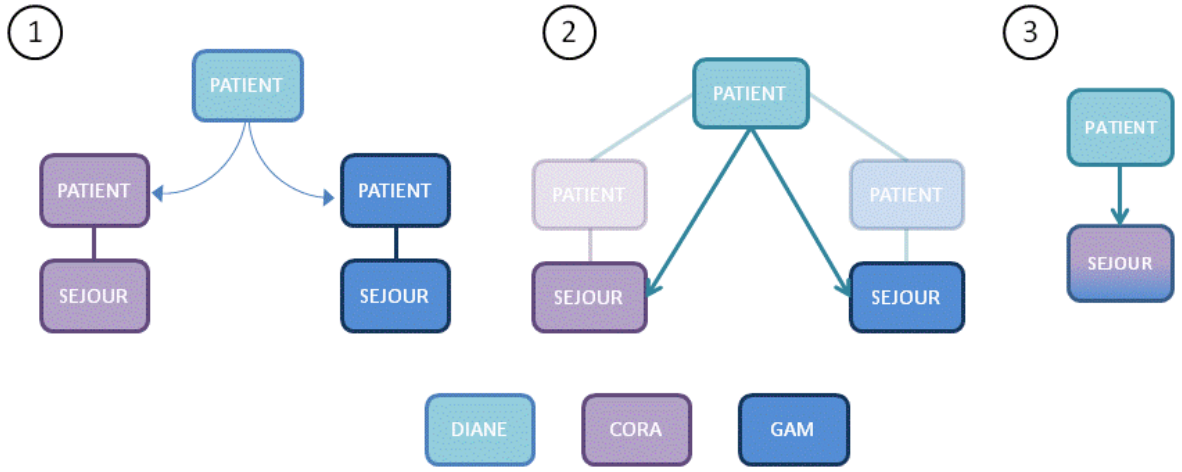


Figure 35 : Etapes de l'intégration des séjours de CORA et GAM avec les patients de DIANE

Tableau 12 : Nombre de séjours liés à des patients

2010-2012	Nombre de séjours sélectionnés (%)
GAM	1 347 849 (70,9%)
CORA	296 733 (19,1%)
Total	1 902 021

RSS

Les RSS des 296 733 séjours en provenance de CORA sont intégrés. L'identifiant de GHM/GHS est mis à jour afin de respecter la contrainte d'intégrité référentielle. Au total 399 497 RSS ont été intégrés.

Mouvement, actes médicaux et diagnostics

Les mouvements liés au RSS sont également intégrés. Ceux-ci comportent les identifiants des modes d'entrée, modes de sortie, unité de soins ainsi que sa qualification (soins intensifs, réanimation, ...).

Lors de cette étape, plusieurs informations sont enrichies grâce à des sources externes. La hiérarchie des unités de soins (du pôle à l'unité fonctionnelle), la classification des actes médicaux (du chapitre à l'acte, dans la CCAM) et la classification des diagnostics (du chapitre au diagnostic dans la CIM10) sont intégrées, ce qui permettra de filtrer ou de regrouper différentes interventions en fonction de la granularité de ces trois dimensions. Il sera ainsi possible de sélectionner les interventions par services, ou par type de chirurgie.

Au total, 502 721 mouvements ont été intégrés, 2 211 038 actes médicaux et 1 569 546 diagnostics.

Lien Séjour - Intervention

Les interventions et les séjours étant maintenant liés aux patients correspondants (Etape 1, figure 36), un patient peut cependant être lié à plusieurs séjours (tableau 13) : il est donc nécessaire de faire le lien entre le séjour et l'acte médical.

Tableau 13 : Récapitulatif du nombre de séjours par patient

Lien séjour	Nombre de patients (%)
Patient sans lien	743 (0.6%)
Patients liés à 1 séjour	3041 (2.8%)
Patients liés à 2 séjours ou plus	106614 (96.6%)

Ce lien est facilement établi puisque la date où l'acte est réalisé est nécessairement comprise entre la date de début et la date de fin du séjour.

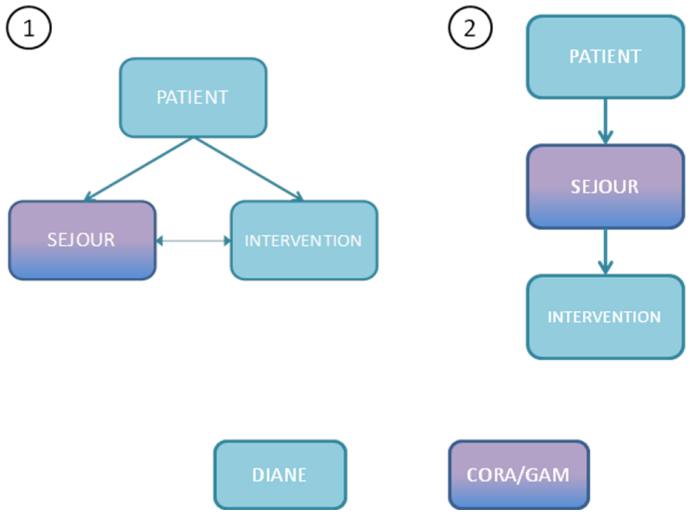


Figure 36 : Etapes de l'intégration des séjours de CORA et GAM et des interventions de DIANE

Si malgré tout plusieurs séjours sont concurrents pour une même intervention, la règle de priorité suivante est appliquée :

1) séjour en provenance de CORA : les autres informations (RSS, Mouvements, etc...) sont disponibles contrairement à GAM.

2) séjour le plus court : Si un patient bénéficie de soins extérieurs, et que lors de cette période il est hospitalisé pour une intervention, deux séjours seront répertoriés. Le séjour le plus long, correspond aux soins extérieurs, englobant le séjour lié à l'intervention. Dans ce cas, le séjour le plus court est sélectionné car il correspond à l'hospitalisation en Médecine-Chirurgie-Obstétrique pour l'intervention.

3) séjour avec l'IEP le plus récent.

En suivant cette règle, 155 899 (96,8%) interventions réalisées entre 2010 et 2012 ont été liées à un séjour unique (voir tableau 13 et étape 2 figure 36). Comme le précise le tableau 14, les interventions sont majoritairement liées à des séjours CORA.

Tableau 14 : Origine des séjours liés aux interventions

Origine du séjour	Nombre d'interventions (%)
Séjour CORA	153,370 (95,3%)
Séjour GAM	2 529 (1,5%)
Pas de séjour	5 103 (3,2%)
Total	161002

Chargement des données

La figure 37 représente le schéma logique de répartition des données des séjours après intégration depuis les sources CORA et GAM dans l'entrepôt d'anesthésie.

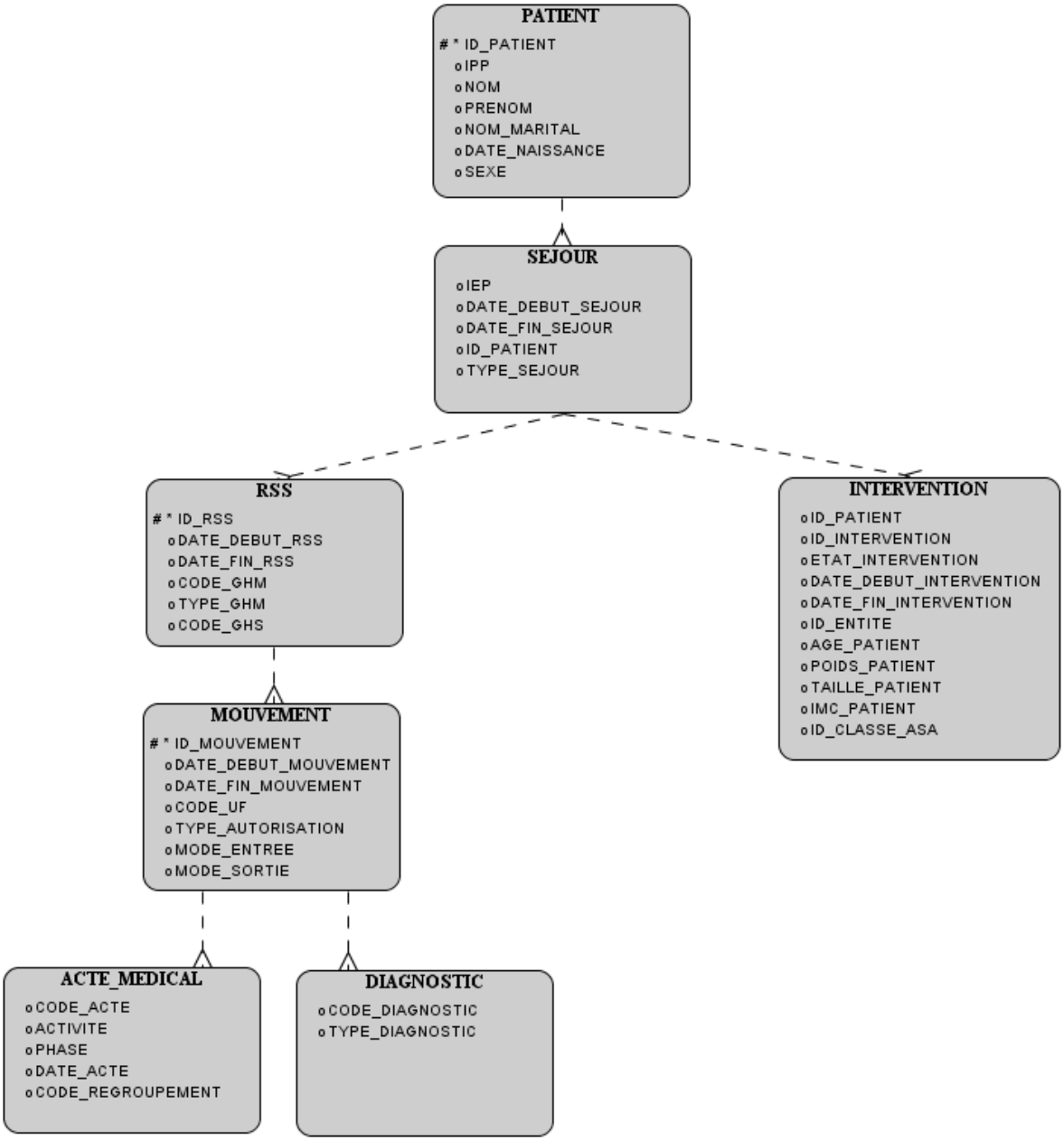


Figure 37 : Schéma logique de données - Séjours

3.2 Intégration des mesures

Le travail d'évaluation de la qualité des données (chapitre 2) a mis en évidence plusieurs problèmes de qualité concernant les mesures. Dans le tableau 15, le travail d'intégration des mesures est développé en plusieurs étapes. Les problèmes de qualité traités à chacune de ces étapes sont précisés, ainsi que les méthodes mises en œuvre.

Tableau 15 : Méthodes utilisées pour le nettoyage et l'intégration utilisée des données de mesures

Etape	Méthode	Problèmes de qualité
Extraction	Sélection des enregistrements et des champs - Mise au format	Format inapproprié
Transformation et intégration	Mise en application des contraintes d'intégrité	Violation d'une contrainte d'intégrité
	Consolidation des données	Synonymes, valeurs imprécises d'unités et paramètre
	Conversion pour obtenir des unités homogènes	Violation de dépendance fonctionnelle
	Identification des valeurs incohérentes	Violation du domaine de valeurs
	Dédoublonnage	Violation de contrainte d'unicité globale
Chargement	Alimentation d'une table partitionnée pour gérer la volumétrie des données	-

Extraction des données sources

Lors de la phase d'extraction des données, plusieurs filtres sont appliqués pour sélectionner les enregistrements en fonction de plusieurs critères :

- Sélection des enregistrements : filtre sur la date de la mesure. En raison de la volumétrie, le flux d'alimentation des mesures est de deux semaines.
- Mise au format des champs (partie entière du champ VALEUR) : longueur de la partie entière inférieure à 12.
  Mise au format des champs (partie décimale du champ VALEUR): arrondissement de la partie décimale à deux chiffres afin de limiter la volumétrie de ce champ. L'application Diane enregistre les mesures avec une précision pouvant aller jusqu'à 10 chiffres après la virgule, alors que pour certains paramètres, la précision suffisante est de deux chiffres après la virgule.

Transformation et intégration

La phase de transformation et d'intégration des mesures correspond à l'application de plusieurs méthodes de nettoyage des données :

- Mise à jour des identifiants pour respecter les contraintes d'intégrité référentielle (intervention, unité, paramètre).
- Consolidation des unités et paramètres des mesures pour prendre en compte les synonymes et les valeurs imprécises. Génération d'un identifiant unique propre à l'entrepôt de données (l'identifiant source est conservé).
- Vérification de la cohérence entre le paramètre et l'unité de la mesure. Les valeurs de certaines mesures peuvent être converties si plusieurs unités différentes sont possibles pour le paramètre.
- Vérification de la cohérence entre la valeur de la mesure et la plage de valeurs possibles associée au paramètre.
- Suppression des doublons.

Chargement des données

La table finale de restitution des mesures est partitionnée par valeurs du champ DATE_MESURE pour faciliter le chargement des données (77). Ainsi, à chaque alimentation les partitions correspondant aux deux semaines de chargement sont tronquées (82). Cette opération vide les partitions considérées sans supprimer la structure physique, rendant l'opération efficace en termes de temps de traitement. Les données sont ensuite rechargées dans les deux partitions avec le flux d'alimentation.

Afin d'optimiser les temps de réponse lors des interrogations des mesures, un index est créé sur les champs les plus utilisés pour les requêtes (dans l'ordre, l'identifiant d'intervention, l'identifiant du paramètre mesure, l'identifiant du protocole de l'appareil de mesure, et la date de la mesure). Afin d'éviter la mise à jour de l'index dans son entier à chaque suppression ou insertion de lignes dans la table, l'index est également partitionné sur la date de la mesure.

La figure 38 représente le schéma logique de données la partie MESURE. La table MESURE est liée aux tables INTERVENTION, PARAMETRE et UNITE et stocke la valeur de la mesure, ainsi que la date d'enregistrement et un champ DOMAINE_VALEURS qui précise si la valeur est comprise dans le domaine de valeurs associées au paramètre.

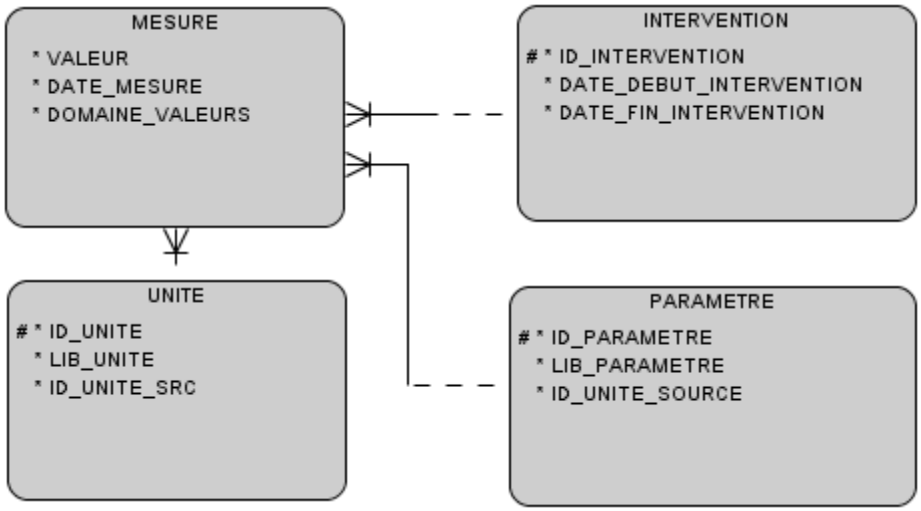


Figure 38 : Schéma logique de données - Mesures

3.3 Intégration des médicaments et des événements

Les enregistrements de médicaments et d'événements (étapes de l'intervention, matériel utilisé, etc.) sont enregistrés dans DIANE dans deux tables distinctes à partir de sélection d'un intitulé dans un menu ou par saisie manuelle dans un champ de texte libre. Ainsi, des informations relatives à des administrations de médicaments peuvent être enregistrées dans la table d'événements. Comme l'a mis en évidence l'évaluation de la qualité des données, les saisies manuelles dans les champs de textes libres génèrent également des violations de contrainte d'intégrité référentielle, des erreurs de saisies ou l'emploi de synonymes.

Le tableau 16 résume l'ensemble des méthodes d'intégration et de nettoyage de données qui seront employées lors de l'alimentation d'un module commun aux enregistrements d'événements et de médicaments.

Tableau 16: Méthodes utilisées pour le nettoyage et l'intégration utilisée des données de médicaments et d'événements

Etape		Problèmes de qualité
Extraction	Intégration des enregistrements d'événements et de médicaments Mise au format	Format inapproprié
Transformation et intégration	Application des contraintes d'intégrité	Violation d'une contrainte d'intégrité
	Consolidation des données	Synonymes, Erreurs de saisies, Valeur imprécise
	Dédoublonnage	Violation de contrainte d'unicité
Chargement	Chargement dans une table unique	

Extraction des données sources

Les données relatives aux enregistrements de médicaments et d'événements sont extraites des deux tables sources de DIANE en fonction de la date de saisie de l'enregistrement et des dates du flux d'alimentation cours (voir figure 34). Un premier filtre permet de ne conserver que les enregistrements présentant des champs documentés. Les données des deux sources sont adaptées à un format commun.

Transformation et intégration

Les données des deux sources sont intégrées et adaptées à un format commun, correspondant à la table EVENEMENT_SRC dans la figure 39. Un identifiant est généré pour identifier de manière unique l'enregistrement d'événement ou de médicament, qu'il ait été saisie manuellement ou sélectionné dans un menu pré-configuré.

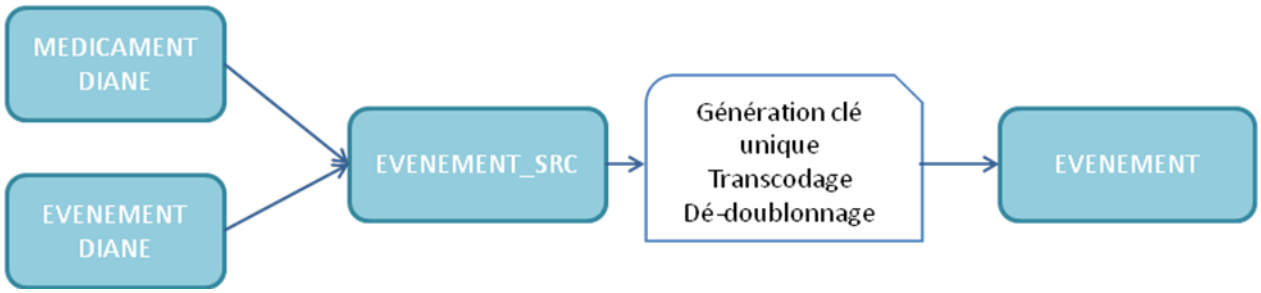


Figure 39 : Chaîne d'alimentation des médicaments et événements

A ce stade, l'information n'est pas encore exploitable parce que les intitulés d'événements sources peuvent comporter des synonymes, des erreurs de saisies ou des informations incomplètes (voir chapitre 2): c'est pourquoi l'événement source est transcodé vers une terminologie de référence, représentant les médicaments et événements liées à la procédure d'anesthésie sans synonymes ou erreurs de saisies. Ce transcodage est réalisé soit manuellement en reliant les intitulés d'EVENEMENT_SRC aux intitulés d'EVENEMENT de la terminologie, soit de façon semi-automatique (83) en recherchant des mots clés dans les intitulés de saisies manuelles et en les reliant automatiquement à des intitulés d'EVENEMENT.

Exemple d'application : le tableau 17 représente des exemples de transcodage entre intitulés d'EVENEMENT_SRC et intitulés d'EVENEMENT issus de la terminologie, pour les interventions réalisées entre 2010 et 2012. Au total, 8 067 713 événements ont été transcodés pour 161 001 interventions, ce qui rend ces informations directement exploitables puisque relier à une la terminologie d'événements. Les données dont le transcodage n'a pas été possible sont conservées et pourront ainsi être transcodés à l'avenir.

Tableau 17 : Table de transcodage entre les événements et médicaments renseignés dans DIANE et la terminologie d'événements définies pour l'entrepôt de données

EVENEMENT_SRC	EVENEMENT
Sufenta 10µg	Sufentanil
Sufantanil	Sufentanil
Sfentanil	Sufentanil
Début de chir	Sufentanil

Chargement des données

L'agrégation par médicament est réalisée dans le chapitre suivant. La figure 40 représente le schéma logique de données pour les événements. La table EVENEMENT est la terminologie utilisée pour représenter les événements liés à la procédure d'anesthésie. La table EVENEMENT_SRC contient tous les intitulés enregistrés dans DIANE, issus des menus ou des champs de saisies libre. La table INTERV_EVENEMENT contient les occurrences d'événements pour chaque intervention. Les champs POSOLOGIE, CONCENTRATION, TOTAL, et les trois champs d'unités correspondants ne sont renseignés que lorsque l'information est une administration de médicament.

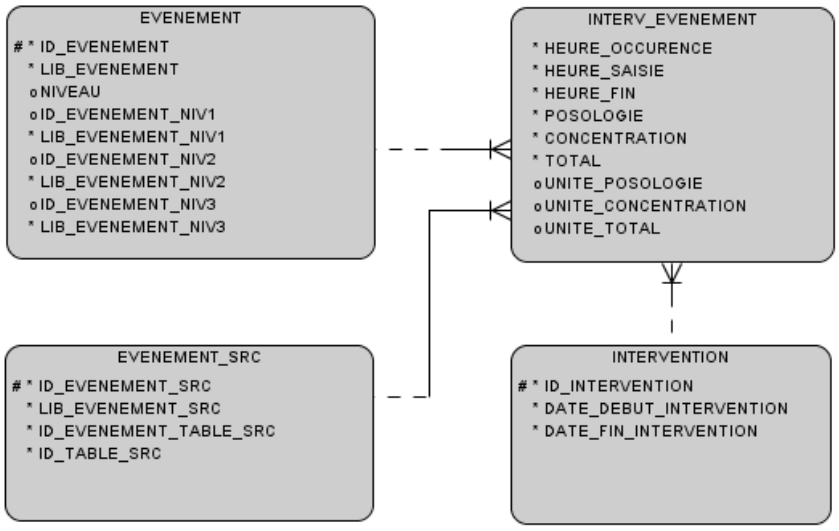


Figure 40 : Schéma logique de données - Evénements

4. Discussion

Dans ce chapitre, les méthodes de nettoyage et d'intégration de données décrites dans la littérature ont été employées en suivant les trois étapes d'extraction, de transformation et de chargements des données.

Les données enregistrées dans les systèmes sources ont pu être intégrées dans une structure commune au sein de l'Infocentre d'anesthésie. Les liens ont été réalisés entre les données déjà présentes dans l'Infocentre et les nouvelles données importées depuis la feuille informatisée d'anesthésie DIANE (mesures, événements et médicaments), le logiciel de PMSI CORA et le logiciel administratif GAM.

Trois types de solutions peuvent être mises en place pour traiter les données présentant des problèmes de qualité :

- Celles-ci peuvent être modifiées si une méthode permet de remédier au problème de qualité (génération d'un identifiant unique, mise au format, application des contraintes d'intégrité ...) ;
- Les données peuvent être conservées et identifiées comme présentant un problème de qualité si aucune méthode ne permet de remédier au problème de qualité mais que la donnée apporte néanmoins une information qui peut être utilisée par une requête. C'est le cas des mesures avec une unité incorrecte : celles-ci sont conservées dans une table distincte et les mesures en dehors du domaine de valeurs sont caractérisées par un champ qualité ;
- Enfin les données peuvent être supprimées si elles n'apportent aucune information (doublons).

Plusieurs éléments doivent être pris en compte lors de cette phase d'ETL pour garantir la robustesse de la chaîne d'alimentation :

- Certains problèmes de qualité peuvent ne pas avoir été détectés lors de l'évaluation de la qualité des données et apparaître à l'avenir avec des changements dans les pratiques d'utilisation des logiciels, le paramétrage des applications, ou la volumétrie des données. La chaîne d'alimentation doit tenir compte de ces éventuels problèmes et permettre l'intégration de nouvelles méthodes de nettoyage ou intégrer dès à présent certains modules même si les problèmes de qualité correspondant n'ont pas été détectés. Dans le cas de l'entrepôt de données d'anesthésie, tous les liens entre les tables sont vérifiés, des identifiants uniques sont générés pour chaque dimension associée aux mesures, aux événements et aux mouvements. Enfin chaque type de données est dédoublonné.
- L'augmentation du temps de chargement peut être exponentielle avec la volumétrie des données. Les traitements d'alimentation doivent donc être indépendants de l'historique des bases de données, afin de conserver des temps d'exécution constants dans le temps. C'est pourquoi l'alimentation des mesures travaille toujours sur deux partitions de la table, indépendantes des autres partitions déjà complètes dans la base. De même, l'index de la table n'est mis à jour que sur les données des deux partitions et ne nécessite pas d'être reconstruit dans son ensemble à chaque alimentation. Ainsi, le temps de chargement de cette table sera toujours constant.

Ce chapitre ne traite que de l'intégration et du nettoyage des données sources au sein de l'entrepôt de données. Le calcul des données agrégées et l'alimentation des magasins de données (voir figure 1) sont abordés dans le chapitre 5. D'autres problèmes pourront être traités lors de cette phase d'agrégation.

5. Conclusion

La phase d'ETL est une étape clé du développement d'un entrepôt de données. En effet, ce processus permet d'intégrer les données enregistrées par plusieurs applications au sein d'une structure commune. Des méthodes de nettoyage sont appliquées pour maîtriser et qualifier la qualité des données intégrées.

Les informations sont représentées en suivant le même formalisme, quel que soit les systèmes qui les ont enregistrées et le schéma résultant de cette intégration est indépendant des systèmes opérationnels sources. Il est ainsi possible d'interroger conjointement les données provenant de plusieurs systèmes.

Cette étape doit tenir compte de la volumétrie des données afin de conserver une durée d'alimentation correcte au fur et à mesure du temps : pour cela, les flux d'alimentation sont sélectionnés pour optimiser les temps d'alimentation : les types de données avec une volumétrie importante mais une fréquence de mise à jour nulle sont alimentés avec un flux temporel court, alors que les types de données avec une faible volumétrie mais une fréquence de mise à plus élevée doivent être alimentés avec un flux temporel plus long.

L'objectif du processus ELT est de disposer d'informations consolidées qui pourront ensuite être exploitées pour calculer des données agrégées ou des indicateurs et alimenter des magasins de données. Ce travail sera abordé lors du prochain chapitre.

Chapitre 4 : Données agrégées

Chapitre 4 : Données agrégées

1. Introduction

Dans le chapitre 3, nous avons présenté l'intégration et le nettoyage des données sources enregistrées dans DIANE, GAM et CORA (mesures, événements, médicaments, séjours) qui constituent le socle de données de l'entrepôt de données: les données y sont stockées sous une forme élémentaire et détaillée, ce qui offre à l'utilisateur une vue transversale sur les informations des systèmes opérationnels intégrés. Ainsi, pour un séjour donné, les tables MESURE et MOUVEMENT répertorient respectivement toutes les mesures réalisées au cours des interventions sous anesthésie au cours de ce séjour, et les différents mouvements du patient entre unités de soins.

