Indicateurs de performance & Qualité des données : Vers une démarche industrielle dans un grand Hôpital Français

Cédric Aubin

Indicateurs de performance & Qualité des données : Vers une démarche industrielle dans un grand Hôpital Français.

Cédric AUBIN

Directeur Conseil Secteur Public / Santé Social du groupe KEYRUS
Enseignant et chercheur associé à l'IFROSS – Université Jean Moulin Lyon 3 – France
Membre de la Société Française d'Évaluation

Tél. : +33 6 65 49 89 32
E-mail : cedric.aubin@keyrus.com / cedric.aubin@univ-lyon3.fr

Actes de la conférence « GISEH 2012 »
Université Laval – Québec – du 30 août au 1er septembre 2012

Résumé. La multiplication des projets Performance pose dans de nombreux hôpitaux la question de la qualité des données et des garanties qu'elles peuvent offrir dans la production d'indicateurs de performance. Le système d'information doit faire l'objet d'une attention systématique afin de mesurer la capacité à restituer une information fiable, de qualité, qui réponde bien aux qualités exigées pour l'élaboration de ces indicateurs.
Dans cet article, nous proposons une démarche industrielle de gestion de la qualité de l'information permettant de mesurer et d'améliorer le contenu de données de santé. Nous expérimentons cette méthode sur des données réelles dans le cadre d'un projet Performance d'un grand Centre Hospitalier Français.

Mots-clés : Performance hospitalière, Qualité des données, Système d'information, Entrepôts de données, Aide à la décision.

Introduction

Les Projets Performance résultent de la volonté du Ministère de la Santé et du Ministère des Comptes publics de mettre en œuvre des contrats avec les Établissements de Santé pour améliorer durablement leur performance et démontrer un impact tangible au profit des patients et des acteurs du système de santé. Ils concrétisent les orientations de la loi « Hôpital, Patients, Santé, Territoires » (HPST), du 21 juillet 2009, qui place la performance au cœur des politiques publiques pour répondre aux défis du système de santé.

L'impact visé porte sur trois dimensions de la performance des Établissements : la qualité de prise en charge des usagers, les conditions de travail et l'attractivité pour les professionnels, et enfin l'efficacité opérationnelle et financière. Par le contrat, les trois parties signataires, à savoir les Directeurs généraux de l'Établissement, l'Agence Régionale de Santé (ARS) et, l'Agence Nationale d'Appui à la Performance (ANAP), s'engagent à mettre en œuvre un plan d'actions et à atteindre des objectifs qui ont été discutés et validés. Ces objectifs, décrits par des indicateurs portant sur les trois dimensions de la performance, constituent un « standard de bonnes pratiques » observées au niveau national.

Or la production d'indicateurs à partir des données du système d'information doit faire l'objet d'une attention systématique afin de mesurer la capacité à restituer une information fiable, de qualité, qui réponde bien aux qualités exigées pour l'élaboration de ces indicateurs de performance.

Le but de cet article est d'examiner dans un tel projet les questions soulevées par la nécessité de fournir des informations de bonne qualité aux utilisateurs finaux dans un environnement où de nombreux systèmes sources de données coexistent et où les règles métiers peuvent être parfois différentes derrière chaque indicateur utilisé par les unités organisationnelles.

Plus précisément, le document examine l'environnement de données architecturé qui comprend diverses sources, *datamarts* et entrepôts de données. Ce document passe en revue la littérature existante sur les définitions de la qualité des données, l'importance du management de la qualité des données, et propose une méthode utilisée pour assurer la qualité des données. Des recommandations sont faites pour appliquer cette démarche industrielle à d'autres organisations de santé et pour s'ouvrir à d'autres recherches.

1. La Qualité des données de santé

1.1 L'environnement informationnel de l'hôpital

Les systèmes d'information hospitaliers (SIH) stockent des volumes de données de plus en plus importants et hétérogènes. On trouve classiquement des systèmes d'information cliniques centrés sur le dossier patient, et un grand nombre d'applications utilisées dans la gestion opérationnelle de l'hôpital. Pour l'essentiel, la gestion administrative des malades (GAM) et la gestion des plateaux techniques (gestion des lits, gestion des blocs etc.), sans oublier la gestion économique et financière (GEF), et la gestion des ressources humaines (GRH).

Pour exploiter ces environnements informationnels hétérogènes, l'approche proposée consiste le plus souvent en l'intégration des données dans un même espace appelé entrepôt de données et la mise en place d'outils utilisés à des fins d'analyse de données ou d'interrogation, d'aide à la décision, d'alerte ou de fouille de données pour générer de nouvelles connaissances [1].

