UNIVERSITE DE LILLE
FACULTE DE MEDECINE HENRI WAREMBOURG

Année 2020

THESE POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE

Mise au point d'une interface de visualisation unifiée de données cliniques temporelles, hétérogènes, hiérarchiques

Présentée et soutenue publiquement le 05/05/2020
à 16h00 au pôle recherche

Par Niels MARTIGNENE

JURY
Président :
Monsieur le Professeur Philippe AMOUYEL
Assesseurs :
Monsieur le Docteur Luc DAUCHET
Monsieur le Docteur Antoine LAMER
Directeur de thèse :
Monsieur le Professeur Emmanuel CHAZARD

Avertissement

La Faculté n’entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Sigles

ASIP Agence Française des Systèmes d'Information Partagés
ATC Anatomical Therapeutic and Chemical classification
AVK Antivitamines K
CCAM Classification Commune des Actes Médicaux
CIM10 Classification internationale des maladies, 10e révision
DAS Diagnostic Associé Significatif
DP Diagnostic Principal
DR Diagnostic Relié
ECG Electrocardiogramme
EHR Electronic health records
GHS Groupe Homogène de Séjours
HAD Hospitalisation à domicile
HL7 Health Level 7
ICD10 International Classification of Disease 10
IEP Identifiant Externe Patient
INR International Normalized Ratio
IPP Identifiant Permanent Patient
LPP Liste des Produits et Prestations
MCO Médecine Chirurgie Obstétrique
NDA Numéro Dossier Administratif
NIP Numéro Identification Patient
PMSI Programme de Médicalisation du Système d'Information
RECIST Response Evaluation Criteria in Solid Tumours
RSS Résumé Standardisé de Séjour
RUM Résumé d'unité médicale
SQL Structured Query Language
SSR Soins de Suite et de Réadaptation
T2A Tarification à l'activité
TALN Traitement Automatique du Language Naturel

Sommaire

Préambule

Le travail scientifique présenté dans cette thèse de médecine a fait l'objet d'une publication d'article international en Anglais (1,2). Conformément aux recommandations de la Faculté, le présent document suit le plan suivant :

- Une introduction longue en Français, qui poursuit deux objectifs :
    - Présenter le contexte avec une orientation principalement pédagogique
    - Présenter le contexte scientifique et l'objectif, comme le fait également l'introduction de l'article en Anglais
- L'abstract en Anglais, tel qu'il a été soumis en complément de l'article reproduit juste après.
- L'article en Anglais, tel qu'il a été publié. Cet article suit le plan classique, dans le format imposé (introduction, matériel et méthodes, résultats, discussion)
- Une discussion en Français, qui reprend pour l'essentiel la discussion en Anglais de l'article

Les références présentées en fin de document, ainsi que les listes de figures et tables, résultent de la fusion des parties en Anglais et en Français.

Introduction

1. Contexte de l'étude

La prise en charge au quotidien des patients dans les établissements de santé repose de plus en plus sur l'utilisation de logiciels informatiques. Ces logiciels sont utilisés pour produire l'information (demandes d'examen, prescription de soins, rédaction de courriers médicaux) et pour y accéder (observations et comptes-rendus d'autres praticiens, résultats de biologie, etc.). Par rapport au dossier papier auxquels ils se substituent, les dossiers numériques sont plus faciles à consulter et à partager au sein des établissements et entre établissements, et peuvent être enrichis par des outils d'aide à la décision.

À notre connaissance, la plupart des logiciels médicaux se sont jusqu'à aujourd'hui efforcés de reproduire numériquement le fonctionnement « papier » qu'ils visent à remplacer. Ils proposent des outils de traitement de texte pour permettre de créer et de lire les courriers médicaux et les comptes-rendus. Ils affichent des tableaux de données pour les informations quantitatives comme les paramètres hémodynamiques, la température, les dosages de biologie. Les observations médicales sont saisies dans des champs de saisie libre ou dans des formulaires (données structurées). Les techniques de visualisation et d'automatisation spécifiques en informatique sont généralement peu exploitées au sein de ces systèmes (3).

Les données issues de ces logiciels, les données administratives et les données de facturation sont rassemblées au sein de grandes bases de données clinico-administratives qui peuvent être utilisés en recherche. La mise en évidence d'associations et de relations causales cachées dans ces données temporelles pourrait permettre la découverte de nouvelles pistes de recherche clinique. Par ailleurs, des visualisations adaptées pourraient faciliter la sélection de patients à partir des entrepôts de santé (« pre-screening »). Il existe peu d'outils et de publications (4) dédiés à la visualisation de ces données temporelles complexes.

Ce travail s'intéressera en particulier à la visualisation d'informations temporelles d'un individu dans le cadre du soin (visualisation transactionnelle), mais nous aborderons également la consultation « en groupe », c'est-à-dire la consultation des informations de plusieurs patients simultanément (visualisation décisionnelle).

2. Données de santé

2.1. Généralités

2.1.1. Typologie des données de santé

On distingue schématiquement plusieurs activités en rapport avec le soin :

- Activités de préadmission, d'admission et de sortie administrative
- Activités de facturation ou valorisation T2A
- Activités de soins dans le service hébergeur
- Activités dans les blocs opératoires et les plateaux techniques
- Activités de laboratoire
- Activités d'imagerie

Pour chaque activité, différents types de données sont générées. Ces données peuvent être structurées (exploitables directement par un algorithme) ou non structurées (elles sont stockées sans format prédéfini, comme le texte des compte-rendus ou des courriers médicaux, et sont interprétées par des humains). Les machines produisent généralement des informations structurées brutes (par exemple des mesures de biologie médicale), alors que les professionnels de santé échangent des informations non structurées à haute valeur interprétative (par exemple un diagnostic).

Les dossiers médicaux sont constitués par l'agrégation d'informations de différentes origines : le **Tableau 1** donne un aperçu de ces données.

***Tableau 1**. Origine, nature et structure des données du dossier médical.*

Catégorie	Nature	Structure
Données administratives	Valeurs, texte	Structurée
Communications entre soignants (Transmissions et observations médicales, courriers médicaux, comptes-rendus)	Texte	Non structurée
Données gérées par les pharmacies centralisées	Valeurs	Structurée
Résultats de biologie médicale	Valeurs, texte	Structurée
Données issues d'appareils de surveillance	Valeurs	Structurée
Données d'imagerie	Images, texte	Non structurée
Données et codes du PMSI	Codes, valeurs	Structurée

Ces données sont produites par une multitude de logiciels et sont reliées par des identifiants uniques spécifiques pour chaque patient et pour chaque séjour. La grande majorité d'entre elles porte par ailleurs une information temporelle, qui permet de retracer en partie l'histoire clinique d'un patient.

La **Figure 1** illustre le cas d'une femme de 88 ans. Dans son dossier médical on retrouve des codes diagnostiques CIM10 (5) (par exemple I10 qui correspond à une hypertension artérielle), des codes d'actes CCAM (6) (comme FELF001 qui correspond à la transfusion de concentré de globules rouges), des courriers médicaux, des prescriptions et/ou administrations médicamenteuses et des dosages de biologie médicale.

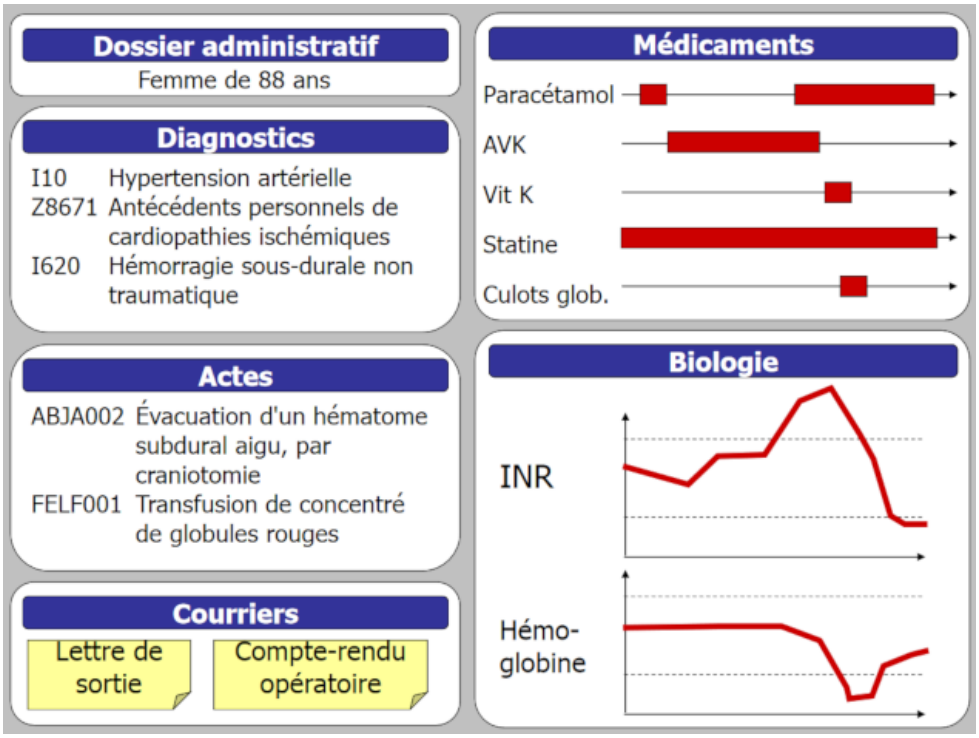

***Figure 1**. Représentation schématique de certaines données structurées du dossier patient.*

Ces informations sont résumées sur une pancarte simple, mais en réalité et dans la plupart des systèmes d'information hospitaliers, toutes ces données sont dispersées dans différents logiciels ainsi que dans le dossier papier, encore très souvent utilisé dans les hôpitaux.

Ci-dessous, nous détaillerons les principaux types de données. Toutes ces données sont reliées par des identifiants numériques partagés par les différents logiciels, identifiants que nous présenterons tout d'abord.

2.1.2. Notions d'interopérabilité

Le concept d'interopérabilité correspond à la capacité de deux systèmes informatiques à échanger des informations. Elle doit être décrite à trois niveaux (7) :

- Interopérabilité technique (« pouvoir communiquer »)
- Interopérabilité syntaxique (« savoir communiquer »)
- Interopérabilité sémantique (« savoir comprendre »)

L'interopérabilité technique correspond à la capacité de deux systèmes à communiquer. C'est le cas de deux machines connectées à un même réseau informatique. Par analogie, deux êtres humains présentent une interopérabilité technique dans la mesure où ils peuvent émettre et recevoir des sons.

L'interopérabilité syntaxique correspond à la capacité de deux systèmes à comprendre un même « langage » ou protocole. Le protocole HL7 est une norme de communication très utilisée pour la communication entre logiciels médicaux. Par analogie, deux êtres humains présentent une interopérabilité syntaxique lorsqu'ils parlent le même langage (par exemple le français).

L'interopérabilité sémantique nécessite l'utilisation de nomenclatures et d'échelles de valeurs communes pour décrire les données. Par analogie, deux êtres humains sont sémantiquement interopérables s'ils connaissent le même jargon (par exemple deux médecins peuvent échanger des informations médicales).

En médecine, les problèmes d'interopérabilité (notamment sémantique) sont très fréquents, pour des raisons historiques et aussi commerciales, et il est donc souvent nécessaire de construire des tables de correspondance (*mappings*) pour que deux systèmes différents puissent échanger des données. C'est le cas par exemple en biologie médicale : en effet les laboratoires utilisent souvent des libellés non standardisés pour les résultats (par exemple un laboratoire peut utiliser « Na » pour la natrémie alors qu'un autre utiliser « Sodium »). Ces

mappings sont généralement créés manuellement mais il existe également des techniques automatisées qui peuvent aider dans les cas complexes (8).

L'ASIP (Agence Française des Systèmes d'Information Partagés) Santé est l'agence française en charge de promouvoir l'interopérabilité des logiciels et technologiques de santé. Elle définit des cadres d'interopérabilité physique, syntaxique et sémantique que doivent respecter les éditeurs.

2.1.3. Identification des patients et des épisodes

Les données sont identifiées par des identifiants uniques qui sont générés par les systèmes d'information. Deux identifiants très importants sont utilisés par les systèmes d'information hospitaliers :

- L'IEP (pour Identifiant Externe du Patient) : il identifie le séjour hospitalier de manière unique. Un nouvel IEP est généré à chaque séjour hospitalier, y compris pour un même patient. On parle également de NDA (Numéro du Dossier Administratif).
- L'IPP (pour Identifiant Permanent du Patient), qui identifie un patient de manière unique. Une fois généré, il est réutilisé pour tous les séjours du même patient dans le même hôpital. On parle également de NIP (Numéro d'Identification Patient).

Ces identifiants permettent de joindre les informations de différentes natures à un épisode (comme un séjour hospitalier) et in-fine à un patient précis. En termes techniques, ces identifiants peuvent être utilisés comme des clés de jointure.

D'autres identifiants uniques sont disponibles dans différentes bases. Par exemple, le numéro « ano » est un identifiant unique présent dans la base de données du PMSI, il permet de chainer les séjours d'un même patient dans différents hôpitaux.

2.2. Catégories de données

2.2.1. Données administratives

Elles sont constituées avant tout de données démographiques qui sont variées et de différentes natures : date de naissance (et âge), sexe, code postal, etc. Certaines sont fixées

mais d'autres peuvent changer au cours du temps, comme par exemple l'adresse du patient, ou bien encore l'âge.

Des mouvements sont générés pour chaque passage dans une unité hospitalière. Ces données permettent de tracer le parcours d'un patient au cours d'un séjour hospitalier. Dans le PMSI, chaque mouvement entraine la production d'un RUM (Résumés d'Unité Médicale) et l'ensemble des RUM correspond à un RSS (Résumé de Sortie Standardisée), qui résume le séjour hospitalier dans son ensemble.

2.2.2. Communications entre soignants

La continuité des soins entre soignants à l'intérieur d'un établissement et entre établissements repose sur la rédaction de courriers et de comptes-rendus (texte libre). Ces données peuvent être catégorisées en fonction de leur temporalité (date et horizon d'utilisation), et par la nature et le profil de l'émetteur et du récepteur. On peut schématiquement classer ces documents dans plusieurs groupes :

- Les transmissions et observations médicales sont rédigées à l'entrée et quotidiennement pendant le séjour du malade, elles sont utilisées pour suivre l'évolution du patient.
- Les courriers médicaux (ou courriers de sortie) sont rédigés lorsque le patient quitte un établissement ou une unité médicale et résument l'histoire du patient à son arrivée et pendant le séjour.
- Les comptes-rendus d'examens sont rédigés par les spécialistes qui réalisent ou interprètent des examens complémentaires, et destinés à être lus par le médecin référent d'un patient.

Ces documents écrits contiennent l'information la plus riche et précise sur l'état de santé du patient, mais il s'agit d'une information non structurée difficile à exploiter de manière automatisée. L'extraction d'informations structurées à partir du texte (Traitement Automatique du Language Naturel ou TALN), à l'aide de systèmes experts ou de techniques d'intelligence artificielle, est un domaine actif de recherche. Des techniques simples de text mining, comme les approches de type *bag of words* (9) peuvent être utilisées, elles sont

améliorées par le zonage des courriers et la prise en compte des négations. On peut également extraire des informations utiles à partir des métadonnées (date, type, auteur).