L'entrepôt de données semble être *a priori* le bon endroit pour rechercher un lien entre un paramètre donné (par exemple le volume courant en cas de ventilation mécanique) et l'augmentation de la durée du séjour ou le passage les soins intensifs. Mais le grand nombre de paramètres rend les différentes phases de filtrage et d'agrégation des données très chronophages ; c'est pourquoi les différentes données de l'entrepôt gagneront à être filtrées et agrégées vers des *magasins de données* qui permettent de les regrouper selon des sous-ensembles définis par l'utilisateur en fonction de la question clinique posée (fig. 41)(26,27). Dans l'exemple de l'étude de l'impact d'un grand volume courant sur la morbi-mortalité, le volume courant et l'unité où est hébergé un patient en post-opératoire sont regroupés dans le même magasin, facilitant ainsi la recherche d'une relation statistique entre les deux.

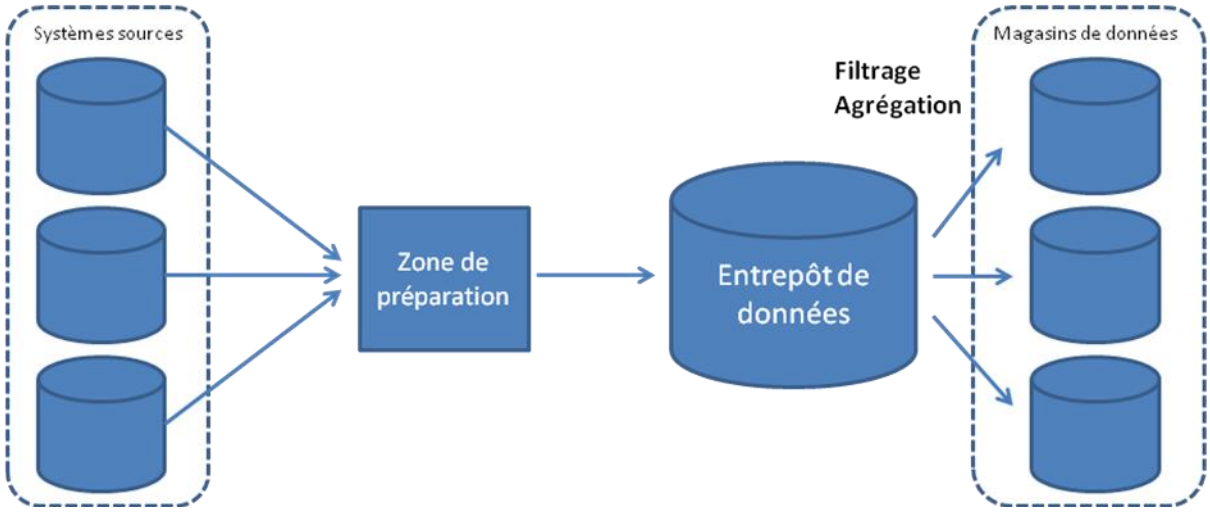


Figure 41 : Agrégation des données de l'entrepôt de données vers des *magasins de données*

Dans ce chapitre nous présentons le développement de quatre modules de *magasins de données* qui permettent de (i) filtrer les informations sur des plages temporelles prédéfinies de l'anesthésie et de la chirurgie, (ii) proposer une agrégation des mesures, (iii) détecter des événements indésirables et mesurer leur durée, et enfin (iv) calculer les doses cumulées des différents médicaments administrés au cours d'une période d'intérêt donnée.

Ces données agrégées au sein de *magasins* propres permettront de faciliter et accélérer les réponses aux requêtes des utilisateurs.

2. Fenêtre d'étude

Afin d'étudier les variations hémodynamiques du patient au cours de l'intervention ou autour d'un événement particulier, il est nécessaire de déterminer des plages temporelles précises (figure 42), permettant de calculer divers indicateurs comme l'hypotension (46). Dans la suite de ce travail, chaque plage temporelle d'intérêt sera appelée "fenêtre d'étude". Ces "fenêtres d'étude" sont délimitées par des événements définis *a priori*. Par exemple, une fenêtre d'étude "anesthésie" peut être définie entre l'événement "induction" et l'événement "fin d'anesthésie". Une définition plus générale peut être faite lorsqu'il s'agit de mesurer l'effet de l'administration d'un médicament : la détection d'une administration entraîne la création d'une fenêtre d'étude qui peut "encadrer" l'administration de ce médicament, et permettre ainsi de comparer l'évolution de certains paramètres avant et après cette administration. Un exemple d'application est donné figure 42 pour l'administration de sufentanil. En cas d'administrations multiples, il y aura autant de fenêtres d'étude que d'administrations. Les événements sélectionnés pour déterminer la borne de début et la borne de fin d'une fenêtre d'étude sont qualifiés d'éléments "déclencheurs". Figure 42, les événements "Entrée au bloc", "Induction", "Incision", "Administration de sufentanil" sont les éléments déclencheurs des fenêtres d'étude "Bloc", "Anesthésie", "Chirurgie", "Sufentanil".

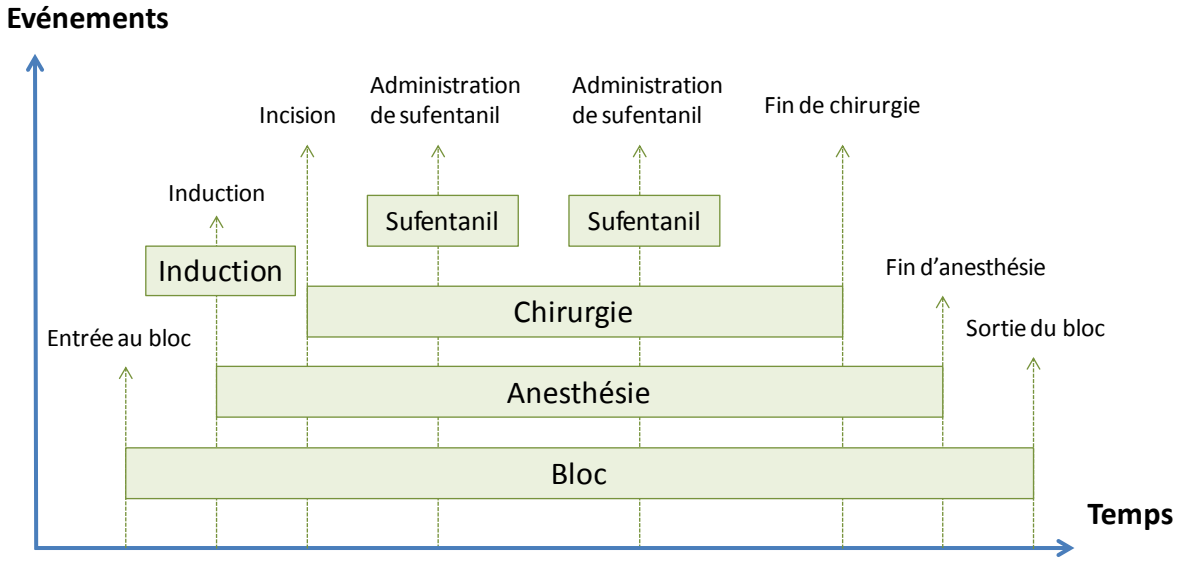


Figure 42 : Exemples des périodes d'intérêt au cours d'une intervention

Dans le cas de l'administration d'un médicament, la fenêtre d'étude afférente est définie *a priori* : elle débute quelques minutes avant l'administration du médicament, et termine quelques minutes après.

La limite principale des fenêtres d'études est que certains événements clés des procédures réalisées sous anesthésie ne sont parfois par disponibles dans la feuille d'anesthésie informatisée, et donc dans l'entrepôt (cf Evaluation de la qualité des données, chap. II). Lorsque les événements déterminant une fenêtre d'étude ne sont pas présent dans un enregistrement, on peut recourir à d'autres événements afin d'estimer la position probable de l'événement manquant, et éviter ainsi de devoir exclure l'enregistrement de l'analyse considérée.

2.1 Méthode

Deux méthodes de calcul sont mises en place pour fixer les bornes de début et de fin de chaque fenêtre d'étude, selon qu'elles nécessitent un ou plusieurs événements déclencheurs.

Le premier type de fenêtre suit la méthode développée dans (81). Les bornes de début et de fin de la fenêtre sont déterminées par un ensemble d'éléments déclencheurs, agrégés en fonction d'un indicateur de priorité. Un seul élément est conservé pour la borne marquant le début de la fenêtre ; idem pour de fin de fenêtre. Ce type de fenêtre ne peut le plus souvent survenir qu'une fois au cours d'une intervention (figure 42).

Le second type de fenêtre ne fait référence qu'à un seul élément déclencheur et ne réalise pas d'agrégation. Ce type de fenêtre temporelle peut se répéter plusieurs fois au cours d'une même intervention, à chaque fois que l'élément déclencheur est enregistré, par exemple l'administration de sufentanil figure 42.

Pour les deux méthodes de calcul, la date d'occurrence des événements déclencheurs peut être décalée d'une durée prédéfinie ce qui permet de définir une fenêtre "encadrant" un événement. Un exemple d'application serait l'étude des effets de l'induction anesthésique : l'événement "induction" est alors encadré par une fenêtre d'étude débutant à T+0 et terminant à T+15min (fig. 43). Un autre exemple d'application est l'étude de l'effet de l'atropine, avec une fenêtre d'étude débutant à T-10min et terminant à T+30min (fig 43).

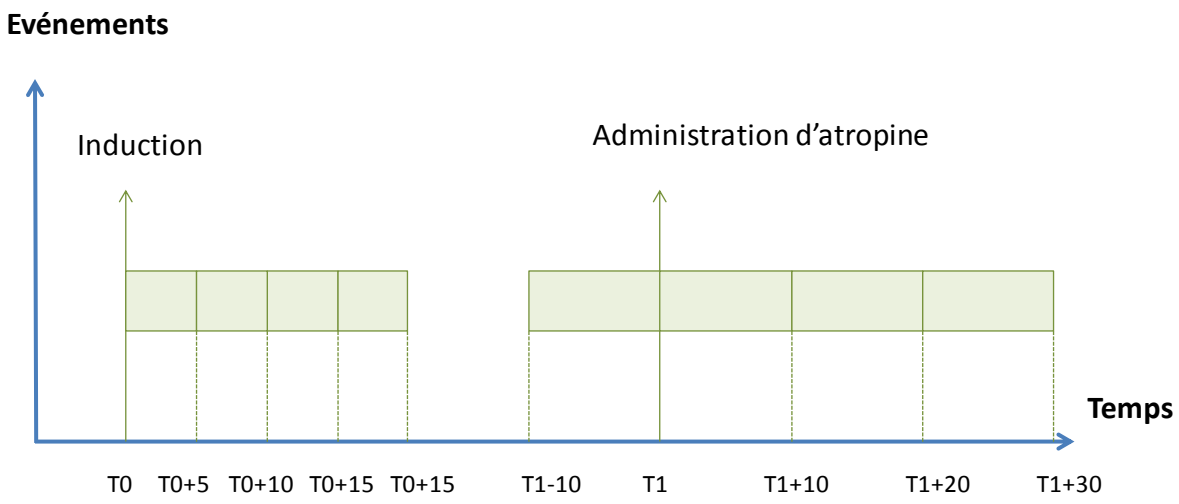


Figure 43 : Définition de fenêtres d'étude autour des événements "Induction" et "Administration d'atropine"

Le tableau 18 présente plusieurs exemples de fenêtres d'étude et les événements déclencheurs associés à ces fenêtres.

Tableau 18 : Exemples de fenêtres d'étude et des événements déclencheurs qui y sont associés

Fenêtre d'étude	Evénements déclencheurs
Bloc	1er enregistrement au bloc opératoire Dernier enregistrement au bloc opératoire
Anesthésie	Induction, 1ère administration d'hypnotiques, ... Fin d'anesthésie, sortie du bloc opératoire, ...
Chirurgie	Incision Fin de chirurgie
Début d'anesthésie - Début de chirurgie	Induction Incision
Induction [-10 min ; 10 min]	Induction
Incision [-10 min; 10 min]	Incision
Administration de sufentanil [-10 min; 0]	Administration d'atropine
Administration de sufentanil [0 ; 10 min]	Administration d'atropine

2.2 Résultats

Modèle de données

La figure 44 représente le modèle logique de données utilisé pour calculer les fenêtres d'étude. Il comporte deux tables de dimensions : FENETRE_ETUDE et BORNE_FENETRE_ETUDE, cette dernière modélisant les différents éléments déclencheurs d'une fenêtre d'étude. Les résultats des calculs sont modélisés par la table INTERV_FENETRE_ETUDE.

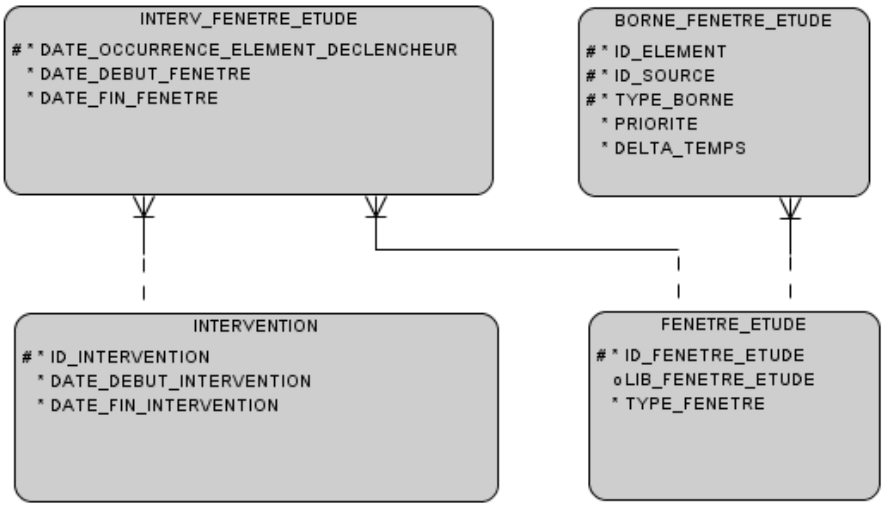


Figure 44 : Modèle logique de données - Fenêtres d'étude

Applications

Le tableau 19 présente le nombre de fenêtres d'étude définies dans le tableau 18, qui ont pu être déterminées pour les interventions réalisées du 01/01/2010 au 31/12/2012.

Tableau 19 : Nombre de fenêtres d'étude - Interventions de 2010 à 2012

Libellés des fenêtres d'étude	Nombre de fenêtres d'étude
Bloc	158284
Anesthésie	156499
Chirurgie	124569
Début d'anesthésie - début de chirurgie	123722
Induction [-10 ; 10]	156755
Incision [-10 ; 10]	156477
Administration d'atropine	21805
Administration de sufentanil	86972

Une analyse détaillée par spécialité chirurgicale permet d'utiliser les fenêtres d'études ainsi définies pour affiner l'analyse des durées des différentes fenêtres: le tableau 20 présente les résultats obtenus pour la durée médiane [1° quartile - 3° quartile] d'anesthésie et de chirurgie.

Tableau 20 : Durée d'anesthésie et de chirurgie par spécialité (présentation des données en médiane [1° - 3° quartile] ; minutes)

Service	Durée d'anesthésie	Durée de chirurgie
Chirurgie viscérale	164 [98;256]	117 [59;199]
Centre de Traitement des Brûlés	68 [36;144]	41 [26;70]
Chirurgie cardio vasculaire	204 [119;273]	148 [76;213]
Chirurgie Pédiatrique	64 [37;109]	31 [15;63]
Chirurgie Thoracique	191 [112;287]	118 [57;204]
Centre médico chirurgical ambulatoire	34 [20;61]	22 [13;39]
Chirurgie gynécologique	73 [35;145]	44 [15;108]
Neurochirurgie	177 [123;265]	96 [54;165]
Neuroradiologie	68 [46;129]	49 [31;89]
Obstétrique	89 [63;347]	45 [34;58]
Chirurgie ophtalmologique	61 [44;85]	33 [20;55]
Chirurgie ORL	92 [56;137]	52 [24,75;93]
Procréation médicalement assistée	16 [11;23]	9 [6;12]
Chirurgie maxillo faciale et reconstructrice	112 [72;177]	72 [39;129]

3. Mesures agrégées

Chaque intervention enregistre plusieurs milliers de mesures pour plusieurs dizaines de paramètres. Ces mesures ne sont pas exploitables en l'état, mais leur agrégation permet de caractériser l'évolution de chaque paramètre en fonction du temps ou d'un événement (par exemple analyse des variations hémodynamiques autour d'une administration de médicament).

3.1 Méthode

L'agrégation des mesures d'un paramètre nécessite que soient définis : la fenêtre d'étude, le paramètre à mesurer et la fonction mathématique à appliquer. Différentes fonctions mathématiques sont facilement disponibles, comme par exemple avec le SGBD Oracle (72) qui propose différents types de fonctions d'agrégation (84) ou analytiques (85) (Annexe 10).Seules les valeurs interprétables (cf. section Intégration des mesures, chapitre II) sont retenues lors de l'analyse.

La figure 45 illustre les étapes de traitement des données pour leur agrégation : (i) les mesures sont sélectionnées en fonction de paramètres prédéterminés dans des fenêtres d'études prédéfinies, puis (ii) les mesures de qualité suffisante sont agrégées et enfin, (iii) une requête pivot (86) met en forme le résultat afin d'obtenir une ligne par intervention et une colonne par mesure agrégée.

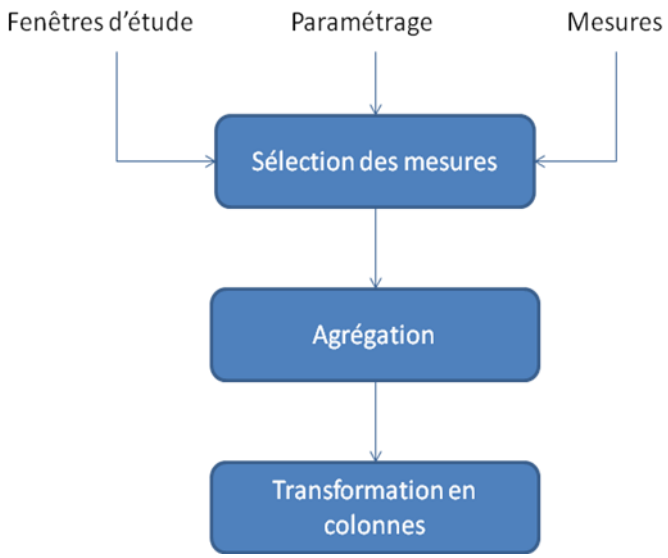


Figure 45 : Etapes du calcul des mesures agrégées

3.2 Résultats

Modèle de données

Le modèle de données utilisé pour calculer les mesures agrégées fait appel à trois nouvelles tables: la table PARAMETRE_MESURE_AGREGEE est la table de dimensions listant les paramètres de calcul de chaque mesure agrégée (paramètre mesuré, fonction, fenêtre d'étude) (27). La table INTERV_MESURE_AGREGEE permet de stocker les résultats des calculs d'agrégation pour chaque mesure

agrégée. La table MESURES_AGREGEES réalise simplement la transformation de la table précédente de lignes en colonnes afin de permettre une interrogation plus rapide.

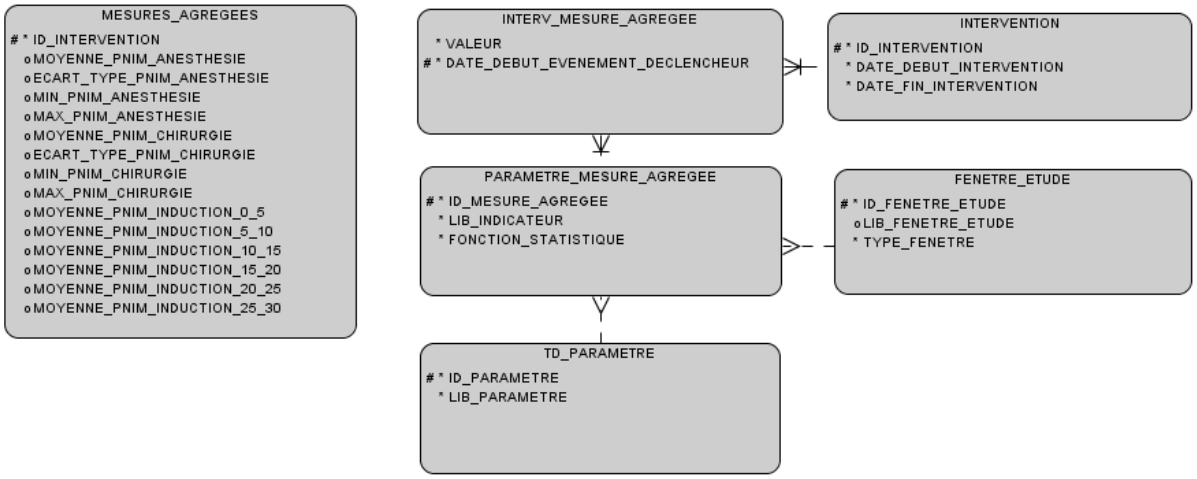


Figure 46 : Modèle logique de données - Mesures agrégées

Exemples d'applications

Exemple 1 : Mesure de la moyenne per opératoire de la Pression Artérielle Moyenne

Nous nous proposons d'évaluer la moyenne de la Pression Artérielle Moyenne (PAM) au cours de chirurgie sous anesthésie générale pour les patients âgés d'au moins 18 ans opérés entre le 01/01/2010 et le 31/12/2012.

Un total de 97 629 interventions correspond aux critères d'inclusion. La moyenne de la PAM mesurée sur l'ensemble la fenêtre d'étude "chirurgie" définie entre les événements "début de chirurgie" et "fin de chirurgie" est de 82,2 (14,6) mmHg. Le tableau 21 présente la fréquence des l'événement "hypotension" pour des seuils de PAM de 55, 60 et 65 mmHg. La figure 47 représente le nombre d'interventions par valeurs de PAM moyenne.

Tableau 21 : Fréquence des hypotensions per-opératoires selon différents seuils

PAM moyenne	Nombre d'interventions (%)
< 55 mmHg	1037 (1,06%)
< 60 mmHg	3635 (3,72%)
< 65 mmHg	9752 (9,99%)

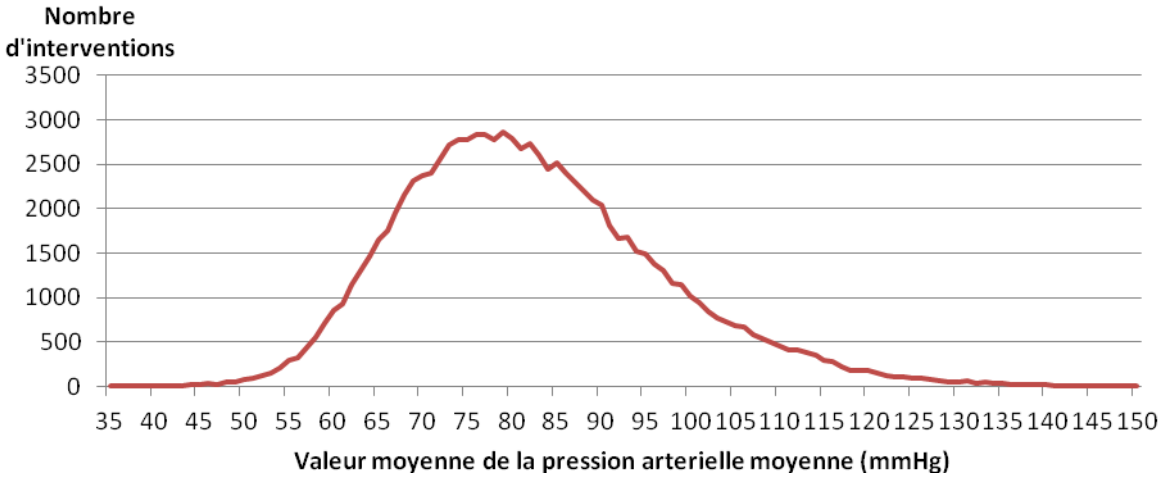


Figure 47 : Distribution de la moyenne de PAM per-opératoire pour les interventions ayant eu lieu entre le 01/01/2010 et le 31/12/2012

Exemple 2 : Effet de l'administration intra veineuse d'atropine sur la fréquence cardiaque

Chaque administration d'Atropine est détectée grâce à la table EVENEMENT ; les mesures d'intérêt sont celles de la fréquence cardiaque (FC), enregistrées dans la table MESURE. La figure 48 présente les variations de FC typiques menant à l'administration d'un bolus intraveineux d'atropine. Plusieurs fenêtres d'études sont définies avant et après l'administration d'atropine (T0) (fig. 49) : la valeur minimale de FC est calculée sur la fenêtre d'étude précédant l'administration d'atropine. Cinq plages de 15 minutes sont définies pour étudier l'évolution de FC après l'administration d'atropine : la valeur maximale de FC est détectée sur la première fenêtre temporelle [T0;T0+15] et sur chacune des quatre fenêtres temporelles suivantes, ce qui permet de suivre l'évolution de FC au cours du temps.

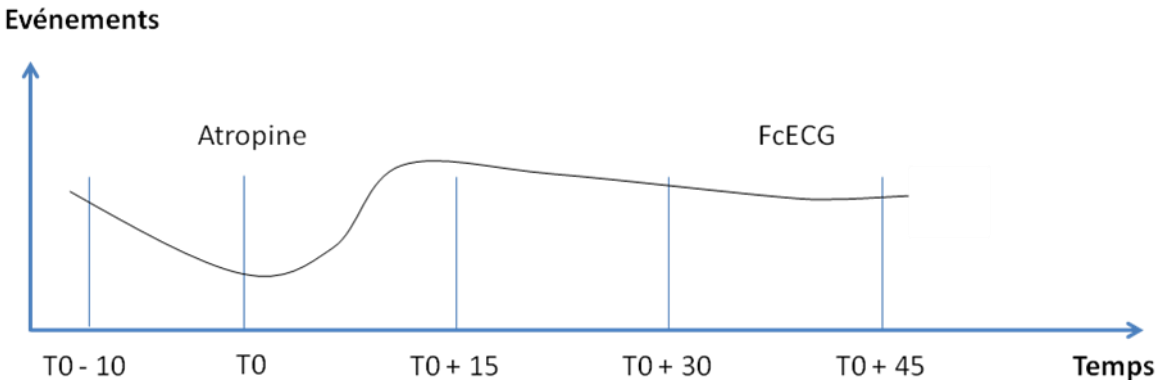


Figure 48 : Evolution de la fréquence cardiaque autour d'une administration d'atropine

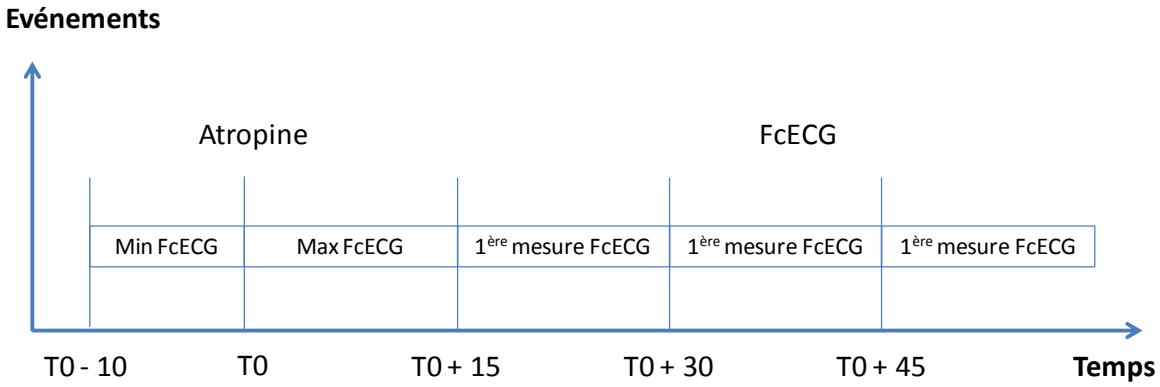


Figure 49 : Fenêtres d'étude et mesures agrégées autour de l'administration d'atropine

Les critères d'inclusions sont :

- Adultes opérés sous anesthésie générale entre le 31/01/2010 et le 31/12/2010 ;
- au moins une administration d'atropine.