Cependant lors de la mise en œuvre de ces entrepôts de données, il convient d'évaluer la qualité des données qui y sont stockées et de rendre explicite et non ambigu le sens de ces données [2]. Cette dernière tâche est une opération difficile car requiert l'alignement de données du SIH avec les objets sémantiques de référence entrant dans la composition des dimensions d'analyse et des indicateurs utilisés. A ce jour, la mise en œuvre de telles architectures d'analyse de données est une bonne opportunité pour les institutions de santé d'améliorer la qualité de leurs données, et de replacer leurs données dans leur contexte sémantique [3].

1.2 Définition de la Qualité des données

La qualité des données est un domaine de recherche qui suscite depuis longtemps un vif intérêt, mais qui émerge tout juste comme champ de recherche à part entière, tel que peuvent l'indiquer Strong et al. [4], Jarke et Vassiliou [5], et Berti [6]. Depuis la fin des années 60, la notion de qualité des données est en effet étudiée par les statisticiens [7] mais c'est dans les années 90 que les sciences de l'information ont commencé à formaliser la problématique de la mesure et de l'amélioration de la qualité des données.

Wang [8] propose de définir la qualité d'une donnée en fonction de l'usage attendu par son utilisateur. Redman [9] dépasse ce concept d'utilité attendue en caractérisant le concept de qualité des données selon 4 dimensions : *l'exactitude, la perfection, la fraîcheur et l'uniformité*. D'autres auteurs ont proposés de mesurer la qualité des données en fonction de processus [10] et de leur but [11]. Wang propose une méthode itérative d'amélioration de la qualité des données nommée TDQM[1] basée sur les 4 phases (*définir, réaliser, contrôler, agir*) de la roue de Deming [12]. D'autres approches similaires visant à mesurer, améliorer et surveiller la qualité des données ont été proposées [13].

En parallèle de ces travaux, l'ingénierie des modèles apporte des méthodes de mesure de la qualité des modèles d'information [14] [23]. Moody propose une méthode d'évaluation subjective (ou empirique) d'un modèle d'information sur la base des 7 critères suivants : *l'exactitude, l'applicabilité, la complétude, l'intelligibilité, l'intégration, la flexibilité et la simplicité* [15]. Kerr et al [16] présente un cadre de mesure de la qualité des données basé sur une classification de 69 critères regroupés en 6 dimensions : *précision, ponctualité, comparabilité, utilisabilité, pertinence et sécurité*.

Naumann et Roker [10] identifient trois approches d'analyse des critères de la qualité des données :
• *Approche orientée sémantique*, basée uniquement sur la signification des critères. Cette approche est la plus intuitive (il s'agit d'une approche où les critères sont examinés de façon générale, c'est-à-dire séparés de tout cadre d'information).
• *Approche orientée traitement*, classant les critères de qualité de l'information selon leur déploiement dans les différentes phases du traitement de l'information.
• *Approche orientée objectif*, caractérisée par une définition des objectifs de la qualité à atteindre et un classement des critères selon les objectifs définis.

Tout le monde s'accorde en résumé sur le fait que la qualité des données peut se décomposer en un certain nombre de dimensions, catégories, critères, facteurs, paramètres ou attributs, mais aucune définition ne fait aujourd'hui l'unanimité [Tableau 1].

[1] *Total Data Quality Management, enseigné au Massachusetts Institute of Technology (web.mit.edu/tdqm).*

Auteurs	Caractérisation de la qualité des données
Strong, Lee et Wang [4]	Approche orientée sémantique : 4 Catégories, 13 Dimensions qualité de données
Jarke et Vassiliou [5]	Approche orientée objectif : 5 Facteurs qualité des entrepôts de données
Calabretto, Pinon, Poullet et Richez [17]	Approche orientée sémantique : 3 Critères de qualité d'information, 8 Critères de qualité des documents
Berti [6]	Approche orientée sémantique : 4 Catégories, 32 Critères de qualité des données multi-sources
Naumann et Rolker [10]	Approche orientée traitement : 3 Classes d'évaluation des critères, 11 Critères qualité de données
Zhu et Gauch [18]	Approche orientée sémantique : 5 Critères de qualité des pages web
Marotta [19]	Approche orientée traitement : 2 points de vue : système et utilisateur, 6 Catégories, 31 Critères

Tableau 1 - Quelques approches de modélisation de la qualité des données (tiré de Harrathi et Calabretto [20]).

L'inconvénient de ces approches ne laisse que relativement peu de choix à l'utilisateur, sans pour autant l'aider à construire un ensemble cohérent et minimal de critères de qualité ou bien l'assister dans leur spécification. La qualité est représentée comme une collection de critères. Elles sont applicables seulement dans le domaine pour lequel elles ont été conçues. Ces méthodes se basent essentiellement sur des méthodes computationnelles qui prennent peu en compte le point de vue des utilisateurs et leur connaissance associée des données. Il conviendra donc de se tourner vers une véritable démarche industrielle de gestion de la qualité des données, associant systématiquement les utilisateurs dans les phases de spécifications des indicateurs et de leurs règles de construction.

