Analyse et visualisation de trajectoires de soins par l'exploitation de données massives hospitalières pour la pharmacovigilance

Thibault Ledieu

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : *Analyse et Traitement de l'Information et de l'Image Médicale*

Par
Thibault LEDIEU

Analyse et visualisation de trajectoires de soins par l'exploitation de données massives hospitalières pour la pharmacovigilance

Thèse présentée et soutenue à Rennes, le 19 Octobre 2018
Unité de recherche : INSERM UMR 1099 LTSI

Rapporteurs avant soutenance :

Régis Beuscart	PU-PH Université de Lille
Catherine Duclos	PU-PH Université Paris 13

Composition du Jury :

Président :	Jean-Daniel FEKETE	DR INRIA
Examinateurs :	Emmanuel OGER	PU-PH Université de Rennes 1
	Christian VILHELM	MCF Université de Lille
Dir. de thèse :	Marc CUGGIA	PU-PH Université de Rennes 1
Co-dir. de thèse :	Frantz THIESSARD	MCU-PH Université de Bordeaux

Invitée
Catherine PLAISANT University of Maryland Institute for Advanced Computer Studies

Remerciements

Je remercie le Professeur Marc Cuggia de m'avoir accueilli dans son équipe pendant cette thèse.

Je remercie le Docteur Frantz Thiessard, pour les échanges constructifs et les bons conseils prodigués pendant cette thèse.

Merci à Catherine Plaisant et à Peter pour les 3 mois passés dans le Maryland et au sein du HCIL.

Un grand merci à Elisabeth Polard pour le soutien et l'énergie qu'elle m'a donnés pendant ces 3 longues années de thèse.

Un grand merci à ma tutrice de thèse, Cécile Chevrier, qui m'a toujours soutenu pendant les moments difficiles.

Merci au Professeur Lotfi Sehnadji pour son écoute et ses conseils.

Je remercie Eric Renault à qui je dois mes bases en programmation et qui continue à me faire découvrir des technologies innovantes.

Merci à Camille Morival pour toute l'aide et les idées nouvelles qu'elle a pu m'apporter pendant ces années.

Merci à Adeline Degremont pour son soutien infaillible et aux bonnes tranches de rigolade pendant ces années passées au laboratoire.

Merci à l'équipe Données Massives en Santé pour tous les bons moments passés pendant ces cinq années parmi vous : Canelle, Pierre, Guillaume, Françoise, Gwenaëlle, Pascal, Véronique et Christian.

Je remercie les personnes appartenant à l'équipe REPERES avec qui j'ai eu la chance de travailler : Emmanuel, André, Erwan, Frédéric.

Merci également à toute l'équipe du Centre de Pharmacovigilance de Rennes : Lucie-Marie, Louise, Marie-Noëlle, Sylvie.

Et enfin le plus important, un immense merci à mes parents et à ma famille.

Table des matières

Introduction : Pharmacovigilance et trajectoires de soins

1. Définition et contexte de la pharmacovigilance

Pendant des siècles, les traitements ont été utilisés en médecine de façon empirique. Néanmoins, déjà à cette époque, les médecins étaient conscients que le risque d'un traitement était un élément fondamental. Hippocrate (460 av JC-370 av JC) considéré comme le fondateur de la médecine, exerçait cette science sur un grand principe : « *Primum non nocere* », qui signifie : « *en premier ne pas nuire* ».

Plus tard au cours du XXème siècle, plusieurs drames sanitaires sont survenus avec de nouveaux médicaments :

- Cancers du foie et des voies biliaires, leucémies chez les personnes ayant reçu un produit de contraste contenant du thorium (Thorotrast®) pour les examens radiologiques. (1)
- Malformations et thalidomides dans les années 1960 (2)
- Cancérogénèses, malformations et diéthylstilbestrol (DES, Distilbène®) avec effet transgénérationnel dans les années 70 (3)
- Maladie de Creutzfeld-Jakob et hormone de croissance dans les années 90 (4)
- Valvulopathie et Benfluorex (Mediator®) en 2009 (5)
- Pilules contraceptives et accidents thromboemboliques veineux en 2012 (6)
- Acide valproïque et malformations congénitales et troubles neuro-comportementaux (7)