La saisie directe d'informations structurées par les soignants est une autre possibilité. Elle pourrait dans certains cas réduire le risque d'erreurs (10), notamment en ce qui concerne la prescription médicamenteuse, mais elle impose une certaine rigidité qui n'est pas toujours facile à articuler avec la grande complexité et la variabilité des prises en charge médicales (11). La demande croissante de saisie d'informations par les soignants observée depuis de nombreuses années conduit d'ailleurs à une augmentation de charge « administrative » des soignants qui est perçue comme excessive (12).

2.2.3. Données gérées par les pharmacies centralisées

Pour chaque médicament prescrit à un patient, différentes informations sont produites par différents acteurs et logiciels, on parle de circuit du médicament. Ce circuit (illustré sur la **Figure 2**) comprend dans l'ordre :

- La prescription (obligatoire) par une profession médicale
- L'analyse des prescriptions par un pharmacien
- La préparation par le pharmacien et parfois l'infirmier (doses unitaires en hospitalisation)
- L'administration du médicament au patient, parfois sous responsabilité d'une infirmier

La délivrance peut être nominative (par exemple les médicaments onéreux comme les chimiothérapies) ou globale (stocks dans la pharmacie du service hébergeur).

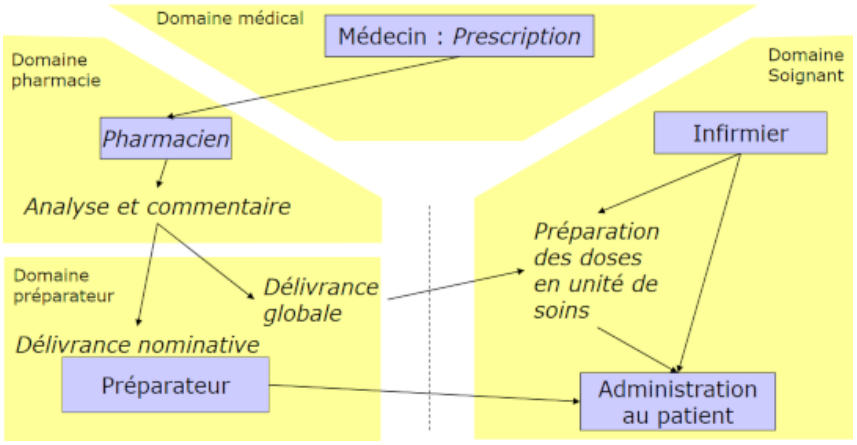

***Figure 2**. Illustration du circuit du médicament dans un hôpital.*

Il existe diverses terminologies pour désigner et décrire les médicaments : nom commercial (par exemple DOLIPRANE®), dénomination commune internationale ou DCI (par exemple paracétamol), nom chimique, etc. Cette diversité de terminologies pose d'importants problèmes d'interopérabilité sémantique.

La classification ATC (Anatomical Therapeutic and Chemical classification) est une terminologie de médicaments très utilisée dont la hiérarchisation est principalement basée sur l'indication thérapeutique.

Les pharmacies gèrent également des dispositifs médicaux (exemples non exhaustifs : pansements, aiguilles, gels à appliquer sur la peau, prothèses articulaires, pacemakers cardiaques). Tous les dispositifs médicaux ont un code LPP (Liste des Produits et Prestations) pour lesquels il existe plus de 350 000 codes. Les dispositifs médicaux implantables (DMI) sont généralement facturés « en sus » de l'hospitalisation, et les codes LPP correspondants sont codés dans le PMSI (liste « en sus » d'environ 2000 codes à ce jour).

Les pharmacies gèrent également les produits de la Liste des Produits et Prestations (13) (LPP), parmi lesquels on retrouve des dispositifs médicaux. Il s'agit de traitements et matériels d'aide à la vie, d'aliments diététiques, d'articles pour pansements, d'orthèses et prothèses externes, de dispositifs médicaux implantables ou encore de véhicules pour handicapés physiques.

2.2.4. Résultats de biologie médicale

La biologie médicale implique plusieurs étapes :

- Prescription par un professionnel médical
- Prélèvement par un infirmier
- Analyse par des automates
- Production et validation des résultats par un biologiste médical

Les données de biologie sont en grande partie produites par des machines et automates. Elles sont relues par des spécialistes puis accessibles par les soignants qui les exploitent souvent directement. Certains examens biologiques complexes ou spécifiques sont interprétés par des biologistes spécialisés, qui rédigent des comptes-rendus accessibles par les cliniciens.

Ces paramètres appartiennent à différents champs : biochimie, hématologie, bactériologie, virologie, immunologie. Il s'agit généralement de valeurs quantitatives suivies dans le temps. On parle alors de dosage biologique, c'est par exemple le cas de la glycémie à jeun qui peut être utilisée pour diagnostiquer et suivre le diabète.

Même s'il existe des terminologies officielles (LOINC, UPAC, etc.) pour décrire les données de biologie, en pratique la plupart des établissements utilisent des nomenclatures « maison » non interopérables. La réutilisation de ces données nécessite de créer des correspondances (*mappings*) ad-hoc pour pouvoir mettre en relations les données de biologie de différentes origines.

Les données « omiques » (génomiques, protéomiques, etc.) peuvent être rattachées à cette catégorie. Elles ne sont pas disponibles en routine, et sont caractérisées par un volume de données beaucoup plus important que les dosages de biologie réalisés au quotidien.

2.2.5. Données issues des appareils de surveillance

Il s'agit de signaux, c'est à dire de mesures répétées avec une granularité temporelle fine et régulière. Ces mesures sont généralement quantitatives, comme par exemple la saturométrie artérielle.

Ces données peuvent être produites momentanément (par exemple lors d'un ECG réalisé au lit du patient) ou mesurées en continu chez les patients « scopés ». C'est le cas notamment des patients sous anesthésie ou des patients hospitalisés en service de réanimation pour lesquels certains paramètres vitaux (comme l'ECG pour suivre l'activité cardiaque) sont monitorés en permanence.

Comme pour la plupart des données issues de machines, ces données peuvent être explorées de manière fine (évolution à la seconde) ou bien résumées de manière globale par leur caractère normal ou anormal, par leur amélioration ou dégradation, ou même encore par des diagnostics suspectés.

2.2.6. Données produites par des appareils d’imagerie médicale

La production des données d’imagerie suit une chronologie simple :

- Prescription médicale (bon de commande)
- Production de l'image
- Réalisation d'un compte-rendu d'interprétation par un radiologue
- Ajout de codes CCAM de réalisation de l'imagerie (parfois automatiquement)
- Eventuellement, ajout de codes CIM10 au dossier en fonction des résultats

Les données d'imagerie sont avant tout constituées par des séries d'images, qui sont des données non structurées. La radiographie, la tomodensitométrie, l'IRM, l'échographie, la scintigraphie et la photographie en sont les principales modalités mais il en existe d'autres.

Les clichés sont parfois interprétés par des spécialistes en radiologie qui rédigent des comptes-rendus (par exemple les scanners abdomino-pelviens), d'autres sont lues directement par les soignants (par exemple la radiographie de thorax).

Des informations structurées peuvent être adossées aux images : codes diagnostiques suspectés, caractère normal ou anormal de l'examen, évolution d'un score dans le temps. Par exemple le score RECIST (14) (Response evaluation criteria in solid tumors) est utilisé pour caractériser l'évolution d'une tumeur après une chimiothérapie.

De plus en plus, des techniques de *deep learning* sont utilisées pour interpréter automatiquement les images obtenues par différents examens, avec des performances qui peuvent être proches voire supérieures à celles des humains dans certaines indications précises (15–17).

2.2.7. Données du PMSI

Le PMSI (Programme de Médicalisation du Système d'Information) est divisé en 5 champs : Médecine-Chirurgie-Obstétrique (MCO), hospitalisation à domicile (HAD), secteur psychiatrique (PSY), soins de suite et de réadaptation (SSR) et soins externes.

Chaque séjour dans le champ MCO est résumé par un RSS, qui rassemble plusieurs RUM comprenant différentes informations structurées :

- Des données administratives et démographiques : âge, sexe, mode d'entrée et de sortie, unité médicale, etc.

- Des codes diagnostiques CIM-10 (5) qui peuvent être saisis en position de Diagnostic Principal (DP), de Diagnostic Relié (DR) ou de Diagnostic Associé Significatif (DAS)
- Des codes d'actes CCAM (6), qui peuvent correspondre à des actes chirurgicaux, des examens d'imagerie, etc.

Un RUM est produit pour chaque unité médicale dans laquelle le patient passe au cours d'un séjour hospitalier. Les RUM produits lors d'un séjour dans une même entité géographique sont regroupés dans un RSS, qui est ensuite traité par un algorithme de groupage complexe pour classer le séjour dans un Groupe Homogène de Séjours (GHS). Depuis la mise en place de la Tarification à l'Activité (T2A) c'est ce GHS qui détermine en grande partie le paiement du séjour au montant forfaitaire correspondant au GHS.

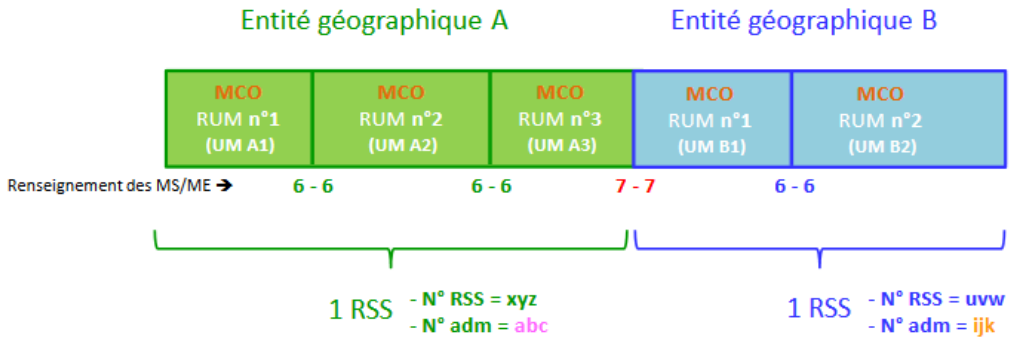

***Figure 3**. Succession de RUM et de RSS au cours d'un séjour hospitalier. ME/MS = Mode d'entrée / Mode de sortie.*

Les modes d'entrée et de sortie des différents RUM indiquent le motif de transfert. Le mode de sortie '9' correspond au décès, c'est l'information qui est utilisée pour analyser la survenue de décès intra-hospitaliers dans les analyses conduites sur les données du PMSI. Comme beaucoup des informations contenues dans le PMSI, la fiabilité de cette information est imparfaite (18).

Les informations du PMSI sont utilisées avant tout pour permettre l'allocation des ressources économiques au travers de la Tarification à l'Activité (T2A). Ces informations sont de qualité variable car elles sont avant tout utilisées à des fins de facturation et non de soins. Malgré tout, les données anonymisées sont compilées au sein de volumineuses bases médico-administratives qui peuvent être réutilisées à des fins de pilotage et de recherche. En 2016, la base nationale comportait ainsi près de 30 millions de séjours pour près de 12 millions de

patients uniques (19). Les données sont disponibles de 2008 à 2019, et un identifiant « ano » unique pour chaque patient permet de suivre les hospitalisations des patients, et in fine d'analyser des parcours hospitaliers.

3. Contextes de visualisation des données de santé

La visualisation des données de santé intervient dans deux contextes :

- *Visualisation transactionnelle* : il s'agit de la visualisation d'un dossier unique par le soignant dans le cadre du soin. Dans ce cas l'utilisateur s'intéresse au patient ou au séjour dans son ensemble, et commence généralement par une visualisation globale pour repérer les anomalies et travaille ensuite sur les différents problèmes en rassemblant mentalement des informations spécifiques à chaque problème. Cette phase intellectuelle n'est pas du tout évidente, et entre dans le cadre de l'extraction de caractéristiques (*feature extraction* (20)), que nous définirons par la suite.
- *Visualisation décisionnelle* : dans ce cas l'utilisateur cherche à mettre en évidence ou à analyser un problème particulier chez une multitude de patients. La visualisation décisionnelle peut être utilisée à des fins de pilotage, par exemple pour comprendre les parcours de soins des patients d'un établissement (21), d'un territoire ou atteints d'une maladie particulière. Elle sert aussi à des fins de recherche, par exemple en faisant de la réutilisation de données (22).

Les outils utilisés dans ces deux contextes ne sont pas tout à fait les mêmes, mais la visualisation de données a le même objectif dans les deux cas, c'est celui de favoriser la *feature extraction*. Il s'agit d'utiliser des techniques de visualisation pour permettre une compréhension de haut niveau de l'information à partir de données brutes de bas niveau.

Pour illustrer simplement ce concept, 3 informations correspondant à un séjour hospitalier sont représentées de manière superposée sur la **Figure 4** :

- Prescription d'un anticoagulant de type antivitamine K jusqu'au 5ème jour
- Réalisation d'une transfusion sanguine le 6ème jour

- Evolution de l'INR (marqueur utilisé pour suivre l'activité anticoagulante des antivitamines K ou AVK), qui dépasse la borne supérieure au 4ème jour et se normalise au 6ème jour, puis passe en zone sous-thérapeutique après 7 jours

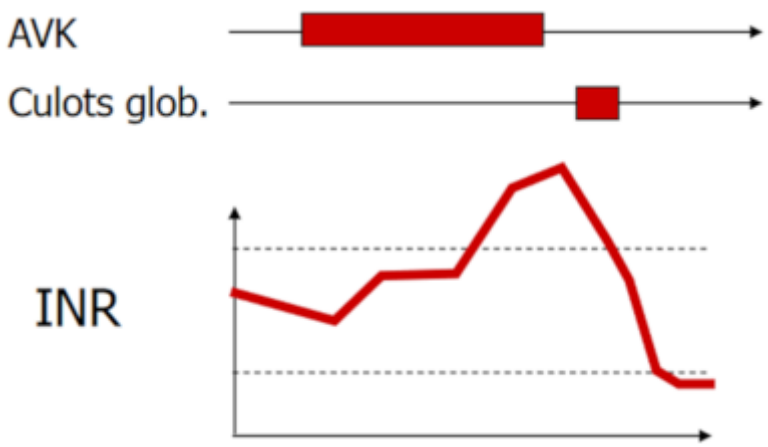

***Figure 4**. Exemple de données brutes issues du système de santé suite à une hémorragie liée aux AVK.*

Même s'il n'y a pas de donnée spécifique « hémorragie liée aux AVK », c'est une information qui peut être reconstruite (ou interprétée) à partir de ces 3 données avec une assez bonne probabilité que l'information construite soit correcte. Ce travail d'abstraction est réalisé automatiquement par un humain expérimenté, à partir du moment où les différentes informations lui sont communiquées de manière rapprochée soit dans l'espace (les données sont alignées l'une en dessous de l'autre), soit dans le temps et idéalement les deux. Une des finalités poursuivies par les outils de visualisation est de rapprocher dans l'espace visuel des données brutes qui relèvent du même problème médical. Ce rapprochement, s'il est bien fait, doit permettre spontanément au lecteur des données de réaliser sans effort l'extraction de caractéristiques.

En analyse décisionnelle, et plus particulièrement dans le cadre de la réutilisation de données, il est nécessaire d'identifier toutes les combinaisons de données, toutes les déviations et les seuils de tolérance (par exemple sur le délai entre deux évènements) lorsqu'on souhaite extraire des traits à partir d'un grand jeu de données brutes. Ce travail est indispensable car il n'existe pas de méthode statistique capable d'analyser simultanément plusieurs tables de données (i.e. plusieurs types d'individus statistiques en même temps). Il aboutit à la

construction de règles qui vont permettre de construire une information de haut niveau sur des bases de données de manière automatisée.