Au total, 5909 interventions ont répondus aux critères d'inclusions retenus. Le tableau 22 présente les résultats obtenus sur chacune des cinq plages temporelles définies dans les figures 48 et 49. La valeur minimum de FC est très basse dans la fenêtre [T0-10;T0], ce qui justifie l'administration d'atropine. La valeur maximale de FC mesurée dans les 10 minutes suivant l'administration d'atropine illustre l'effet tachycardisant de l'atropine. Les valeurs de FC obtenues sur les trois fenêtres suivantes montrent la persistance prolongée de l'effet tachycardisant.

Tableau 22 : Fréquence cardiaque observée sur chacune des cinq fenêtres d'études analysées avant et après une injection d'atropine

Fenêtre d'étude et fonction mathématique appliquée	Fréquence cardiaque Moyenne (Ecart-type) min^{-1}
Min [T0-10 ; T0]	44,7 (9,1)
Max [T0 ; T0+10]	81,6 (20,1)

FC à [T0+15]	74,7 (17,4)
FC à [T0+30]	74,5 (17,0)
FC à [T0+45] min	73,4 (16,6)

D'un point de vue technique, les mesures agrégées stockées dans les magasins de données permettent d'extraire directement les résultats et permettent ainsi d'éviter de recalculer les agrégations à chaque interrogation. De part leur formalisme, elles sont bien adaptées aux questions relatives au déroulement de l'anesthésie et de la chirurgie, ainsi qu'aux différents événements qui s'y produisent.

4. Temps passé hors seuil

Différentes publications rapportent un lien statistique entre certains événements indésirables survenant au cours de l'anesthésie et une augmentation de la durée de séjour ou de la mortalité: hypotension artérielle, valeur d'indice bispectral (BIS) abaissée et faible dosage en hypnotiques (44–46,48). Ces événements indésirables peuvent être définis selon un formalisme similaire à celui présenté ci-dessus lors de l'agrégation des mesures: un paramètre diminue en dessous d'une valeur seuil prédéterminée, ce qui déclenche un "chronomètre" qui mesure le temps passé en dessous de ce seuil. Ainsi, Bijker et al. ont défini quarante huit indicateurs d'hypotension, obtenus selon que la durée d'hypotension artérielle systolique – définie par divers seuils allant de 40 à 100 mmHg – est de 1, 5 ou 10 min. Les méthodes utilisées pour gérer le problème des données manquantes et calculer le temps "hors seuil" effectif sont très variables (44–46,48). L'objectif de cette partie de notre travail est d'automatiser la détection d'événements indésirables grâce à des méthodes simples et des paramètres prédéfinis.

Dans la suite de ce travail, le temps "hors seuil" correspond à une durée pendant laquelle un paramètre présente des valeurs en dehors d'un seuil prédéfini. Il peut s'agir de la durée d'un seul épisode "hors seuil", ou du cumul des durées de plusieurs épisodes "hors seuil" au cours d'une même intervention. Sur la figure 50, le temps hors seuil est représenté par les zones orangées ; un épisode débute lorsqu'une première mesure passe en dessous d'un seuil prédéterminé et se termine lorsqu'une mesure repasse au dessus de ce seuil. Un premier épisode de temps hors seuil débute à T1 et finit à T2, le deuxième épisode débutant à T3 et se terminant à T4.

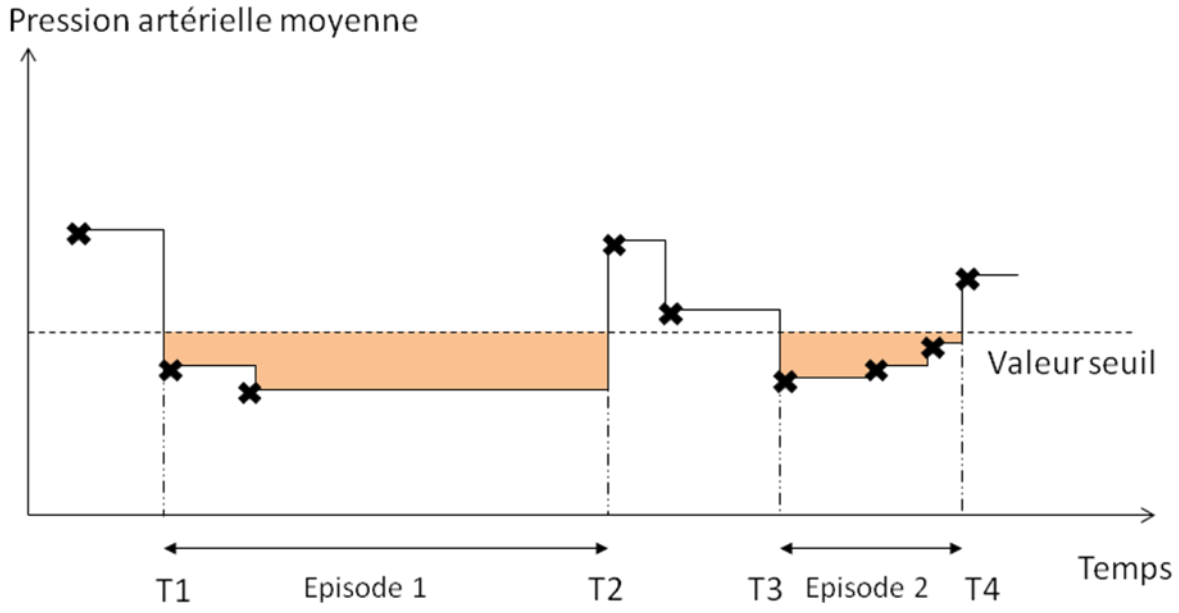


Figure 50 : Répétition de la valeur entre deux mesures successives pour estimer la valeur d'un paramètre au cours du temps

Différentes méthodes d'interpolation utilisées dans la littérature permettent de traiter le problème des valeurs manquantes. Sessler *et al.* ont choisi de répéter la valeur entre deux mesures successives (46) : cette méthode est employée figure 50. Kertai *et al.* utilisent l'interpolation linéaire pour compléter les données entre deux mesures successives (48) (figure 51).

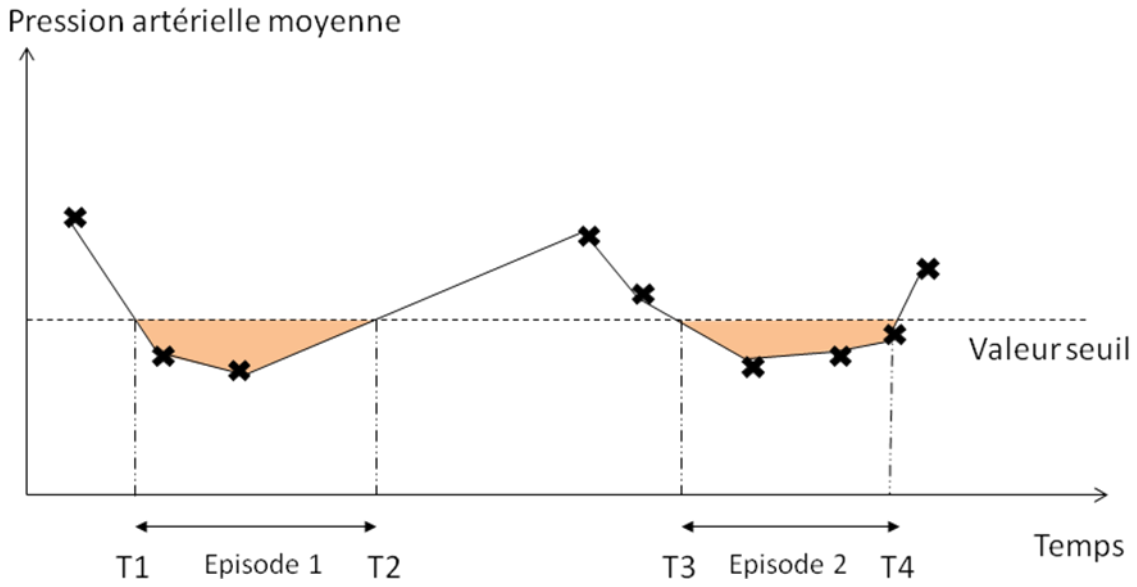


Figure 51 : Interpolation linéaire entre deux mesures successives pour estimer la valeur d'un paramètre au cours du temps

La période de temps *maximale sans mesure (Dmax)* est définie a priori comme la durée maximale acceptable entre deux mesures d'un paramètre donné; elle peut varier d'un paramètre à l'autre. Lorsque deux mesures sont séparées par un intervalle de temps *D* supérieur à *Dmax*, la période entre *D* et *Dmax* est considérée comme un intervalle de temps sans mesure et n'est pas comptabilisée dans le calcul du temps hors seuil afin de ne pas biaiser le résultat. Pour l'intervention représentée figure 52, le temps entre la 3ème et la 4ème mesure est supérieur à *Dmax* et n'est pas comptabilisé comme temps hors seuil pour l'épisode 1. La *proportion maximale sans mesure (Pmax)* correspond à la somme des *Dmax* sur le temps totale de la fenêtre étudié. Kertai *et al.* ne conservent pas les interventions pour lesquelles ce rapport est supérieur à 25%.

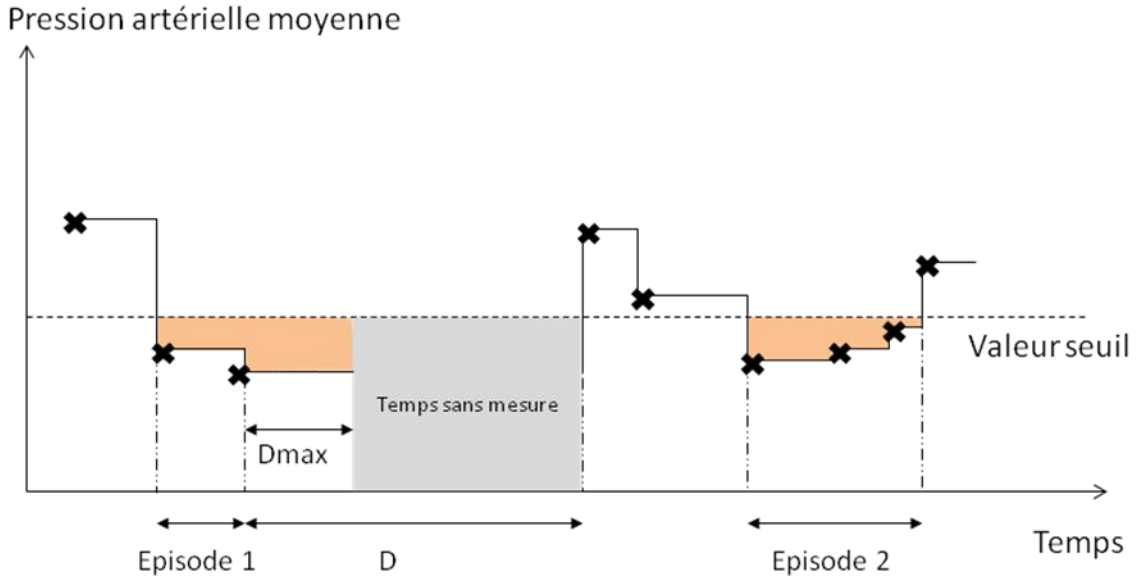


Figure 52 : Estimation des données manquantes pour un enregistrement de pression artérielle moyenne

L'algorithme de calcul du temps hors seuil doit donc tenir compte de plusieurs contraintes :

- être paramétrable pour permettre le calcul du temps hors seuil en fonction de paramètres de mesure, de plages d'étude et de valeurs seuils (fixe ou relative à l'intervention) différents.
- pouvoir faire la distinction, dans le flux de données, entre les interventions et les seuils.
- prendre en compte les valeurs manquantes
- détecter les dates d'occurrence des épisodes, la valeur extrême au cours de l'épisode ainsi que sa date d'occurrence
- calculer le nombre d'épisodes et la durée totale en dehors du seuil par intervention et par seuil

4.1 Méthode

Afin de calculer le temps hors seuil en suivant les différentes méthodes définies par la littérature, l'algorithme doit être indépendant des paramètres comme la valeur seuil ou *Dmax*. Ces paramètres sont présentés tableau 23.

Tableau 23 : exemples de paramètres prédéfinis pour le calcul du temps hors seuil

Paramètre de calcul	Exemples	Intérêt
Paramètre mesuré	FcECG, PNIm, SpO2	Sélection des mesures
Plage d'étude	Induction-Incision, Anesthésie	Sélection des mesures
Groupe de patient	Patient adultes, Chirurgie cardiaque	Sélection des patients
Valeur seuil, absolue ou relative	PNIm < 65 mmHg PNIm < 80% de la première mesure au bloc FcECG > 100 bpm	Calcul du temps hors seuil
Temps minimal entre deux mesures (Dmax)	60 pour la FcECG 360 pour la PNIm	Gestion des données manquantes
Méthode d'interpolation entre deux valeurs	Interpolation linéaire ou répétition	Calcul du temps hors seuil
Proportion de temps sans mesure acceptable (Pmax)	25%	Gestion des données manquantes

L'algorithme de calcul des événements indésirables comporte plusieurs étapes, l'objectif étant d'optimiser le temps de calcul lors de la phase de détection des épisodes d'événements indésirables. La figure 53 représente ces différentes étapes.

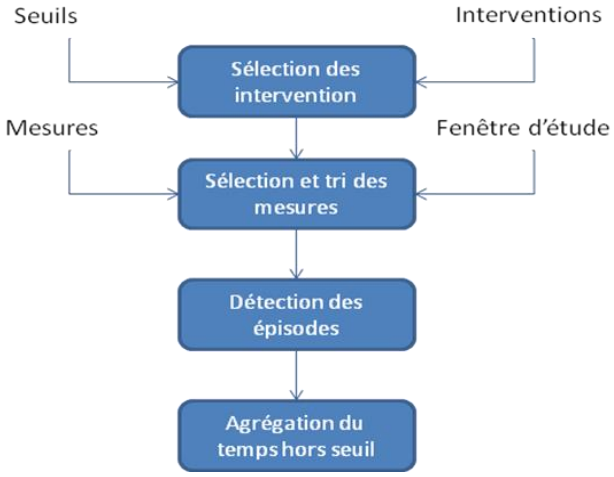


Figure 53: Processus de calcul du temps hors seuil

Préparation des données

La première étape consiste à préparer les données nécessaires au calcul: les interventions et seuils à calculer sont sélectionnés, ainsi que les dates de début et de fin des périodes associés à chaque intervention et à chaque seuil (fig. 54). Puis, les mesures sont intégrées lorsqu'elles sont comprises dans les plages de temps. Elles sont ensuite classées par intervention, par seuil, par plage d'étude et par heure de mesure (fig. 55).

ID_INTERVENTION	SEUIL	PARAMETRE	DATE_DEBUT_FENETRE	DATE_FIN_FENETRE
110456	1	1	10:12:25	12:16:23
110456	2	1	10:12:25	12:16:23
245684	3	6	13:25:11	14:45:41

SEUIL	LIB_SEUIL
1	FcECG < 50 bpm [anesthésie]
2	FcECG > 100 bpm [anesthésie]
3	PNIm < 60 mmHg [anesthésie]

Figure 54 : (Haut) Table des interventions, seuils et plages temporelles sélectionnés pour la détection des événements indésirables ; (Bas) Table correspondant aux seuils d'événements indésirables

ID_INTERVENTION	SEUIL	PARAMETRE	VALEUR	DATE_MESURE
110456	1	1	67	10:12:36
110456	1	1	68	10:13:01
...				
110456	1	1	72	12:15:58
110456	2	1	67	10:12:36
...				
110456	2	1	72	12:15:58
245684	3	6	82	13:26:24
245684	3	6	74	13:29:58
...				
245684	3	6	81	14:42:14

Changement de seuil

Changement d'intervention

Figure 55 : Mesures filtrées et triées par intervention, seuil, fenêtre d'étude et date de mesure pour la détection des événements indésirables

Calcul des épisodes hors seuil

Les mesures filtrées et triées dans les étapes précédentes sont ensuite parcourues une à une. Lorsqu'une mesure est en dehors de la valeur seuil, un épisode est déclenché. Celui-ci est clôt lorsqu'une mesure est à nouveau dans la plage normale. Si l'intervalle entre deux mesures (*D*) est supérieur à *Dmax* et qu'un épisode était déclenché, celui est fermé. La différence entre *D* et *Dmax* est comptabilisée comme temps sans mesure. L'algorithme tient compte des changements d'intervention et de seuil. Tout épisode en cours est clôturé et le temps sans mesure associé à l'intervention précédente est enregistré. Le fonctionnement de l'algorithme est illustré figure 56.

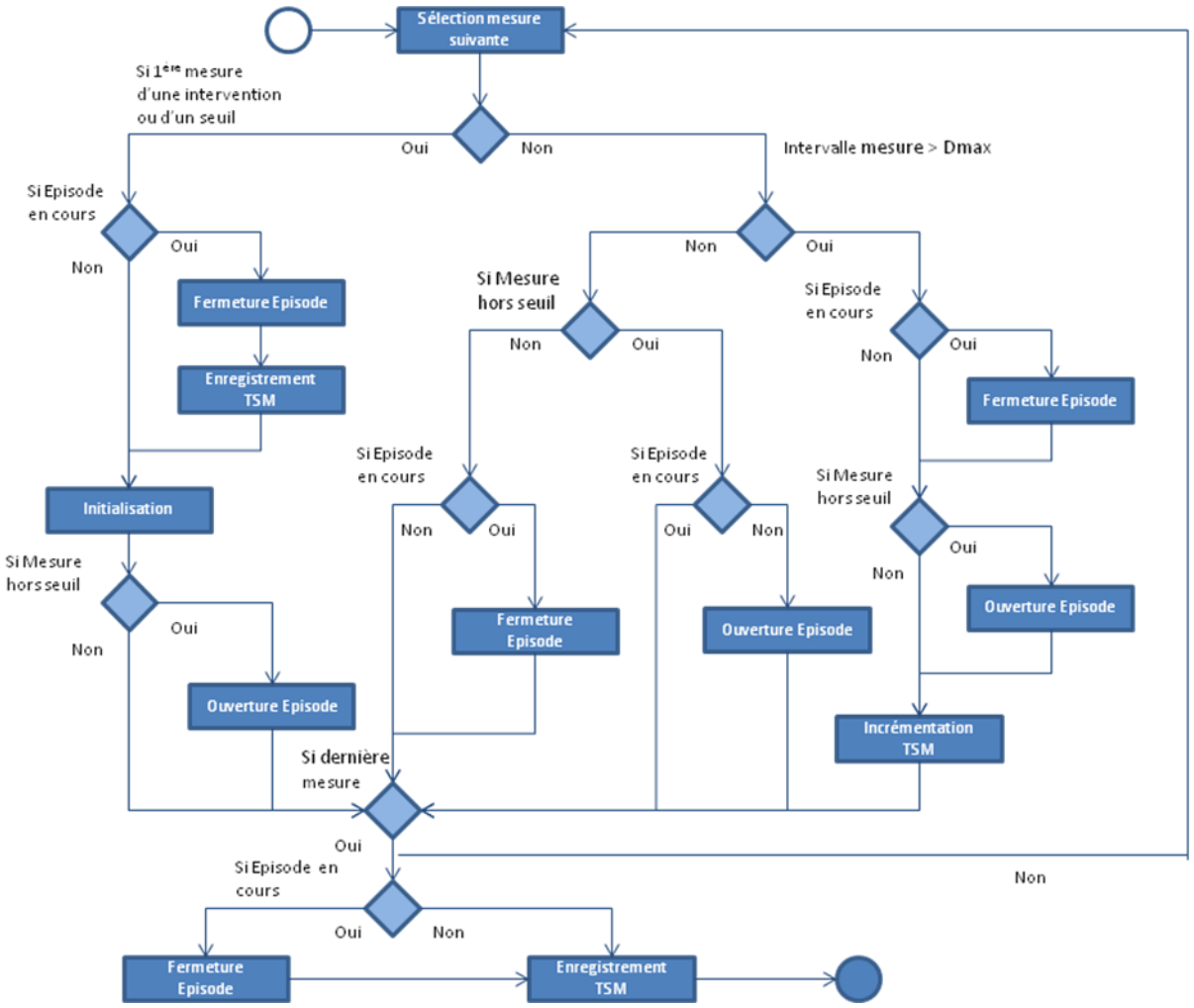


Figure 56 : Algorithme de calcul des épisodes hors seuil

Calcul du temps hors seuil

Pour chaque intervention et chaque seuil, les épisodes sont agrégés afin de calculer le temps total passé en dehors du seuil, le nombre d'épisodes, la valeur moyenne en dehors du seuil, la valeur extrême et sa date d'occurrence.

Pour chaque enregistrement, un indice de qualité est calculé, qui prend la valeur de 0 ou de 1 : 1 si la proportion de temps sans mesure est inférieure à *Pmax*, 0 sinon. Il permettra par la suite de n'utiliser lors de l'analyse statistique que les interventions pour lesquelles l'indice de qualité vaut 1.

Les résultats (temps hors seuil, nombre d'épisodes, valeur extrême, date valeur extrême, ...) sont transformés en colonnes, par seuil grâce à la requête pivot 21. Le résultat obtenu est une ligne par intervention, plusieurs colonnes pour chaque seuil. Les statistiques peuvent ensuite être réalisées par intervention.

(21)

```
SELECT *
FROM
    (SELECT ID_INTERVENTION,      ID_SEUIL,      B_TEMPS_HORS_SEUIL,
    TEMPS_HORS_SEUIL FROM    TEMPS_HORS_SEUIL)
PIVOT
    (MIN(B_TEMPS_HORS_SEUIL)  AS  B,  MIN(TEMPS_HORS_SEUIL)  AS  THS
    FOR (ID_SEUIL)
    IN
    (230 AS PNIM_I_50_INDUC_INC, 231 AS PNIM_I_55_INDUC_INC, 232
    AS PNIM_I_60_INDUC_INC, 233 AS PNIM_I_65_INDUC_INC,
    234 AS PNIM_I_70_INDUC_INC, 235 AS PNIM_I_75_INDUC_INC))
```

4.2 Résultats

Modèle de données

La figure 57 représente le modèle logique de données du module TEMPS HORS SEUIL. Celui-ci comporte une table permettant le paramétrage des différents seuils, la table SEUIL. La table EPISODE stocke les différents épisodes d'un seuil et d'une intervention donnés. La table TEMPS_HORS_SEUIL enregistre les agrégations des épisodes.

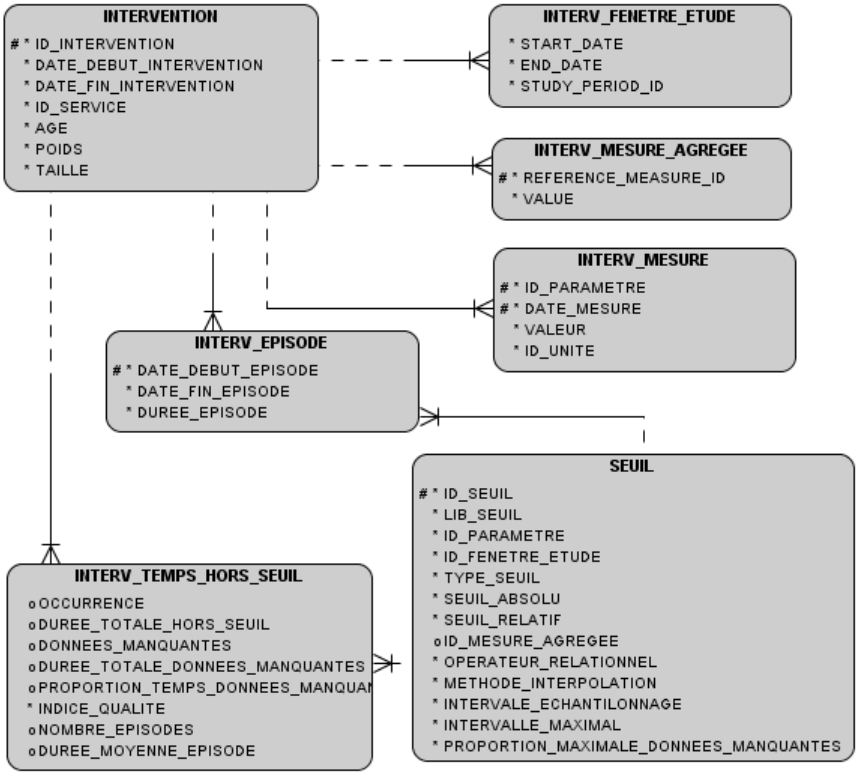


Figure 57 : Modèle logique de données - Temps hors seuil

Exemple d'application : temps passé en hypotension

La méthode de calcul du temps hors seuil ainsi définie a été appliquée aux anesthésies générales réalisées chez des adultes entre le 01/01/2010 et le 31/12/2012. Seules les procédures ayant utilisé du propofol pour l'induction de l'anesthésie ont été analysées. Les calculs ont été réalisés pour cinq seuils différents de Pression Artérielle Moyenne (PAM), échelonnés entre 50 et 70 mmHg. Les valeurs manquantes ont été estimées par interpolation linéaire; l'intervalle de temps maximal admissible *Dmax* entre deux mesures consécutives était fixé à 360 secondes.

Le tableau 24 présente la distribution fréquentielle des interventions ayant présenté un épisode d'hypotension artérielle en fonction des différents seuils de PAM retenus, ainsi que la distribution des durées de ces épisodes (1° quartile, médiane et 3° quartile).

La majorité (80.2%) des procédures analysées a présenté une diminution de la PAM à moins de 70 mmHg : ce résultat n'est pas cliniquement relevant, mais il permet de vérifier le bon fonctionnement de la méthode d'analyse mise en œuvre. Les quatre autres seuils ont davantage de sens clinique, en particulier ceux de 55 et de 50 mmHg. Il est logique de constater que plus l'hypotension est sévère, plus la durée passée dans cet état est brève: 4,2 [2,3 – 7,6] min passées avec une PAM<50mmHg, 9,7 [4,8-20,8] min passées avec une PAM<60mmHg.