Dans le domaine de la santé, le besoin croissant d'utiliser les données du dossier patient à des fins d'analyses épidémiologiques ou de santé publique, est souvent freiné par la mauvaise qualité des données [21]. Les causes de défaillance sont caractérisées de systématiques et hasardeuses [22] aux différents niveaux du processus de saisie. La mauvaise qualité des données est due principalement aux erreurs de saisie de l'information à la source. Fautes d'orthographes, codes incorrects, abréviations erronées, saisies dans un mauvais champ sont autant de sources de dégradation de la qualité qui peuvent avoir des conséquences néfastes pour l'institution. De même, la coexistence d'une multitude d'applications opérationnelles entraine une duplication des données. On retrouve ainsi souvent des données enregistrées plusieurs fois dans les systèmes informatiques sous des identifiants différents. De plus, des données correctes à un moment donné, peuvent devenir erronées à la suite d'un changement de système.

1.3 Importance du management de la qualité de données

La qualité de données ne se limite donc pas seulement à aider les organisations à charger des données correctes dans leurs systèmes d'information ; elle permet également de se débarrasser des données erronées, corrompues ou dupliquées. Le nettoyage des données devient une étape essentielle dans l'intégration d'informations dans les systèmes.

La gestion de la qualité des données ou *Data Quality Management (DQM)* est la capacité à fournir des données fiables capables de répondre aux besoins métiers et techniques des utilisateurs. Elle se mesure en termes d'exactitude, de cohérence, d'unicité, d'intégrité, de disponibilité, et d'homogénéité. C'est une méthode de gestion des informations ayant pour objectif de gérer et de comparer des données entre différents systèmes d'information ou bases de données d'une entreprise. En règle générale, il s'agit de transformer des données de qualité en renseignements utiles qui sont essentiels à l'entreprise.

1.4 Le processus de gestion de la qualité des données

Pour faire face à la problématique de gestion de la qualité des données, on décompose le processus selon quatre étapes clés, à savoir :

• **Le profilage des données** permet d'évaluer le niveau courant de la qualité des données et d'en mesurer l'évolution dans le temps. Cette étape étudie la structure des tables et les relations entre les tables, la pertinence des données (colonnes utilisées, poids des colonnes vides…) et la validité de formats (adresses, informations d'identification…) ;

• **Le redressement** a pour objectif de redresser la qualité des données en utilisant des techniques d'analyse syntaxique ou grammaticale (*parsing)*, de transformation des données, de dédoublonnage et de consolidation, et des méthodes statistiques (écarts types, analyse de bornes…). Il est exécuté en respectant un workflow démarrant par une détection des anomalies qui ne respectent pas les règles définies, effectuant une correction des erreurs et une consolidation de la donnée, pour terminer par une validation des données. Les données ne respectant pas les conditions de validations sont isolées et traitées manuellement au travers d'une interface homme-machine (IHM) ou de fichiers Excel.

• **L'enrichissement** est la phase d'amélioration de la complétude des données, à présent corrigées et validées. Cette étape apporte une plus-value à l'information.

• **La supervision** permet d'analyser et de piloter l'évolution de la qualité des données, à l'aide d'un tableau de bord synthétisant la qualité à travers les dimensions identifiées et son évolution dans le temps. Cette dernière étape permet alors d'orienter les nouveaux traitements et d'établir des objectifs adaptés.

La mise en place d'un tel processus basé sur des mesures chiffrées permet de gérer la qualité des données dans la durée et pas seulement de manière ponctuelle.

2. Cas étudié et méthodes

2.1. Le terrain mobilisé

Un grand Centre Hospitalier Français a signé début 2011 un contrat Performance avec l'ANAP et l'ARS dont l'objectif est de lui permettre de dégager d'ici à 2012 les ressources nécessaires pour consolider sa situation économique et investir au profit des patients, y compris les plus fragiles.

Ce Centre Hospitalier est un pôle de référence hospitalo-universitaire en France. Il regroupe un peu plus de 6800 professionnels répartis sur 3 sites hospitaliers, 61 services hospitaliers et 15 pôles d'activité médicale et médico-technique. Avec ses 1900 lits et places, il réalise en 2009 environ 60 000 entrées, 330 000 consultations médicales et 482 000 journées d'hospitalisation en court séjour. Son budget d'exploitation est de 444 M€.

Cependant, sa situation économique l'inscrit dans un contexte particulier et l'engage sur la voie d'un retour à l'équilibre financier pérenne. Le Centre Hospitalier lance en effet en 2009 un Plan de retour à l'équilibre (PRE) préparé en concertation avec le corps médical et composé d'une centaine de projets. Ce PRE prévoit un engagement sur dépenses et recettes permettant la réduction de son déséquilibre financier à horizon 2012. Puis de retrouver une marge d'exploitation lui permettant de maintenir son équilibre avec un retrait des aides publiques et à terme, et de relancer son processus d'investissement.