La qualité et l'exhaustivité des données collectées en routine sont très variables ce qui rend très compliquée la réutilisation de ces données en recherche (23–25). La **Figure 5** illustre plusieurs variations possibles de la figure précédente, dans lesquelles une ou plusieurs informations peuvent manquer. Malgré tout, la même information de haut niveau (« Hémorragie liée aux AVK ») peut être construite dans ces trois cas, avec cependant un niveau de confiance variable.

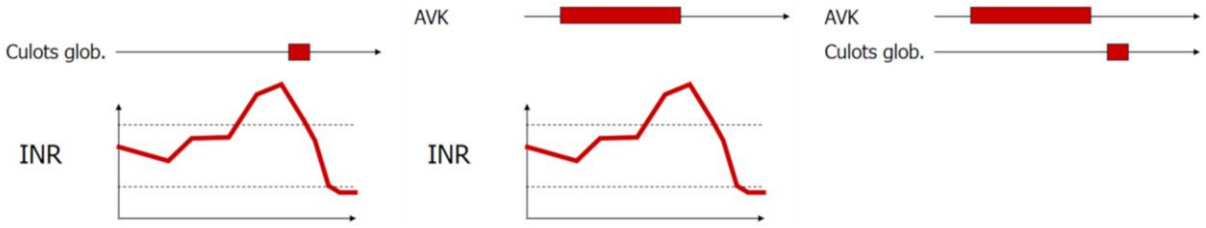

***Figure 5**. Variations possibles des données disponibles suite à une hémorragie liée aux AVK.*

4. Visualisation transactionnelle

4.1. Problèmes spécifiques en santé

La visualisation transactionnelle correspond à la consultation d'un dossier patient par le soignant dans le cadre du soin. Le soignant utilise différents outils, pertinents à différents moments, pour assurer la prise en charge :

- Il utilise une vue globale ou synthèse qui lui permet de repérer les problèmes posés par le patient
- Il consulte les informations produites (courriers médicaux, comptes-rendus d'examens) à visée diagnostique et thérapeutique
- Il commande la réalisation de nouveaux examens, il change les prescriptions et rédige des notes d'observations
- Il utilise des vues spécifiques pour explorer en détail des problèmes particuliers posés par le patient

L'échelle de temps adaptée à la visualisation est variable selon le contexte, il peut s'agir de la vie entière du patient comme d'une fenêtre de quelques secondes pour un signal instantané.

Tout ceci représente un grand volume de données et d'alertes qui entraine une charge mentale importante chez le soignant, en particulier lorsque les interfaces ne sont pas pratiques, sont mal rangées ou dispersées sur plusieurs onglets, sont lentes ou boguées. Il s'agit d'une source de fatigue et d'erreur possible par les professionnels de santé (26).

4.2. Projets de recherche et prototypes

Le logiciel *LifeLines* (27) (**Figure 6**) a été développé en 1996 dans l'université du Maryland aux Etats-Unis, au sein du HCIL (Human-Computer Interaction Lab).

Ce logiciel est capable de représenter différentes données (antécédents, examens, diagnostics, traitements, etc.) au sein d'une même ligne de vie en utilisant un seul type de représentation : des rectangles colorés caractérisés par une date de début et une date de fin. Des couleurs peuvent être utilisées pour caractériser ces évènements (par exemple les antécédents sont en gris). Les évènements sont superposés dans une même ligne de vie et regroupés au sein de catégories qui peuvent être repliées pour masquer les informations qui ne sont pas pertinentes.

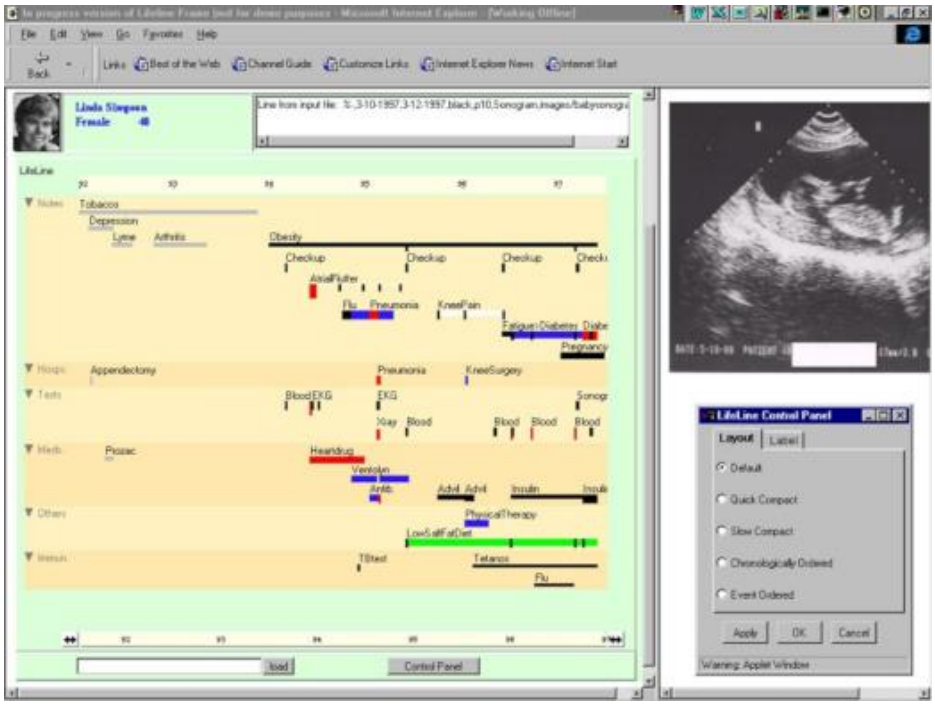

***Figure 6**. Interface du logiciel LifeLines (1996).*

Le logiciel *MultiMedia Stream System (MMSS) TimeLine* (28) (**Figure 7**) a été développé au début des années 2000 pour aider la visualisation d'histoires médicales et de données

géophysiques. Le but de ce travail était avant tout de développer un langage de requêtage visuel et textuel adapté aux données temporelles. Ce logiciel est capable de représenter les données sous la forme d'une arborescence terminologique qui peut être repliée. Une vue condensée des évènements est affichée pour les différentes catégories de l'arborescence.

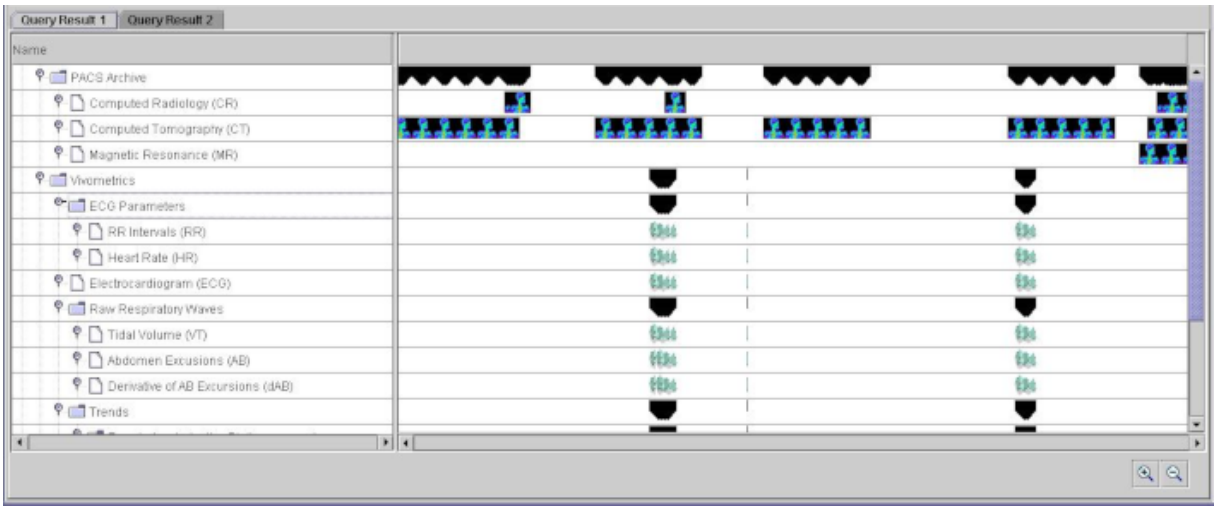

***Figure 7**. Inteface du logiciel MMSS TimeLine (2000).*

L'équipe de *Bade et al.* (29) a publié en 2004 un article qui compile différents types de visualisations de données combinant les couleurs, les formes et des techniques interactives pour coder les différentes dimensions de données temporelles hétérogènes. Ils présentent la capture d'un prototype de logiciel pour visualiser les données de patients de réanimation (**Figure 8**).

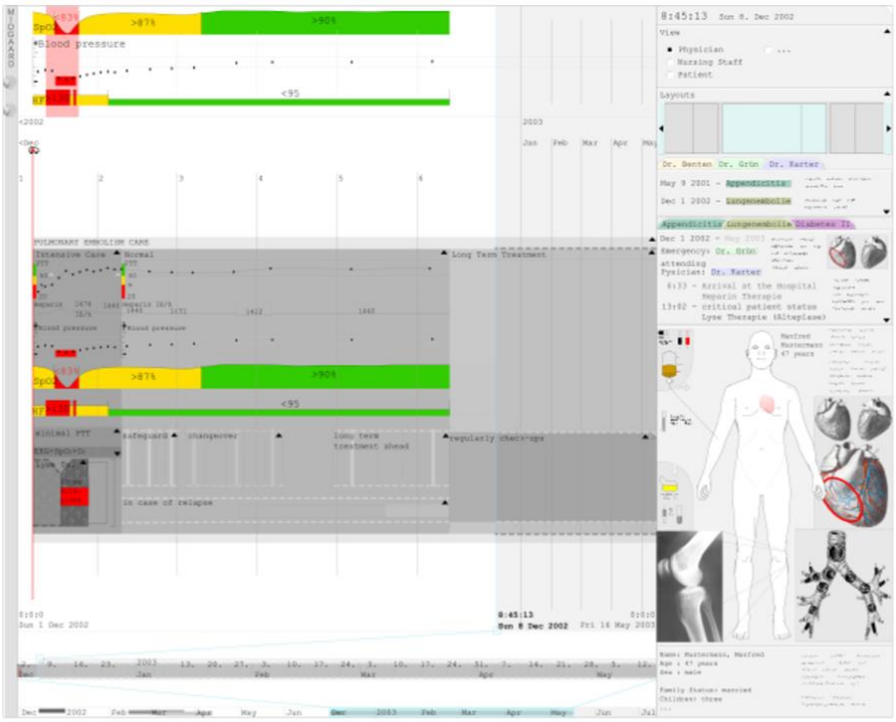

***Figure 8**. Interface du prototype présenté dans l'article de Bade et al. (2004).*

Le logiciel *KNAVE-II* (30) (**Figure 9**) a été développé en 2004. Les données sont organisées au sein d'un arbre, et l'utilisateur peut créer ses propres vues en glissant-déposant les données depuis cet arbre vers le panneau de visualisation. Il est possible de changer l'échelle temporelle en « zoomant » sur les données, et des statistiques simples sont calculées sur les différentes données affichées et en fonction de l'échelle temporelle utilisée.

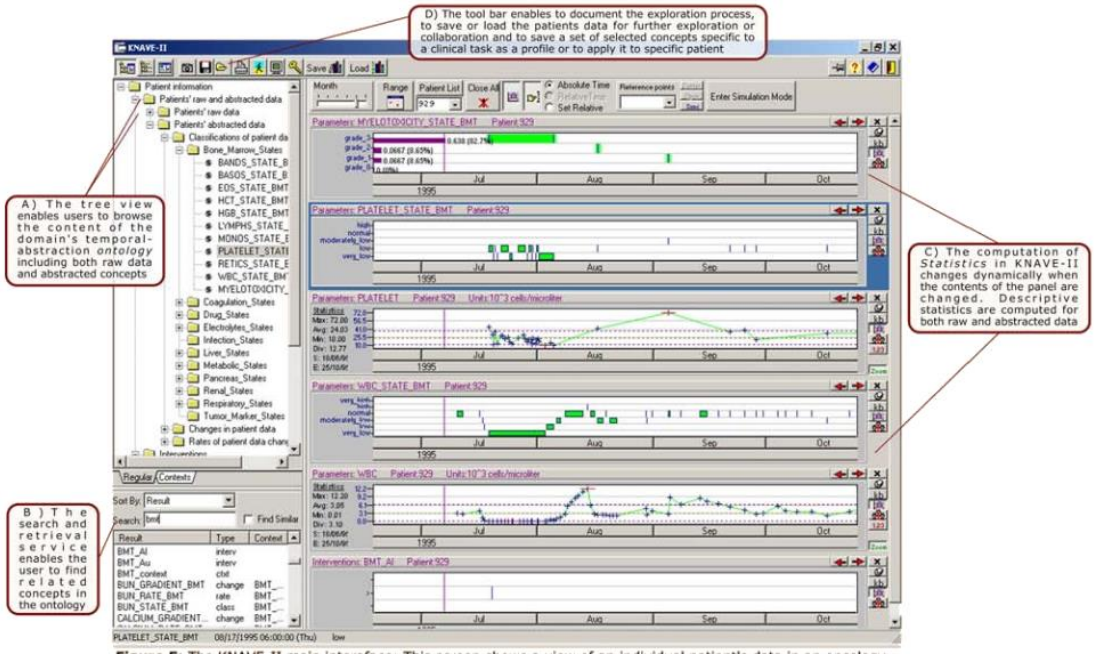

***Figure 9**. Interface du logiciel KNAVE-II (2004).*

Le logiciel *TimeLine* (31) (**Figure 10**) a été développé en 2007. Comme les précédents, il permet de naviguer dans la timeline à différents niveaux de zoom. Un arbre hiérarchique à gauche permet d'étendre et de replier les nœuds pour contrôler le détail de l'information affichée. La vue en ligne de vie peut incorporer des éléments multimédias. Un panneau en haut à droite affiche des informations détaillées sur les évènements sélectionnés dans la timeline, comme par exemple des tableaux, des courriers ou des images.