Tableau 24: Calcul de l'hypotension pour les patients de 18 ans et plus entre 2010 et 2012

Seuil	Nombre d'interventions (%)	Durée (min) 1er quartile	Durée (min) Médiane	Durée (min) 3ème quartile
< 50 mmHg	16078 (15,4%)	2,3	4,2	7,6
< 55 mmHg	28528 (27,3%)	3,0	5,5	11,9
< 60 mmHg	44053 (42,1%)	4,8	9,7	20,8
< 65 mmHg	59304 (56,7%)	6,0	15,4	35,1
< 70 mmHg	83891 (80,2%)	9,6	25,4	58.0

5. Administration de médicaments

Chaque administration d'un médicament est enregistrée dans la table EVENEMENT de l'entrepôt de données (cf. chap. 3). Différentes analyses sont possibles selon la question posée: dose cumulée d'une molécule donnée sur l'ensemble de l'anesthésie ou une sur une période restreinte. Le terme "médicament" est utilisé au sens large, puisque tous les agents médicamenteux utilisés au cours de l'anesthésie sont considérés comme des "médicaments", c'est-à-dire par exemple les morphiniques, les hypnotiques, les curares, les médicaments à visée hémodynamique, les solutés de remplissage, les produits dérivés du sang, etc.

Les différentes étapes de l'agrégation des administrations de médicaments sont présentées dans la figure 58.

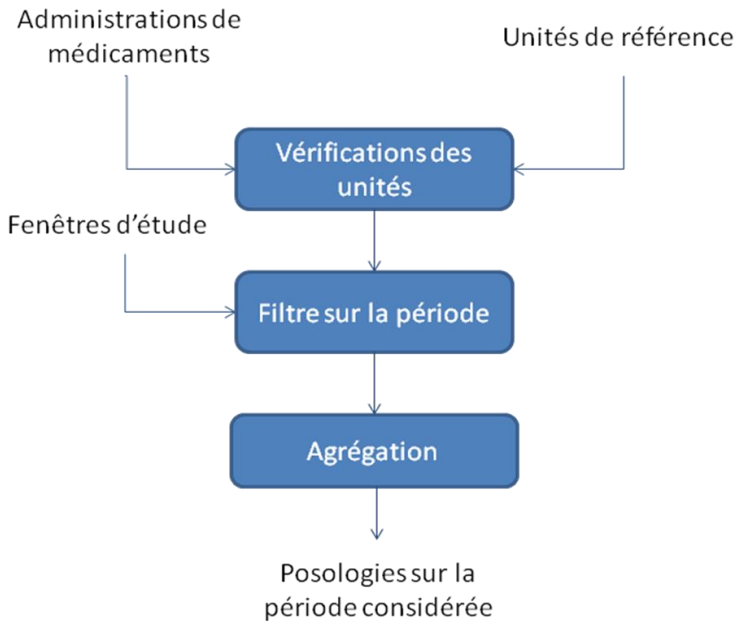


Figure 58 : Etapes d'agrégation des administrations médicamenteuses pour chaque intervention

5.1 Méthode

Vérifications des unités enregistrées

La diversité des pratiques anesthésiques avec une même molécule entraîne une diversité de posologies et d'unités pour les administrations médicamenteuses enregistrées dans l'entrepôt. Par exemple, le propofol utilisé en bolus peut être renseigné en millimètres (ml) ou en milligrammes (mg). Quand il est administré au pousse seringue électrique (PSE), sa concentration peut être de 1% ou de 2%, et son débit peut être renseigné en $mg.H^{-1}$, en $ml.H^{-1}$, en $mg.kg^{-1}.min^{-1}$ ou en $\mu g.kg^{-1}.min^{-1}$ Il est parfois administré par un dispositif d'administration à objectif de concentration (AIVOC) qui communique automatiquement à la feuille informatisées DIANE les concentrations cible et effet en $\mu g.ml^{-1}$. Dans notre entrepôt, dans cette situation, les informations sur le débit massique de propofol sont perdues. Des exemples de molécules utilisées pour l'anesthésie ainsi que leurs posologies en fonction du mode d'administration choisi sont présentés dans le tableau 25 (cf. chapitre II). Dans certains cas, les unités enregistrées dans l'entrepôt sont erronées; elles sont automatiquement supprimées lors de l'analyse.

Tableau 25 : Exemples de médicaments et des unités qui peuvent y être associées

Médicament	Unités correctes de référence	Utilisation
Alfentanil	µg	Bolus
Clonidine	µg	Bolus
Remifentanil	µg	Bolus
	ng.ml plasma	AIVOC
	$ng.ml^{-1}$	AIVOC
	$ml.h^{-1}$	SAP
Propofol	mg	Bolus
	$\mu g.ml^{-1}$ plasma	AIVOC
	$\mu g.ml^{-1}$	AIVOC
	$ml.h^{-1}$	SAP
Sufentanil	$\mu g.ml^{-1}$	AIVOC
	$ml.h^{-1}$	SAP
	ml	Bolus
	$\mu g/h^{-1}$	SAP
	$\mu g.kg^{-1}.h^{-1}$	AIVOC plasmatique ou cérébrale

Analyse sur une période temporelle prédéfinie

Pour une question clinique spécifique portant sur une période temporelle bien définie, les données relatives aux administrations médicamenteuses ne sont agrégées que sur la période d'intérêt.

On peut par exemple calculer les doses des différents médicaments utilisés pour l'induction, ou entre le début et la fin de la chirurgie, ou encore spécifiquement pendant la période passée en salle de surveillance post interventionnelle (SSPI). On peut également concentrer l'analyse sur les quinze minutes précédant ou suivant un événement prédéterminé.

A titre d'illustration, la figure 59 montre différentes applications : l'administration de propofol peut n'être comptabilisée que pour la période de l'induction anesthésique. A l'inverse, cette analyse permet de sélectionner les interventions où le propofol a été utilisé pour l'induction. L'administration de sufentanil peut n'être comptabilisée que sur la période chirurgicale. L'éphédrine peut n'être comptabilisée qu'après un épisode d'hypotension défini par exemple par un seuil de PAM donné, et pendant une durée prédéterminée. Enfin, l'administration de morphine peut n'être comptabilisée que sur la période où le patient est présent en salle de réveil.

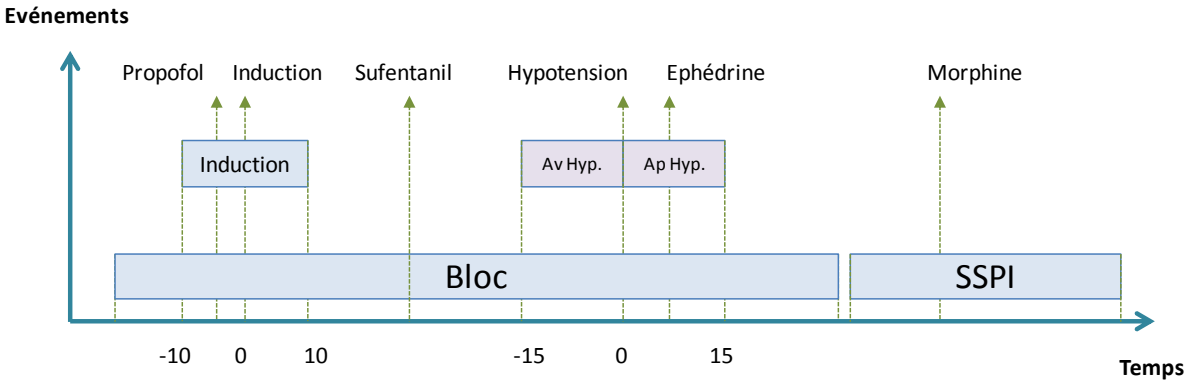


Figure 59 : Exemple d'administrations de médicaments au cours de plages temporelles définies

Agrégation et requête pivot

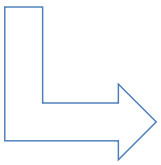
Une fois l'agrégation réalisé intervention par intervention, la requête pivot 22 (86) réalise l'extraction des informations pour un grand nombre d'interventions, en fournissant des résultats du type présent/absent pour chaque médicament, ainsi que la dose administrée sur la période d'intérêt. Le résultat est présenté en colonnes (figure 60).

(22)

```
SELECT *
FROM (
        SELECT ID_INTERVENTION, ID_MEDICAMENT, POSOLOGIE,
    ID_UNITE
        FROM INTERV_MEDICAMENT_BLOC)
PIVOT (
        SUM(POSOLOGIE) AS TOTAL, MAX(ID_UNITE) AS ID_UNITE
        FOR (ID_MEDICAMENT)
        IN (1 AS PROPOFOL, 2 AS REMIFENTANIL, ...,      98
    AS MORPHINE)
    );
```

Les enregistrements d'administration de propofol et de sufentanil concernant deux interventions sont agrégés afin de fournir le total de produit administré lors du passage au bloc opératoire (colonnes TOTAL_PROPOFOL et TOTAL_SUFENTANIL). Les deux colonnes PROPOFOL et SUFENTANIL sont les champs binaires précisant si chaque médicament est administré au moins une fois au cours de la plage étudiée.

MEDICAMENT_BLOC_SRC			
ID_INTERVENTION	MEDICAMENT	POSOLOGIE	UNITE
123456	PROPOFOL	250	mg
123456	SUFENTANIL	25	µg
123456	SUFENTANIL	10	µg
987654	PROPOFOL	200	mg
987654	SUFENTANIL	30	µg
987654	SUFENTANIL	10	µg
987654	SUFENTANIL	10	µg

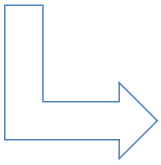


MEDICAMENT_BLOC				
ID_INTERVENTION	PROPOFOL	TOTAL _PROPOFOL	SUFENTANIL	TOTAL _SUFENTANIL
123456	1	250	1	35
987654	1	200	1	50

Figure 60 : Illustration d'une requête pivot. Le premier tableau contient les enregistrements d'administrations de médicaments. Après pivot des lignes en colonnes, le second tableau contient les doses totales administrées ainsi qu'un champ binaire indiquant si le médicament a été administré ou non.

Quand un médicament est administré plusieurs fois selon des modalités d'administration différentes (unités différentes), les doses correspondant à chaque unité sont agrégées indépendamment les unes des autres. Un champ binaire fournit également l'information présent/absent, quel que soit le nombre d'unités différentes enregistrées. Ainsi, la figure 61 illustre le cas de deux interventions différentes (colonne de gauche, ID 123456 et ID 456789) au cours desquelles les administrations de propofol sont renseignées avec deux unités différentes, mg et $\mu g.ml^{-1}$. La requête pivot agrège donc ces administrations selon deux colonnes distinctes, TOTAL_PROPOFOL en mg et MOY_PROPOFOL_CIBLE en $\mu g.ml^{-1}$.

MEDICAMENT_BLOC_SRC			
ID_INTERVENTION	MEDICAMENT	POSOLOGIE	UNITE
123456	PROPOFOL	250	mg
456789	PROPOFOL	3	µg/ml
456789	PROPOFOL	4	µg/ml

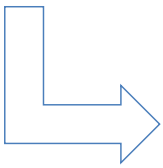


MEDICAMENT_BLOC					
ID_INTERVENTION	PROPOFOL_COMPOSITE	PROPOFOL	TOTAL_PROPOFOL	PROPOFOL_CIBLE	MOY_PROPOFOL_CIBLE
123456	1	1	250	0	-
456789	1	0	-	1	3,5

Figure 61 : exemple de requête pivot dans le cas où le propofol est enregistré avec deux unités différentes : ml et µg.ml^{-1}

Si une administration médicamenteuse est renseignée avec une unité incorrecte, l'information de l'administration est conservée mais la dose n'est pas exploitée. Ainsi, figure 62, l'intervention ID 654321 présente un enregistrement de propofol avec une unité erronée (µg.ml^{-1}.kg^{-1}): la dose de propofol n'est pas conservée lors de l'analyse, et le champ ERREUR_PROPOFOL conserve l'information de cette unité erronée, ce qui sera utilisable pour évaluer la qualité des données.

MEDICAMENT_BLOC_SRC			
ID_INTERVENTION	MEDICAMENT	POSOLOGIE	UNITE
123456	PROPOFOL	250	mg
456789	PROPOFOL	3	µg/ml/kg



MEDICAMENT_BLOC						
ID_INTERVENTION	PROPOFOL_COMPOSITE	PROPOFOL	TOTAL_PROPOFOL	PROPOFOL_CIBLE	MOY_PROPOFOL_CIBLE	ERREUR_PROPOFOL
123456	1	1	250	0	-	0
456789	1	0	-	0	-	1

Figure 62 : exemple de requête pivot en cas d'unité erronée de propofol

5.2 Exemple d'application : détermination automatique des médicaments utilisés lors de l'induction anesthésique

L'événement temporel qui marque le début de l'anesthésie est présent dans 95% des feuilles informatisées d'anesthésie. Il est donc parfaitement exploitable pour rechercher quels médicaments sont utilisés pour l'induction anesthésique, comme illustré figure 63.

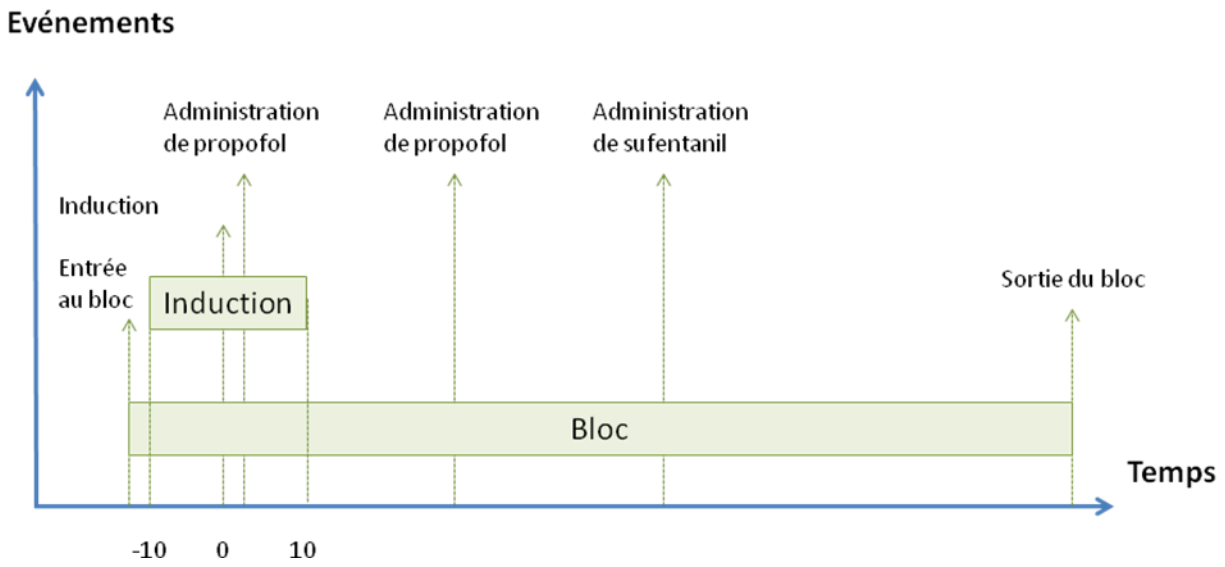


Figure 63 : Administration de médicaments lors de l'induction anesthésique

La posologie des médicaments hypnotiques utilisés pour l'induction est calculée sur une fenêtre d'étude de 20 minutes, débutant 10 min avant l'événement "induction" et poursuivie jusqu'à 10 min après. Lorsque plusieurs médicaments sont administrés, un algorithme détermine lequel a probablement joué le rôle prépondérant. Par exemple, l'induction au sévoflurane est fréquemment suivie par une administration de propofol; la technique d'induction utilisée n'en demeure pas moins "inhalatoire": on cherche donc à déterminer si le sévoflurane est utilisé pendant la période de 20 minutes autour de l'induction. L'algorithme suivant a été déterminé selon les "règles métier" de l'anesthésie, et a été testé ensuite sur les enregistrements de l'entrepôt :

```
Si Max(Sévoflurane[-10;10] Induction) > 3
     Alors induction = Sévoflurane
Sinon Si Administration(Propofol [-10;10] Induction)
     Alors induction = Propofol
Sinon Si Administration(Etomidate[-10;10] Induction)
     Alors induction = Etomidate
Sinon Si Administration(Ketamine[-10;10] Induction)
     Alors induction = Kétamine
Sinon Si Administration(Clonidine[-10;10] Induction)
     Alors induction = Clonidine
Sinon Méthode d'induction non disponible.
```

Analyse et critères d'inclusion

Les patients adultes ayant bénéficié d'une anesthésie générale avec ventilation mécanique entre le 01/01/2010 et le 31/12/2012 ont été inclus dans l'analyse. L'analyse automatisée des enregistrements consistait à recueillir l'ensemble des médicaments hypnotiques administrés dans la période [-10 - 10] min autour de l'événement composite "induction anesthésique", et à déterminer ensuite selon l'algorithme présenté ci-dessus de quelle "technique d'induction anesthésique" il s'agissait. Le formalisme suivant était utilisé :

- induction inhalatoire (sevoflurane)

- induction IV (propofol)
- induction IV (étomidate)
- induction IV (kétamine)
- induction IV utilisant la clonidine

Le tableau 26 présente les données présentées par année, sous forme de nombre (%).

Tableau 26: Méthode d'induction de 2010 à 2012

Année	Induction inhalatoire (sevoflurane)	Induction IV (propofol)	Induction IV (étomidate)	Induction IV (kétamine)	Induction IV avec clonidine	Auncun hypnotique détecté	Total
2010	4386 (8.4%)	31503 (60.3%)	407 (0.8%)	669 (1.3%)	820 (1.6%)	14465 (27.7%)	52250
2011	4902 (9.2%)	32704 (61.1%)	504 (0.9%)	505 (0.9%)	568 (1.06%)	14340 (26.8%)	53523
2012	5099 (9.2%)	33547 (60.7%)	580 (1.0%)	664 (1.2%)	499 (0.9%)	14840 (26.9%)	55229

Les résultats présentés indiquent une proportion importante et stable d'induction IV (propofol) en 2010, 2011 et 2012. Les inductions inhalatoires (sevoflurane) semblent augmenter légèrement, de même que les inductions IV (étomidate). Les inductions IV (kétamine) restent l'exception, de même que celles qui utilisent la clonidine. Enfin, les inductions anesthésiques pour lesquelles aucun médicament hypnotique n'est retrouvé représentent une proportion non négligeable de ces interventions. Un défaut de détection des médicaments administrés, ou des administrations médicamenteuses renseignées avec retard par rapport à l'induction pourraient expliquer ce résultat.

6. Discussion

Dans ce chapitre, nous avons présenté plusieurs méthodes d'agrégation des données pour l'alimentation de *magasins de données*. Les exemples d'applications donnés en illustration montrent qu'il est possible de détecter automatiquement certaines **techniques anesthésiques** (technique d'induction par exemple) ou des **événements indésirables (survenue, profondeur et durée d'une hypotension par exemple**). Les données agrégées permettent également de suivre l'**évolution de paramètres physiologiques** en fonction d'un événement particulier comme l'administration d'atropine.

Le tableau 27 résume ces différentes informations, les données utilisées pour leur calcul, et l'intérêt qu'elles présentent.

Tableau 27 : Récapitulatif des données agrégées dans l'entrepôt de données

Nouvelle information	Information source	Intérêt
Fenêtre d'étude	Mesures Evénements	Déterminer des plages temporelles précises Compléter des informations manquantes
Mesures agrégées	Mesures Fenêtres d'étude	Représenter une tendance
Temps hors seuil	Mesures Mesures agrégées Fenêtres d'étude	Détecter des événements indésirables
Administration de médicaments	Médicaments Fenêtres d'étude	Connaître la dose totale administrée d'un médicament sur une plage temporelle précise

Ces nouvelles données, déjà calculées et **disponibles**, pourront être utilisées facilement dans le cadre d'analyses statistiques et reflètent un des avantages de l'entrepôt de données : proposer des **indicateurs métier agrégés et calculés à partir de données hétérogènes issues de multiples systèmes opérationnels**.

Plusieurs contraintes sont prises en compte, en particulier en définissant des règles métier et en permettant le paramétrage des calculs :

- Définir plusieurs niveaux de qualité, utiles en fonction de la requête. En effet, chaque requête n'ayant pas besoin de données de même qualité, l'entrepôt de données peut conserver les données dont la qualité est moindre. Par exemple, les enregistrements d'administrations de médicaments dont l'unité est incorrecte sont conservés : même si la posologie ne peut pas être utilisée, l'heure d'administration peut être intégrée pour des analyses ;
- Tenir compte de l'évolution des besoins des utilisateurs et faciliter le paramétrage des calculs de données agrégées (nouveaux médicaments, fenêtres d'études, mesures agrégées et temps hors seuils) ;
- Améliorer la qualité des données et détecter des périodes d'intérêt (fenêtre d'étude).

La documentation et le paramétrage des méthodes d'agrégation tels que présentés dans ce chapitre permettent de reproduire les calculs, d'une année sur l'autre et de pouvoir les comparer avec les résultats d'autres structures qui implémenteraient les mêmes méthodes. A notre connaissance, les travaux similaires à partir des données de la feuille informatisée d'anesthésie sont peu documentés et présentent des méthodologies différentes qui rendent peu reproductibles et comparables les études entre elles (43,46,48).

7. Conclusion

Ce chapitre a présenté les méthodes d'agrégation de données à partir des données intégrées dans l'entrepôt. Ces données alimentent les *magasins de données* et présentent une vue métier des données.

L'avantage des magasins de données est de proposer des **indicateurs métier** déjà calculés qui faciliteront les analyses. Le calcul de ces indicateurs tient compte de la qualité des données sources et est configurable afin de tenir compte de l'évolution des besoins des utilisateurs au cours du temps.

La méthode de calcul du temps hors seuil présentée dans ce chapitre permet de détecter de manière reproductible les événements indésirables telles que l'hypotension, la bradycardie ou le BIS bas et d'évaluer les méthodes de prise en charge de l'anesthésie.

Ces indicateurs seront utilisés lors des deux prochains chapitres qui présenteront deux études de cas liées aux méthodes de prise en charge de l'anesthésie au bloc opératoire.

Chapitre 5 : Hypotension liée à l'induction

Chapitre 5 : Hypotension liée à l'induction

1. Introduction

La survenue d'une hypotension suite à l'induction anesthésique est un effet secondaire fréquent, en particulier chez les patients âgés ou ceux dont le score ASA est élevé, ceux qui présentent des antécédents cardiovasculaires et des traitements hypotenseurs. Au cours de l'anesthésie générale, un lien statistique a été établi entre la survenue d'une hypotension alors que la fraction en halogénés est basse et que l'index bispectral est bas, et une augmentation de la durée de séjour et de la mortalité (1).

L'hypotension induite par l'anesthésie générale est multifactorielle : l'administration de dérivés morphiniques entraîne un bloc sympathique qui pouvait éventuellement masquer une hypovolémie relative, d'autant plus marquée que le patient était à jeun. Les médicaments hypnotiques induisent une vasodilatation qui induit une hypotension par inadéquation contenant/contenu. Un effet inotrope négatif de certains hypnotiques est possible également. Les traitements les plus fréquemment utilisés pour traiter l'hypotension induite par l'anesthésie sont l'adaptation des doses d'antalgiques et d'hypnotiques, l'administration de médicaments vasoconstricteurs comme l'éphédrine, la néosynéphrine voire la norépinéphrine.

Le but de cette étude rétrospective est de caractériser la fréquence, la sévérité et la durée de l'hypotension induite par l'induction au propofol d'une anesthésie générale au CHRU de Lille. L'objectif secondaire est de rechercher les facteurs prédictifs d'une hypotension prolongée ou d'une résistance aux traitements habituellement employés.

2. Méthode

Pour cette étude, les patients inclus répondent aux critères suivants :

- âge $\geq$ 18 ans ;
- anesthésie générale (intervention caractérisée par la présence de l'événement "Intubation" et d'au moins dix mesures de volume courant expiré supérieures à 0 avec une moyenne comprise entre 100 et 800 millilitres) ;
- date d'intervention entre le 01/01/2012 et le 31/12/2014 ;
- induction au propofol.

Pour cette étude, l'hypotension est caractérisée par la présence et la durée du premier épisode d'hypotension commençant entre l'induction et l'incision, afin de ne pas tenir compte des épisodes d'hypotension liés à d'autres facteurs que l'induction. Les seuils de 50, 55, 60, 65, 70 et 75 mmHg sont sélectionnés. Les intervalles de temps entre l'induction et le début l'épisode ainsi qu'entre l'induction et la valeur minimale sont également calculés. Les paramètres sélectionnés pour la détection des épisodes sont présentés en annexe 11.

Le tableau 28 présente les différentes informations disponibles dans l'entrepôt de données qui seront exploitées dans le cadre de cette étude.

Tableau 28: Informations disponibles dans l'entrepôt de données et utilisées dans le cadre de cette étude.

Type d'information	Intérêt
Mesures agrégées et événements	Sélectionner les patients ayant bénéficier d'une anesthésie générale
Fenêtre d'étude	Détecter la période entre le début de l'anesthésie et le début de la chirurgie
Temps hors seuil	Détecter le premier épisode d'hypotension
Médicament	Déterminer la dose de propofol utilisée à l'induction et les traitements éventuels de l'hypotension
Séjour	Mortalité et durée de séjour

Les données des patients inclues ont été décrites selon la moyenne (écart-type) ou médiane (1° quartile ; 3° quartile) pour les variables continues et pour les variables discrètes, nombre (%) ou proportion selon les cas. Un test t de Student était appliqué à la comparaison de durée de séjour entre les patients ayant présenté une hypotension et ceux n'ayant pas présenté d'hypotension pour chaque seuil de PAM. Un $p<0,05$ était considéré significatif; de façon similaire, un test du Chi^2 était appliqué pour comparer les taux de mortalité de ces deux populations.

3. Résultats

Entre 2012 et 2014, 171 042 interventions ont été enregistrées dans l'entrepôt, dont 74 245 correspondaient aux critères d'inclusion et 60122 présentent les informations nécessaires à l'étude, c'est-à-dire le médicament utilisé pour l'induction et les données relatives à la pression artérielle non invasive durant la période d'induction. Parmi ces interventions, 57 947 ont eu recours au propofol ont donc été exploitées. (fig. 64).