Dans ce contexte, le contrat Performance est au cœur de l'organisation des soins et s'articule autour de 5 pistes de travail :

1. Optimiser la gestion des lits et les organisations de travail
2. Améliorer la performance de l'imagerie
3. Améliorer la performance des blocs opératoires et de l'anesthésie
4. Améliorer la performance de la biologie et la pertinence des prescriptions
5. Réduire les coûts de LGG (Logistique et Gestion Générale)

Ce plan d'actions fait l'objet d'un suivi rigoureux qui repose sur la production d'indicateurs de performance construits à partir des données recueillies dans le système d'information de production de soins et décisionnel de l'établissement. Les données à extraire proviennent de logiciels opérationnels, qui n'ont pas été conçu à l'origine dans un but décisionnel. Chaque indicateur décisionnel produit à partir de ces données opérationnelles doit donc être évalué au regard de la qualité des données utilisées.

Pour atteindre cet objectif important, la Direction du Système d'information (DSI) a initialisé un projet dénommé « Indicateurs de performance & Qualité des données » dont l'objectif était double :

• Produire les indicateurs de performance des projets entrant dans le cadre de son contrat performance à partir des bases de données des logiciels utilisés dans les services de soins et médico-technique,

• Mettre en place des outils de mesure de la qualité qui permettent de garantir la cohérence des indicateurs de performance qui sont produits.

Ce projet s'appuie sur notre intervention en tant que directeur de mission d'une société de conseil en système d'information, qui accompagne depuis longtemps ce Centre Hospitalier dans la mise en place de son système d'information de pilotage de la performance. Il alimente par ailleurs notre réflexion dans le cadre de nos travaux de recherche sur l'appropriation des outils d'aide à la décision.

Un projet d'infocentre décisionnel a effectivement été lancé en 2005 afin d'améliorer le pilotage opérationnel et stratégique de l'établissement. Les objectifs du projet se sont traduit en pratique sur l'exploitation des données stratégiques, prévisionnelles, opérationnelles, avec mise à disposition d'outils de type tableaux de bord, simulation, prévision, aide à la décision, requêtes à la demande.

Les indicateurs et les tableaux de bord sont développés et mis en œuvre au travers d'un entrepôt de données, en s'appuyant sur les outils suivants :

• un extracteur de données (ETL[2]) pour la procédure d'alimentation de l'entrepôt de données,

• un outil de *reporting* permettant la réalisation d'états et la diffusion de tableaux de bord.

Dans le cadre spécifique de ce projet, un outil de gestion de la qualité des données est rajouté à cette architecture pour réaliser le profilage et les analyses mono-sources et multi-sources des données. Il est couplé à une base de données stockant des tables contenant pour chaque analyse les données invalides et les indicateurs.

[2] *Extract, Transform and Load.*

2.2. Méthodologie

Nous avons mis en œuvre une démarche industrielle en 6 étapes [Figure 1] pour la production de ces indicateurs en corrélation avec une étude qualitative de la donnée.

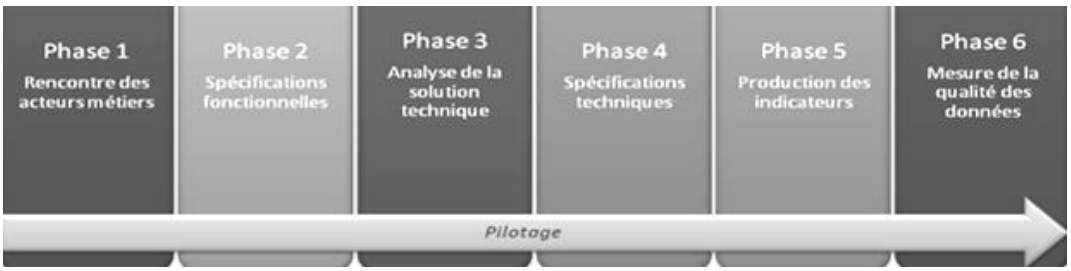

Figure 1 : La démarche proposée

Phase 1 : La rencontre des acteurs métiers

L'objectif de cette phase est d'acquérir la connaissance métier nécessaire à la compréhension des processus opérationnels afin d'être capable de conceptualiser les indicateurs avec l'accord des équipes métiers.

La conduite de cette phase a consisté en la réalisation de **2 types de réunions** :

• **« Réunions Spécifications »** : le chef de projet initie la spécification de chaque indicateur par rapport aux documents annexes transmis par les utilisateurs référents de l'établissement. Le point de vue utilisateur est ici essentiel.