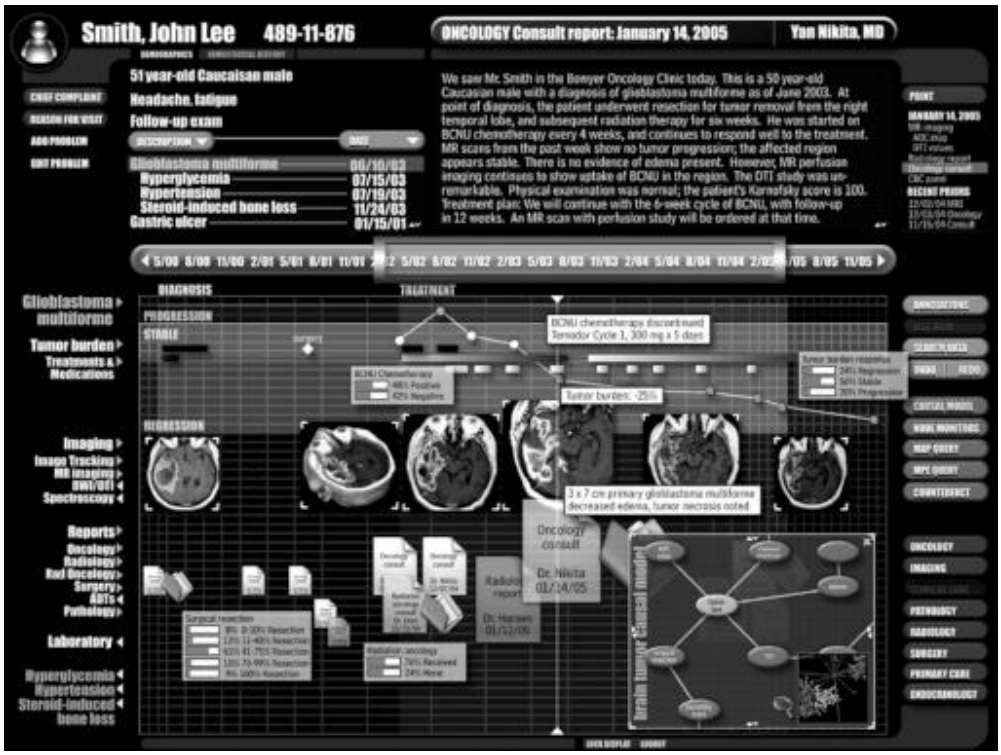

***Figure 10**. Interface du logiciel TimeLine (2007).*

Plus récemment, le logiciel HARVEST (32) a été développé en 2015. Il fournit une timeline simple qui représente les évènements par des traits verts. Un panneau « nuage de mots » résume les informations importantes extraites automatiquement à partir des courriers du patient, qui sont listés et visualisables en intégralité dans un panneau dédié. Lorsqu'un concept / problème est sélectionné (dyspnée sur la capture) la timeline affiche en violet toutes les visites pour lesquels le courrier mentionne le problème.

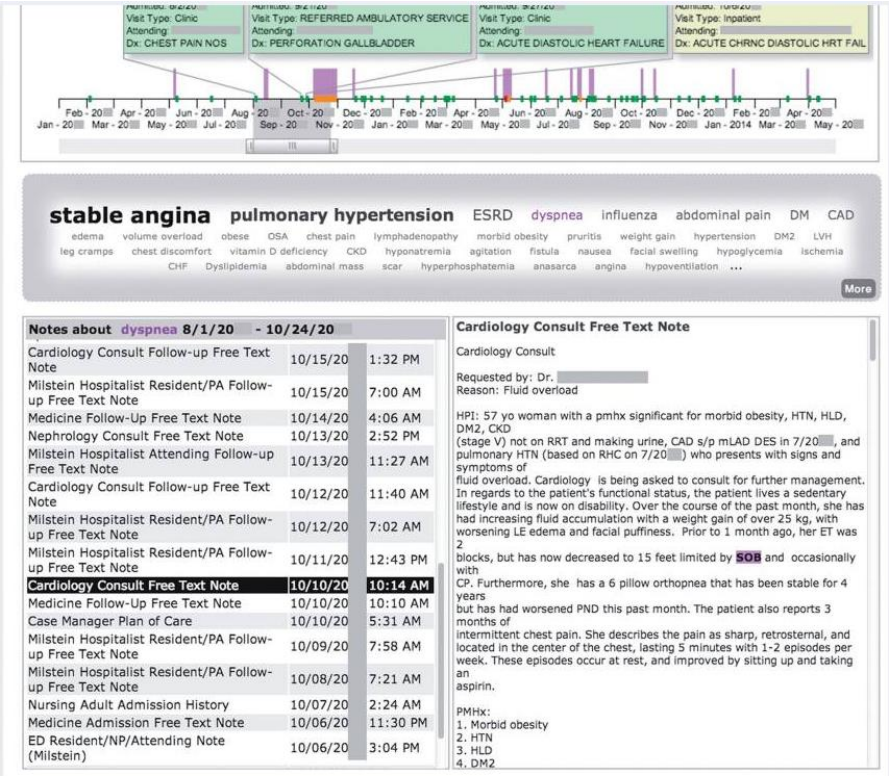

***Figure 11**. Interface du logiciel HARVEST (2015).*

D'autres prototypes (33–35) ont été développés pendant cette période, mais il y a eu peu de nouvelles propositions en visualisation transactionnelle au cours des dernières années. Les systèmes d'informations hospitaliers incorporent quelques-unes de ces idées dans leurs interfaces.

4.3. Systèmes d'information hospitaliers

Pendant longtemps et encore aujourd'hui, la plupart des logiciels médicaux qui constituent les SIH se sont efforcés de reproduire numériquement le fonctionnement « papier » qu'ils visent à remplacer. Les informations saisies dans ces logiciels peuvent être libres ou structurées. La saisie d'information libre (non structurée) est utilisée notamment pour l'observation médicale ou le courrier médical. Les données sont saisies dans des champs de texte, qui sont en fait le substitut numérique direct de la feuille de papier. L'intérêt de l'informatisation dans ce contexte est de faciliter la production, la consultation et le partage des données d'un même patient au sein d'un ou de plusieurs établissements.

La saisie de données dites structurées est réalisée à l'aide de formulaires qui sont composés de différents types de champs : texte court, cases à cocher, radiobox, listes déroulantes, etc. Ces formulaires permettent d'enregistrer des informations qui sont exhaustives (notamment lorsque les champs sont obligatoires) et directement exploitables par des logiciels. C'est par exemple souvent le cas des prescriptions informatisées.

Cependant, ces formulaires sont contraignants pour le clinicien. Ils sont généralement plus lents à remplir que le texte libre (interfaces dites « clic-bouton »), ils obligent la saisie d'informations parfois sans intérêt pour la prise en charge, et ils limitent la liberté du clinicien dans les cas atypiques, qui sont en réalité routiniers dans le domaine de la santé. La **Figure 12** illustre un formulaire structuré créé dans le logiciel DxCare® de Medasys.

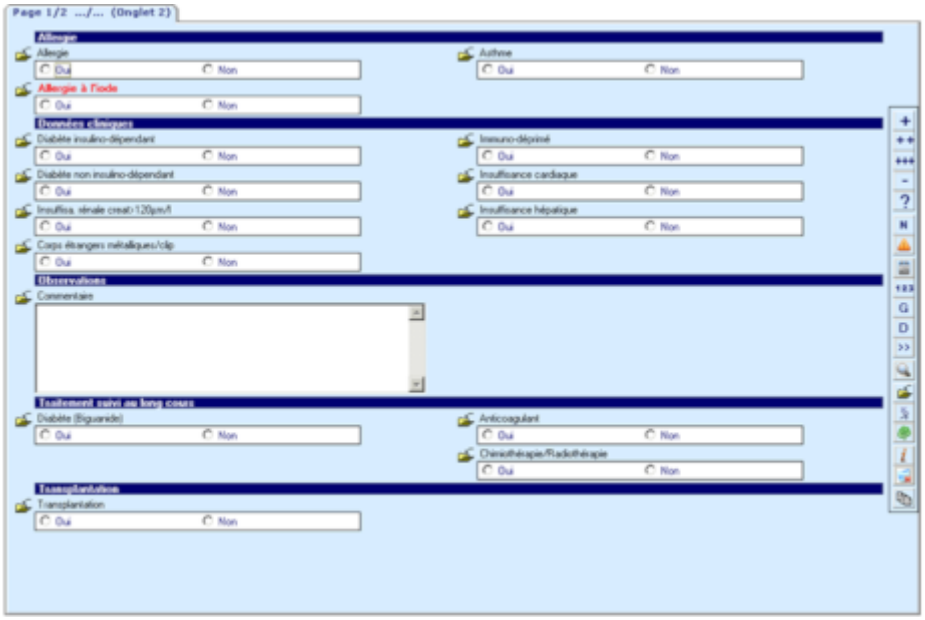

***Figure 12**. Formulaire de saisi personnalisé dans le logiciel DxCare®.*

Dans ces systèmes, une partie des informations existe de manière redondante sous forme structurée et non structurée. Avec l'imagerie par exemple, l'information existe sous forme de donnée brute (image) et sous une forme interprétée dans le compte-rendu du radiologue. La synthèse des séjours hospitaliers est rédigée dans des courriers de sortie (non structurés), et saisie au moyen de codes CIM-10 et CCAM dans le ou les RUM du patient. Ces données peuvent être en désaccord.

Par ailleurs les données sont souvent accessibles par des modules différents, qui ne permettent pas de facilement relier les données cliniques pertinentes pour un problème donné. Il est souvent difficile ou impossible de superposer l'évolution des mesures biologiques, des constantes hémodynamiques et les administrations de médicaments au sein d'une même ligne de vie.

Le logiciel *DxCare®* intègre depuis 2015 une timeline (**Figure 13**) qui permet de visualiser les éléments du dossier patient sur de longues périodes, et peut mettre en exergue des données médicales pertinentes dans un ou plusieurs contextes pathologiques particulier.

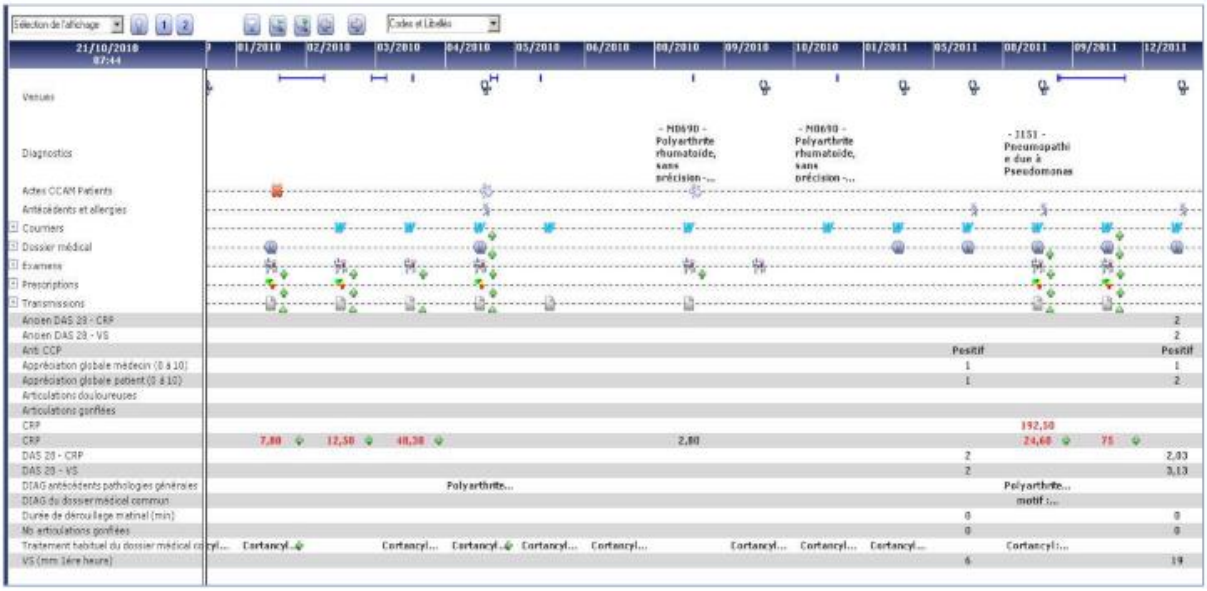

***Figure 13**. Pancarte de synthèse médicale affichée dans le logiciel DxCare®.*

Cet outil permet notamment d'afficher les séjours et venues, les diagnostics, les actes CCAM, les courriers, questionnaires, résultats d'examens ainsi que d'autres données d'intérêt saisies dans un tableau de données. L'échelle de temps est ajustable de l'année à la journée. Des vues associant des filtres et des tableaux de données pertinentes peuvent être paramétrées pour répondre aux besoins de cas d'usages cliniques spécifiques. En revanche, l'organisation des informations et notamment leur hiérarchisation est assez peu flexible dans cet outil.

5. Visualisation décisionnelle

5.1. Utilisations en santé

L'analyse décisionnelle peut être utilisée à des fins de pilotage ou de recherche.

La visualisation décisionnelle ou peut-être plutôt « en groupe » permet de représenter les résultats de requêtes sur des bases de données médicales :

- Pour mettre au point et vérifier rapidement les requêtes de sélection de patients, de séjours ou bien de toute autre unité
- Pour vérifier la compatibilité d'une sélection de patients avec des critères de sélection d'une étude (pre-screening), et pourquoi pas affiner ces critères

Elle peut aussi être utilisée pour détecter des patterns et des tendances générales, comme par exemple des parcours médicaux types. Dans ce cas d'utilisation, il faut remarquer tout d'abord que la visualisation brute d'un grand nombre de données très hétérogènes afin de visualiser des patterns tend à produire des images très complexes et totalement inexploitables. Cette hétérogénéité peut porter par exemple sur la variabilité des parcours pour une même prise en charge, sur l'ordonnancement des évènements ou sur les temps qui les séparent, etc.

Dans ce contexte, le travail sur la représentation visuelle individuelle n'est que rarement suffisant et il faut y adjoindre d'autres outils pour homogénéiser les données afin de pouvoir les exploiter :

- Techniques de visualisation : Le filtrage des évènements permet de cacher toutes les données sans intérêt lorsqu'on s'intéresse à un problème précis (par exemple on peut omettre les données clinico-biologiques lorsqu'on s'intéresse au parcours global du patient), et l'alignement des entités permet de s'intéresser à l'avant-après autour d'un évènement précis (par exemple l'administration d'un médicament)
- Traitements statistiques : Différentes techniques statistiques peuvent permettre de simplifier les données afin d'identifier semi-automatiquement des *patterns* récurrents dans les données. Ceci nécessite de faire abstraction de la variabilité inter-individuelle

(variabilité temporelle, d'ordre, etc.) qui est toujours très importante dans le domaine médical

5.2. Outils génériques existants

Le logiciel *OutFlow* (36) (**Figure 14**) développé par IBM en 2012 permet de créer des diagrammes de Sankey qui peuvent être utilisés pour la visualisation décisionnelle. Quelques réglages simples sont proposés pour simplifier / homogénéiser les parcours suivis par les différents sujets. Le logiciel identifie différents parcours, qui sont séparés en fonction de l'outcome (défini comme favorable ou non) et des différents états des individus au cours du temps (représentés par des rectangles colorés). Des barres horizontales sont utilisées à chaque étape ou nœud du parcours pour représenter l'effectif. Les transitions entre états sont représentés par les bandes colorées, dont la hauteur est proportionnelle au nombre de transitions. Un réglage de « seuil de simplification » permet d'influer sur l'algorithme de simplification des parcours, et de choisir la balance entre les « grands patterns » et les singularités individuelles. On parle de diagramme de Sankey pour ce type de visualisation décisionnelle. Ces diagrammes peuvent être intéressants pour représenter des flux, et ils permettent plus généralement de suivre une variable au cours du temps.