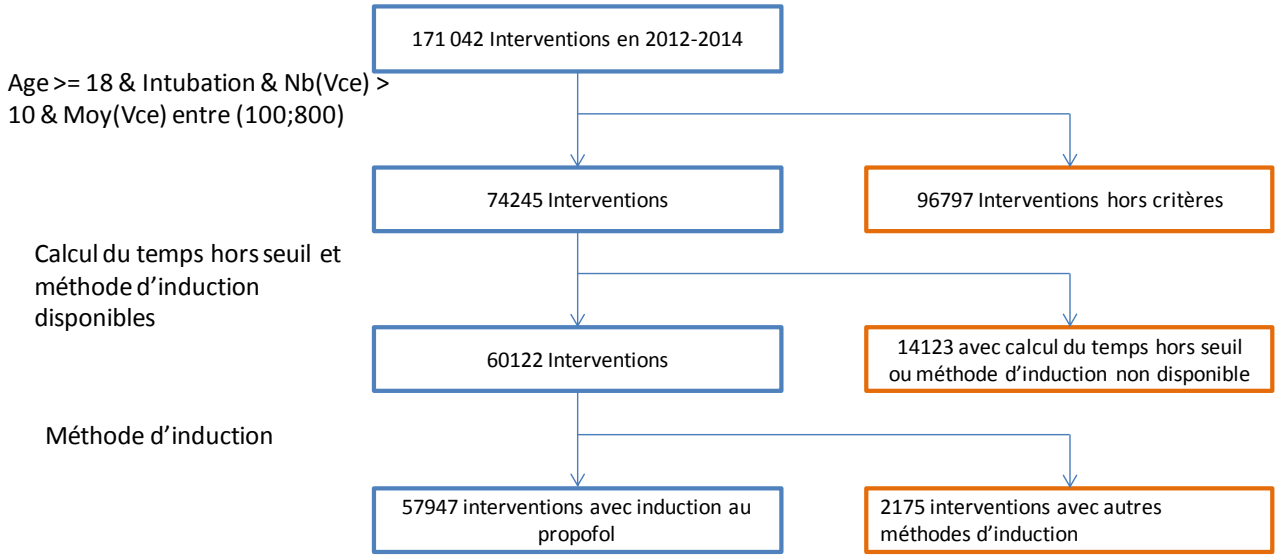


Figure 64 : Sélection des patients

3.1 Statistiques descriptives

La population incluse était âgée de 52 (17) ans, avec un poids de 76 (19) kg et un IMC de 26 (10) kg/m². La proportion d'hommes et de femmes est de 50/50. La proportion d'interventions par classe ASA 1, 2, 3, 4 and 5 est respectivement de 36,0, 42,2, 20,3, 1,4 et 0,1%. Les interventions réalisées en urgence représentaient 13% de l'ensemble des interventions.

La dose médiane de propofol utilisée pour l'induction était de 200 (150;250) µg. La PAM moyenne durant l'anesthésie est de 80 (11) mmHg.

Tableau 29 : Statistiques descriptives

Variable	Moyenne/Médiane/Proportion
Age	52 (17) ans
ASA1	36,0 %
ASA2	42,2%
ASA3	20,3%
ASA4	1,4%
ASA5	0,1%
Poids	76 (19) kg
Taille	169 (10) cm
IMC	26 (10) kg/m²
Sexe (H/F)	50/50
Urgence	13%

Mortalité	1,0%
Durée de séjour	5 (3-10) jours

3.2 Fréquence de l'hypotension

L'hypotension après l'induction survient dans 12,1% des interventions pour un seuil de 50 mmHg jusqu'à 80,4% pour un seuil de 75 mmHg. La durée médiane du premier épisode d'hypotension est de 3,12 minutes jusqu'à 16,8 minutes et l'intervalle entre l'induction et le début du premier épisode est de 11,21 minutes et 5,52 pour respectivement les seuils de 50 mmHg et 75 mmHg.

Tableau 30 : Statistiques descriptives de la survenue d'hypotension après l'induction

Seuil	Nombre d'interventions (%)	Durée médiane en dessous du seuil (min)	Durée médiane entre induction et début de l'hypotension (min)
<50	7034 (12,1%)	3,12	11,21
<55	13915 (24,0%)	4,72	9,50
<60	23122 (39,9%)	5,78	8,12
<65	32695 (56,4%)	8,23	7,00
<70	40752 (70,3%)	11,48	6,18
<75	46586 (80,4%)	16,80	5,52

3.3 Classes ASA, Age

Les tableaux 31 et 32 présentent la proportion des classes ASA 3, 4 et 5 et la moyenne (écart-type) de l'âge pour les patients présentant de l'hypotension à l'induction. Les patients présentant de l'hypotension aux seuils les plus bas (<50 mmHg, <55 mmHg et <60 mmHg) sont plus âgés et la proportion de classe ASA 3, 4 et 5 est plus importante ($p < 0,001$).

Tableau 31 : Survenue d'hypotension et classe ASA – Proportion de classes ASA 3-4-5

Seuil	Pas d'hypotension	Survenue d'hypotension	p
<50	11,3%	15,1%	<0,001
<55	23,1%	27,4%	<0,001
<60	39,3%	42,1%	<0,001
<65	56,6%	55,7%	0,017
<70	71,3%	66,7%	<0,001
<75	81,8%	75,5%	<0,001

Tableau 32 : Survenue d'hypotension à l'induction et âge

Seuil	Pas d'hypotension	Survenue d'hypotension	p
<50	50,73 (17,57)	59,25 (16,15)	<0,001
<55	50,02 (17,06)	57,31 (16,65)	<0,001
<60	49,36 (17,57)	55,39 (17,08)	<0,001
<65	49,49 (17,45)	53,53 (17,56)	<0,001
<70	50,39 (17,33)	52,35 (17,71)	<0,001
<75	51,83 (17,08)	51,75 (17,75)	0,681

3.4 Mortalité et durée de séjour

Le tableau 33 présente le taux de mortalité en fonction de la survenue d'hypotension pour les différents seuils. Ce taux est plus important pour les patients présentant une hypotension inférieure à 50 mmHg ($p < 0,01$) et inversement est plus important pour les patients présentant une hypotension au-dessous des seuils de 70 et 75 mmHg.

Tableau 33 : Survenue d'hypotension après l'induction et mortalité

Seuil	Pas d'hypotension (%)	Survenue d'hypotension (%)	p	OR
<50	0,9	1,5	< 0,001	1,68 (1,35-2,06)
<55	0,9	1,1	0,54	1,20 (0,99-1,44)
<60	1	1	0,415	1,07 (0,91-1,27)
<65	1,1	0,9	0,098	0,87 (0,74-1,02)
<70	1,3	0,8	< 0,001	0,64 (0,54-0,76)
<75	1,6	0,8	< 0,001	0,51 (0,42-0,60)

Le tableau 34 présente la durée de séjour en fonction de la survenue d'hypotension, pour les différents seuils. Lorsque le patient présente une hypotension à l'induction inférieure à 50, 55 et 60 mmHg, la durée de séjour est plus élevée ($p < 0,01$). Inversement, lorsque le patient présente une hypotension inférieure à 70 et 75 mmHg, la durée de séjour est plus courte.

Tableau 34 : Survenue d'hypotension après l'induction et durée de séjour

Seuil	Hypotension	Durée de séjour	p
<50	Non	8,87 (14,02)	<0,001
	Oui	11,30 (16,30)	
<55	Non	8,77 (14,03)	<0,001
	Oui	10,43 (15,22)	
<60	Non	8,77 (13,99)	<0,001
	Oui	9,77 (14,84)	
<65	Non	9,06 (14,68)	0,126
	Oui	9,25 (14,07)	

<70	Non	9,70 (15,13)	<0,001
	Oui	8,94 (13,99)	
<75	Non	10,68 (16,28)	<0,001
	Oui	8,80 (13,80)	

4. Discussion

Dans cette étude, nous avons caractérisé la fréquence, la profondeur et la durée de l'hypotension survenue après une induction au propofol lors d'une anesthésie générale au CHRU de Lille.

Le résultat principal de cette étude rétrospective sur 57 947 anesthésies générales induites avec au moins une administration de propofol est de mesurer la fréquence de l'hypotension induite par l'induction anesthésique au propofol à 12,1% pour une PAM < 50 mmHg. La durée médiane de cette hypotension sévère est de 3,12 min. Elle survient 11,2 min (médiane) après l'induction. La survenue de cette hypotension présente un lien statistique avec une mortalité augmentée (1,5% vs 0,9%, p < 0,001) ou avec une durée de séjour prolongée (8,8 vs 11,3 jours, p <0 ,001). L'âge plus élevé des patients dans le groupe avec hypotension est statistiquement significatif (59,2 vs 50,7 ans, p < 0,001). De même, la proportion plus importante d'hypotension lorsque le score ASA est de 3 ou 4 vs 1 ou 2 est statistiquement significative (p < 0,001).

D'autre part, l'intérêt de cette étude est de démontrer la faisabilité d'une analyse automatisée des effets indésirables induits par l'induction anesthésique avec une molécule prédéfinie, pendant une période de temps prédéterminée comme elle s'étendant de l'induction jusqu'au début de la chirurgie. Seul le premier épisode d'hypotension a été retenu pour l'analyse. L'intérêt d'analyser seulement les patients chez qui un événement "intubation" a été enregistré est d'éviter toutes les anesthésies "légères" sans intubation, essentiellement les sédations en ventilation spontanée où les doses d'hypnotiques sont plus faibles que lorsque les patients sont intubés, ou des patients déjà intubés nécessitant une reprise chirurgicale, habituellement plus grave que les autres, et chez qui la mortalité attendue est de fait augmentée.

On constate que l'influence de l'âge sur la fréquence de survenue de l'hypotension est constante pour l'ensemble des seuils de PAM analysés, avec une différence des médianes d'âge de 8,5 ans pour le seuil de PAM le plus bas, à 2 ans pour le seuil de PAM de 70 mmHg, différence qui malgré une significativé statistique n'est cependant probablement pas cliniquement relevante.

De même, la différence de durée du séjour entre le groupe hypotension et le groupe sans hypotension est de 2,4 jours pour le seuil de PAM de 50 mmHg (p < 0,001), et se réduit progressivement lorsque le seuil retenu pour définir l'hypotension augmente, devenant non significatif au seuil de 65 mmHg, ce qui a cliniquement du sens puisqu'une PAM > 60 ne peut sans doute pas être considérée comme une hypotension chez les patients en bon état général.

En ce qui concerne la mortalité, on constate qu'il faut un seuil de PAM très bas (à 50 mmHg) pour observer une augmentation de la mortalité dans le groupe hypotension (p<0.001), alors qu'il n'y a pas de différence pour les seuils de PAM supérieurs. Un effet paradoxalement délétère pourrait même être évoqué pour les seuils de PAM de 70 et 75 mmHg (p<0.001), où une analyse plus poussée des facteurs confondants doit être réalisée.

5. Conclusion

Cette étude illustre un exemple d'exploitation de l'entrepôt d'anesthésie afin de mesurer un effet indésirable fréquent de l'induction de l'anesthésie générale, permettant ainsi de rechercher des variables explicatives qui pourront être analysées par régression logistique et fournir des indices sur les paramètres et les seuils d'hypotension à éviter afin d'améliorer l'évolution postopératoire intra-hospitalière des patients pris en charge.

Chapitre 6 : Volume courant expiré

Chapitre 6 : Volume courant expiré

1. Introduction

La technique de ventilation mécanique au cours de l'anesthésie générale a évolué au cours des dernières décennies. Les recommandations des sociétés savantes ont progressivement évolué de recommandations de volume courant de l'ordre de 7 à 10 ml.kg-1 avec une fréquence ventilatoire de 8 à 12 cycles.min-1 vers une diminution du volume courant à 6 à 8 ml.kg-1 et l'application d'une PEEP afin d'éviter les atélectasies. Aujourd'hui, l'impact d'un volume courant élevé sur la morbidité et la mortalité postopératoire est souligné par de nombreux auteurs (87–89), même si des données contradictoires existent(90,91). Un élément clé, bien que technique, de cette question est l'appréciation du poids idéal du patient, car le poids réel ne peut évidemment pas être utilisé comme référence pour fixer le volume courant de référence.

Dès que des recommandations formelles émanant des sociétés savantes d'anesthésie réanimation existeront, un suivi automatisé des pratiques de chaque établissement pratiquant les soins sous anesthésie générale sera nécessaire afin de s'assurer de la bonne qualité de la ventilation. C'est dans ce contexte que l'étude suivante portant sur l'entrepôt d'anesthésie du CHRU de Lille a été réalisée : montrer la faisabilité du recueil du volume courant effectif utilisé pour la ventilation de chaque patient opéré sous anesthésie générale, et calculer patient par patient le poids idéal afin d'y rapporter le volume courant administré.

2. Méthode

Les interventions réalisées entre le 01/01/2010 et le 31/12/2014 au CHRU de Lille qui présentaient les caractéristiques suivantes ont été analysées :

- patient âgé de 18 ans ou plus ;
- poids renseigné et inférieur à 250 kg ;
- taille renseignée ;
- ventilation artificielle avec un volume courant (VCE) moyen compris entre 200 et 1000 ml.

Les interventions réalisées dans le bloc de chirurgie pédiatrique, de chirurgie thoracique et de chirurgie cardio-vasculaire ont été exclues afin de ne pas recueillir d'interventions chez des enfants (formules de poids idéal non adaptées), lors de ventilations unipulmonaires (chirurgie thoracique) et lors d'apnées prolongées au cours des CEC (chirurgie cardiaque).

Pour chaque intervention analysée, le poids idéal était calculé selon la formule *Poids idéal = 50 + 2,3 x ((taille / 2,54) – 60)* pour les hommes et *Poids idéal = 45,5 + 2,3 x ((taille / 2,54) – 60)* pour les femmes. Le volume courant expiré était moyenné sur la période de chirurgie, c'est-à-dire entre les bornes "début de chirurgie" et "fin de chirurgie" afin de n'avoir à colliger qu'une valeur par intervention, VCE_m. Le poids idéal ($poids_i$) du patient était calculé à partir des formules présentées ci-dessus. VCE_m était rapporté au $poids_i$ de façon à obtenir le volume courant pondéré idéal (VCE_i). La pression de pic (P_{peak}) était également recueillie et moyennée sur la même période de temps. A titre informatif, la durée de séjour et la mortalité ont été recueillies.

Seules des statistiques descriptives ont été réalisées: description épidémiologique de la population incluse (âge, poids, taille, IMC, score ASA, urgence ou non, spécialité chirurgicale). Les résultats ont été séparés par sexe en raison des différences liées aux formules de poids idéal entre les sexes. Les différences de durée de séjour et de mortalité ont été analysées en séparant les VCEi selon quatre intervalles prédéfinis :

- $VCE_i < 6$ ml.kg^{-1} (groupe 1)
- $6 < VCE_i < 8$ ml.kg^{-1} (groupe 2)
- $8 < VCE_i < 10$ ml.kg^{-1} (groupe 3)
- $VCE_i > 10$ ml.kg^{-1} (groupe 4)

Un test du Chi2 a été appliqué aux taux de décès mesurés pour chaque intervalle, en séparant les hommes et les femmes. Un $p<0.05$ a été considéré significatif. Les données des variables continues sont présentées sous forme *moyenne (écart type)*, celles des variables discrètes sont présentées sous forme *nombre (%)*.

3. Résultats

Parmi l'ensemble des interventions enregistrées dans l'entrepôt, 93509 correspondaient aux critères d'inclusion de cette étude. Les données épidémiologiques sont présentées tableau 1.

Variables	Homme	Femme	Total
Sexe	46468	47041	93506
Urgence	4832 (10,4%)	3826 (8,1%)	8658 (9,2%)
Age	52,2 (17,0)	50,1 (17,3)	51,2 (17,2)
ASA 1-2	33809 (72,8%)	39919 (84,9%)	73728
ASA 3-4-5	12659 (27,2%)	7122 (15,1%)	19781
Taille (cm)	175,4 (8,0)	163,2 (7,3)	169,3 (9,8)
Poids (kg)	80,7 (17,7)	70,6 (18,6)	75,6 (18,8)
IMC (kg.m-2)	26,1 (5,3)	26,5 (6,9)	26,3 (6,2)

La figure 65 présente l'évolution du VCE_i année par année, en séparant hommes/femmes.

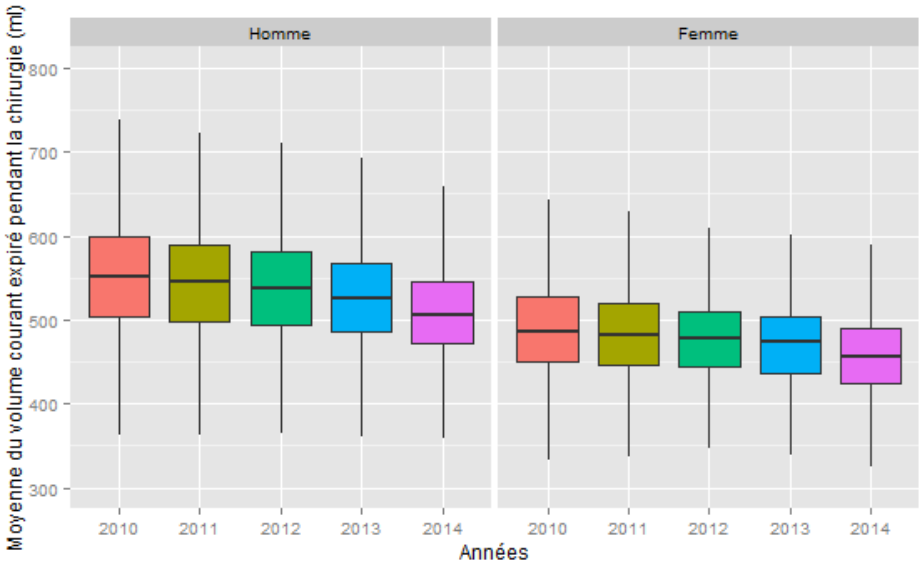


Figure 65 : Evolution du Volume courant expiré en fonction des années, par sexe

Les tableaux 35 et 36 représentent respectivement l'évolution du nombre de patients par groupe.

Tableau 35 : Evolution de la moyenne du volume courant expiré en fonction de l'année pour les hommes

Année	Groupe 1	Groupe 2	Groupe 3	Groupe 4
2010	390 (5,2%)	3888 (52,3%)	2924 (39,3%)	235 (3,1%)
2011	500 (5,8%)	4813 (56,5%)	3024 (35,5%)	176 (2,0%)
2012	535 (5,4%)	6004 (60,9%)	3178 (32,2%)	139 (1,5%)
2013	665 (6,9%)	6373 (66,0%)	2533 (26,2%)	92 (0,9%)
2014	907 (8,7%)	7713 (73,6%)	1803 (17,2%)	55 (0,5%)

Tableau 36 : Evolution de la moyenne du volume courant expiré en fonction de l'année pour les femmes

Année	Groupe 1	Groupe 2	Groupe 3	Groupe 4
2010	106 (1,5%)	1735 (23,9%)	4017 (55,4%)	1394 (19,2%)
2011	114 (1,3%)	2334 (26,1%)	5029 (56,3%)	1456 (16,3%)
2012	133 (1,4%)	2699 (27,4%)	5634 (57,3%)	1367 (13,9%)
2013	184 (1,9%)	3031 (31,0%)	5456 (55,7%)	1121 (11,4%)
2014	210 (2,0)	4147 (39,7%)	5271 (50,5%)	818 (7,8%)

4. Discussion

La population étudiée entre 2010 et 2014 est également répartie entre les hommes et les femmes, présente des scores ASA faibles (ASA 1-2) en majorité (73% pour les hommes, 85% pour les femmes), et présente un IMC similaire dans les deux sexes, à 26,3 (6,2) $kg.m^{-2}$. Le résultat principal de cette étude observationnelle sur 93509 interventions sous anesthésie générale avec intubation et ventilation contrôlée est que la proportion de femmes dans les groupes 3 et 4 est largement supérieure à celle des hommes (58.3% vs 17.7% en 2014).

Un deuxième résultat important est que la tendance globale entre 2010 et 2014 est à la diminution du volume courant (figure 65), avec un volume courant médian qui passe de 551,3 ml à 504,5 ml chez les hommes, et de 484,4 ml à 455,0 ml chez les femmes sur cette période. Mais cette diminution ne suffit pas à diminuer la proportion de femmes "surventilées".

Différentes explications peuvent être avancées pour expliquer ces observations:

- une erreur dans le calcul automatisé du poids idéal ; cette hypothèse à été vérifiée et éliminées lors des tests préliminaires à la réalisation de l'étude. La détection automatique du sexe du patient, de même que sa taille, pourraient avoir conduit à des erreurs de formules, conduisant à une surestimation des volumes courants administrés par rapport au poids idéal. En pratique, toutes les mesures de taille aberrantes sont automatiquement éliminées, et ces enregistrements ne sont pas traités.

- une formule de poids idéal adaptée pour l'homme mais pas pour la femme. La formule de poids idéal utilisée ici est celle de Devine (1974). Il serait intéressant de tester si les mêmes résultats sont obtenus avec d'autres formules.

Quoi qu'il en soit, la puissance de calcul de l'entrepôt de données d'anesthésie permet à l'évidence de "surveiller" l'évolution des volumes courants administrés aux patients sous ventilation mécanique, ce qui

fournit une aide indéniable aux responsables qualité des pôles d'anesthésie pour vérifier si les messages émanant des sociétés savantes sont appliqués correctement sur le terrain. Le lien avec une éventuelle morbidité n'a pas été étudié ici. La difficulté principale sera d'établir la causalité entre un volume courant élevé et une augmentation de ma morbidité post opératoire, car les facteurs confondants sont nombreux.

5. Conclusion

Cette étude démontre une fois de plus la puissance et la relative simplicité de mise en œuvre de l'entrepôt d'anesthésie. Son utilité pour l'amélioration des pratiques devrait être relativement simple à démontrer en reproduisant cette étude d'année en année.

Discussion

Discussion

De plus en plus de données sont enregistrées quotidiennement dans les structures de soins par des systèmes opérationnels, peu connectés entre eux. Dans ce travail nous nous posons la question de l'utilisation secondaire de ces données, et en particulier celles enregistrées par la feuille informatisée d'anesthésie et les logiciels administratifs et de PMSI gérant les séjours des patients.

L'intérêt principal de l'utilisation secondaire des données est de pouvoir réutiliser des volumes très importants de données enregistrées par des applications non ou peu interconnectées. Les problèmes rencontrés lors de la réutilisation de ce type de données sont liés à la variabilité de la qualité des données, la volumétrie et l'hétérogénéité entre les différents systèmes.

L'utilisation secondaire de données de santé nécessite plusieurs étapes, présentées dans ce travail :

La première étape consiste à comprendre les flux d'informations au sein du système d'information et à identifier les applications opérationnelles en lien avec la problématique à traiter (cf. chap. 1). En effet, le SIH compte un nombre important d'applications et chacune de ces applications enregistre un important volume de données, dont une partie est destinée à la gestion quotidienne de l'application. Lors de cette étape, il s'agit de déterminer quelles données présentent un intérêt pour répondre à une problématique de recherche clinique, d'aide la décision ou d'analyse médico-économique.

La seconde étape consiste à évaluer la qualité des données enregistrées par les systèmes opérationnels sélectionnés (cf. chap. 2). Pour cela, un ensemble de méthodes d'évaluation de la qualité peut être mis en œuvre en fonction des problèmes de qualité recherchés et des caractéristiques des données (volumétrie, type de données, ...).

Lors de **la troisième étape**, les données sont intégrées depuis les systèmes sources vers une structure commune : l'entrepôt de données (cf. chap. 3). Au cours de l'alimentation de l'entrepôt de données, les enregistrements sont dédoublonnées, nettoyés et transformés en fonction des problèmes de qualité détectés précédemment. Ils sont ensuite chargés dans un modèle de données commun, optimisé pour pouvoir interroger facilement un volume important de données. Lors de ce processus appelé ETL (pour Extract, Transform, Load), la qualité des données est caractérisée. En effet, les besoins en termes de qualité de données vont différer en fonction de l'utilisation.

Après l'intégration des données sources, un travail d'agrégation permet de résumer les informations pertinentes à travers des indicateurs métier (cf. chap. 4). Ainsi, dans ce travail, nous avons proposé plusieurs méthodes d'agrégation pour calculer des mesures agrégées et détecter automatiquement des techniques d'anesthésie ou la survenue d'événements indésirables. Ces données n'étaient pas présentes dans les systèmes sources et facilitent les analyses statistiques en proposant des indicateurs calculées à partir de données consolidées et en résumant les informations contenues au départ dans un volume de données important.

1. Bilan

1.1 Méthodes

Nous avons travaillé sur toutes les étapes nécessaires à la réutilisation de données enregistrées par des applications opérationnelles. Pour cela, différents types de méthodes ont été employées. Celles-ci étaient issues de la littérature ou ont été développées par nous même lorsque c'était nécessaire.

Ainsi, lors de l'étape d'évaluation de la qualité des systèmes sources, nous avons utilisé plusieurs méthodes définies par la littérature comme l'analyse de colonne, l'analyse de domaine et l'analyse de clé primaire. La majorité de ces méthodes peuvent être mises en œuvre automatiquement et permettent ainsi d'évaluer un grand volume de données hétérogènes. En revanche, plusieurs méthodes comme le *gold standard* ou le *data element agreement*, dont la littérature fait référence (61), n'ont pas pu être appliquées en raison de l'absence de base de données de référence pour contrôler les enregistrements. Woodall *et al.* rapportent également plusieurs manques pour détecter certains problèmes de qualité (67). Les méthodes automatiques d'évaluation de la qualité présentent surtout des lacunes pour détecter les enregistrements manquants ou certifier de l'exactitude des informations enregistrées : dans le cas de la feuille informatisée d'anesthésie, aucune autre source de données n'est disponible.

Lors de la phase d'intégration, deux types de méthodes sont employées pour optimiser la qualité des données disponibles pour répondre aux analyses statistiques. Tout d'abord, les méthodes d'intégration et de nettoyage (26,27) visent à stocker des informations détaillées dans l'entrepôt de données, en conservant les données avec une qualité moindre (le niveau de qualité requis dépend de l'utilisation que l'on veut faire des données). Lors de cette étape, seules les données inutilisables sont supprimées. Puis les méthodes d'agrégation de données (26,27) permettent de retravailler et résumer l'information contenue dans les données détaillées pour proposer à l'utilisateur des indicateurs métier (cf. chap. 4).

Les différentes méthodes utilisées dans ce travail pourront être réutilisées lors de l'extension du projet pour intégrer des données en provenance d'autres systèmes (cf. Perspectives).

1.2 Données disponibles

L'entrepôt de données collige les informations enregistrées par les systèmes sources depuis 2010. Le tableau 38 présente la volumétrie enregistrée pour chaque type de données. L'entrepôt de données continuera à être alimenté de manière hebdomadaire à la fin de ce projet de thèse. Ainsi, environ 55 000 nouvelles interventions seront intégrées chaque année et pourront également servir de support aux futures études.

Tableau 37 : Données disponibles

Données	Nombre d'enregistrements
Patients	175 214
Interventions	276 812
Evénements	43 314 015
Mesures	1 545 582 585
Séjours	2 377 129
Mouvements	1 880 072
Actes CCAM	8 830 944
Diagnostics	6 290 712

1.3 Etudes réalisées et à venir

L'exploitation des données collectées au sein de l'entrepôt de données permet de réaliser des études cliniques, mais aussi des travaux d'informatique médicale (Annexe 12).

En anesthésie, grâce à la volumétrie et aux types d'informations stockées dans l'entrepôt de données, plusieurs types d'études peuvent être réalisées rétrospectivement (29) : recherche clinique (recherche de cas pour une étude, analyse exploratoire, ...), évaluation de la qualité (évaluation des complications en fonction de techniques de prise en charge ou de traitements), management des activités (optimisation, ...).