• **« Validation des tableaux de bord »** : le chef de projet présente chaque extraction en expliquant les calculs effectués afin de les valider avec les personnes concernées.

Phase 2 : Les spécifications fonctionnelles

La phase de rédaction des spécifications fonctionnelles permet de synthétiser les différents échanges avec les acteurs métiers. Cette étape permet de détailler pour chaque indicateur sa définition, son calcul, son périmètre, sa source probable, sa maille d'extraction, et sa fréquence d'extraction.

Ces spécifications fonctionnelles permettent ainsi la validation d'un consensus, point de départ de toute phase de réalisation. Cette étape permet également d'accroitre l'implication des différents acteurs, facilitant la conduite du changement.

Phase 3 : Analyse de la solution technique

Durant cette phase, il s'agit d'analyser la solution technique pour définir les méthodes d'extraction des données et de production des indicateurs. Les objectifs de cette phase sont doubles : d'une part, s'appuyer sur l'existant de manière à produire plus rapidement l'indicateur et à éviter deux sources de données contradictoires ou deux calculs pour le même indicateur générant habituellement des écarts difficiles à expliquer. Ce qui entrainerait naturellement un problème de crédibilité du projet. Intégrer d'autre part les indicateurs ANAP dans l'entrepôt de données existant afin de bénéficier rapidement des indicateurs existants mais aussi des informations structurantes (ex : la structure des pôles), de réutiliser les moyens déjà en place (normes, mode de mise en production, script de lancement automatique des extractions…).

Phase 4 : Les spécifications techniques

Une fois les indicateurs définis et les méthodes d'extraction établies, la phase de rédaction des spécifications techniques peut être démarrée. Il s'agit alors pour le chef de projet de spécifier les sources, la correspondance entre données (*mapping*), la modélisation, les extractions. D'autre part, il s'agit pendant cette phase de spécifier également la conception des interfaces pour la correction des données dans les applications sources. Cette étape est nécessaire car ces interfaces peuvent contenir des degrés de complexité très différents.

Phase 5 : La production des indicateurs

Cette étape concerne la phase de mise en œuvre des indicateurs. Il s'agit de réaliser les flux d'extraction des données depuis les sources spécifiées avec l'extracteur, les flux de correction et d'enrichissement des données avec l'outil de DQM, et la production de l'information facilement consultable sur un portail à accès restreint. Cette étape intègre également le développement des IHM de corrections des données saisies au préalable. Ces écrans auront un impact majeur dans l'amélioration de la qualité des données. Tout démarrage d'un processus de gestion de la qualité de données implique la mise en place d'un accès aux données sources pour les corriger.

Phase 6 : Mesure de la qualité des données

Il s'agit enfin ici de définir un ensemble de rapports concernant l'analyse des sources des indicateurs et des indicateurs produits. Cette analyse des sources et des cibles, nœud de chacune des étapes du projet, renforce la prise de conscience des utilisateurs dans l'amélioration des données saisies, affirme l'utilité d'une méthodologie mise en œuvre pour améliorer la qualité des données, et consolide la crédibilité des indicateurs mis à disposition des utilisateurs.

3. Résultats

Nous relatons dans cette section uniquement les résultats obtenus dans la dernière étape de la démarche, à savoir ceux portant sur la mesure de la qualité des données.

Les fonctions de qualité de données (profilage, redressement, enrichissement et supervision) abordées en 1.4, sont entièrement intégrées à l'outil de DQM. Cependant, nous constatons que seules les deux premières fonctions ont été utilisées dans les exemples traités dans l'étude de cas. L'essentiel des analyses concerne en effet la première étape de profilage pour envisager par la suite les corrections nécessaires. L'outil propose à la fois une vue graphique et fonctionnelle de ce processus d'intégration, et permet une approche accessible et non technique des données. Il offre une prise en main rapide et un niveau de réutilisation élevé.

3.1. La gestion des lits

Le nombre de lits installés et le nombre de lits occupés proviennent de deux applications distinctes. Pour chacune, la structure hospitalière (Pôle, Service, UF[3], CAC[4]) est actualisée de façon autonome depuis la structure de référence (outil GAM). Nous pouvions donc avoir des taux d'occupation incohérents simplement à cause d'un problème d'incohérence de structure sur l'une des deux sources.

[3] *Unité Fonctionnelle.*
[4] *Code d'Activité.*

Nous avons mis en place dans la phase profilage les analyses suivantes :
• Analyse de la cohérence du nombre de lits indisponibles
• Analyse d'incohérence de structure entre les différentes applications (plusieurs analyses)

Nous avons donc effectué la comparaison des 2 applications par rapport à la structure de référence, ce qui nous a permis de valider la cohérence des taux d'occupation [Tableau 2].