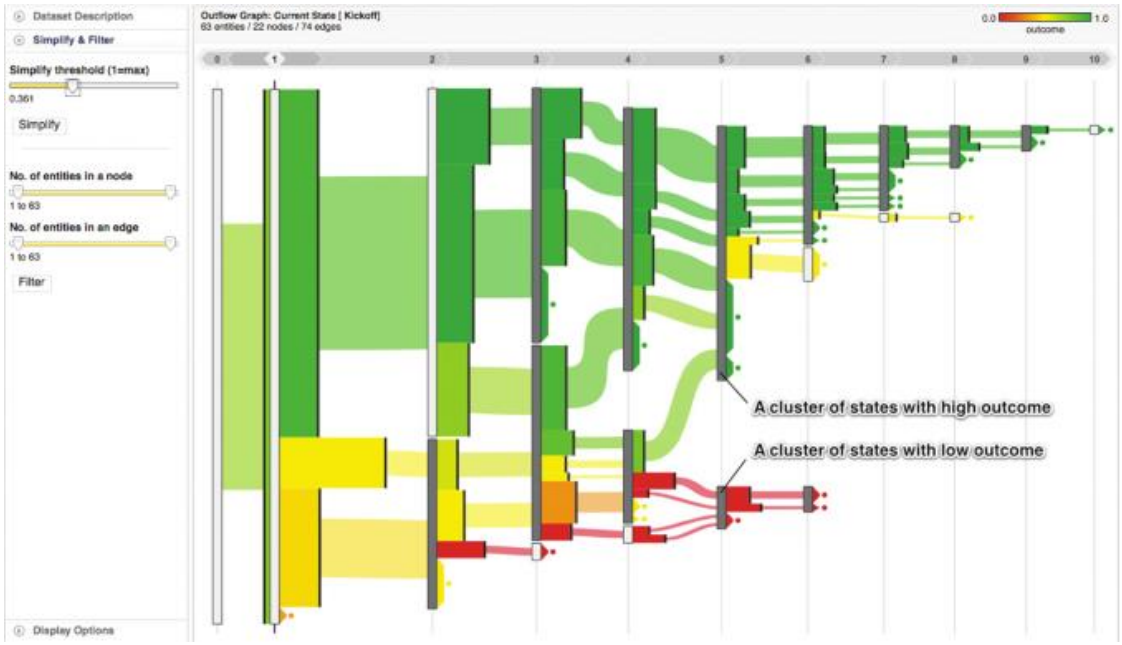

***Figure 14**. Interface du logiciel OutFlow (2012).*

Ce type de diagramme peut être intéressant pour représenter l'évolution d'un paramètre particulier au cours du temps. À titre d'exemple, les diagrammes de Sankey sont généralement utilisés pour des flux d'énergie, contexte dans lequel ils fonctionnent bien puisqu'on s'intéresse au parcours d'une seule donnée fongible (énergie). Ils sont beaucoup plus difficiles à exploiter lorsqu'on s'intéresse à des données hétérogènes, de qualité variable qui évoluent au cours du temps.

Le logiciel *Care Pathway Explorer* (37) (**Figure 15**) a été développé en 2015. Il utilise un algorithme « SPAM » pour faire du *pattern mining*, et les séquences les plus fréquentes identifiées dans les données sont affichées dans un tableau et représentées sous la forme d'un diagramme de Sankey. Les flux sont colorés en fonction de leur corrélation avec l'outcome du patient.

Sequence	Outcome
Vital → Diagnosis → Medication	positive
Lab → Diagnosis → Medication	positive
Lab → Diagnosis → Lab	negative
Diagnosis → Medication	neutral

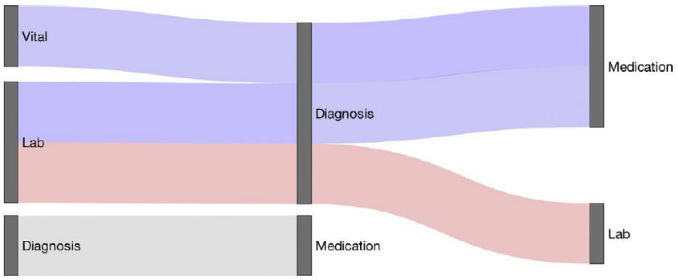

***Figure 15**. Interface du logiciel Care Pathway Explorer (2015).*

Le logiciel *EventFlow* (38) est un outil payant développé en 2013 par le HCIL (Human-Computer Interaction Lab) de l'université du Maryland, équipe qui avait déjà travaillé sur le logiciel LifeLines discuté précédemment. Il permet d'explorer les évènements temporels d'une cohorte de patients et de les simplifier de manière itérative pour mettre en évidence des séquences communes (**Figure 16**). La hauteur de chaque pattern affiché dans l'interface est proportionnelle à sa fréquence dans les données.

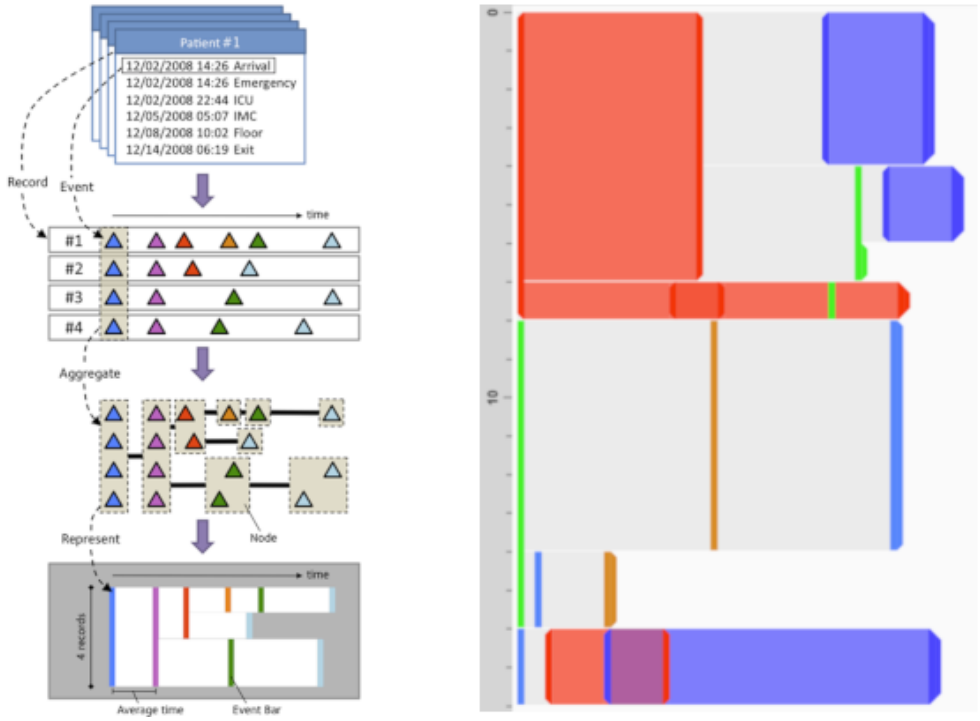

***Figure 16**. Algorithme de simplification des séquences temporelles utilisé dans EventFlow (2013).*

Cet outil fournit également un langage visuel de requêtage qui permet de préciser des séquences à rechercher et des séquences d'évènements à exclure. L'outil affiche une vue simplifiée de chaque entité correspondante ou exclue par la requête afin d'aider l'utilisateur à mettre au point sa requête (**Figure 17**). Les évènements de type « période » sont représentées par une barre horizontale qui marque le début et la fin (ici pour les médicaments), et les évènements ponctuels sont représentés par un triangle colorés.

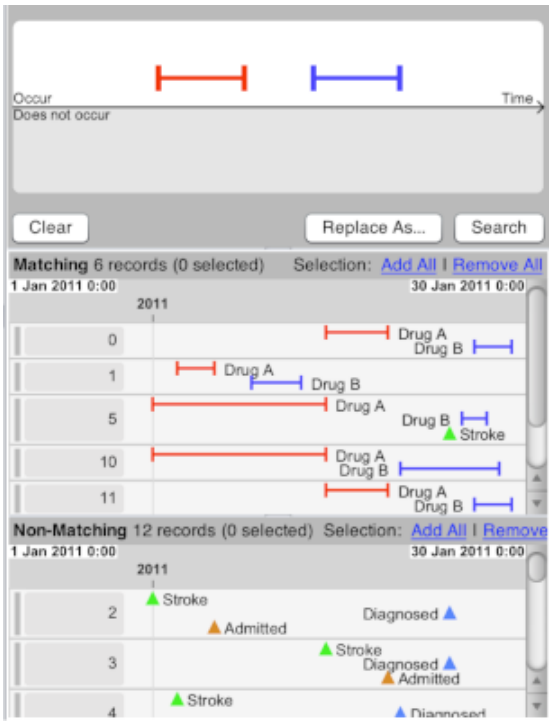

***Figure 17**. Visualisation des évènements individuels dans le logiciel EventFlow.*

Il existe un besoin encore peu comblé de créer un langage simple, traduisible sous forme visuelle et textuelle, qui soit adapté au requêtage de données temporelles. Les autres outils de requêtage existants, comme la langage SQL ou bien les scripts (R, etc.) sont difficiles à utiliser dans ce contexte et sont une source d'erreurs fréquente.

Il existe de nombreux autres prototypes d'outils dédiés à la visualisation et à l'analyse décisionnelle (39–46). Certains sont génériques et d'autres sont spécifiques de situations particulières.

6. Objectif de ce travail

L'objectif stratégique est de participer à l'amélioration de la conduite du soin et de certaines activités de recherche en santé, en rendant les interfaces numériques plus acceptables, plus agréables et plus performantes. Il s'agit de faciliter l'exploration dynamique de données du dossier patient et d'en faciliter la compréhension par le soignant ou le chercheur, en rapprochant les informations pertinentes dans le temps et dans l'espace visuel afin de favoriser le processus d'extraction de caractéristiques (*feature extraction*) (20).

Ce travail comporte 4 objectifs opérationnels qui sont dans l'ordre :

- Abstraction des types et représentations de données
- Spécification d'un prototype intégrant ces types de données
- Développement d'un prototype mettant en œuvre ces spécifications
- Evaluation sommaire du prototype développé

Abstract en Anglais

1. Introduction

Electronic health records (EHR) contain structured and unstructured data. This data has a temporal component which can be used to create timeline visualizations, as several prototypes and EHR software already do. However, medical data is very heterogeneous, and based on complex and terminologies which can be difficult to display without inducing information overload. It is essential to find new ways to visualize this data that can help users find relevant information quickly, with minimal cognitive overhead, and help them extract hidden features from raw data. This has applications for both usual care and biomedical research. Our goal was to devise simple visualization techniques for handling medical data in both contexts.

2. Methods

An abstraction layer for structured EHR data was devised after an informal literature review and discussions between authors. Several programming languages and toolkits were explored, and a prototype was developed. Two experts were recruited to evaluate it on simple cases.

3. Results

Temporal data was abstracted in three simple types: events, states and measures, with appropriate visual representations for each type. A prototype, called ”Heimdall” was developed. It can load and display complex heterogeneous structured temporal data in a straightforward way. The main view can display events, states and measures along a shared timeline. Users can summarize data using temporal, hierarchical compression and filtering techniques. Default and custom views can be used to work in problem-oriented ways. A simple evaluation with two experts has found conclusive results.

4. Conclusion

The “Heimdall” prototype provides a comprehensive and efficient graphical interface for EHR data visualization. It is open source, and the current version can be used with an R package (out of CRAN).

Article en Anglais

1. Introduction

Electronic health records data usually contain structured data, such as administrative data (patients' identity, patient flow, billing, and demographics), diagnoses encoded in ICD10 (47), procedures encoded in French CCAM (48), drug encoded in ATC (49), medical devices encoded in French LPP (13), laboratory results, and non-structured data, such as free-text medical records and medical imaging. All of them are characterized by their temporal aspects (e.g. the date of a measurement, the period of a diagnosis, or the validity period of a prescription), and relate to one patient. Structured data is often based on terminologies, and is encoded with multivalued qualitative variables and supported by a hierarchy of codes.

Data visualization is interesting within 2 contexts. "Transactional" visualization relates to visualization of one patient's data, in order to provide the patient with appropriate care. "Decisional" visualization relates the visualization of several patient's data, for management purposes (21) or research purposes (22), notably in the context of data reuse. Structured data can be directly used, sometime after the use of simple data management. Unstructured data, such as free-text records, is harder to visualize. Information can be extracted using "bag of words" (9) techniques (with negation identification), or deep learning for medical imaging (15–17).

In both contexts, the main objective of data visualization is to provide the user with a comprehensive view, and to enhance the mental process of feature extraction (20), by making it easy to detect temporal associations. This process may in some cases overpass data quality issues (23–25), and enable the construction of variables inferred from actual data (e.g. in case of INR elevation and red cells decrease, an expert can infer the patient encountered an hemorrhage). However, data complexity may overload induce fatigue and overload the user, leading to interpretation errors (26).

Many visual interfaces have been developed for transactional EHR data visualization (29), such as *LifeLines* (27), *MultiMedia Stream System (MMSS) TimeLine* (28), *KNAVE-II* (30), *TimeLine* (31), *HARVEST* (32) , or other prototypes (33–35). However, such representation methods are

rarely implemented in commercial hospital information systems. Some visual interfaces have also been developed for decisional EHR data visualization, such as *OutFlow* (36), *Care Pathway Explorer* (37), *EventFlow* (38), or other prototypes (39–46). Many of them are dedicated for specific tasks.

The objective of this work is to propose a visualization method that could either be used for transactional or decisional visualization of EHR data. For that purpose, we will propose an abstraction EHR data, and then develop and evaluate a prototype called Heimdall.

2. Method

2.1. EHR data abstraction

The first step of this study consisted in proposing an abstraction layer for EHR data. This step relied on an informal literature review (including the software above-mentioned), and a discussion between authors, taking their data expertise and their clinical experience into account.

2.2. Prototype development

The development language was chosen in order to get a fast prototype, with advanced user interaction, and that could be interfaced with R, the programming statistical language (50)

2.3. Prototype evaluation

In order to evaluate the tool, two medical experts (not involved in the development process) were asked to retrieved medical information over 24 cases of acute kidney injury (AKI, defined as an increase by at least 26.5 µmol/l in 48 hours, or an increase of more than 50% within 7 days, without any N18* diagnostic code). The 24 records were represented through 3 interfaces: the “classical interface” was designed to look like classical EHR software (with 1 page per type of information, containing tables without charts), the “manually optimized interface” proposed a single page with similar information but the information was first manually filtered by an expert, in order to present only interesting data, and the “Heimdall interface” was the interface developed in this project, without any prior data filtering. The experts were briefly taught how to use the interfaces.

Expert A used the classical interface to evaluate records #1-8, then the manually optimized interface for records #9-16, and then the Heimdall interface for records #17-24. Expert B used the same interfaces respectively for records #17-24, then #9-16 then #1-8. The experts were asked to answer 4 yes or no questions:

- Q1: Was a nephrotoxic treatment stopped by the physician, at day+0 or day+1 after AKI?
- Q2: Was there a urinalysis (sodium, potassium), at day+0 or day+1 after AKI?
- Q3: Was there a urinalysis (proteinuria), at day+0 or day+1 after AKI?
- Q4: Was urinary tract imaging performed at day+0, day+1 or day+2 after AKI?
- Q5: Was a N17* ICD10 code finally encoded as diagnosis?

The number of errors and the time required to answer the questions were compared using Student t tests.

3. Results

3.1. EHR data abstraction

We identified 3 types of information:

- Events, which occur at a single point in time, and have a label and a date (e.g. medical procedure, document)
- States, which are characterized by a label, a start date and a stop date (e.g. hospital stay, stay in a specific unit)
- Measures, which are characterized by a date, a label and a value (with or without normality range) (e.g. laboratory result, clinical parameter)

For each of kind of information, the label can be included into a hierarchy. The folding of a category from a hierarchy must result in the merging of the symbols corresponding to the pieces of data. Moreover, additional information can be provided, e.g. the content for a document, which is a sort of event, or the unit for a laboratory result, which is a sort of measure.