D'un point de vue universitaire, l'entrepôt de données peut être interrogé pour sélectionner et extraire automatiquement des informations au lieu de parcourir les dossiers patients manuellement. Ainsi, une dizaine d'études ont pu être réalisées dans le cadre de thèses de médecine, de mémoires d'internes d'anesthésie et de mémoires d'élèves Infirmiers-Anesthésistes Diplômés d'Etat (IADE). Certaines de ses études ont fait l'objet de présentation au congrès de la Société Française d'Anesthésie Réanimation et de l'European Society of Anesthesia (92).

Plusieurs études ont également été menées pour évaluer la qualité des soins (voir Chapitre 5 et 6) et les techniques de prises en charge.

Plusieurs thématiques d'informatique médicale peuvent également être abordées dans ce travail : qualité dans les bases de données opérationnelles, méthode d'évaluation de la qualité, méthode d'intégration et de nettoyages des données dans un entrepôt de données, méthodes de calcul des données agrégées : fenêtre d'étude (81), temps hors seuil.

1.4 Difficultés rencontrées

La mise en place d'un projet d'informatique médicale, et plus particulièrement d'entrepôt de données rencontre plusieurs types de difficultés, de l'acceptation du projet jusqu'à l'exploitation des données (31) (voir tableau 39).

Les premières difficultés rencontrées sont d'ordre politique et réglementaire, et sont inhérentes à la mise en place d'un projet informatique au sein d'une structure de soins : difficulté d'accès aux bases de données sources, acceptation du projet par les utilisateurs finaux (anesthésistes), interface relationnelle entre les différents acteurs ou encore gestion de la confidentialité des données. Pour maîtriser ces difficultés nous avons créé un groupe de travail avec les interlocuteurs clés (anesthésistes, informaticiens, statisticiens…). La gouvernance du projet par le comité de pilotage permet également de choisir communément les plans d'actions à mettre en œuvre. Enfin, ce travail a été inscrit au projet d'établissement du CHRU de Lille et bénéficie ainsi du support de la hiérarchie institutionnelle.

Lors de l'étape d'analyse des besoins et de création du modèle de données, les principales difficultés ont été de comprendre les besoins des cliniciens (leur implication dans ce projet étant souvent la première dans un projet informatique) et de trouver un consensus entre eux (leur niveau d'expérience et la différence des pratiques entre les services résultant dans des points de vue différents). Pour surmonter ces difficultés, un groupe de travail a été défini et regroupe des experts de chaque métier. La participation au projet des élèves infirmiers anesthésistes lors de leur stage recherche a également été l'occasion de bénéficier de leur expertise quant à l'utilisation du logiciel DIANE. A chaque étape du développement, des chiffres clés illustrant l'activité sont proposés aux cliniciens afin de valider avec eux la validité des données.

Enfin, plusieurs difficultés d'ordre technique sont généralement rencontrées sur ce type de projet, lorsque qu'il s'agit de manipuler de grands volumes de données hétérogènes. L'étape d'intégration et de nettoyage des données est essentielle puisqu'elle collige les données enregistrées par des systèmes distincts et d'optimiser leur qualité. Puis la phase d'agrégation des données permet de résumer les informations pertinentes à travers des indicateurs pré-calculés. L'optimisation et le *tuning* des bases de données avec les technologies adéquates et proposées par les éditeurs de bases de données permettent également de diminuer les temps de calcul.

Tableau 38 : Difficultés rencontrées au cours du projet

Type de problèmes	Difficultés	Solutions
Politique et réglementaire	Mise en place du projet - Acceptation par les utilisateurs finaux.	Identifier et convaincre les interlocuteurs clé. Mise en place d'un comité de pilotage du projet.
	Obtenir l'accès aux bases de données sources.	Convaincre les décideurs de l'intérêt du projet
	Faire l'interface relationnelle entre les différents acteurs / services impliqués dans la gestion ou l'utilisation des données hospitalières.	Partager les informations. N'utiliser les données que pour les objectifs définis.
	Hébergement et confidentialité des données.	Déclaration CNIL. Anonymisation des données pour les analyses.
Analyse des besoins	Comprendre les processus métier et les flux d'informations (56) : pas de consensus entre les utilisateurs.	Identifier les experts métier et définir un groupe de travail
	Comprendre l'architecture et le fonctionnement des applications sources (31) : les applications sont souvent mal documentées.	Tests et collaboration avec l'éditeur ou les utilisateurs de l'application
Qualité des données	Données incohérentes – Doublons.	Mettre en place des méthodes d'intégration et de nettoyage de données. Améliorer la qualité des données dans les systèmes sources.

Création du modèle de données (56)	Dimensions définies différemment dans les systèmes.	Développer un modèle de données orienté métier et indépendant des applications opérationnelles.
	Prévoir les requêtes les plus courantes.	
	Définir la granularité des données.	
Intégration des données (56)	Optimiser le temps d'alimentation.	Partitionner les tables pour travailler sur quelques partitions liées au flux d'alimentation.
	Mettre à jour les données nécessaires lors de l'alimentation.	Paramétrage du flux d'alimentation et marquage des données (date de création et date de mise à jour).
	Maintenance de la chaîne d'alimentation.	Documenter les différentes processus et modules de la chaîne d'alimentation. Paramétrer les processus pour pouvoir les exécuter indépendamment. Développer des indicateurs d'alimentation.
Exploitation des données	Analyser un grand volume de données.	Agrégation des données. Optimisation et tuning de la base de données. Utilisation de logiciels appropriés (R).

2. Perspectives

2.1 Axe multicentrique

Tout d'abord, il est envisageable d'intégrer les données enregistrées par d'autres centres de soins avec leur feuille informatisée d'anesthésie respective. En effet, le schéma de l'entrepôt de données est indépendant du système source. Dans le cas où ceux-ci étaient différents de ceux intégrés dans ce travail, seuls les premiers modules du processus ETL devraient être adaptés afin de correspondre aux schémas des systèmes sources. Cette évolution permettrait de comparer des types de populations et des pratiques différentes, mais aussi de bénéficier d'une volumétrie encore plus conséquente.

2.2 Axe métier

Il est également possible d'étendre le domaine métier de l'entrepôt de données, ou de le connecter à d'autres entrepôts de données. Les logiciels de réanimation, en particulier, présentent des caractéristiques proches de celles de la feuille informatisée d'anesthésie. Cela serait l'occasion de réutiliser les méthodes employées dans le cadre de cette thèse. L'intégration de ce type d'informations permettrait également de suivre les patients dans le temps, et de répondre à d'autres problématiques.

2.3 Améliorations des logiciels sources

L'évaluation de la qualité des données a permis de mettre en évidence des problèmes de qualité liés aux applications logicielles. Ces problèmes de qualité n'ont pas toujours été détectés lors du développement des logiciels et se révèlent après plusieurs années d'utilisation. Ce travail pourrait servir de support aux éditeurs des applications évalués.

2.4 Identito-vigilance

L'entrepôt de données met en commun des informations issues de systèmes sources différents. Certaines informations comme les identités des patients sont partagées par les trois systèmes, CORA, GAM et DIANE. Elles sont enregistrées à deux moments du séjour hospitalier, respectivement lors de l'admission et lors de l'entrée au bloc opératoire. Grâce à l'entrepôt de données, il est possible de comparer les trois sources d'informations afin de corriger l'une d'elles si des erreurs sont détectées ou encore de compléter des données manquantes en s'appuyant les autres sources.

2.5 Aide à la décision

Les données agrégées (voir chapitre 4) peuvent être intégrées dans des solutions d'aide à la décision comme des tableaux ou des rapports statistiques et permettraient aux cliniciens d'avoir une vision globale de leur activité.

Conclusion

Conclusion

Ce travail aborde la thématique de la réutilisation des données de santé. Un travail en plusieurs étapes a été réalisé et a permis d'intégrer les données enregistrées par des applications opérationnelles hétérogènes au sein d'un entrepôt de données. Ce système, optimisé pour interroger des volumes de données importants, a été employé pour réaliser plusieurs analyses rétrospectives. Ces études n'auraient pas pu être réalisées directement avec les systèmes sources en raison de l'hétérogénéité, de la qualité et de la volumétrie des données.

L'intérêt clinique de ce travail est de pouvoir mettre en rapport des informations enregistrées par des applications différentes ; en l'occurrence les événements indésirables apparus lors de l'intervention ainsi que les techniques de prise en charge avec la survenue de complications au cours du séjour hospitalier. Plusieurs études ont par exemple mis en évidence des liens statistiques entre la survenue d'hypotension et une augmentation de la durée de séjour ou de la mortalité (46–48).

La première étape de ce travail a permis de sélectionner 4 systèmes déployés au CHRU de Lille : 3 applications opérationnelles et un entrepôt de données, l'objectif étant d'intégrer les informations enregistrées par les applications opérationnelles au sein de l'entrepôt de données. Dans un second temps, la qualité des données de ces systèmes a été évaluée et a permis de mettre en évidence différents problèmes de qualité dus à l'hétérogénéité des systèmes, à des erreurs de développements des applications et bases de données sources, ainsi qu'à une variabilité dans les habitudes d'utilisation des interfaces ou à un mauvais renseignement des informations par les utilisateurs. Cela nous a permis de mettre en application les méthodes d'évaluation de la qualité des données détaillée dans la littérature ou développées pour ce travail.

Le troisième temps de ce travail avait pour objectif d'intégrer les données opérationnelles au sein d'un modèle de données commun, rendant possible le croisement d'informations enregistrées par des systèmes sources distincts. Cette étape est décisive puisqu'elle fournit un socle de données consolidées, permettant par la suite de calculer des données agrégées qui résument l'information et permettent de réaliser facilement des études sur des grands volumes de données.

L'entrepôt de données a permis de réaliser plusieurs études rétrospectives, dont deux ont été présentées ici. Le premier visait à déterminer la fréquence de l'hypotension suivant l'induction au propofol dans le cadre d'une anesthésie générale. La seconde avait pour objectif d'évaluer le respect des recommandations en termes de ventilation et l'impact sur le devenir du patient.

Ce travail a permis de mettre en avant plusieurs points :

- Les données enregistrées par les applications opérationnelles du SIH peuvent être réutilisées pour réaliser des études rétrospectives. La feuille informatisée d'anesthésie, en particulier, permet d'évaluer la qualité de prise en charge et l'impact des événements indésirables au cours de la procédure d'anesthésie ;

- Les problèmes de qualité présents dans ces enregistrements doivent être évalués pour déterminer dans quelle mesure ces enregistrements peuvent être réutilisés, et quelles méthodes de nettoyage de données doivent être employées pour obtenir une qualité de données optimale ;

- Même si certains problèmes de qualité de données peuvent être résolus ou maîtrisés a posteriori lors de l'intégration des données sources dans l'entrepôt de données, les résultats de l'évaluation de la

qualité des données pourraient être mis à profit pour améliorer les systèmes sources et optimiser la qualité des données en amont.

Les données analysées par les professionnels de l'anesthésie du CHRU de Lille doivent permettre de définir de « nouvelles règles » pour la prise en charge anesthésique peropératoire. Le respect de ces règles devrait permettre à terme de voir diminuer la morbidité et la mortalité des anesthésies telles qu'elles sont observées aujourd'hui.

Suite à ce travail, il est envisagé d'intégrer d'autres systèmes afin d'élargir le type d'études possible.

Références

Références

1. Haux R. Health information systems – past, present, future. Int J Med Inf. 2006 Mar;75(3–4):268–81.

2. Reichertz PL. Hospital information systems—Past, present, future. Int J Med Inf. 2006 Mar;75(3–4):282–99.

3. Beck T, Gollapudi S, Brunak S, Graf N, Lemke HU, Dash D, et al. Knowledge engineering for health: a new discipline required to bridge the "ICT gap" between research and healthcare. Hum Mutat. 2012 May;33(5):797–802.

4. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inf. 2013 Jan;82(1):1–9.

5. Edsall DW, Deshane P, Giles C, Dick D, Sloan B, Farrow J. Computerized patient anesthesia records: less time and better quality than manually produced anesthesia records. J Clin Anesth. 1993 Aug;5(4):275–83.

6. Sockolow PS, Bowles KH, Adelsberger MC, Chittams JL, Liao C. Impact of homecare electronic health record on timeliness of clinical documentation, reimbursement, and patient outcomes. Appl Clin Inform. 2014;5(2):445–62.

7. Burke HB, Sessums LL, Hoang A, Becher DA, Fontelo P, Liu F, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc. 2015 Jan 1;22(1):199–205.

8. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff Proj Hope. 2005 Oct;24(5):1103–17.

9. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc JAMIA. 2007;14(1):1–9.

10. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Ann Intern Med. 2009 Sep 1;151(5):359–60.

11. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38–44.

12. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Summits Transl Sci Proc. 2010 Mar 1;2010:1–5.

13. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med. 2013 Dec 1;274(6):547–60.

14. Serguei Pakhomov P, Susan A. Weston MS, Steven J. Jacobsen MD, Christopher G. Chute MD, Ryan Meverden BS, and V??ronique L. Roger MD. Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure. Am J Manag Care [Internet]. 2007 Jun 1 [cited 2015 Apr 6];13(June 2007 - Part 1 6 - Pt 1). Available from: http://www.ajmc.com/publications/issue/2007/2007-06-vol13-n6-pt1/Jun07-2488p281-288

15. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 Jun;13(6):395–405.

16. Miller JL. The EHR solution to clinical trial recruitment in physician groups. Health Manag Technol. 2006 Dec;27(12):22–5.

17. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc JAMIA. 2009 Dec;16(6):869–73.

18. Pivette M, Mueller JE, Crépey P, Bar-Hen A. Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review. BMC Infect Dis. 2014 Nov 18;14(1):604.

19. Heffernan R, Mostashari F, Das D, et al. System descriptions: New York City syndromic surveillance systems. Morb Mortal Wkly Rep. 2004;53 (Suppl):23–7.

20. Howley MJ, Chou EY, Hansen N, Dalrymple PW. The long-term financial impact of electronic health record implementation. J Am Med Inform Assoc JAMIA. 2015 Mar;22(2):443–52.

21. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM. Bridging the inferential gap: the electronic health record and clinical evidence. Health Aff Proj Hope. 2007 Apr;26(2):w181–91.

22. Ancker JS, Shih S, Singh MP, Snyder A, Edwards A, Kaushal R. Root Causes Underlying Challenges to Secondary Use of Data. AMIA Annu Symp Proc. 2011;2011:57–62.

23. Holzer K, Gall W. Utilizing IHE-based Electronic Health Record systems for secondary use. Methods Inf Med. 2011;50(4):319–25.

24. Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, Davidson KW, et al. Using EHRs to integrate research with patient care: promises and challenges. J Am Med Inform Assoc. 2012 Sep 1;19(5):684–7.

25. Dentler K, ten Teije A, de Keizer N, Cornet R. Barriers to the reuse of routinely recorded clinical data: a field report. Stud Health Technol Inform. 2013;192:313–7.

26. Inmon WH. Building the Data Warehouse. Wiley; 1992. 320 p.

27. Kimball R. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons; 1998. 801 p.

28. Myers DL, Burke KC, Burke JD Jr, Culp KS. An integrated data warehouse system: development, implementation, and early outcomes. Manag Care Interface. 2000 Mar;13(3):68–72.

29. Einbinder JS, Scully KW, Pates RD, Schubart JR, Reynolds RE. Case study: a data warehouse for an academic medical center. J Healthc Inf Manag JHIM. 2001;15(2):165–75.

30. Silver M, Sakata T, Su HC, Herman C, Dolins SB, O'Shea MJ. Case study: how to apply data mining techniques in a healthcare data warehouse. J Healthc Inf Manag JHIM. 2001;15(2):155–64.

31. Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers M, Weinstein RA. Development of a Clinical Data Warehouse for Hospital Infection Control. J Am Med Inform Assoc. 2003 Sep 1;10(5):454–62.

32. Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF. Atlas – a data warehouse for integrative bioinformatics. BMC Bioinformatics. 2005 Feb 21;6(1):34.

33. Lyman JA, Scully K, Harrison Jr. JH. The Development of Health Care Data Warehouses to Support Data Mining. Clin Lab Med. 2008 Mar;28(1):55–71.

34. De Mul M, Alons P, van der Velde P, Konings I, Bakker J, Hazelzet J. Development of a clinical data warehouse from an intensive care clinical information system. Comput Methods Programs Biomed. 2012 Jan;105(1):22–30.

35. Alshawi S, Saez-Pujol I, Irani Z. Data warehousing in decision support for pharmaceutical R&D supply chain. Int J Inf Manag. 2003 Jun;23(3):259–68.

36. Chazard E. Automated detection of adverse drug events by data mining of electronic health records [Internet]. Université du Droit et de la Santé - Lille II; 2011 [cited 2014 Feb 5]. Available from: http://tel.archives-ouvertes.fr/tel-00637254

37. Cao X, Wong STC, Hoo KS, Tjandra D, Fu JC, Lowenstein DH. A web-based federated neuroinformatics model for surgical planning and clinical research applications in epilepsy. Neuroinformatics. 2004;2(1):101–18.

38. ISO/TR 22221:2006 - Health informatics - Good principles and practices for a clinical data warehouse [Internet]. [cited 2015 Apr 7]. Available from: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40783

39. Décret no 94-1050 du 5 décembre 1994 relatif aux conditions techniques de fonctionnement des établissements de santé en ce qui concerne la pratique de l'anesthésie et modifiant le code de la santé publique (troisième partie: Décrets). 94-1050 décembre, 1994.

40. Lienhart A, Auroy Y, Péquignot F, Benhamou D, Warszawski J, Bovet M, et al. Survey of anesthesia-related mortality in France. Anesthesiology. 2006 Dec;105(6):1087–97.

41. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, et al. Mortality after surgery in Europe: a 7 day cohort study. Lancet. 2012 Sep 22;380(9847):1059–65.

42. Hamilton MA, Cecconi M, Rhodes A. A systematic review and meta-analysis on the use of preemptive hemodynamic intervention to improve postoperative outcomes in moderate and high-risk surgical patients. Anesth Analg. 2011 Jun;112(6):1392–402.

43. Reich DL, Bennett-Guerrero E, Bodian CA, Hossain S, Winfree W, Krol M. Intraoperative tachycardia and hypertension are independently associated with adverse outcome in noncardiac surgery of long duration. Anesth Analg. 2002 Aug;95(2):273–7, table of contents.

44. Bijker JB, van Klei WA, Vergouwe Y, Eleveld DJ, van Wolfswinkel L, Moons KGM, et al. Intraoperative hypotension and 1-year mortality after noncardiac surgery. Anesthesiology. 2009 Dec;111(6):1217–26.

45. Kertai MD, Pal N, Palanca BJA, Lin N, Searleman SA, Zhang L, et al. Association of perioperative risk factors and cumulative duration of low bispectral index with intermediate-term mortality after cardiac surgery in the B-Unaware Trial. Anesthesiology. 2010 May;112(5):1116–27.

46. Sessler DI, Sigl JC, Kelley SD, Chamoun NG, Manberg PJ, Saager L, et al. Hospital stay and mortality are increased in patients having a “triple low” of low blood pressure, low bispectral index, and low minimum alveolar concentration of volatile anesthesia. Anesthesiology. 2012 Jun;116(6):1195–203.

47. Walsh M, Devereaux PJ, Garg AX, Kurz A, Turan A, Rodseth RN, et al. Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension. Anesthesiology. 2013 Sep;119(3):507–15.

48. Kertai MD, White WD, Gan TJ. Cumulative duration of “triple low” state of low blood pressure, low bispectral index, and low minimum alveolar concentration of volatile anesthesia is not associated with increased mortality. Anesthesiology. 2014 Jul;121(1):18–28.

49. Lesser JB, Sanborn KV, Valskys R, Kuroda M. Severe bradycardia during spinal and epidural anesthesia recorded by an anesthesia information management system. Anesthesiology. 2003 Oct;99(4):859–66.

50. Sanborn KV, Castro J, Kuroda M, Thys DM. Detection of intraoperative incidents by electronic scanning of computerized anesthesia records. Comparison with voluntary reporting. Anesthesiology. 1996 Nov;85(5):977–87.

51. SFAR - Dossier anesthésique (SFAR 2001) [Internet]. [cited 2015 Feb 14]. Available from: http://www.sfar.org/article/54/dossier-anesthesique-sfar-2001

52. BOW Médical [Internet]. [cited 2014 Jul 5]. Available from: http://www.bowmedical.com/

53. Présentation | Publication ATIH [Internet]. [cited 2015 Mar 19]. Available from: http://www.atih.sante.fr/mco/presentation

54. CCAM en ligne - CCAM [Internet]. [cited 2015 Feb 14]. Available from: http://www.ameli.fr/accueil-de-la-ccam/index.php

55. Aide a la classification avec la CIM 10 (+PMSI) [Internet]. [cited 2015 Feb 14]. Available from: http://taurus.unine.ch/icd10

56. Gray GW. Challenges of building clinical data analysis solutions. J Crit Care. 2004 Dec;19(4):264–70.

57. Hogan WR, Wagner MM. Accuracy of Data in Computer-based Patient Records. J Am Med Inform Assoc. 1997;4(5):342–55.

58. Brennan PF, Stead WW. Assessing Data Quality From Concordance, through Correctness and Completeness, to Valid Manipulatable Representations. J Am Med Inform Assoc. 2000 Jan 1;7(1):106–7.

59. Piprani B, Ernst D. A Model for Data Quality Assessment. In: Meersman R, Tari Z, Herrero P, editors. On the Move to Meaningful Internet Systems: OTM 2008 Workshops [Internet]. Springer Berlin Heidelberg; 2008 [cited 2015 Feb 9]. p. 750–9. Available from: http://link.springer.com/chapter/10.1007/978-3-540-88875-8_99

60. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013 Oct;46(5):830–6.

61. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.

62. Kim W, Choi B-J, Hong E-K, Kim S-K, Lee D. A taxonomy of dirty data. Data Min Knowl Discov. 2003;7(1):81–99.

63. Müller H, j. Problems, Methods and Challenges in Comprehensive Data Cleansing. Humboldt-Universität zu Berlin, Institut für Informatik; 2003. Report No.: HUB-IB-164.

64. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000;23(4):3–13.

65. Oliveira P, Rodrigues F, Henriques PR. A Formal Definition of Data Quality Problems. IQ [Internet]. 2005 [cited 2014 Jul 18]. Available from: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf

66. Johnson T, Dasu T. Data Quality and Data Cleaning: An Overview. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data [Internet]. New York, NY, USA: ACM; 2003 [cited 2015 Feb 24]. p. 681–681. Available from: http://doi.acm.org/10.1145/872757.872875

67. Woodall P, Oberhofer M, Borek A. A Classification of Data Quality Assessment and Improvement Methods. [cited 2015 Jan 28]; Available from: http://www.researchgate.net/profile/Philip_Woodall/publication/266030806_A_Classification_of_Data_Quality_Assessment_and_Improvement_Methods/links/5423f6cb0cf238c6ea6e8035.pdf

68. Hainaut J-L. Bases de données: concepts, utilisation et développement : cours et exercices corrigés. Paris: Dunod; 2011.

69. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. VLDB J. 2001 Dec;10(4):334–50.

70. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969 Dec;64(328):1183.

71. Christen P. Data Matching [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 [cited 2014 Jul 18]. Available from: http://link.springer.com/10.1007/978-3-642-31164-2

72. Oracle | Hardware and Software, Engineered to Work Together [Internet]. [cited 2015 Feb 24]. Available from: http://www.oracle.com/index.html

73. UTL_MATCH [Internet]. [cited 2015 Feb 24]. Available from: https://docs.oracle.com/cd/E18283_01/appdev.112/e16760/u_match.htm

74. Fox C, Levitin A, Redman T. The notion of data and its quality dimensions. Inf Process Manag. 1994 Jan;30(1):9–19.

75. Loshin D. Master data management. Amsterdam ; Boston: Elsevier/Morgan Kaufmann; 2009. 274 p.

76. Weil G, Motamed C, Eghiaian A, Guye ML, Bourgain JL. The use of a clinical database in an anesthesia unit: focus on its limits. J Clin Monit Comput. 2014 May 17;1–5.

77. Partitions, Views, and Other Schema Objects [Internet]. [cited 2015 Mar 17]. Available from: http://docs.oracle.com/cd/E11882_01/server.112/e40540/schemaob.htm#CNCPT88859

78. Managing Indexes [Internet]. [cited 2015 Mar 26]. Available from: http://docs.oracle.com/cd/B19306_01/server.102/b14231/indexes.htm

79. Maletic JI, Marcus A. Data Cleansing: Beyond Integrity Analysis. 2000.

80. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ [Internet]. 2009 [cited 2015 Mar 26];338. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2714692/

81. Lamer A, De Jonckheere J, Marcilly R, Tavernier B, Vallet B, Jeanne M, et al. A substitution method to improve completeness of events documentation in anesthesia records. J Clin Monit Comput. 2015 Jan 30;

82. Maintaining Partitions [Internet]. [cited 2015 Mar 26]. Available from: http://docs.oracle.com/cd/E11882_01/server.112/e25523/part_admin002.htm#i1008226

83. Lamer A, Marcilly R, Jeanne M, Logier R. Automatic scanning of free-text entries. Stud Health Technol Inform. 2014;205:1196.

84. Database SQL Language Reference [Internet]. [cited 2015 Mar 18]. Available from: https://docs.oracle.com/database/121/SQLRF/functions003.htm#SQLRF20035

85. Analytic Functions [Internet]. [cited 2015 Mar 11]. Available from: http://docs.oracle.com/cd/E11882_01/server.112/e41084/functions004.htm#SQLRF06174

86. 11g-pivot.html [Internet]. [cited 2015 Mar 11]. Available from: http://www.oracle.com/technetwork/articles/sql/11g-pivot-097235.html

87. Futier E, Constantin J-M, Paugam-Burtz C, Pascal J, Eurin M, Neuschwander A, et al. A Trial of Intraoperative Low-Tidal-Volume Ventilation in Abdominal Surgery. N Engl J Med. 2013 Jul 31;369(5):428–37.

88. Severgnini P, Selmo G, Lanza C, Chiesa A, Frigerio A, Bacuzzi A, et al. Protective mechanical ventilation during general anesthesia for open abdominal surgery improves postoperative pulmonary function. Anesthesiology. 2013 Jun;118(6):1307–21.

89. Gu W-J, Wang F, Liu J-C. Effect of lung-protective ventilation with lower tidal volumes on clinical outcomes among patients undergoing surgery: a meta-analysis of randomized controlled trials. CMAJ Can Med Assoc J J Assoc Medicale Can. 2015 Feb 17;187(3):E101–9.

90. Fernandez-Bustamante A, Klawitter J, Repine JE, Agazio A, Janocha AJ, Shah C, et al. Early effect of tidal volume on lung injury biomarkers in surgical patients with healthy lungs. Anesthesiology. 2014 Sep;121(3):469–81.