Ces analyses permettent de vérifier que la structure (UF, CAC) utilisée dans \"Gestion des Lits\" est conforme à la structure de référence GAM.Si les structures ne sont pas conformes, les indicateurs \"Nb lits installés\", \"Taux d'occupation\" peuvent être impactés.
Si des données sont invalides il faut effectuer les modifications dans l'outil \"Gestion des Lits installés\".
Les mises à jour seront visibles dès le lendemain.

Date de dernière MAJ	Champs Analysés	Description Analyse	Nb Lignes Total	Nb Lignes Invalides	Valides (%)	Invalides (%)
21/02/12	Code UF	On vérifie que les UF utilisées dans l'application 'Gestion des lits installés' sont bien présentes dans la structure de référence (GAM) de l'année en cours.	190	0	100,00%	0,00%
21/02/12	CAC - Type de CAC	On vérifie que les couples CAC-Type de CAC utilisés dans l'application 'Gestion des lits installés' sont bien présents dans la structure GAM de l'année en cours.	175	15	92,11%	7,89%
21/02/12	UF-CAC-Type de CAC	On vérifie que les triplets UF-CAC-Type de CAC utilisés dans l'application 'Gestion des lits installés' sont bien présents dans la structure GAM de l'année en cours.	173	17	91,05%	8,95%
21/02/12	Code CAC	On vérifie que les CAC utilisés dans l'application 'Gestion des lits installés' sont bien présents dans la structure de référence (GAM) de l'année en cours.	187	3	98,42%	1,58%
21/02/12	I_ANAP_LITS_DISP_UF_CAC_JOUR_LITS_INST_INF_LITS_IND	Cette analyse met en évidence les UF pour lesquelles le nombre de lits indisponibles est supérieur au nombre de lits disponibles pour un jour donné.	11 902	91	99,24%	0,76%

Tableau 2 : Résultats des analyses Gestion des lits

Ces analyses portent ici sur des statistiques simples telles que le nombre de lignes et le nombre de valeurs invalides. En vérifiant régulièrement ces métriques et en observant leur tendance et leur évolution, nous avons pu constater l'amélioration ou la dégradation de la qualité de ces données. Les lignes invalides ont pu être affichées pour chaque analyse et ainsi permettre de redresser et réparer toutes les données non conformes directement dans l'outil de Gestion des lits installés.

3.2. L'imagerie

Les analyses sur l'imagerie permettent de s'assurer de la cohérence des informations saisies dans le logiciel d'imagerie, source principale des indicateurs. Le profilage analyse les examens sans nature, les examens sans radiologues, les examens sans salle, les examens sans type [Tableau 3].

Ces analyses permettent de vérifier que les données remontées pour le calcul des indicateurs d'imagerie ne sont pas faussées.
Si des données sont invalides il faut effectuer les modifications dans l'outil d'imagerie.
Les mises à jour seront visibles dès le lendemain.

Date de dernière MAJ	Champs Analysés	Description Analyse	Nb Lignes Total	Nb Lignes Invalides	Valides (%)	Invalides (%)
18/10/11	Examen_sans_nature	Cette analyse met en évidence les examens d'imagerie sans nature	124 345	0	100,00%	0,00%
18/10/11	Examen_sans_radiologue	Cette analyse met en évidence les examens d'imagerie sans radiologue	124 345	0	100,00%	0,00%
18/10/11	Examen_sans_salle	Cette analyse met en évidence les examens d'imagerie sans salle	124 345	0	100,00%	0,00%
18/10/11	Examen_sans_type	Cette analyse met en évidence les examens d'imagerie sans type	124 345	0	100,00%	0,00%

Tableau 3 : Résultats des analyses Imagerie

Aucune action correctrice n'a été envisagée dans l'outil d'imagerie en raison des bons résultats des analyses.

3.3. Les blocs opératoires et l'anesthésie

Afin d'être le plus précis possible, les indicateurs blocs et anesthésie sont basés sur les données des feuilles de salles et de l'application de gestion des blocs. Ainsi le nombre d'IADE[5] et d'IBODE[6] par exemple est exhaustif pour une opération donnée.

Les données de chaque personne présentes dans la feuille de salle sont croisées avec les données RH et planning, ce qui nous assure une cohérence des indicateurs remontés.

Nous avons mis en place dans le profilage les analyses suivantes [Tableau 4] :
• Matricule intervenant : cette analyse met en évidence les intervenants qui n'ont pas de correspondance dans le fichier intervenant- par rapport à leur matricule. On compare donc des données issues de l'application des blocs avec les données RH.
• Intervention sans anesthésiste ou chirurgien opérateur : on vérifie qu'un anesthésiste ou chirurgien est bien déclaré comme « responsable » pour chaque intervention. Cette information est importante car de nombreux indicateurs sont basés sur ces listes de responsables.
• Intervention sans date de fin : information importante dans le calcul du taux d'occupation des salles.