4. Prototype development

4.1.1. Technical aspects

Through the R package, Heimdall incorporates several ready-to-use functions, enabling data loading (from different formats) and terminology handling (e.g. loading codes, labels and hierarchy).

Heimdall is a read-only visualization tool, it does not enable to input data nor to edit existing data.

There around 2,500 lines of code for the prototype, in addition to 6,000 more in utility code that is shared with other software. Several libraries are used: OpenGL (51) provides low-level graphical primitives, dear ImGui (52) provides higher-level base graphical functions and widgets.

It is available for Microsoft Windows and GNU/Linux distributions, and will soon be usable in Apple macOS and web browsers through Emscripten (53).

4.1.2. From data types to symbol

Pieces of data relating to the same data type are grouped into a component (e.g. 3 values of kalemia are represented through a unique curve of kalemia). Each component is defined with a time axis ranging from the left to the right. Components are designed to be stacked vertically. In each component, the symbol used depend on the data nature.

Events are represented as small triangles. Combination of triangles (e.g. in case of terminology folding or temporal compression) result in a new triangle, with the number of initial symbols in its middle (left part of Figure 18).

States are represented as transparent rectangles. Combination of rectangles result in opacification of the background color (center part of Figure 18).

Measures are represented as curves. A first folding level transforms the curves into events, where abnormal values (i.e. outside the normality bounds) are represented in red, and normal values in blue. Then combination of symbols leads to the same treatment as for events (third

part of Figure 18). During the utilization, the user can choose the type of interpolation (54): LOCF (Last Observation Carrier Forward), linear or (cubic) spline (55).

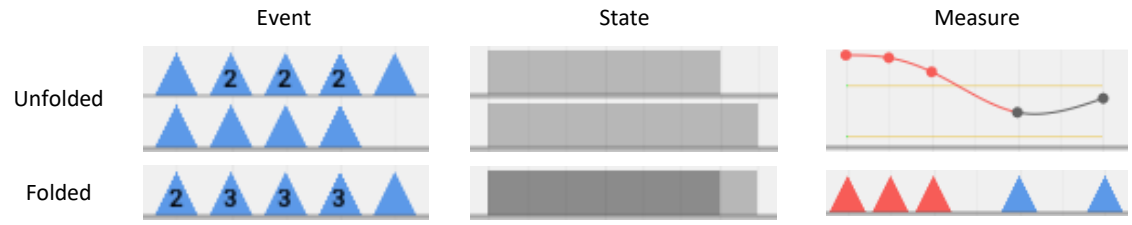

Figure 18. *Unfolded and folded representation of 3 type of components.*

4.1.3. Window organization

The main window principally looks like a stack of components, with a unique time axis, from the left to the right (56). The components are stacked vertically and organized according to their hierarchy. The first level of the hierarchy is the patient. The second level is linked to the data type (e.g. Diagnoses, laboratory results, etc.). Next levels are directly retrieved from the terminologies that are relevant (e.g. ICD10 (47), CCAM (48), ATC (49), LPP (13), etc.). This kind of organization enables Heimdall to serve both for transactional (one patient) and decisional purposes (several patients).

The representation of big amount of data is made easier by 3 mechanisms. First, the time axis can be compressed by pressing the Control key and using the mouse wheel. Temporal compression is handled by merging temporally close symbols (Figure 19). Second, the top-bottom axis can be compressed by clicking on specific arrows, which enable components folding. Vertical reduction is handled by merging symbols which are close with respect to their hierarchical position in the terminology. Third, a dropdown menu enables to fast select filtered view, corresponding to medical problems. For instance, the “renal” view enables to display only diagnoses, procedures, laboratory results and drugs in relation with the kidneys or the renal function. Some views are available by defaults, and the user can create custom views.

In case of multi-patient view, the time axis of the patients is aligned by default with real dates. However, it is possible to select a given common event (e.g. a surgical procedure) and automatically align all the patients according to this event.

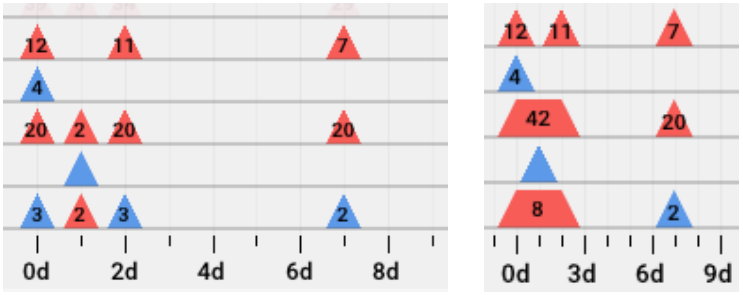

Figure 19. *Example of temporal compression of events. Individual events (left) are merged when zooming out (right).*

4.1.4. Overview of a case

The following images (Figure 20) illustrate a medical record containing medical procedures, biological measures, drug prescriptions and care flow data. The first part shows a partially unfolded display, with visible individual coagulation measures, the second part shows a fully folded record, and the third part shows a filtered view, in which only coagulation measures are visible.

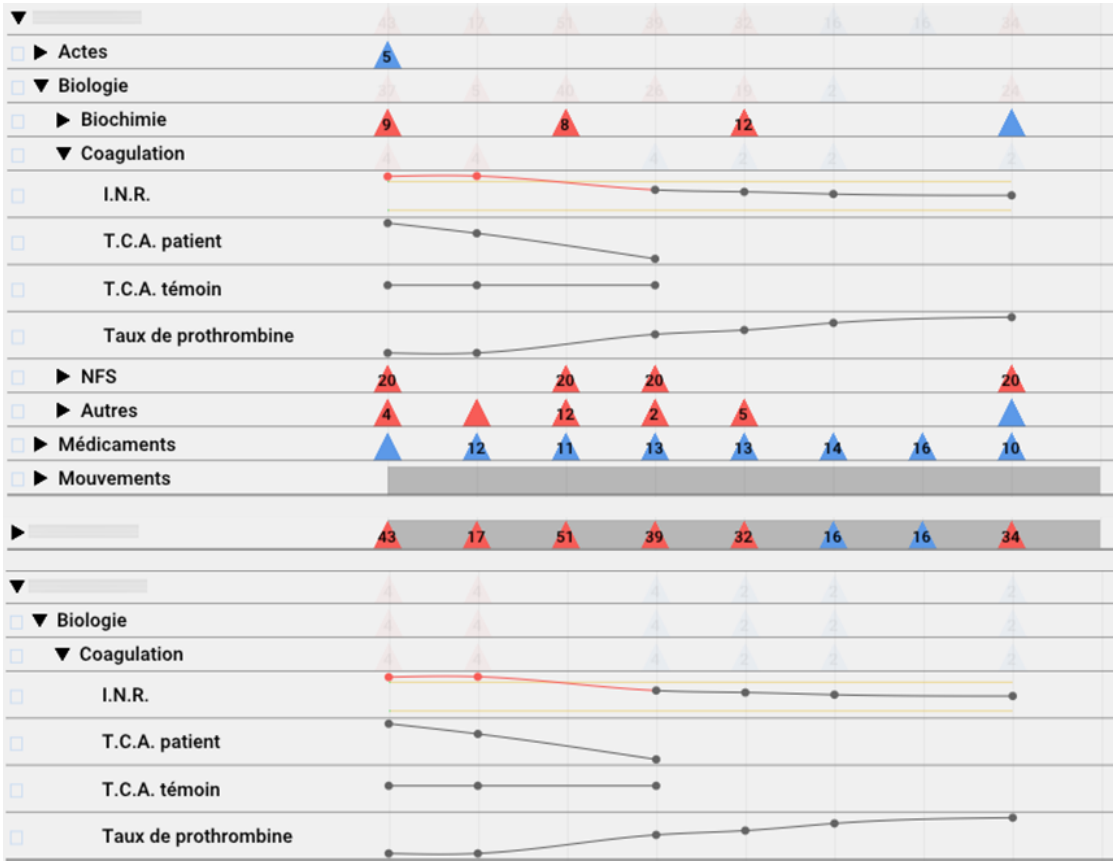

Figure 20. *Medical record with partially unfolded data (top part), fully folded view (middle part) and filtered view (bottom part)*

4.2. Prototype evaluation

The loading of data from 3,500 patients, including 360,000 pieces of data in Heimdall from R took about 1 second. It requires about 50 MB of memory (20 for numeric or temporal data, 15 for character data, and 15 for the application, including terminologies, graphic interface, polices, and graphical vertices).

The results of the evaluation are presented in **Table 2**. For both users, the manually optimized interface and the Heimdall interface were faster to use than the classical interface. The number of errors does not significantly differ.

***Table 2**. Results of the evaluation: time to answer the 5 questions, and number of errors. In each cell, the numbers (mean and standard deviation) relate to the evaluation of 8 clinical cases.*

	Classical interface	Manually optimized interface	Heimdall interface	p val.
Expert #1				
Time required (sec)	54.25 (5.42)	35.00 (2.62)	40.88 (5.22)	*< 0.001*
Number of errors	0.38 (0.74)	0.12 (0.35)	0.25 (0.46)	*0.662*
Expert #2				
Time required (sec)	57.75 (6.82)	42.25 (5.23)	45.38 (4.47)	*< 0.001*
Number of errors	0.38 (0.74)	0.12 (0.35)	0.12 (0.35)	*0.546*

5. Discussion

Medical data is very heterogeneous both in form and in nature, and most of it has a temporal component. We believe there is an opportunity for new visualization techniques that could take these specificities into account.

After reviewing the literature dedicated to medical temporal data visualization, we have a made a data visualization prototype called “Heimdall”, and we have tested it on simple cases.

5.1. Strengths and weaknesses

5.1.1. Strengths

Combining a foldable hierarchical view with a timeline is a powerful technique that allows easy yet understandable data summarization without losing track of the temporal information. Furthermore, the interface we have implemented allows the user to explore the data with several simple tools: alignment, filtering and custom views.

The software is relatively fast and reacts instantly to user commands even with thousands of entities, each containing several hundred pieces of data. The source code is open, which allows other tools to use provide the visualization tools implemented in Heimdall.

After a short introduction, the two experts that have tested it managed to navigate the data relatively easily despite its foreign nature compared to more traditional EHR software.

5.1.2. Weaknesses

Some data cannot be well visualized in the current iteration of our prototype. This includes non-temporal data (e.g. birthdate), and data with imprecise timing (e.g. ancient medical history). Non-structured data cannot be visualized very well in Heimdall, though the associated structed metadata can be used with some success (e.g. creation date, type of document).

The evaluation was done with a simple artificial case and was performed by two experts only. A more thorough evaluation remains to be done.

Heimdall is a prototype, and the source code needs to be improved before it can be reliably used and integrated in other software. Furthermore, unlike other software such as EventFlow (38), it only provides rudimentary tools for querying and filtering based temporal relations, which is a major limit for more complicated decisional applications.

5.2. Conclusion

Heimdall provides a comprehensive and efficient graphical interface for EHR data visualization. It is open source, and the current R package (alpha version) can be downloaded at https://koromix.dev/files/R.

Discussion en Français

Il nous semble nécessaire de créer de nouveaux outils de visualisation pour représenter de gros ensembles de données hétérogènes temporelles :

- Dans le contexte transactionnel, nous pensons que la capacité de rapprocher spatialement et temporellement des données médicales d'un patient peut faciliter la compréhension des cas complexes.
- Dans le contexte décisionnel, la création de règles pour simplifier les données (*feature extraction*).

Ceci pourrait faciliter la compréhension des caractéristiques temporelles de ces données, préalable indispensable à la création de règles pour permettre l'identification des *patterns* temporels. C'est un préalable indispensable à la simplification des données pour permettre des analyses statistiques.

Au terme d'une revue de la littérature et de différents outils utilisés en production, nous avons identifié les fonctionnalités fondamentales de visualisation de données médicales temporelles que nous désirions voir au sein d'un même logiciel. Nous avons développé un premier prototype baptisé *Heimdall* qui rassemble une partie de ces fonctionnalités, et nous l'avons évalué sommairement à l'aide de quelques scénarios de tests simples à l'aide d'un échantillon de données de séjours hospitaliers.

1. Forces et faiblesses

1.1. Forces

La conjonction d'une représentation temporelle des données et le support des terminologies hiérarchiques est un paradigme puissant lorsqu'il est utilisé avec des données structurées, ou bien encore avec les métadonnées de données non structurées (par exemple la date et le type d'un courrier médical). Heimdall permet à la fois d'avoir une vue synthétique de données temporelles, et d'explorer en détails ces données à l'aide de représentations facilement généralisables.

Il s'agit d'une application versatile et facile à utiliser, dans laquelle de nombreux types de données temporelles peuvent être chargés. Les quelques représentations fournies permettent de représenter la majorité des données rencontrées en médecine. Les visualisations et les outils de navigation disponibles permettent à la fois de créer des vues synthétiques des données, mais aussi d'en explorer rapidement les détails et les relations temporelles (concordance, succession dans le temps, etc.).

Par ailleurs, le code source est ouvert et relativement simple d'accès. Il peut donc être adapté pour supporter de nouveaux types de représentations, ou être embarqué directement dans des applications natives ou web.

1.2. Faiblesses

Certaines données ne bénéficient pas des représentations de Heimdall et/ou ne sont pas affichées de manière optimale. Il peut s'agir notamment :

- Des données non temporelles (par exemple la date de naissance)
- Des données mal localisées dans le temps (par exemple les antécédents médicaux anciens)

Les données non structurées comme les courriers médicaux ne peuvent pas non plus être bien exploitées (sauf les métadonnées qui sont généralement structurées). Des développements basés sur le TALN pourraient permettre de représentations les informations contenues dans ces courriers.

Nous avons fait le choix d'une grande généricité dans la modélisation et la représentation des données. Ceci nous permet de visualiser des types de données qui n'ont pas été prévus à l'avance pour être affichés dans l'application, à l'inverse de la majorité des logiciels utilisés dans le domaine médical. Cette orientation a ses inconvénients. Il est certain que pour des workflows bien identifiés et routiniers, cette approche reste inférieure à la création d'outils métiers dédiés avec une interface spécifique. Les résultats de l'évaluation sont en faveur de ce fait, puisque les dossiers évalués sur un « DPI optimisé » étaient évalués plus rapidement par le premier évaluateur. Cependant, l'interface du logiciel Heimdall est moins facile à

appréhender pour des utilisateurs déjà habitués à des présentations classiques de dossiers informatisés.

L'utilisabilité de l'outil a été évaluée sur un cas d'utilisation simple et relativement artificiel, et nous avons testé et développé l'application en nous appuyant sur des jeux de données que nous connaissions déjà largement. Pour le moment, aucune évaluation de l'utilisabilité de l'application par des utilisateurs finaux n'a été effectuée. Cette étape sera nécessaire pour déterminer l'intérêt réel de l'application et dans la perspective de la diffusion de Heimdall à plus grande échelle.

Heimdall fournit quelques outils d'aide à l'analyse décisionnelle, mais ils sont embryonnaires et des développements supplémentaires sont probablement nécessaires dans ce contexte.

2. Comparaison avec d'autres outils

Heimdall reprend les éléments fondementaux de la visualisation en timeline de données médicales inaugurée par le logiciel LifeLines (27) en 1996, tout en y ajoutant différentes évolutions :

- Interface modernisée, avec notamment deux versions : « claire » ou « sombre »
- Pluralité des représentations visuelles : évènements, périodes et mesures
- Hiérarchisation des données, avec des hiérarchies flexibles et ajustables par l'utilisateur
- Outils basiques de visualisation décisionnelle : Filtre et alignement
- Navigation simplifiée dans le temps grâce à une interface simple et performante

L'application peut être utilisée seule (chargement de fichiers CSV), depuis R ou bien intégrée à d'autres solutions comme par exemple des sites web. Il s'agit cependant d'un prototype, et des développements supplémentaires sont probablement nécessaires pour qu'il puisse être utilisé en routine.