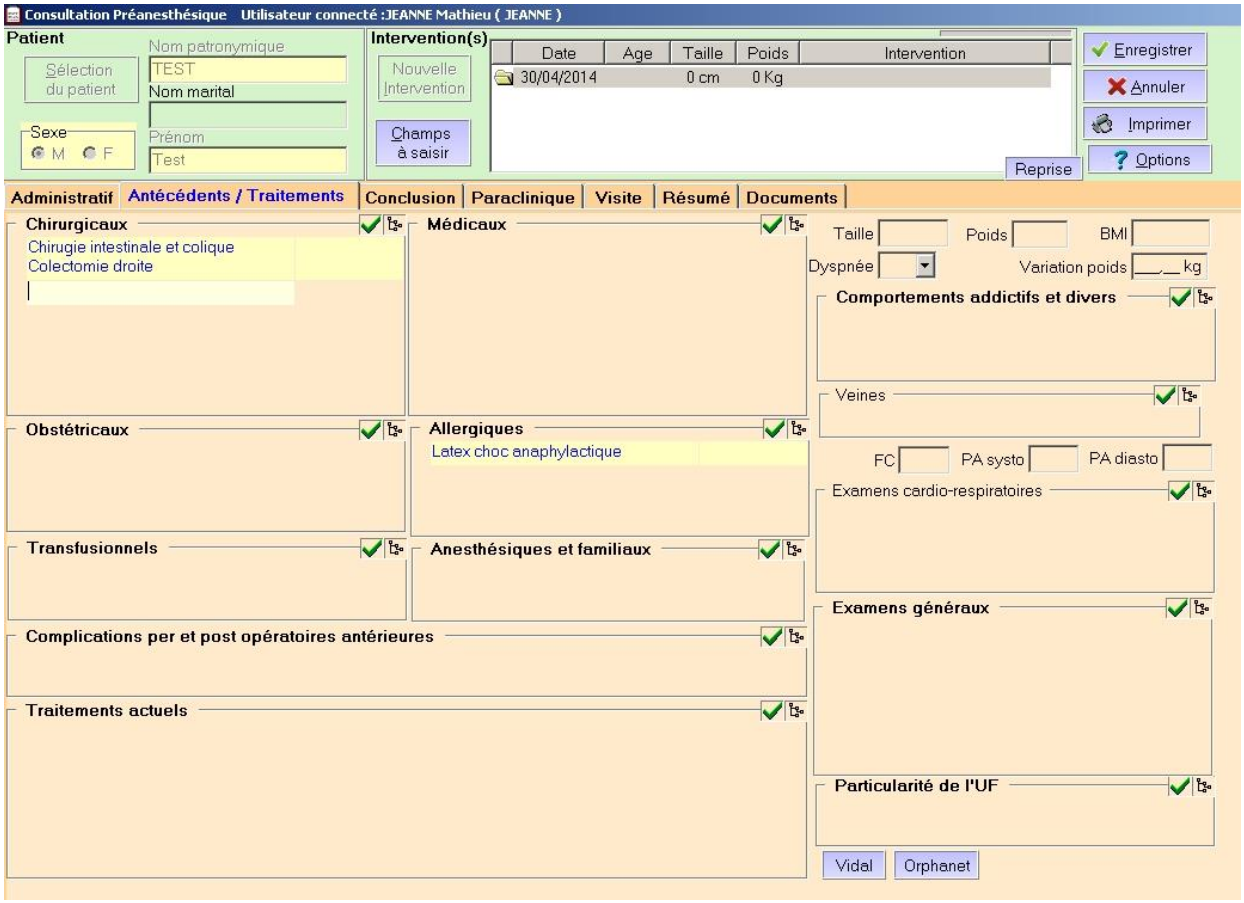
91. Hansen JK, Anthony DG, Li L, Wheeler D, Sessler DI, Bashour CA. Comparison of Positive End-Expiratory Pressure of 8 versus 5 cm H2O on Outcome After Cardiac Operations. J Intensive Care Med. 2014 Jan 31;

92. Jaspar J, Lallemant F, Jeanne M, Lamer A, Vallet B, Tavernier B. Caractérisation de l'anesthésie générale : analyse rétrospective des patients opérés d'une fracture du col fémoral au CHU de Lille. Ann Fr Anesth Réanimation. 2014 Sep;33, Supplement 2:A358.

Annexes

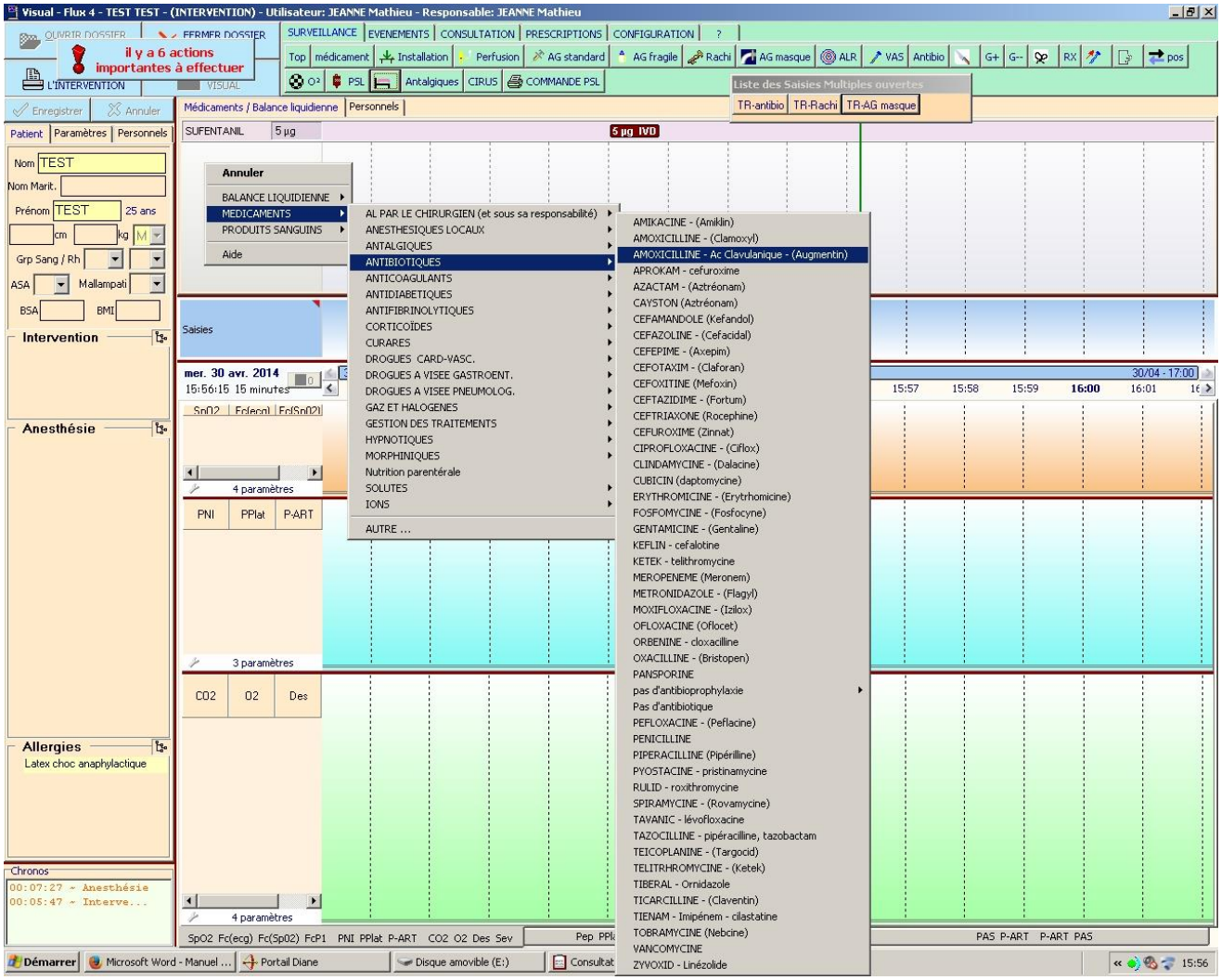
Annexe 1 : Module de consultation pré-opératoire

Le module de la consultation préanesthésique propose à l'utilisateur plusieurs catégories d'informations à renseigner : antécédents chirurgicaux, antécédents médicaux, allergies, traitements en cours etc...

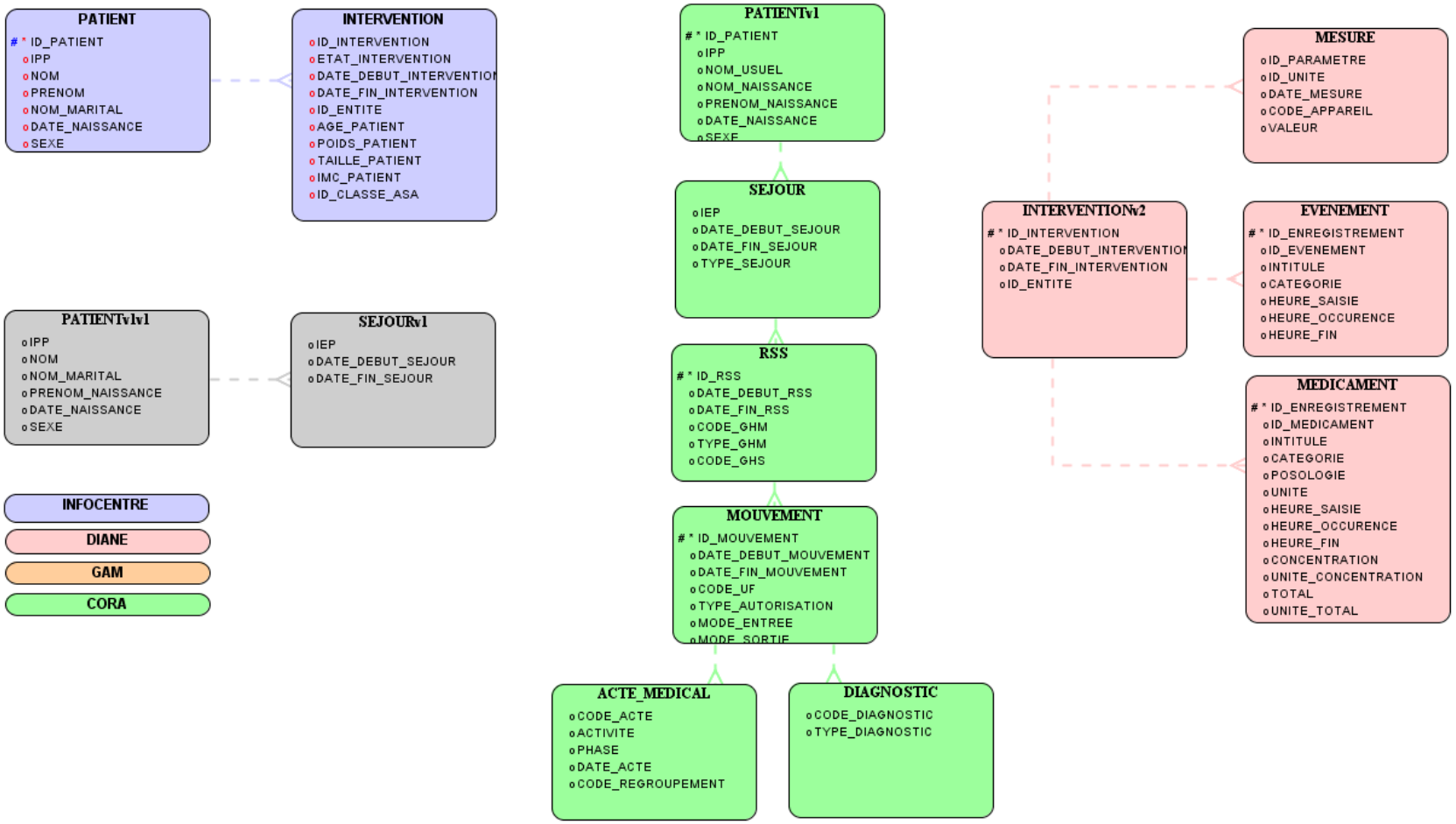


Annexe 2 : Menu pré-configuré pour le renseignement des médicaments dans le module per-opératoire.

La figure présente l'arborescence des listes pour les administrations de produits. Le premier niveau propose les catégories "balance liquidienne", "médicaments" et "produits sanguins". Après avoir sélectionné "Médicaments", puis "Antibiotiques", 47 antibiotiques sont proposés à l'utilisateur.



Annexe 3 : Modèle de logique de données des systèmes sources



Annexe 4 : Problèmes de qualité étudiés pour les données des systèmes sources DIANE, GAM, CORA et Infocentre d'anesthésie

Problèmes de qualité	Données
Un champ, un enregistrement	
Valeur manquante	Patient Infocentre IPP
	Patient Infocentre Nom
	Patient Infocentre Prénom
	Patient Infocentre Date de naissance
	Patient Infocentre Sexe
	Intervention Infocentre Date début intervention
	Intervention Infocentre Date fin intervention
	Intervention Infocentre Id entité
	Intervention Infocentre Age patient
	Intervention Infocentre Poids patient
	Intervention Infocentre Taille patient
	Intervention Infocentre IMC patient
	Intervention Infocentre Classe ASA
	Médicament DIANE Intitulé
	Médicament DIANE Heure de saisie
	Médicament DIANE Heure d'occurrence
	Médicament DIANE Posologie
	Médicament DIANE Unité posologie
	Médicament DIANE Concentration
	Médicament DIANE Unité Concentration
	Médicament DIANE Total
	Médicament DIANE Unité Total
	Evénement DIANE Intitulé
	Evénement DIANE Heure d'occurrence
Valeur incorrecte	Mesure DIANE Date d'occurrence
	Médicament DIANE Posologie
	Médicament DIANE Date d'occurrence
	Evénement DIANE Date d'occurrence
	Acte Médical CORA Date Acte
Violation du domaine de valeurs	Patient Infocentre Date de naissance
	Patient Infocentre Sexe
	Intervention Infocentre Date début d'intervention
	Intervention Infocentre Date fin d'intervention
	Intervention Infocentre Poids
	Intervention Infocentre Taille
	Intervention Infocentre IMC
Erreur de saisie	Mesure DIANE Paramètre
	Médicament DIANE Intitulé
	Evénements DIANE Intitulé

Valeur imprécise	Mesure DIANE Paramètre
	Médicament DIANE Intitulé
	Evénements DIANE Intitulé
Un champ, plusieurs enregistrements	
Violation de contrainte d'unicité	Patient DIANE IPP
	Patient CORA IPP
Synonymes	Intervention Infocentre Classe Asa
	Intervention Infocentre Entité
	Mesure DIANE Paramètre
	Mesure DIANE Unité
	Médicament DIANE Intitulé
	Evénements DIANE Intitulé
	RSS CORA GHM GHS
	Mouvement CORA Autorisation
	Mouvement CORA Mode d'entrée
	Mouvement CORA Mode de sortie
	Acte médical CORA Acte Médical
	Diagnostic CORA Diagnostic
Format inapproprié	Patient Infocentre
	Séjour Infocentre
	Patient GAM
	Patient CORA
	Sejour GAM
	Sejour CORA
	RSS CORA
	Mouvement CORA
	Acte Médical CORA
	Diagnostic CORA
	Mesure DIANE
	Evénement DIANE
	Médicament DIANE
Violation d'une règle métier	Evénement DIANE Date Anesthésie (début > fin)
	Evénement DIANE Date Chirurgie (début > fin)
Plusieurs champs, un seul enregistrement	
Violation de dépendance fonctionnelle	Mesure DIANE Unité
	Médicament DIANE Unité
Violation du domaine de valeurs	Mesure DIANE Valeur
	Médicament DIANE Posologie
Violation d'une règle métier	Intervention Infocentre Date intervention (début > fin)
	Age - Poids, Taille
Une seule table	
Doublons similaires	Patient DIANE
Doublons incohérents	Patient DIANE
Enregistrements manquants	Mesure DIANE
	Evénement DIANE
	Médicament DIANE

	Séjour GAM - CORA
	Séjour RSS
	Mouvement
	Diagnostic
	Acte CCAM
Violation de contrainte d'unicité globale	Mesure DIANE
	Médicament DIANE
	Evénement DIANE
	Séjour GAM
	Séjour CORA
	Mouvement CORA
	Diagnostic CORA
Plusieurs tables	
Violation d'intégrité référentielle	Evénement DIANE
	Médicament DIANE
	Mesure DIANE (identifiant paramètre)
	Mesure DIANE (identifiant d'unité)
	Mouvement CORA (identifiant du mode d'entrée)
	Mouvement CORA (identifiant du mode de sortie)
	Acte médical CORA
	Diagnostic CORA
Différence de structure	Médicament Diane Unité - Mesure DIANE Unité
Plusieurs bases de données	
Différence de structure	Patient DIANE - Patient GAM
	Patient DIANE - Patient CORA
	Séjour GAM - Séjour CORA
Différence de syntaxe	Patient DIANE - Patient GAM
	Patient DIANE - Patient CORA
	Séjour GAM - Séjour CORA
Différence de représentation	Patient DIANE - Patient GAM
	Patient DIANE - Patient CORA
	Séjour GAM - Séjour CORA
Absence de lien entre deux systèmes sources différents	Patient DIANE - Patient GAM
	Patient DIANE - Patient CORA
Doublons incohérents entre deux systèmes sources différents	Patient DIANE - Patient GAM
	Patient DIANE - Patient CORA
Doublons similaires entre deux systèmes sources différents	Patient DIANE - Patient GAM
	Patient DIANE - Patient CORA
Enregistrements manquants	Séjour GAM - Séjour CORA

Annexe 5 : Résultats de l'évaluation - Problème de qualité de niveau Schéma

Problèmes de qualité de données	Data	Résultats Nombre de colonnes avec problème détecté / Nombre total de colonnes (%)
Format inapproprié	Patient Infocentre	0 (0%)
	Intervention Infocentre	0 (0%)
	Patient GAM	0 (0%)
	Patient CORA	0 (0%)
	Sejour GAM	0 (0%)
	Sejour CORA	0 (0%)
	RSS CORA	3 (42,86%)
	Mouvement CORA	3 (33,33%)
	Acte Medical CORA	1 (16.67%)
	Diagnostic CORA	1 (33,33%)
	Mesure DIANE	5 (83,33%)
	Evénement DIANE	1 (12,50%)
	Médicament DIANE	7 (50,00%)
Différence de structure	Médicament DIANE Unité - Mesure DIANE Unité	1 (100%)
Différence de structure	Patient Infocentre - Patient GAM	5 (83.33%)
	Patient Infocentre - Patient CORA	5 (83.33%)
	Séjour GAM - Séjour CORA	0 (0%)
Hétérogénéité de syntaxe	Patient Infocentre - Patient GAM	0 (0%)
	Patient Infocentre - Patient CORA	0 (0%)
	Séjour GAM - Séjour CORA	0 (0%)
Hétérogénéité de représentation	Patient Infocentre - Patient GAM	0 (0%)
	Patient Infocentre - Patient CORA	2 (33.33%)
	Séjour GAM - Séjour CORA	0 (0%)

Annexe 6 : Résultats des problèmes de qualité liés aux enregistrements des bases de données

Problèmes de qualité de données	Données	Résultats Nombre d'enregistrements avec problème détecté / Nombre total d'enregistrements
Un champ, un enregistrement		
Valeur manquante	Patient Infocentre IPP	557 (0,50%)
	Patient Infocentre Nom	1 (<0,001%)
	Patient Infocentre Prénom	1 (<0,001%)
	Patient Infocentre Date de naissance	0 (0%)
	Patient Infocentre Sexe	0 (0%)
	Intervention Infocentre Date début intervention	0 (0%)
	Intervention Infocentre Date fin intervention	0 (0%)
	Intervention Infocentre Id entité	1115 (0,69%)
	Intervention Infocentre Age patient	78 (0,05%)
	Intervention Infocentre Poids patient	36275 (22,53%)
	Intervention Infocentre Taille patient	54878 (34,08%)
	Intervention Infocentre IMC patient	56536 (35,12%)
	Intervention Infocentre Classe ASA	1212 (0,75%)
	Médicament DIANE Intitulé	0 (0%)
	Médicament DIANE Heure de saisie	0 (0%)
	Médicament DIANE Heure d'occurrence	0 (0%)
	Médicament DIANE Heure de fin	1754292 (64,86%)
	Médicament DIANE Posologie	0 (0%)
	Médicament DIANE Unité posologie	0 (0%)
	Médicament DIANE Concentration	2053451 (75,92%)
	Médicament DIANE Unité	2054993 (75,98%)

	Concentration	
	Médicament DIANE Total	2302245 (85,12%)
	Médicament DIANE Unité Total	2299192 (85,00%)
	Evénement DIANE Intitulé	1309 (0.01%)
	Evénement DIANE Heure d'occurrence	0 (0%)
Violation du domaine de valeur	Patient Infocentre Date de naissance	223 (0.23%)
	Patient Infocentre Sexe	0 (0%)
	Intervention Infocentre Date début d'intervention	0 (0%)
	Intervention Infocentre Date fin d'intervention	0 (0%)
	Intervention Infocentre Poids	692 (0.43%)
	Intervention Infocentre Taille	1103 (0.68%)
	Intervention Infocentre IMC	1159 (0.72%)
Erreur de saisie	Mesure DIANE Paramètre	0 (0%)
	Evénement DIANE Intitulé	3 (75%)
	Médicament DIANE Intitulé	8 (100%)
	Mesure DIANE Unité	4 (9,75%)
Valeur imprécise	Evénement DIANE Intitulé	0 (0%)
	Médicament DIANE Intitulé	0 (0%)
	Paramètre DIANE	9 (64,28%)
Un champ, plusieurs enregistrements		
Violation de contrainte d'unicité	Patient Infocentre IPP	6 (0.005%)
	Patient CORA IPP	0 (0%)
Synonymes	Intervention Infocentre Classe Asa	0 (0%)
	Intervention Infocentre Entité	0 (0%)
	Mesure DIANE Paramètre	5 (35.71%)
	Mesure DIANE Unité	0 (0%)
	Médicament DIANE Intitulé	8 (100%)
	Evénements DIANE Intitulé	3 (75%)
	RSS CORA GHM GHS	0 (0%)
	Mouvement CORA Autorisation	0 (0%)
	Mouvement CORA Mode d'entrée	0 (0%)

	Mouvement CORA Mode de sortie	0 (0%)
	Acte médical CORA Acte Médical	0 (0%)
	Diagnostic CORA Diagnostic	0 (0%)
Violation d'une règle métier	Evénement DIANE Date Anesthésie (début > fin)	1543 (1,19%)
	Evénement DIANE Date Chirurgie (début > fin)	1303 (1,06%)
Plusieurs champs d'un seul enregistrement		
Violation de dépendance fonctionnelle	Mesure DIANE Unité	5714 (<0,01%)
	Médicament DIANE Unité	174 (0,01%)
Violation du domaine de valeurs	Mesure DIANE Valeur	883709 (0,31%)
	Médicament DIANE Posologie	2394 (0,75%)
Violation d'une règle métier	Intervention Infocentre Date intervention (début > fin)	0 (0%)
Une table		
Doublons incohérents	Patient Infocentre	0 (0%)
Doublons similaires	Patient Infocentre	82 (0,7%)
Enregistrements manquants	Mesure DIANE	62150 (38,36%)
	Evénement DIANE	54048 (34.03%)
	Médicament DIANE	21097 (13.28%)
Violation de contrainte d'unicité globale	Mesure DIANE	40062 (0.03%)
	Médicament DIANE	14246 (0.67%)
	Evénement DIANE	57668 (0.58%)
	Séjour GAM	51584 (5.40%)
	Séjour CORA	26 (0.0006%)
	Mouvement CORA	0 (0%)
	Diagnostic CORA	0 (0%)
Plusieurs tables		
Violation d'intégrité référentielle	Intervention Infocentre Id classe ASA	0 (0%)
	Evénement DIANE	1875355 (18.78%)
	Médicament DIANE	577605 (21.35%)
	Mesure DIANE (identifiant paramètre)	0 (0%)
	Mesure DIANE (identifiant d'unité)	0 (0%)
	Mouvement CORA (identifiant du mode d'entrée)	2 (< 0.001%)
	Mouvement CORA (identifiant du mode de	1 (< 0.001%)

	sortie)	
	Acte médical CORA	0 (0%)
	Diagnostic CORA	0 (0%)
Différence de structure	Médicament Diane Unité - Mesure DIANE Unité	1 (100%)
Plusieurs bases de données		
Absence de lien entre deux systèmes sources différents	Patient DIANE - Patient GAM	5810 (3,64%)
	Patient DIANE - Patient CORA	5728 (3,64%)
Doublons incohérents entre deux systèmes sources différents	Patient S1 - Patient S2	469 (0.29%)
	Patient S1 - Patient S3	907 (0.58%)
	Séjour GAM - Séjour CORA	447 (0.10%)
Doublons similaires entre deux systèmes sources différents	Patient S1 - Patient S2	2977 (1.86%)
	Patient S1 - Patient S3	2931 (1.86%)

Annexe 7 : Causes des problèmes de qualité

Plusieurs causes des problèmes de qualité ont été définies empiriquement et sont détaillées ci-dessous :

- Hétérogénéité des systèmes : modèles de données et des représentations des univers métier hétérogènes en fonction des applications
- Conception de la base de données : contraintes métier non prises en compte lors de la conception de la base de données
- Conception de l'application : champs de saisie ouverts
- Paramétrage de l'application : définition de listes d'objets avec des synonymes, des fautes de frappe, des intitulés imprécis, des doublons.
- Utilisation de l'application : mauvaise documentation de l'utilisateur, hétérogénéité dans les habitudes d'utilisation des applications
- Artefacts appareils : les problèmes liés aux capteurs de mesure et les problèmes de connexion entre l'appareil de mesure et l'application

Problèmes de qualité	Causes					
	Hétérogénéité des systèmes	Conception de la base de données	Conception de l'application	Paramétrage de l'application	Utilisation de l'application	Artefacts appareils
Un champ, un enregistrement						
Valeur manquante			x		x	
Valeur incorrecte					x	x
Violation du domaine de valeurs			x		x	x
Erreur de saisie					x	
Valeur imprécise					x	
Un champ, plusieurs enregistrements						
Violation de contrainte d'unicité		x	x		x	
Synonymes				x	x	
Format inapproprié		x				
Violation d'une règle métier					x	x
Plusieurs champs, un enregistrement						
Violation de dépendance		x	x	x	x	x

fonctionnelle						
Violation du domaine de valeurs		x	x	x	x	x
Violation d'une règle métier			x			
Une table						
Doublons similaires			x		x	
Doublons incohérents			x		x	
Enregistrements manquants*				x	x	
Violation de contrainte d'unicité globale			x			
Plusieurs tables						
Violation d'intégrité référentielle			x			
Différence de structure		x				
Plusieurs bases de données						
Différence de structure	x					
Différence de syntaxe	x					
Différence de représentation	x					
Absence de lien entre deux systèmes sources différents	x					
Doublons incohérents entre deux systèmes sources différents	x				x	

Doublons similaires entre deux systèmes sources différents	x				x	
Enregistrements manquants	x				x	

Annexe 8 : Incidence des problèmes de qualité sur l'utilisation secondaire des données

Problème de qualité	Incidence			
	Constitution de l'entrepôt (architecture)	Utilisation des données		
		Pas d'information	Information fausse	Information inutilisable
Un champ, un enregistrement				
Valeur manquante		x		
Valeur incorrecte			x	
Violation du domaine de valeurs			x	
Erreur de saisie				x
Valeur imprécise				x
Un champ, plusieurs enregistrements				
Violation de contrainte d'unicité	x		x	
Synonymes				x
Format inapproprié	x		x	
Violation d'une règle métier				
Plusieurs champs, un enregistrement				
Violation de dépendance fonctionnelle	x			
Violation du domaine de valeurs			x	
Violation d'une règle métier			x	
Une table				
Doublons incohérents	x		x	
Doublons similaires	x			
Enregistrements manquants		x		
Violation de contrainte d'unicité globale	x			
Plusieurs tables				
Violation d'intégrité référentielle				x
Différence de structure	x			
Plusieurs bases de données				
Différence de structure	x			
Différence de représentation	x			

Différence de syntaxe	x			
Absence de lien entre deux systèmes sources différents	x	x		
Doublons incohérents entre deux systèmes sources différents	x		x	
Doublons similaires entre deux systèmes sources différents	x			

Annexe 9 : Article publié dans le Journal of Clinical Monitoring and Computing

A substitution method to improve completeness of events documentation in anesthesia records

Antoine Lamer[1], Julien De jonckheere[1], Romaric Marcilly[1], Benoît Tavernier[2], Benoît Vallet[2], Mathieu Jeanne[1, 2], Régis Logier[1].