Le drame de la thalidomide a joué un rôle majeur dans la prise de conscience et la nécessité de surveiller les médicaments post-commercialisation. La pharmacovigilance, est définie par l'OMS comme «*la science et les activités relatives à la détection, l'évaluation, la compréhension et la prévention des effets indésirables et de tout autre problème lié au médicament* ». (8)

Historiquement, la pharmacovigilance est la première des vigilances sanitaires avec la création de l'Agence Nationale du Médicament en 1993 (9). En France, cette vigilance est d'abord née du terrain et mise en place au sein des établissements de soins par les pharmacologues. Les autorités sanitaires ont par la suite organisé une veille sanitaire spécifique au médicament qui a donné naissance aux centres régionaux de pharmacovigilance. On dénombre aujourd'hui un ensemble de 31 centres répartis sur l'ensemble du territoire et pour la plupart hébergés par les centres hospitaliers universitaires. Ces 31 centres sont sous la tutelle de l'Agence nationale de sécurité du médicament et des produits de santé. Cette autorité dépend elle-même du ministère de la santé. Mensuellement, les directeurs des centres de pharmacovigilance sont réunis lors d'un comité technique où une expertise des signaux de pharmacovigilance de l'ensemble du territoire est conduite.

La iatrogénie médicamenteuse correspond selon l'Organisation Mondiale de la Santé à toute réponse néfaste et non recherchée à un médicament survenant à des doses utilisées chez l'homme à des fins de prophylaxie, de diagnostic et de traitements

(1969). La iatrogénie médicamenteuse comprend tous les effets indésirables pouvant survenir suite à la prise d'automédication ou à la mauvaise observance d'un traitement. A l'inverse, la iatrogénie n'intègre pas les cas suivants : intoxication médicamenteuse volontaire ou accidentelle et les toxicomanies.

La loi du 9 août 2004 (10), relative à la politique de santé publique plaçait déjà la réduction de la iatrogénie médicamenteuse comme une priorité de santé publique avec un objectif de réduction des évènements iatrogènes d'origine médicamenteuse, entrainant à cette époque 130 000 hospitalisations par an sur le territoire, à moins de 90 000 par an. (11)

En 2013, le rapport sur la surveillance et la promotion du bon usage du médicament en France des Professeurs Bégaud et Costagliola confirmait l'importance des effets indésirables et de la iatrogénie médicamenteuse dans la population générale. Ce rapport mentionnait aussi l'existence d'une iatrogénie médicamenteuse importante chez les personnes âgées, expliqué par la présence d'une importante polymédication. (3)

La stratégie nationale de santé 2018-2022, adoptée officiellement par le gouvernement d'Edouard Philippe en décembre 2017, fait aussi de la iatrogénie une priorité. En effet, dans le 3ème axe de la stratégie nationale : « Garantir la qualité, la sécurité et la pertinence des prises en charge à chaque étape du parcours de santé », on peut trouver les objectifs suivants : la prévention de la polymédication et de la iatrogénie médicamenteuse justifient de poursuivre les travaux visant à promouvoir le bon usage des médicaments. (12)

La pharmacovigilance repose sur la déclaration spontanée des effets indésirables. L'OMS (2000) précise qu'un EIM peut résulter d'un usage normal ou non comme un mésusage, un usage abusif, un syndrome de sevrage, une pharmaco dépendance, une erreur médicamenteuse, une inefficacité thérapeutique, un effet sur le produit de conception ou un produit défectueux ou de mauvaise qualité. Les EIM sont classés par l'OMS par fréquence, nature, mécanisme de survenue, prévisibilité et gravité.