Ces analyses permettent de vérifier que les données dans les fichiers interventions-k, interventions-intervenants-k. intervenants-ipop sont correctes. S'il existe des données invalides elles peuvent impacter le calcul des indicateurs tel que le temps de travail, taux de mobilisation PNM,taux de mobilisation PM, NB IADE-IBODE / Salle, Nb interventions par chirurgiens, anesthésistes
Si les données sont invalides, il faut corriger directement dans les fichiers et les redéposer sur le serveur
Les mises à jour seront visibles dès le lendemain.

Date de dernière MAJ	Champs Analysés	Description Analyse	Nb Lignes Total	Nb Lignes Invalides	Valides (%)	Invalides (%)
18/10/11	Matricule intervenant	Cette analyse permet en évidence les intervenants qui n'ont pas de correspondances dans le fichier intervenant-ipop par rapport à leur matricule	99 818	0	100,00%	0,00%
18/10/11	Interventions_sans_anesthesiste_operateur	Cette analyse met en évidence les interventions pour lesquelles il n'y a pas d'anesthésistes opérateurs	18 996	7 077	62,74%	37,26%
18/10/11	Interventions_sans_chirurgien_operateur	Cette analyse met en évidence les interventions pour lesquelles il n'y a pas de chirurgiens opérateurs	18 996	40	99,79%	0,21%
18/10/11	Interventions_sans_date_de_fin	Cette analyse met en évidence les interventions pour lesquelles les colonnes etat_sortie, sortie1 et sortie2 sont pas renseignées	18 996	18	99,91%	0,09%

Tableau 4 : Résultats des analyses Blocs opératoires

Le redressement des données invalides s'opère directement dans les fichiers sources que l'on redépose sur le serveur. L'effet est visible le lendemain. Aucun rapprochement de données ni d'arbitrage n'a été nécessaire.

3.4. La biologie

Dans les analyses de biologie [Tableau 5], nous utilisons des données issues de l'ATIH[7] comparées aux données de biologie afin de s'assurer de leur complétude de part et d'autre.

Nous avons mis en place les analyses suivantes :
• Analyse des analyses de biologie sans prescription.
• Analyse de la cohérence des IEP-GHM[8] de l'application de biologie par rapport aux données de l'ATIH.

[5] *Infirmier Anesthésiste Diplômé d'État.*
[6] *Infirmier de Bloc Opératoire Diplômé d'État.*
[7] *L'Agence Technique de l'information sur l'Hospitalisation met en œuvre le programme de médicalisation du système d'information (PMSI) devenu un outil de pilotage contribuant à mesurer la performance des établissements de santé.*
[8] *Identifiant externe du patient (IEP) – Groupe homogène de malades (GHM).*

Ces analyses permettent de vérifier les IEP contenant un GHM non présent dans le fichier FF_ANAP_ICR_GHM.csv. On vérifie aussi les prescriptions pour lesquelles le code prescription est vide.
Si les données sont invalides, il faut corriger directement dans les fichiers et les redéposer sur le serveur
Les mises à jour seront visibles dès le lendemain.

Date de dernière MAJ	Champs Analysés	Description Analyse	Nb Lignes Total	Nb Lignes Invalides	Valides (%)	Invalides (%)
18/10/11	IEP-GHM	Cette analyse permet en évidence les IEP avec GHM non présent dans le fichier des GHM	129 239	3	100,00%	0,00%
18/10/11	Code_prescription_vide	Cette analyse met en évidence les prescriptions pour lesquelles le code prescription est vide	639 968	714	99,89%	0,11%

Tableau 5 : Résultats des analyses Biologie

Même opération que dans le cas précédent : le redressement s'effectue directement dans les fichiers sources que l'on redépose sur le serveur. L'effet est visible le lendemain. Aucun rapprochement de données ni d'arbitrage n'a été nécessaire.

Conclusions et recommandations

Un nombre important de projets Performance sont mis en œuvre dans les hôpitaux et implique désormais l'intégration de données à travers des approches visant à produire des indicateurs fiabilisés et des analyses garanties. Se pose cependant dans chacun de ces projets la problématique de la qualité de l'information. Afin de réduire la distance entre les données issues des systèmes d'information opérationnels et la construction partagée d'indicateurs à partir de ces données, nous proposons une démarche industrielle d'amélioration de la qualité de l'information, prenant en compte le point de vue de l'utilisateur (étapes 1 à 3) pour définir l'information stockée dans la base de données, et instrumentée par des outils spécifiques de gestion de la qualité des données (étapes 4 à 6).