[1]INSERM CIC-IT 807, University Hospital, Lille.
[2]Pôle d'Anesthésie Réanimation, University Hospital, Lille, France.

Introduction

For the last several years, Hospital Information Systems (HIS) deployment has become more and more important in healthcare organizations. HIS have been developed mainly for administrative, billing or medico legal purposes, but they also constitute large volume databases which could be sources of new clinical knowledge through retrospective analyses as data mining [1-3]. An Anesthesia Information Management System [4] (AIMS) stores clinical information from the pre-intervention anesthetic consultation up to the discharge from the recovery room. Despite the high volume of collected data, such a database does not make statistical analysis easy [5], as its architecture is organized in a transactional scheme. A transactional database is organized in order to be fed and queried through a specific associated software. It is developed in a way to provide fast answers regarding day-to-day operations or data and curves display. This organization usually limits queries to one particular record at a time. Furthermore, in most cases, this kind of scheme does not adhere to standardized formats and can only be read and not modified. As a result, the ability to add real time functionalities as well as to extract and treat data for decision support or research projects is limited with most AIMS.

In order to overcome this limitation, one solution is to develop a data warehouse [6-8], which is a common repository fed with data extracted from various sources. In a data warehouse, the data model is organized and optimized to query a high volume of data and is independent from source systems. Solutions for optimizing response time as partitioning and indexing are used and could be adapted over the time in order to answer new user's needs. An anesthesia data warehouse may be used to carry out research projects and allow answering questions about the occurrence of intraoperative incidents (e.g., hypotension), the effects of treatments, and their relation with postoperative outcome [9-11]. It may also be used to provide decision makers and clinicians with information relevant to practice improvement [12].

Such a solution should allow obtaining cleaned and structured data but the quality of the recorded data may still be questioned. Even if AIMS is better documented than paper record [13, 14], the quality of data registered may vary. Data quality refers to different dimensions [15-18]: completeness, correctness, concordance, plausibility and currency. Data completeness includes documentation, breadth, density and predictive documentation [17]. In the case of the anesthesia record, data documentation often depends on user's training with the AIMS and on the operative room workload (during the surgical procedure, clinicians cannot record events while performing action on the patient). Yet, several studies have shown that during anesthesia, poor documentation due to omitted items impairs the overall quality of the entered data [19, 20]. Missing or poor quality information may substantially affect the results of a retrospective statistical analysis.

In order to reliably analyze such databases, it is for example important to precisely identify start and end times of each period of interest, especially of anesthesia. However, in our hospital's AIMS (DIANE®, Bow Medical, Amiens, France)[21], as in most other available AIMS, such events are directly documented by the clinicians through the AIMS menus. We recently observed in our database for the year 2012 that the two events proposed by the AIMS menus for start and end times of anesthesia were documented only in 44987 and 44791 out of the 55229 anesthesia procedures,

respectively. Only 44523 procedures (80.6%) included both events. In summary, 19.4 % of the surgical interventions performed under anesthesia in 2012 could not be used for retrospective studies requiring the anesthesia procedure duration.

In this study, we propose a method for replacing those missing events by estimating the start and end times of anesthesia using other available anesthesia events based both on manual and automatic entrances. We first evaluated the nearness of the other anesthesia events with the documented "start of anesthesia" and "end of anesthesia" events. In a second step, we defined substitution rules in order to estimate anesthesia start and end times as accurately as possible. Then, the delay between each documented event and the substitution one was evaluated. Finally, the substitution rules were applied on a 3-year dataset in order to estimate the improvement in term of completeness due to the introduction of the proposed method.

Materials and Methods

1 - Databases description

DIANE® AIMS has been used at the Lille University Hospital since 2005. During a surgical procedure, it continuously records around fifty parameters (heart rate, arterial pressure, oxygen saturation, expired carbon dioxide, ...) issued from the anesthesia monitor and ventilator. These measurements are stored in the AIMS database every 30 sec depending on data availability. On average, 4000 measurements are recorded for each intervention. Clinicians (anesthesiologists and anesthetic nurses) also enter information about the stages of the intervention (patient installation, start of anesthesia, ...), drugs administered and applied anesthesia techniques (tracheal intubation,...) through the DIANE® interface. Around 80 events are recorded for each intervention.

The Lille University Hospital has been using the billing-software CORA® (McKesson, San Francisco, United States) since 2010. This system records information such as duration of hospital stay, diagnosis and medical procedures.

The Lille University Hospital has developed a clinical data warehouse [22] which collects data from DIANE® and from CORA®. The data warehouse is fed with data organized in an optimized structure through a three-step process named ETL (Extract, Transform, Load). The first step consists in extracting relevant data from DIANE® and CORA®. Second, cleaning and transformation processes deduplicate data, associate free-text entries with predefined items or convert measurements in a reference unit when necessary. Finally, cleaned data are loaded in the data warehouse and stored in a common scheme. Therefore, analysis can be conducted on data issued from both DIANE® and CORA® and this allows cross-checking per-operative events (originally stored in DIANE®) with post-operative outcomes (originally stored in CORA®).

The data warehouse has been fed with data since 2010, so that more than 225 000 interventions are potentially available for retrospective analysis.

2 - AIMS record period of interest

In this study, we focus on surgical procedure conducted between 2010 and 2012 and more precisely on events and measurements registered during the anesthetic procedure.

Typically, four main periods can be distinguished during any surgical procedure under general anesthesia (GA):

- The patient care period, which is defined as the period between admission of the patient in the operating room and his discharge from the Post Anesthesia Care Unit (PACU).
- The anesthesia period, which is typically defined as the period starting with the first anesthetic agent administration until awakening of the patient / weaning of mechanical ventilation.
- The surgery, which is typically defined as the period starting with first skin incision until last surgical suture.
- The recovery period, which is defined by the period starting with admission in the PACU until the patient is discharged to either ward, home or intensive care.

In this study, we focused on the anesthesia period limits defined as "start of anesthesia" and "end of anesthesia".

2 - Definition of substitution events

Non-recorded events may be replaced by other events or measurements registered by the AIMS that are closely related in time. A panel of experts, composed of three anesthesiologists and two anesthetic nurses defined several substitution events closely related in time to the documented “start of anesthesia” and the “end of anesthesia” events. Table 1 presents the substitution events chosen for each documented event and their origin (i.e., automatically retrieved by the AIMS, for instance "1st non null expired carbon dioxide", or entered manually by the clinician, for instance " tracheal intubation").

Table 1: Defined substitution events for each initial event and their origin

Initial event	Substitution event	Origin of the substitution event	
Start of anesthesia	Start of record in the operative room	User input	Step of intervention
	Intubation	User input	Step of intervention
	Start of surgery (Skin incision)	User input	Step of intervention
	1st non null expired carbon dioxide	Automatic	Measurement
	1rst measurement of non null tidal volume	Automatic	Measurement
	1st hypnotics administration	User input	Drug administration
	1st analgesics administration	User input	Drug administration
	1st myorelaxants administration	User input	Drug administration
End of anesthesia	Extubation	User input	Step of intervention
	End of surgery	User input	Step of intervention
	Transfer in PACU	User input	Step of intervention
	Arrival in PACU	User input	Step of intervention
	End of intervention (last physiological record in the operative room)	User input	Step of intervention
	Last hypnotics administration	User input	Drug administration
	Last myorelaxants administration	User input	Drug administration

For the year 2012 data warehouse records, the time between documented events ("start of anesthesia" or "end of anesthesia") and each defined substitution events was measured in all AIMS records where the documented and substitution events were both available. For each pair of events (documented and substitution), the median (1st quartile ; 3rd quartile) time delay was measured.

To select substitution events that were close enough to their documented events, only those with a median time delay between -5 and +5 minutes were kept. Then, the selected substitution events were reordered by median time delay to the documented event. A priority score was associated to each substitution event: the substitution event with the smallest delay ranked highest, while the substitution event with the greatest delay ranked lowest priority.

For each documented event, a substitution rule has been developed in order to replace the missing event by the available substitution event with the highest priority.

3 - Evaluation of substitution rules

In order to evaluate the ability of substitution events to replace the documented events without changing significantly the total duration of the anesthesia procedure, we analyzed data recorded during the years 2010 to 2012 that included the documented “start of anesthesia” and the “end of anesthesia” events. We then computed the mean time delay between each event selected by substitution rules and the two documented events.

Finally, we applied the validated decision rules on the 2010, 2011 and 2012 year in order to assess whether the method applied actually improved the completeness of the documentation.

Data are expressed as median (1st quartile ; 3rd quartile).

Results

1 - Substitution rules determination

In 2012, 49498 AIMS records presented the documented events and at least one of the substitution events. For each substitution event, the tables 2 and 3 present the time interval between this event and the documented one.

Table 2: Interval between documented “start of anesthesia” and events selected as substitution events

Substitution event	Median (minutes)	1st quartile - 3rd quartile (minutes)	Number of records where the event is documented
Start of record in the operative room	-13.23	[-20.32 ; -7.00]	49699
1st non null expired carbon dioxide	1.05	[-2.28 ; 5.50]	37219
1rst analgesics administration	1.00	[0.00 ; 2.00]	48837
1rst hypnotics administration	2.00	[0.00 ; 3.00]	38963
1rst measurement of non null tidal volume	2.55	[-2.03 ; 8.15]	49674
1rst myorelaxants administration	4.00	[3.35 ; 4.77]	13047
Tracheal intubation	5.00	[3.05 ; 7.00]	30710
Bronchial intubation	7.00	[5.08 ;8.87]	12
Start of surgery (Skin incision)	28.58	[16.25 ; 43.28]	41221

Table 3: Interval between documented “end of anesthesia” and events selected as substitution events

Substitution event	Median (minutes)	1st quartile - 3rd quartile (minutes)	Number of records where the event is documented
Last myorelaxants administration	-93.71	[-148.25 ; -31.65]	10734
Last hypnotics administration	-65.47	[-132.03 ; -23.48]	36252
End of surgery	-7.88	[-16.22 ; -0.88]	36619
Extubation	0.00	[-2.04 ; 4.08]	27316
Transfer in PACU	0.00	[0.00 ; 2.00]	2316
End of intervention (last physiological record in the operative room)	0.08	[0.00 ; 0.32]	44519
Arrival in PACU	5.30	[3.12 ; 8.35]	35888

For the “start of anesthesia” event, the events "1rst non null measurement of expired carbon dioxide ", "1rst administration of analgesics", "1rst administration of hypnotics", "1rst measurement of non null tidal volume ", "1rst administration of myorelaxing drugs" and "tracheal intubation" all presented a median time delay between -5 and +5 minutes. These events were selected as substitution events to replace the “start of anesthesia” event. The table 4 presents the validated substitution events priority for the “start of anesthesia” event replacement.

Table 4: Priority table for "start of anesthesia" substitution events

Events	Priority	Median [1st quartile - 3rd quartile]
1st administration of analgesia drugs	1	1.00 [0.00 ; 2.00]
1st non null expired carbon dioxide	2	1.05 [-2.28 ; 5.50]
1rst administration of hypnotic drugs	3	2.00 [0.00 ; 3.00]
1rst measurement of non null tidal volume	4	2.55 [-2.03 ; 8.15]
1rst administration of curare	5	4.00 [3.35 ; 4.77]
Tracheal intubation	6	5.00 [3.05 ; 7.00]

For the "end of anesthesia" event, substitution events "extubation", "transfer to PACU" and "end of intervention (last physiological record in the operative room)" had a median time delay between -5 and +5 minutes. These events were selected to replace the documented event "end of anesthesia" in case this event was not recorded. Table 5 presents the validated substitution events priority for the “end of anesthesia” event replacement.

Table 5: Priority table for "end of anesthesia" substitution events

Events	Priority	Median [1st quartile - 3rd quartile]
End of intervention (last physiological record in the operative room)	1	0.08 [0.00 ; 0.32]
Transfer in PACU	2	0.00 [0.00 ; 2.00]
Extubation	3	0.00 [-2.04 ; 4.08]

2 - Substitution rules validation

Table 6 and 7 present statistics associated with the time interval between selected events and documented events, respectively for the beginning and the end of anesthesia. For the year 2010 to 2012, we obtained event detection with a precision of 0.00 (-2.22 ; 2.00) minutes for the start of anesthesia and 0.10 (0.00 ; 0.35) minutes for the end of anesthesia.

Table 6: Time delay (in minutes) between documented "start of anesthesia" and possible substitutive event selected according to priority rules (Table 4)

Year	Number of records	Median [1st quartile - 3rd quartile]
2010	45791	0.00 [-2.03 ; 2.00]
2011	47461	0.00 [-2.28 ; 2.00]
2012	49289	0.00 [-2.33 ; 2.00]
Total	142541	0.00 [-2.22 ; 2.00]

Table 7 : Time delay (in minutes) between documented "end of anesthesia" and possible substitutive event selected according to priority rules (Table 5)

Year	Number of records	Median [1st quartile - 3rd quartile]
2010	40659	0.15 [0.02 ; 0.40]
2011	41934	0.08 [0.00 ; 0.32]
2012	43660	0.08 [0.00 ; 0.32]
Total	126253	0.10 [0.00 ; 0.35]

Finally, the substitution rules' ability to increase the number of valid records was evaluated on the whole dataset of records. Results for the years 2010 to 2012 are presented in table 8. On the whole dataset, both documented events were documented in 129 281 records (80.3 %) of the 161 002 records. After application of the substitution method, 156 504 records (97.2 %) were validated.

Table 8: Evolution of the number of valid records before and after substitution

Year	Total number of records	Number of valid records (%)	
		Before substitution	After substitution
2010	52250	41747 (79.9)	50412 (96.5)
2011	53523	43011 (80.4)	52166 (97.5)
2012	55229	44523 (80.6)	53926 (97.6)
Total	161002	129281 (80.3)	156504 (97.2)

Discussion

In this study, we tested a method that would allow overcoming the absence of time markers of the anesthetic procedures in an AIMS, thus impairing e.g. the post hoc computation of anesthetic duration. We show that missing events of interest can be automatically replaced by substitution events with minimal time-delay. When applied to the particular case of detecting start and end times of the anesthetic procedure, in our 2010-2012 database, data completeness showed on absolute increase of 21.1%.

In our study, around 20% of the records lacked information about the beginning or the end of anesthesia. In a recent study, Weil *et al.* [19] found a much lower rate of around 3% of missing "beginning of anesthesia" event. The difference may be explained by the AIMS architecture and user interface differences, as a software can make some events mandatory, or not.

Even if for each variable the rate of missing data does not exceed 20%, the accumulation of unavailable data for all the variables may exclude a very large number of cases. This may be detrimental to the statistical analysis, in particular when each variable is required, in the case of a multivariate analysis for example. The method presented in this paper allowed us to maximize the total number of validated records, which may improve the reliability of retrospective statistical analyses conducted on the data warehouse.

There are few studies about automated replacement of missing information in anesthesia databases. Spring *et al.* [23] proposed a decision-rule based software that automatically detects anesthesia documentation errors and alerts clinicians about missing information. However, even if the rate of completed record is up to 99%, the records are not completed automatically and require the intervention of a clinician. In our case, the system is able to automatically substitute the missing information by other available data. Sandberg *et al.* [24] developed an algorithm which detects after the first fifteen minutes of an intervention if the user documented patient allergies. Otherwise, a one-time prompt is sent to the user via pager. This method has not been implemented in Lille because we do not have a real time access to the database. Moreover, we are working on retrospective data, registered up to 4 years ago for which it is impossible to complete missing information. Another simple solution would have consisted in defining mandatory information. This solution was put in place in Lille for some information as ASA status or anesthetist identity. We do not find that this solution was optimal for temporal event as users documented them just before closure of record or force the system to close improperly.

Our study presents some limitations. First, even if this technique is able to replace missing information, we have no information about the substitution quality for one particular record. However, we have evaluated the global quality of the method by computing the time delay between the documented event of interest and the substituted one and found a median interval of less than 1 minute, which is acceptable and demonstrate the quality of the method. Moreover, we are currently working on a quality index based on the number of substitution events available together with the time delay between events. This index will allow to determinate the quality of substitution events for each record. Another limitation of this study is that we did not test the influence of the improvement of validated record on a particular statistical analysis. Finally, the method needs the use of a data warehouse and cannot be directly implemented in the AIMS due to the system's architecture. Therefore this limitation does not allow real time event substitution but is still of interest for retrospective statistical analysis. This study underlines some of the benefits of the use of a data warehouse. Initial data registered in the AIMS may be used to compute and provide new consolidated information to

deliver data to the data warehouse. Moreover, the flexible architecture of the data warehouse allows the implementation of new data processing methods or the improvement of existing ones.

This method could also be generalized to other missing anesthesia or surgical events: as an example, substitution events could be defined to detect automatically start and end of surgery (data not shown). This could show some benefits in other databases for other clinical settings as long as these databases include time events.

Financial Disclosure

There was no financial support outside of the University Hospital

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Haux, R. (2006) Health information systems – past, present, future. International Journal of Medical Informatics 75, 268–281.

[2] Pitt, E.A. (2009) Application of data mining techniques in the prediction of coronary artery disease : use of anaesthesia time-series and patient risk factor data (Thesis). Queensland University of Technology.

[3] Chazard, E., Ficheur, G., Bernonville, S., Luyckx, M., Beuscart, R. (2011) Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed 15, 823–830.

[4] Douglas, J.R., Ritter, M.J. (2011) Implementation of an Anesthesia Information Management System (AIMS). Ochsner J 11, 102–114.

[5] Nunez, C.M. (2004) Advanced techniques for anesthesia data analysis. Seminars in Anesthesia, Perioperative Medicine and Pain 23, 121–124.

[6] Kimball, R. (1998) The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons.

[7] Wisniewski, M.F., Kieszkowski, P., Zagorski, B.M., Trick, W.E., Sommers, M., Weinstein, R.A. (2003) Development of a Clinical Data Warehouse for Hospital Infection Control. J Am Med Inform Assoc 10, 454–462.

[8] De Mul, M., Alons, P., van der Velde, P., Konings, I., Bakker, J., Hazelzet, J. (2012) Development of a clinical data warehouse from an intensive care clinical information system. Comput Methods Programs Biomed 105, 22–30.

[9] Taffé, P., Sicard, N., Pittet, V., Pichard, S., Burnand, B., ADS study group (2009) The occurrence of intra-operative hypotension varies between hospitals: observational analysis of more than 147,000 anaesthesia. Acta Anaesthesiol Scand 53, 995–1005.

[10] Walsh, M., Devereaux, P.J., Garg, A.X., Kurz, A., Turan, A., Rodseth, R.N., Cywinski, J., Thabane, L., Sessler, D.I. (2013) Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension. Anesthesiology 119, 507–515.

[11] Komatsu, R., You, J., Mascha, E.J., Sessler, D.I., Kasuya, Y., Turan, A. (2013) Anesthetic induction with etomidate, rather than propofol, is associated with increased 30-day mortality and cardiovascular morbidity after noncardiac surgery. Anesth. Analg. 117, 1329–1337.

[12] Bréant, C., Borst, F., Nkoulou, R., Irion, O., Geissbuhler, A. (2007) Closing the loop: bringing decision support clinical data at the clinician desktop. Stud Health Technol Inform 129, 890–894.

[13] Jang, J., Yu, S.H., Kim, C.-B., Moon, Y., Kim, S. (2013) The effects of an electronic medical record on the completeness of documentation in the anesthesia record. Int J Med Inform 82, 702–707.

[14] Sanborn, K.V., Castro, J., Kuroda, M., Thys, D.M. (1996) Detection of intraoperative incidents by electronic scanning of computerized anesthesia records. Comparison with voluntary reporting. Anesthesiology 85, 977–987.

[15] Weiskopf, N.G., Weng, C. (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20, 144–151.

[16] Fox, C., Levitin, A., Redman, T. (1994) The notion of data and its quality dimensions. Information Processing & Management 30, 9–19.

[17] Weiskopf, N.G., Hripcsak, G., Swaminathan, S., Weng, C. (2013) Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 46, 830–836.

[18] Müller, H. (2003) Problems, Methods and Challenges in Comprehensive Data Cleansing (Technical Report No. HUB-IB-164). Humboldt-Universität zu Berlin, Institut für Informatik.

[19] Weil, G., Motamed, C., Eghiaian, A., Guye, M.L., Bourgain, J.L. (2014) The use of a clinical database in an anesthesia unit: focus on its limits. J Clin Monit Comput 1–5.

[20] Devitt, J.H., Rapanos, T., Kurrek, M., Cohen, M.M., Shaw, M. (1999) The anesthetic record: accuracy and completeness. Can J Anesth 46, 122–128.

[21] BOW Médical [WWW Document], n.d. URL http://www.bowmedical.com/ (accessed 7.5.14).

[22] Lamer, A., Jeanne, M., Vallet, B., Ditilyeu, G., Delaby, F., Tavernier, B., Logier, R. (2013). Development of an anesthesia data warehouse: Preliminary results. IRBM 34, 376–378.

[23] Spring, S.F., Sandberg, W.S., Anupama, S., Walsh, J.L., Driscoll, W.D., Raines, D.E. (2007) Automated documentation error detection and notification improves anesthesia billing performance. Anesthesiology 106, 157–163.

[24] Sandberg, W.S., Sandberg, E.H., Seim, A.R., Anupama, S., Ehrenfeld, J.M., Spring, S.F., Walsh, J.L. (2008) Real-time checking of electronic anesthesia records for documentation errors and automatically text messaging clinicians improves quality of documentation. Anesth. Analg. 106, 192–201.

Annexe 10 : Fonctions d'agrégations

Fonction	Type de fonction
Moyenne	Agrégation
Ecart-type	Agrégation
Médiane	Agrégation
1er Quartile	Analytique
3ème Quartile	Analytique
Minimum	Agrégation
Maximum	Agrégation
1ère mesure	Analytique
Moyenne des deux premières mesures	Analytique
Dernière mesure	Analytique
Moyenne des deux dernières mesures	Analytique

Annexe 11 : Définition des seuils d'hypotension

Seuils	PAM < 50	PAM < 55	PAM < 60	PAM < 65	PAM < 70
ID_SEUIL	3	3	3	3	3
LIB_SEUIL	Hypotension avec PAM < 50	Hypotension avec PAM < 50	Hypotension avec PAM < 50	Hypotension avec PAM < 50	Hypotension avec PAM < 50
ID_PARAMETRE	PAM	PAM	PAM	PAM	PAM
ID_FENETRE_ETUDE	Induction-Incision	Induction-Incision	Induction-Incision	Induction-Incision	Induction-Incision
ID_GROUPE_PATIENT	Adulte	Adulte	Adulte	Adulte	Adulte
TYPE_SEUIL	Absolue	Absolue	Absolue	Absolue	Absolue
SEUIL_ABSOLU	50	55	60	65	70
SEUIL_RELATIF	-	-	-	-	-
ID_MESURE_AGREGEE	-	-	-	-	-
OPERATEUR_RELATIONNEL	<	<	<	<	<
METHODE_INTERPOLATION	Répétition	Répétition	Répétition	Répétition	Répétition
INTERVALE_ECHANTILLONNAGE	300	300	300	300	300
INTERVALLE_MAXIMALE	360	360	360	360	360
PROPORTION_MAXIMALE	25	25	25	25	25

Annexe 12 : Etudes réalisées grâce à l'exploitation de l'entrepôt de données

Titre de l'étude		Type d'étude
Hypotension après l'induction au propofol dans le cadre d'une anesthéise générale : quelle fréquence, quelle durée, et quelles conséquences pour le patient ?	Article en cours de rédaction	Evaluation de la qualité
Vce et poids idéal	Etude en cours de réalisation	Evaluation de la qualité
Incidence de l'hypotension chez les patients âgés opérés aux urgences pour une fracture du col du fémur (92)	Thèse de médecine (Jérôme Jaspart)	Recherche clinique
Facteurs péri-opératoires prédictifs de mortalité à 6 mois de sujets âgés opérés de chirurgie carcinologique digestive	Thèse de médecine (Léa Sartre Buisson)	Recherche clinique
Comparaison des méthodes d'anesthésie et de leurs conséquences hémodynamiques au cours des cholécystectomies réalisées au service des urgences et en chirurgie programmée	Mémoire (Adrien Berthier)	Recherche clinique
Impact de l'évolution des pratiques anesthésiques en termes de ventilation et de remplissage sur les complications pulmonaires post-oesophagectomies carcinologiques.	Thèse de médecine (Justine Mullie)	Evaluation de la qualité
Influence du niveau de pression artérielle moyenne au cours des CEC (circulation extra corporelle) et insuffisance rénale aigue post opératoire : étude rétrospective au CHRU de Lille	Mémoire (Juliette Masse)	Evaluation de la qualité
Estimation du temps d'occupation du bloc opératoire	Michel Delecroix - Papier accepté au congrès IEEE EMBC	Management
Bénéfices cliniques de l'utilisation de l'ANI	Julien De jonckeere - Article en cours de rédaction	Recherche clinique
Utilisation de l'O2 et N2O au bloc opératoire dans un CHRU	Mémoire élève IADE (Eddie Tamboukti)	Recherche clinique
Ventilation protectrice : état des lieux des pratiques IADE	Mémoire élève IADE (Marion Charlet)	Evaluation de la qualité

Utilisation de l'atropine dans la bradycardie sinusale	Mémoire élève IADE (Aurélien Kozycki)	Recherche clinique
L'ANI, un bénéfice pour les IADE et les patients	Mémoire élève IADE (Gwénaëlle Renard) Abstract accepté au congrès SFAR2015	Recherche clinique
Ventilation des obèses	Mémoire élève IADE (Caroline Destailleurs)	Evaluation de la qualité
Complétude de la feuille informatisée lors de la consultation pré-opératoire	Article en cours de rédaction	Informatique médicale
Evaluation de la qualité des données enregistrées avec la feuille informatisée d'anesthésie : étude rétrospective sur cinq années d'utilisation au CHRU de Lille	Article en cours de rédaction	Informatique médicale
Proposition d'un algorithme et d'un modèle de données pour la détection automatisée d'événements indésirables au cours de l'anesthésie	Article soumis au Computer Methods and Programs in Biomedicine	Informatique médicale
Méthode pour l'amélioration de la complétude des données temporelles dans des enregistrements d'anesthésie	Article publié au Journal of Clinical Monitoring and Computing	Informatique médicale