Les EIM sont un enjeu de santé publique majeur. (13) En effet, une étude parue dans le Journal of the American Medical Association en 1995 estime un nombre moyen par hôpital et par an d'environ 1900 EIM et 1600 EIM potentiels. (14)

Les professionnels de santé ont le devoir de déclarer toute réaction nocive et non voulue suspectée d'être due à des médicaments, y compris en cas de surdosage, de mésusage, d'abus, d'erreurs médicamenteuses et d'une exposition professionnelle. Depuis Mars 2017, les patients ont désormais la possibilité de déclarer les effets indésirables dont ils font l'objet sur un portail commun des vigilances accessible sur une plateforme web[1]. Dans le rapport JY Grall remis à la ministre de la santé en 2013 (15), il était déjà question de la mise en place d'un portail commun des déclarations permettant d'unifier et d'harmoniser le recueil des signaux sanitaires. L'enjeu majeur

[1] https://signalement.social-sante.gouv.fr/psig_ihm_utilisateurs/index.html#/accueil

de ce portail est la facilitation de la déclaration de tous les évènements indésirables comprenant les signaux de pharmacovigilance. Une des limites importantes de la pharmacovigilance réside en la sous-notification (13) (16). Les déclarations de pharmacovigilance sont en effet professionnels-de-santé-dépendantes et maintenant patients-dépendantes. Souvent les effets indésirables sont sous-notifiés. Cette sous-notification varie aussi en fonction du temps, des médicaments et effets concernés. Le portail des vigilances a pour but principal de répondre à cette sous-notification.

Pour répondre à ces attentes nationales, la première des vigilances sanitaires de France, s'organise et adapte les technologies disponibles à la détection et l'évaluation des signaux des effets indésirables. Des départements de pharmacoépidémiologie se développent au sein des grandes institutions nationales comme l'Agence nationale de sécurité des médicaments et des produits de santé et la Caisse nationale de l'Assurance maladie et investiguent la détection des signaux sur les données du SNDS (Système National des Données de Santé). La pharmacoépidémiologie, science qui étudie via l'application de méthodes épidémiologiques l'efficacité, le risque et l'utilisation des médicaments, surveille notamment à partir des bases de données administratives la sécurité de l'utilisation des médicaments. Dans ce contexte, l'ANSM (Agence Nationale de Sécurité du Médicament) a lancé un appel d'offre pour la constitution de deux plateformes de pharmacovigilance, à Bordeaux (DrugSafe) et à Rennes (PEPS). (17) Les travaux de la présente thèse s'inscrivent dans le cadre de cette dernière.

L'apport de la surveillance automatique des signaux de pharmacovigilance sur les données de santé (hospitalières ou ambulatoires) peut représenter une réponse à cette limite majeure de la pharmacovigilance.

Plusieurs équipes en France explorent des méthodes de détection du signal de pharmacovigilance sur bases de données médico-administratives dans le but que ces méthodes et algorithmes de détection puissent être utilisés en routine par les institutions nationales et régionales. (18) L'une des limites importantes à l'utilisation de ce type de base est le manque de certaines informations telles que les valeurs des tests biologiques, le motif précis de l'hospitalisation, qui peuvent être dans certains cas nécessaire à l'étude de l'imputabilité d'un médicament dans la survenue d'un effet indésirable. La mise en œuvre de technologie et de méthodes en routine permettant la réutilisation des données hospitalières incluant des approches d'intégration données (entrepôts de données biomédicaux) ainsi que des algorithmes de fouille de données, et outils de visualisation pourrait être une avancée dans le domaine de la gestion de la sécurité et de la pertinence de l'utilisation des médicaments dans le cadre d'un établissement, d'une région ou sur l'ensemble du territoire, répondant ainsi au quatrième axe de la stratégie nationale de santé 2018-2022 ou la création du « Health Data Hub » un laboratoire d'exploitation des données de santé. (19)

2. Etat de l'art du concept de « trajectoire de soins »

Le raisonnement en pharmacovigilance nécessite une analyse des événements passés chez un ou plusieurs patients, c'est-à-dire l'analyse de trajectoire de soins. Si le terme de « trajectoire » est de plus en plus utilisé dans le domaine de la santé, il reste difficile d'en trouver une définition formelle faisant consensus.