Dans notre expérimentation sur les périmètres fonctionnels retenus dans ce projet Performance, le travail effectué sur la donnée (profilage et redressement pour l'essentiel) a produit des résultats significatifs. L'apport de ce travail est double. Premièrement, ce travail a permis au Centre Hospitalier de contrôler et améliorer son système d'information opérationnel en mesurant la distance entre les indicateurs de performance demandés et les données entrant dans leur composition. Deuxièmement, nous avons repéré et amélioré la qualité de certaines données, et nous disposons aujourd'hui d'un entrepôt de données où l'information est mieux vérifiée et plus consensuelle.

Notre expérimentation a été mise en œuvre avec succès dans le cadre du projet Performance en produisant des indicateurs de performance garantis. Elle a permis ainsi au Centre Hospitalier de mieux inter-opérer avec les institutions comme l'ANAP et les ARS.

Toute l'expérimentation n'est pas automatisée, l'aide des utilisateurs des différents domaines est toujours requise et essentielle. L'entrepôt de données complété d'un outil de gestion de la qualité des données est un pas vers une meilleure confiance des systèmes d'information d'une part, et vers une meilleure qualité des informations extraites d'autre part. Il serait intéressant d'appliquer notre méthode sur d'autres hôpitaux afin de vérifier la véracité de notre approche, et de rechercher les impacts de cette qualité de l'information sur l'appropriation des tableaux de bord par les utilisateurs.

Références

[1] Gainer et al. (2007). Using the i2b2 Hive for Clinical Discovery: an Example. AMIA Annual Symp Proceedings.

[2] Choquet et al. (2010). Un modèle de connaissances pour mesurer la qualité d'une source d'information. Ingénierie de la Connaissance IC2010.

[3] Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers and M Weinstein RA. (2003). Development of a clinical data warehouse for hospital infection control. Journal of the American Medical Informatics Association. vol. 10 (5) pp. 454-462.

[4] Strong D., Lee Y., Wang R. (1997). Data quality in context, Communications of the ACM, vol. 40, no. 5, 103-110.

[5] Jarke M. et Vassiliou Y. (1997). Data warehouse quality design: A review of the DWQ project, In Proceedings of the International Conference on Information Quality (IQ), Cambridge.

[6] Berti L. (1999). Qualité de données multi sources et recommandation multicritère, INFORSID 99.

[7] Fellegi I.P. et Sunter A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, vol. 64.

[8] Wang R.Y. (1998). A product Perspective on Total Data Quality Management. *Communication of the ACM*, vol. 41, no.2.

[9] Redman T.C. (1996). Data Quality for the Information Age. *Artech House*.

[10] Naumann F. et Rolker C. (2000). Assessment Methods for Information Quality Criteria, Proceedings of the International Conference on Information Quality (IQ2000) Cambridge.

[11] Peralta V. (2006). Data quality evaluation in data integration systems. *PhD Thesis*, Université de Versailles (France) and Universidad de la República. Uruguay.

[12] Deming, WE. (1982). Out of the Crisis, *MIT Press*, Cambridge.

[13] Berti-Equille L., Moussouni F. (2005). Quality-aware integration and warehousing of genomic data. In *Proc. of the 10th Intl. Conference on Information Quality* (IQ'05). MIT, Cambridge, U.S.A.

[14] Krogstie J., Lindland O.I., Sindre, G. (1995). Towards a Deeper Understanding of Quality in Requirements Engineering. *Proceedings of the 7th International Conference on Advanced Information Systems Engineering (CAISE)*, Jyvaskyla, Finland.

[15] Moody D.L. (2003). Measuring the quality of data models: an empirical evaluation of the use of quality metrics in practice. *Proceedings of the Eleventh European Conference on Information Systems, ECIS*.

[16] Kerr K., Norris A., Stockdale R. (2007). Data Quality Information and Decision Making: A Healthcare Case Study. *Proc. 18th Australasian Conference on Information Systems.*

[17] Calabretto S., Pinon J.M., Poullet L. et Richez M.A. (1998). De la qualité de l'information à la qualité de la documentation, Document Numérique, vol.12, no.1, 37-52.

[18] Zhu X. et Gauch S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece .288–295.

[19] Marotta A. (2002). Quality Management in MSIS, Technical Report INCO TR-03-03. ISSN 0797-6410.

[20] Harrathi R., Calabretto S. (2006), Un modèle de qualité de l'information, Extraction et gestion des connaissances EGC'2006, Lille, 17-20 janvier 2006.

[21] Goldberg SI, Niemierko A., Turchin A. (2008). Analysis of data errors in clinical research databases. *AMIA Annual Symp Proc*. Nov 6:242-6.

[22] Arts D., De Keizer N., Scheffer G.J. (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association.* vol. 9 (6) pp. 600-611.

[23] Moody D.L., Shanks G.G. (1998). What Makes A Good Data Model? A Framework For Evaluating.