Notre intérêt portait avant tout sur la visualisation des données et non pas sur le traitement automatisé des données. A ce titre, Heimdall ne dispose pas d'outils de détection de

séquences disponibles dans d'autres outils, par exemple dans Care Pathway Explorer (37) ou EventFlow (38).

3. Perspectives

3.1. Réutilisation des données

L'informatisation des systèmes d'information et l'évolution des législations permet de plus en plus d'accéder à de grandes bases de données médicales. On assiste d'ailleurs depuis plusieurs années à une augmentation progressive de la recherche médicale basée sur la réutilisation de données (57).

L'application développée facilite l'exploration des données temporelles présentes au sein de ces bases avec un minimum de manipulations ou d'expertise technique. Elle fournit la plupart des terminologies de base pour permettre de hiérarchiser et simplifier rapidement l'information. Elle fournit également les outils nécessaires pour importer ou créer de nouvelles terminologies. Le paquet R qui accompagne l'application facilite l'intégration de données plus complexes, et permet d'automatiser l'utilisation de l'outil.

Ces grandes bases de données sont caractérisées par des données de bas niveau, d'exhaustivité et de qualité très variable. La réalisation d'analyses statistiques sur ces bases nécessite de construire des règles de nettoyage et de transformation des données. Avant d'effectuer cette gestion des données, il est nécessaire de les comprendre. Les outils disponibles dans Heimdall peuvent aider à comprendre les relations (notamment temporelles) des données contenues dans ces bases.

Ce type d'outil est particulièrement intéressant et d'actualité pour l'analyse des parcours de soin et l'exploitation des entrepôts de données de santé (EDS) qui sont mis en place dans différents centres hospitaliers (58).

3.2. Développements ultérieurs

Ce prototype montre la possibilité de conjuguer des données hétérogènes temporelles au sein d'une même interface à l'aide d'outils simples regroupés au sein d'une interface intuitive :

timeline, représentations visuelles simples, simplification hiérarchique des terminologies. Pour aller plus loin, le logiciel devra être évalué en situation réelle :

- Contexte transactionnel : probablement en partenariat avec un laboratoire d'évaluation comme celui du CIC-IT (Evalab).
- Contexte décisionnel : diffusion de l'outil aux chercheurs et mise en place comme outil (parmi d'autres) de visualisation des données d'un entrepôt de santé.

Heimdall est un logiciel relativement rapide mais il reste assez peu optimisé. Actuellement il peut afficher quelques dizaines de milliers d'unités (patients, séjours, etc.) et quelques millions de mesures tout en restant fluide, mais nous pensons qu'il est possible de faire beaucoup mieux. Avec quelques optimisations algorithmiques, il est probablement possible d'afficher un million d'entités sans difficulté particulière.

Il sera nécessaire de développer et d'implémenter des outils d'abstraction temporelle qui permettront de détecter et de combiner des séquences de données hétérogènes (dans le temps et/ou l'espace) en informations de haut niveau (46,59,60) pour permettre d'utiliser Heimdall à des fins d'analyse décisionnelle.

Enfin, il s'agit pour le moment uniquement d'une interface de visualisation des données. Il pourrait être intéressant, notamment dans une perspective transactionnelle, de permettre la modification des données directement au sein de l'application.

Conclusion

L'informatisation des établissements de santé aboutit à la production de grandes bases de données clinico-administratives. La mise en évidence d'associations et de relations causales cachées dans ces données temporelles pourrait aider à améliorer la prise en charge des patients et permettre la découverte de nouvelles pistes de recherche clinique. Cependant, ces données sont hétérogènes et proviennent de systèmes isolés qui ne permettent pas de les explorer facilement. L'objectif de ce travail était de construire un outil de visualisation unifiée de ces données.

Une revue de la littérature dédiée à la visualisation de données cliniques et plus généralement biographiques nous a permis d'imaginer des représentations qui prennent en compte la temporalité, la hiérarchie et la nature des données médicales. Nous avons développé un prototype de visualisation interactive qui peut être utilisé seul, ou intégré à d'autres logiciels à l'aide d'une librairie C++ ou d'un package R.

L'application développée (« Heimdall ») est capable de représenter un grand nombre de données temporelles hétérogènes au sein d'une même interface. Appliquée à un dossier clinique, la vue principale résume l'histoire des patients tout en offrant un accès direct aux informations sous-jacentes. Différentes techniques de visualisation et d'interaction facilitent l'exploration des données à différents niveaux de détails et d'abstraction, de l'aperçu global à l'analyse fine des données. Les utilisateurs peuvent utiliser des vues personnalisées pour hiérarchiser les informations en fonction des problèmes cliniques à explorer.

Une évaluation sommaire de l'application a été réalisée, mais l'utilisation de cet outil à des fins de recherche ou de soins n'a pas encore été évaluée. Le développement d'options supplémentaires permettra de faciliter la visualisation simultanée des données provenant de plusieurs patients. L'exploration de données non structurées comme les courriers n'a pas été traitée dans le cadre de ce travail.

Liste des tables

Liste des figures

Références

1. Martignene N, Balcaen T, Bouzille G, Calafiore M, Beuscart J-B, Lamer A, et al. Heimdall, a computer program for electronic health records data visualization. MIE 2020 [IN PRESS]. 2020;

2. Martignene N, Balcaen T, Bouzillé G, Calafiore M, Legrand B, Chazard E. Heimdall, logiciel de visualisation des données temporelles des dossiers patients électroniques. Rev DÉpidémiologie Santé Publique. 1 mars 2020;68:S26.

3. Zahabi M, Kaber DB, Swangnetr M. Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation. Hum Factors J Hum Factors Ergon Soc. août 2015;57(5):805-34.

4. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. J Am Med Inform Assoc JAMIA. mars 2015;22(2):330-9.

5. CIM-10 FR 2015 à usage PMSI | Publication ATIH [Internet]. [cité 12 déc 2018]. Disponible sur: https://www.atih.sante.fr/cim-10-fr-2015-usage-pmsi

6. CCAM en ligne - CCAM [Internet]. [cité 12 déc 2018]. Disponible sur: https://www.ameli.fr/accueil-de-la-ccam/index.php

7. Référentiel Général d'Interopérabilité RGI Version 1.0 [Internet]. [cité 13 déc 2018]. Disponible sur: https://references.modernisation.gouv.fr/sites/default/files/RGI_Version1%200.pdf

8. Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: Construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. AMIA Annu Symp Proc. 2011;2011:392-401.

9. Ordonez P, Armstrong T, Oates T, Fackler J. Using Modified Multivariate Bag-of-Words Models to Classify Physiological Data. In: 2011 IEEE 11th International Conference on Data Mining Workshops. 2011. p. 534-9.

10. Bates DW. Using information technology to reduce rates of medication errors in hospitals. BMJ. 18 mars 2000;320(7237):788-91.

11. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface TerminologiesFacilitating Direct Entry of Clinical Data into Electronic Health Record Systems. J Am Med Inform Assoc. 1 mai 2006;13(3):277-88.

12. Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: a survey of internal medicine residents and program directors. Arch Intern Med. 22 févr 2010;170(4):377-80.

13. Liste des produits et prestations - LPP [Internet]. [cité 13 déc 2018]. Disponible sur: https://www.ameli.fr/etablissement-de-sante/exercice-professionnel/nomenclatures-codage/lpp

14. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer. 1 janv 2009;45(2):228-47.

15. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 1 août 2018;29(8):1836-42.

16. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 13 déc 2016;316(22):2402-10.

17. Zhou SK, Greenspan H, Shen D. Deep Learning for Medical Image Analysis. Academic Press; 2017. 460 p.

18. EMOIS Nancy 2011 - Taux de mortalité hospitalier : les données du PMSI sont-elles utilisables ? [Internet]. [cité 13 déc 2018]. Disponible sur: https://www.canal-u.tv/video/canal_u_medecine/emois_nancy_2011_taux_de_mortalite_hospitalier_les_donnees_du_pmsi_sont_elles_utilisables.6835

19. Analyse de l'activité hospitalière 2016 | Publication ATIH [Internet]. [cité 12 déc 2018]. Disponible sur: https://www.atih.sante.fr/analyse-de-l-activite-hospitaliere-2016

20. Emmanuel CHAZARD et al. Secondary use of healthcare structured data: the challenge of domain-knowledge based extraction of features-.

21. Chazard E, Beuscart R. Graphical representation of the comprehensive patient flow through the Hospital. AMIA Annu Symp Proc. 2007;2007:110-4.

22. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int [Internet]. 2015 [cité 12 déc 2018];2015. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4468280/

23. 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.

24. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinforma. 1 mars 2010;2010:1-5.

25. de Lusignan S, Pearce C, Shaw N, Liaw S-T, Michalakidis G, T Vicente M, et al. What are the barriers to conducting international research using routinely collected primary care data? Stud Health Technol Inform. 1 janv 2011;165:135-40.

26. Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Sittig DF. Information Overload and Missed Test Results in Electronic Health Record–Based Settings. JAMA Intern Med. 22 avr 2013;173(8):702-4.

27. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records. In: Bederson BB, Shneiderman B, éditeurs. The Craft of Information Visualization [Internet]. San Francisco: Morgan Kaufmann; 2003 [cité 12 nov 2018]. p. 308-12. (Interactive Technologies). Disponible sur: http://www.sciencedirect.com/science/article/pii/B978155860915050038X

28. Aoyama DA, Hsiao J-TT, Cárdenas AF, Pon RK. TimeLine and visualization of multiple-data sets and the visualization querying challenge. J Vis Lang Comput. 1 févr 2007;18(1):1-21.

29. Bade R, Schlechtweg S, Miksch S. Connecting Time-oriented Data and Information to a Coherent Interactive Visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 2004 [cité 12 nov 2018]. p. 105–112. (CHI '04). Disponible sur: http://doi.acm.org/10.1145/985692.985706

30. Goren-Bar D, Shahar Y, Galperin-Aizenberg M, Boaz D, Tahan G. KNAVE II: The Definition and Implementation of an Intelligent Tool for Visualization and Exploration of Time-oriented Clinical Data. In: Proceedings of the Working Conference on Advanced Visual Interfaces [Internet]. New York, NY, USA: ACM; 2004 [cité 12 déc 2018]. p. 171–174. (AVI '04). Disponible sur: http://doi.acm.org/10.1145/989863.989889

31. Bui AAT, Aberle DR, Kangarloo H. TimeLine: Visualizing Integrated Patient Records. IEEE Trans Inf Technol Biomed. juill 2007;11(4):462-73.

32. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, et al. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc JAMIA. mars 2015;22(2):263-74.

33. Aigner W, Miksch S. CareVis: Integrated visualization of computerized protocols and temporal patient data. Artif Intell Med. 1 juill 2006;37(3):203-18.

34. FlowingMedia. TimeFlow: Timeline visualization application [Internet]. 2018 [cité 18 avr 2018]. Disponible sur: https://github.com/FlowingMedia/TimeFlow

35. Spring N, Plaisant C, Marchand G, Wang TD, Mukherjee V, Roseman D, et al. Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison. IEEE Trans Vis Comput Graph. 2009;15(6):1049-56.

36. Wongsuphasawat K, Gotz D. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. IEEE Trans Vis Comput Graph. déc 2012;18(12):2659-68.

37. Perer A, Wang F, Hu J. Mining and exploring care pathways from electronic medical records with visual analytics. J Biomed Inform. 1 août 2015;56:369-78.

38. Monroe M, Lan R, Lee H, Plaisant C, Shneiderman B. Temporal Event Sequence Simplification. IEEE Trans Vis Comput Graph. déc 2013;19(12):2227-36.

39. A JL, B JL, B DN, A EK. VizTree: a Tool for Visually Mining and Monitoring Massive Time Series Databases.

40. Soulakis ND, Carson MB, Lee YJ, Schneider DH, Skeehan CT, Scholtens DM. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. J Am Med Inform Assoc JAMIA. mars 2015;22(2):299-311.

41. Guerra-Gómez J, Pack ML, Plaisant C, Shneiderman B. Visualizing Change over Time Using Dynamic Hierarchies: TreeVersity2 and the StemView. IEEE Trans Vis Comput Graph. déc 2013;19(12):2566-75.

42. Plaisant Catherine. Visualization of temporal patterns in patient record data. Fundam Clin Pharmacol. 17 oct 2017;32(1):85-7.

43. Bellazzi R, Larizza C, Magni P, Bellazzi R. Temporal data mining for the quality assessment of hemodialysis services. Artif Intell Med. 1 mai 2005;34(1):25-39.

44. Wongsuphasawat K, Guerra Gómez JA, Plaisant C, Wang TD, Taieb-Maimon M, Shneiderman B. LifeFlow: visualizing an overview of event sequences. In ACM Press; 2011 [cité 19 avr 2018]. p. 1747. Disponible sur: http://dl.acm.org/citation.cfm?doid=1978942.1979196

45. Zhao J, Papapetrou P, Asker L, Boström H. Learning from heterogeneous temporal data in electronic health records. J Biomed Inform. 1 janv 2017;65:105-19.

46. Vandromme M, Jacques J, Taillard J, Hansske A, Jourdan L, Dhaenens C. Extraction and optimization of classification rules for temporal sequences: Application to hospital data. Knowl-Based Syst. 15 avr 2017;122:148-58.

47. World Health Organization. International Statistical Classification of Diseases and Related Health Problems [Internet]. 2018 [cité 12 août 2019]. Disponible sur: http://www.who.int/classifications/icd/en/

48. Social Security. French comon classification of medical procedures - CCAM [Internet]. [cité 12 août 2019]. Disponible sur: https://www.ameli.fr/accueil-de-la-ccam/telechargement/index.php

49. World Health Organization. Anatomical Therapeutic Chemical classification [Internet]. 2019 [cité 12 août 2019]. Disponible sur: http://www.whocc.no

50. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2011 [cité 12 août 2019]. Disponible sur: http://www.R-project.org/

51. OpenGL - The Industry Standard for High Performance Graphics [Internet]. [cité 12 nov 2018]. Disponible sur: https://www.opengl.org/

52. omar. Dear ImGui: Bloat-free Immediate Mode Graphical User interface for C++ with minimal dependencies: ocornut/imgui [Internet]. 2018 [cité 12 nov 2018]. Disponible sur: https://github.com/ocornut/imgui

53. Main — Emscripten 1.38.16 documentation [Internet]. [cité 10 déc 2018]. Disponible sur: http://kripken.github.io/emscripten-site/

54. Moritz S, Bartz-Beielstein T. imputeTS: time series missing value imputation in R. R J. 2017;9(1):207–218.

55. Cubic spline interpolation - tools.timodenk.com [Internet]. [cité 12 déc 2018]. Disponible sur: https://tools.timodenk.com/?p=cubic-spline-interpolation

56. van Bemmel, Musen. Handbook of Medical Informatics [Internet]. 1997 [cité 12 déc 2018]. Disponible sur: http://www2.hawaii.edu/~nreed/ics691BMI2/discpapers/handbookMICh7.pdf

57. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 1 janv 2018;77:34-49.