Une des premières utilisations du terme « trajectoire » en santé est de Glaser (1967) (20) qui propose la notion de « *trajectoire de la maladie* ». Cette notion se base sur « *l'analyse du soin en tant que processus temporel* ». Elle permet de mettre en lumière le rôle des interactions entre les patients, l'équipe médicale, et la structure institutionnelle sur l'évolution de la trajectoire. Une deuxième formulation de ce concept apparaît sur la base des maladies chroniques. Chaque maladie engendre des trajectoires imposant des actes médicaux. Le traitement médical décidé à la suite d'un diagnostic se traduit par un plan d'action, le schéma de trajectoire. Cette trajectoire va évoluer en fonction de la réponse aux traitements. Hornbrook définit quant à lui un épisode de soin comme une « *période (avec un début et une fin "théoriquement" identifiable) durant laquelle un problème de santé, une pathologie spécifique, un symptôme perçu par un patient ou un traitement, est présent et donne lieu à un ensemble de prestations délivrées par un ensemble de professionnels ou d'institutions* ». (21) Cette définition de l'épisode de soins le rapproche de la notion de trajectoire.

Lionel Perrier a effectué une revue de la littérature provenant de la base de données Medline (22), avec quatre définitions de la notion de trajectoire :

- « *Une trajectoire de maladie peut être vue comme une imbrication et une succession de tâches dont l'ensemble constitue l'arc de travail» c'est-à-dire «l'ensemble du travail qui aura besoin d'être fait pour maîtriser le cours de la maladie (activité thérapeutique, travail de confort, sécurité clinique, organisation des moyens, etc)* » (23)
- « *On appelle trajectoire une succession dans le temps de volumes de types «prises en charge », à partir d'un instant initial, et pendant une durée donnée* » (24)
- « *Une trajectoire patient est une succession dans le temps d'événements relatifs à un patient ou à un groupe de patients, relevant d'un problème ayant provoqué cette succession d'événements* » (25)
- « *La trajectoire d'un patient est une séquence ordonnée d'événements pathologiques et d'interactions avec le système de soins* » (26)

L'auteur précise qu'il est nécessaire de distinguer trois grands types de trajectoire :

- Les trajectoires dites « simplifiées », destinées à reconstituer à partir du Dossier Patient Informatisé l'itinéraire du patient au sein de l'espace de soins
- Les trajectoires dites « médicales », identifiant le couple « itinéraire produit ». Chaque ressource consommée est relevée pour chaque consultation ou

séjour. Ces ressources sont les actes thérapeutiques, les médicaments, ou les examens.
- Les trajectoires dites « médico-économiques » permettant d'évaluer le coût total de la prise en charge d'une pathologie.

Perrier décrit également deux catégories scindant ces 3 types de trajectoires : les trajectoire « mono établissement de santé », et les trajectoires « multi établissements de santé ». Cette distinction pose le problème du chaînage des séjours. Ce chaînage permettrait de relier les différentes hospitalisations d'un même patient dans des établissements différents. Au delà des séjours en établissement de soins, la reconstitution des trajectoires nécessite d'autres informations (par exemple les actes réalisés en ville, la dispensation de médicament) qui justifierait l'intégration de sources de données d'établissement public et privé, mais aussi de ville (contact avec les médecins généralistes ou les pharmacies).

Dans sa thèse de médecine, Nicolas Jay propose une autre définition du concept de trajectoire de soins. (27) La trajectoire est une séquence d'événements, donc une liste d'états ordonnés dans le temps. Cette liste est bornée à la fois dans l'espace et dans le temps. Certaines définitions de trajectoire de soins intègrent la notion de trajectoire de « maladie » alors que d'autres intègrent tous les événements de santé quel qu'ils soient. D'autres font référence à des groupes de patients, sous-entendant l'existence de profils types de trajectoires. L'auteur en conclut donc que la notion de trajectoire est dépendante du point de vue de l'utilisateur et que plusieurs facteurs vont déterminer ce point de vue :

- La fenêtre d'observation
- La qualité (métier) de l'observateur va déterminer quels types d'événements de santé comptent dans la construction de la trajectoire.
- L'information contenue dans une trajectoire est de caractère polymorphe, à la fois temporelle, géographique, démographique, médicale, et économique
- L'échelle d'observation peut être individuelle ou collective (implique alors l'existence d'une typologie des trajectoires)

Selon Françoise Riou (28), les objectifs d'une trajectoire de soins sont multiples :

- Comprendre les logiques d'orientation
- Modéliser et normaliser les prises en charge
- Evaluer l'efficience des prises en charge

Pour répondre à ces objectifs, la représentation d'une trajectoire doit posséder 3 éléments indispensables :

1. Caractéristique de population concernée
2. Les bornes de la trajectoire :
    - Physique : entre services
    - Temporelle : période d'observation fixé selon la pathologie
3. Contenu :

- Les éléments du parcours en fonction de l'objectif poursuivi :
    - Observation globale : succession ordonnée des déplacements des patients entre entités (typologie de lieux)
    - Observation d'un sous-système local : complétion par la différenciation des lieux de soins sur d'autres critères (proximité, moyen technique, etc…)
    - Observation interne à un hôpital : par exemple, au travers des données issues du PMSI
- Les caractéristiques temporelles

Des travaux plus récents ont été effectués par Huang et al (29). Dans ce rapport, Les auteurs prennent l'hypothèse que l'analyse des trajectoires de soins permet de découvrir quels "comportements médicaux" sont essentiels et/ou critiques pour les parcours de soins, en apportant une information d'ordre temporelle.

Les auteurs proposent une nouvelle approche de la fouille de données dans les trajectoires de soins. Ils définissent l'analyse comme :

1. La découverte de la connaissance sur l'impact des événements cliniques sur les soins du patient
2. L'utilisation de cette connaissance pour réviser/améliorer les parcours de soins

Les auteurs proposent un modèle de trajectoire permettant de tenir compte de l'écart de temps entre les événements. Les éléments cliniques retrouvés comme étant les plus fréquents sont considérés comme les éléments critiques.

A l'aide des différents travaux que nous avons pu étudier dans notre état de l'art des trajectoires de soins, nous avons choisi de modéliser notre trajectoire sous la forme d'une séquence d'événements horodatés organisés de façon chronologique. Cette séquence est bornée par le premier et dernier événement de la séquence. Un événement est caractérisé par une date de début et une date de fin, un type d'événement, et une valeur numérique. Les types d'événements peuvent être un diagnostic, un terme retrouvé dans un document, une administration médicamenteuse, ou encore un changement de service. Il faut cependant souligner que la véracité des informations utilisées pour générer les séquences est très variable : les termes, codes, ou traitements présents dans le système d'information ne sont que traces dont la véracité et la temporalité peuvent rester assez flou es.

Le choix des types d'événements contenus dans une séquence sera à la libre appréciation des utilisateurs, l'idéal étant de ne conserver que l'information pertinente pour vérifier l'hypothèse posée et uniquement celle-ci. La génération de ces événements se fera en fonction de règles métiers.

Cette modélisation simple à l'avantage de pouvoir être appliquée différentes méthodes standards de fouille de données. Cette modélisation permet également

d'être à la fois assez riche en information tout en facilitant les développements des outils de visualisation.

3. Problématiques de la thèse

a. Les hypothèses et questions posées

Dans les travaux effectués au cours de cette thèse, nous avons une approche proposant d'exploiter la richesse et le volume des données intra hospitalières pour des cas d'usage de pharmacovigilance. Cette approche reposera sur la modélisation de trajectoires de soins intra hospitalières adaptées aux besoins spécifiques de la pharmacovigilance. Il s'agira, à partir des données à disposition, de caractériser les évènements d'intêret et d'identifier un lien entre l'administration de ces produits de santé et l'apparition des effets indésirables, ou encore de rechercher les cas de mésusage du médicament.