58. Lille s'engage pour une recherche d'excellence sur son territoire [Internet]. [cité 13 déc 2018]. Disponible sur: https://www.reseau-chu.org/article/lille-sengage-pour-une-recherche-dexcellence-sur-son-territoire/

59. Du F, Shneiderman B, Plaisant C, Malik S, Perer A. Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. 2015;14.

60. Morchen F. A better tool than Allen's relations for expressing temporal knowledge in interval data. 2006.

61. Simple DirectMedia Layer - Homepage [Internet]. [cité 10 déc 2018]. Disponible sur: https://www.libsdl.org/

Annexe 1 : Aspects techniques

1. Architecture et plateformes

Le logiciel est constitué d'environ 3000 lignes de code spécifiques aux logiciels, auxquelles on peut additionner près de 6000 lignes de code utilitaires développées en commun avec d'autres logiciels.

L'interface graphique est générée à l'aide de l'API OpenGL (51), qui est une interface de programmation bas niveau de rendu visuel 2D et 3D. La libraire ImGui (52) fournit des fonctions et des widgets de base pour faciliter le développement de l'interface graphique. La librairie SDL (61) (Software Development Layer) permet d'abstraire l'accès aux principales API d'entrée/sortie pour deux systèmes d'exploitation : macOS et Linux. La **Figure 21** illustre l'architecture générale et l'organisation des briques utilisées par Heimdall.

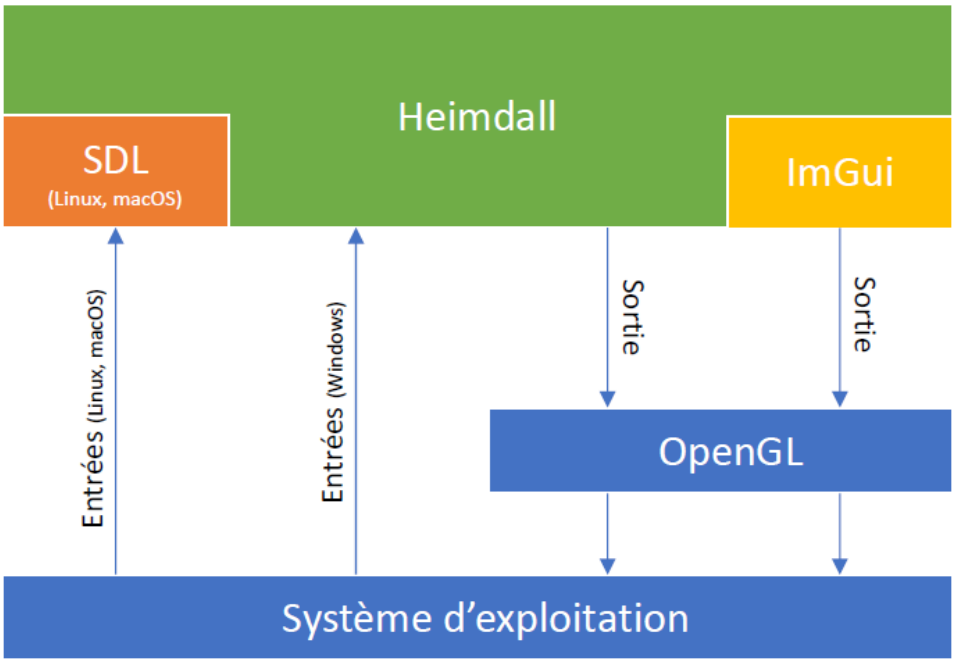

***Figure 21**. Architecture générale du logiciel Heimdall.*

L'application est disponible sur les trois systèmes d'exploitation principaux : Microsoft Windows (Vista et plus), Apple macOS et les distributions GNU/Linux.

Elle est également disponible pour le web à l'aide du SDK Emscripten (53), qui fournit un environnement de compilation vers du Javascript (ou du WebAssembly pour les navigateurs les plus récents). Quelques adaptations du code de Heimdall ont été nécessaires, notamment au niveau des shaders OpenGL pour qu'ils puissent fonctionner avec l'API WebGL version 2.

2. Boucle principale

Heimdall est une application graphique interactive. Contrairement aux scripts d'analyse, les applications graphiques doivent tourner en continu. Pour cela le cœur de Heimdall repose sur une boucle principale simple (**Figure 22**) qui comprend trois étapes à chaque itération : lecture des entrées de l'utilisateur, modification et calcul des données, rendu visuel de toute l'interface. Cette architecture est très proche de celle observée dans les jeux vidéo (on parle de *game loop*).

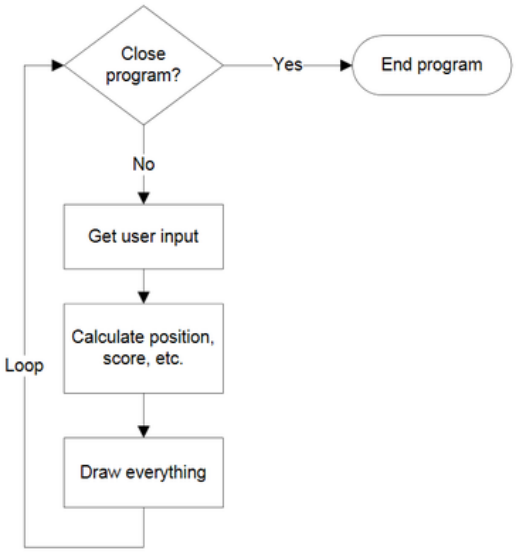

***Figure 22**. Boucle principale de Heimdall.*

Contrairement à la plupart des jeux vidéo (qui doivent gérer des animations en continu, des entités intelligentes ou scriptées, etc.), en l'absence d'entrées Heimdall n'a rien à faire. Dans ce cas et pour économiser des ressources, Heimdall met en pause sa boucle principale et continue à afficher la dernière version de l'interface qui a été dessinée.

3. Tamisage hiérarchique des données

Les données sont représentées par un type somme qui contient toutes les informations nécessaires à chaque représentation visuelle. Les données d'une entité sont stockées en mémoire au sein d'un tableau linéaire homogène trié par le temps.

Pour chaque entité, la transformation des données d'un tableau unique vers un tableau par ligne d'affichage est réalisée à chaque rendu de l'interface à l'aide de l'algorithme suivant :

- Tout d'abord, les données sont parcourues dans l'ordre. Le libellé de la donnée est recherché dans les terminologies activées afin d'en obtenir le chemin (ex : le libellé « Na+ » est relié à un chemin « /Biologie/Biochimie/Na+ »). Si ce libellé n'est pas trouvé, la donnée n'est pas affichée. Sinon, une ligne est créée ou réutilisée pour chaque niveau du chemin. Avec l'exemple du « Na+ », nous aurions donc les lignes suivantes : « / », « /Biologie », « /Biologie/Biochimie », « /Biologie/Biochimie/Na+ ». Chaque ligne correspond à un tableau dans lequel une référence vers la donnée Na+ est stockée. Si une nouvelle donnée (par exemple « /Biologie/Biochimie/K+ ») est présente dans l'entité, les lignes en commun seront réutilisées (ici il y en a 3 : « / », « /Biologie » et « /Biologie/Biochimie ») et une nouvelle ligne « /Biologie/Biochimie/K+ » sera créée. Lorsqu'un niveau est replié, l'algorithme s'interrompt et les lignes filles ne sont pas créées.
- Dans un second temps, toutes les lignes crées à l'étape précédente sont triées par l'ordre alphanumérique du chemin complet correspondant. Avec l'exemple précédent et en supposant que tous les chemins sont déployés il y aurait quatre lignes qui contiennent toutes cette donnée : « / », « /Biologie », « /Biologie/Biochimie », « /Biologie/Biochimie/Na+ ». Cette étape est nécessaire car les données de l'entité ne peuvent pas être prétriées en fonction des chemins complets, qui sont dynamiques car ils dépendent des terminologies activées. De ce fait, les lignes crées après la première étape peuvent être partiellement désordonnées.

Heimdall garde en mémoire la hauteur de chaque entité en fonction des terminologies et de l'état de repliement des différents chemins. Lorsqu'une ligne est dépliée ou repliée, ou bien lorsqu'une terminologie est activée ou désactivée, le logiciel recalcule rapidement la hauteur

de toutes les entités en fonction des paramètres d'affichage. En connaissant la hauteur de chaque entité, il est très facile de savoir quelles sont les entités au moins partiellement visibles dans la fenêtre de l'application. Ainsi, la majorité des entités peut être ignorée complètement à chaque rendu d'image, et celui-ci est très rapide.

4. Calcul des courbes

3 algorithmes d'interpolation (54) sont supportés :

- Interpolation *Last Observation Carried Forward* (LOCF)
- Interpolation linéaire
- Interpolation de type spline cubique

L'interpolation de type LOCF consiste à utiliser la dernière valeur connue pour tracer la courbe jusqu'à la prochaine valeur disponible, à partir de laquelle la courbe s'infléchie brutalement pour rejoindre la nouvelle valeur.

L'interpolation de type linéaire utilise une fonction affine qui passe par les deux valeurs connues les plus proches autour d'un point donné. Chaque paire consécutive de valeurs connues est reliée par un segment de droite.

Dans l'interpolation de type spline cubique, la courbe suit un ensemble de fonctions polynomiales du troisième degré qui lui donnent un aspect lisse, sans discontinuité autour des valeurs relevées. Contrairement à l'interpolation polynomiale simple, l'interpolation de type Spline est réalisée « par morceaux » ce qui limite la complexité des équations à résoudre lorsqu'il y a de nombreuses valeurs sur la courbe.

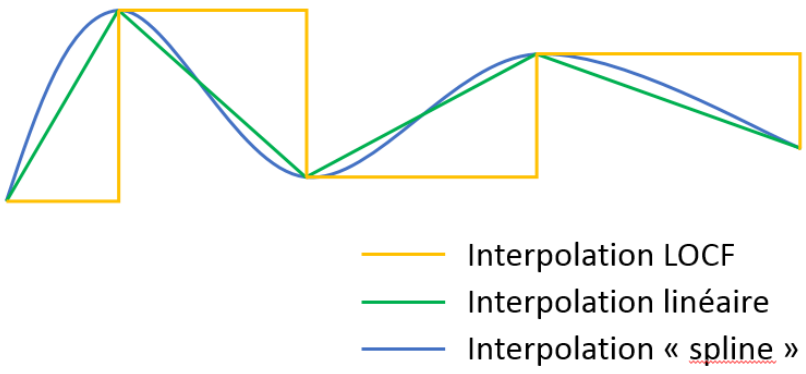

***Figure 23**. Illustration des types d'interpolation de courbes implémentés dans Heimdall.*

Ces méthodes d'interpolation appellent des commentaires spécifiques au monde médical. En termes de mouvement, l'interpolation spline est celle qui minimise à chaque instant l'accélération subie par un objet, donc qui correspond à la dépense d'énergie la plus faible. Physiologiquement, cela correspondrait à une variation la plus progressive possible de chaque paramètre biologique. Cette hypothèse semble donc relativement raisonnable et conservatrice. Pourtant, il existe des contextes dans lesquels elle semble peu adaptée. Ainsi par exemple, si un patient est hospitalisé pour une hyperkaliémie bruyante, il est raisonnable de penser que cette hyperkaliémie n'a débuté que quelques heures ou quelques jours plus tôt. Dans ce cas, l'interpolation LOCF est plus proche de la réalité. Un autre exemple est la recherche clinique prospective des essais contrôlés : l'interpolation LOCF y est notamment utilisée pour l'imputation de données manquantes. Dans le cas notamment où les données ne seraient pas manquantes par hasard (ex : patient quittant le bras intervention en raison d'un effet indésirable), et puisqu'en général le traitement vise à normaliser un paramètre dont la valeur est anormale à l'inclusion, cette imputation tend à diminuer la taille d'effet entre le groupe traitement et le groupe contrôle. Elle est donc considérée comme une bonne pratique, car elle diminue le risque de première espèce et est donc conforme au principe général « primum non nocere ».

L'interpolation linéaire, si elle ne repose sur aucune base théorique, présente l'avantage de donner un résultat proche de l'interpolation spline, mais au prix d'un calcul nettement plus simple, et permettant une meilleure robustesse aux mesures aberrantes, aux mesures trop rapprochées, et aux extrémités des fenêtres.

Annexe 2 : « Heimdall »

« Heimdall est le gardien de l'Ásgard. Sa demeure, Himinbjorg (litt. « le château du ciel »), est située en dehors des murailles d'Asgard, à proximité du pont de Bifröst, qui relie l'Asgard au Midgard. Capable d'entendre l'herbe pousser et une seule feuille tomber, ainsi que de voir jusqu'aux confins du monde, Heimdall n'a pas non plus besoin de dormir. »

AUTEUR : Nom : MARTIGNENE **Prénom :** Niels

Date de Soutenance : 05/05/2020

Titre de la Thèse : Mise au point d'une interface de visualisation unifiée de données cliniques temporelles, hétérogènes, hiérarchiques

Thèse - Médecine - Lille 2020

Cadre de classement : Santé Publique

DES + spécialité : Santé Publique

Mots-clés : Dossier patient électronique, Visualisation de données, Timeline, Extraction de caractéristiques

Résumé :

Contexte : Le dossier patient électronique contient des données temporelles structurées (PMSI, biologie médicale, médicaments, etc.) ou non (documents, images, etc.). Leur utilisation est double : transactionnelle (un seul patient, pour le soin) ou décisionnelle (plusieurs patients, réutilisation de données). Des outils de visualisation existent mais ne couvrent pas ces deux champs, rendent mal l'aspect hiérarchique des terminologies, et décloisonnent mal les données de sources différentes. L'objectif est de concevoir un tel outil.

Méthodes : Conception : nous abstrayons les types de données, puis spécifions les composants graphiques et leur mode de compactage. Développement : le prototype est développé en C++, disponible en librairie R. Evaluation : deux médecins répondent à 5 questions portant sur 24 cas cliniques réels d'insuffisance rénale aiguë, en utilisant 3 interfaces dont Heimdall.

Résultats : Prototype : le temps suit l'axe horizontal, les concepts suivent l'axe vertical. Les "événements" sont représentés sous forme de triangles, les "états" de rectangles, les "mesures" de courbes (avec interpolation LOCF, linéaire ou spline). Ces composants sont embarqués dans une arborescence qui exploite notamment celle des terminologies (CIM10, CCAM, etc.), qui permet un repliement vertical, et avec fusion des composants. Une condensation temporelle est possible. En mode décisionnel, les patients peuvent être alignés sur un événement (ex : appendicectomie). Les vues par problème (ex : fonction rénale, hématologie, etc.) déploient automatiquement des composants et en compactent d'autres. Évaluation : pour charger 3500 patients (360 000 valeurs), il faut 1 seconde et 50 Mo de mémoire. L'interface Heimdall est aussi rapide à utiliser par des médecins qu'une interface filtrée, et plus qu'une interface brute. Le taux d'erreur est identique.

Conclusion : Heimdall est intuitif, entièrement automatisé et interfacé avec R. Les vues par problème font gagner du temps. Actuellement, les données non-temporelles n'y trouvent pas de place, et Heimdall n'embarque pas d'outil de requête. Heimdall est open source et peut être utilisé via un paquet R (hors CRAN).

Composition du Jury :

Président :	Monsieur le Professeur Philippe AMOUYEL
Assesseurs :	Monsieur le Docteur Luc DAUCHET
	Monsieur le Docteur Antoine LAMER
Directeur de Thèse :	Monsieur le Professeur Emmanuel CHAZARD