Ces objectifs nécessiteront de retrouver des motifs (ou ensembles non ordonnés) d'évènements apparaissant fréquemment dans les trajectoires des patients ou de rechercher plus spécifiquement des motifs choisis par l'utilisateur pharmacovigilant.

L'hypothèse posée dans cette thèse est qu'une approche visuelle interactive serait adaptée pour l'exploration des données biomédicales hétérogènes et multi-domaines dans le champ de la pharmacovigilance.

b. Etapes de la thèse et démarche

Les différentes travaux réalisés dans le cadre de cette thèse, visent à répondre à des besoins d'analyse en pharmacoviglance à l'échelle individuelle, et populationnelle. Les différentes étapes des travaux réalisés pour cette thèse seront les suivantes, de manière itérative pour l'approche mono-patient et pour l'approche populationnelle :

1. Analyse des besoins, Etat de l'art
2. Modélisation de la trajectoire de soins
   2.1. Modèle d'information des trajectoires de soin
   2.2. Modèles visuels interactifs
3. Implémentation d'un prototype
4. Evaluation de l'utilisabilité ou de l'apport dans les pratiques avec application sur un cas d'usage

c. Approches méthodologiques et techniques envisagées

La réalisation de ces travaux a nécessité de recourir à différents champs méthodologique et techniques multidisciplinaires dont les principaux sont les suivants :

- Définition formelle et caractérisation d'un parcours de soin intra hospitalier ainsi que de ses cas d'usage
- Développement de méthodes d'extraction automatique d'informations sur données hétérogènes (numériques, symboliques et textuelles) pour l'instanciation du modèle du parcours de soins intra-hospitalier

- Développement de méthodes permettant de caractériser automatiquement ou détecter les parcours d'une population cible, des parcours similaires, ou des parcours remarquables de patients ayant été exposés à un produit de santé
- Développement d'outils de requêtes et de visualisations spatiales et temporelles des parcours intra-hospitaliers, mono- et multi-patients
- Techniques de fouilles de données hétérogènes : analyse formelle de concepts, algorithme de détection de séquences, fouilles de texte
- Traitement automatique du langage pour l'indexation, l'extraction de concepts et de contextes
- Méthodes de visualisation de données (frise chronologique, langage iconique, cartographie, diagramme de Sankey…) et de conception d'IHM
- Méthodes d'évaluation d'IHM

Chapitre 1 : Technologies du Big Data en santé

1. Le nouveau paradigme du big data

Le terme de Big Data est apparu lors de ces dernières années avec l'augmentation des capacités de stockage et de traitement. Le nombre de données numériques a explosé (90% des données du monde furent crées lors des deux dernières années, et chaque jour sont créés 2,4 trillions de gigabits de données). Cette augmentation de la production de données est liée à l'évolution d'internet où chaque internaute devient acteur et producteur de données : les photos, vidéos échangées chaque jour contribuent à faire augmenter significativement le Big Data. Le Big Data est un ensemble de technologies et de méthodes de traitements permettant d'aborder des problématiques de gestion de volume et d'exploitation des données.

a. Bref historique

La collection et la conservation des informations ont toujours fait partie de l'activité humaine, de l'invention de l'écriture à la construction de gigantesques centres de données. En 2013, Gil Press a publié un article dans Forbes (30) retraçant l'histoire du Big Data. Selon lui, l'histoire de la massification des données n'est pas aussi récente qu'on pourrait le penser puisqu'elle remonte au milieu du XXème siècle. En effet, 70 ans auparavant, ce phénomène était déjà connu sous le terme d' « explosion de l'information », terme utilisé pour la première fois en 1941 selon le *Oxford English Dictionnary*. En 1944, Fremont Rider estimait que chaque bibliothèque universitaire américaine doublait en taille tous les 16 ans. Plus tard, en 1967, Marron et Bain ont émis l'hypothèse que la compression des données est nécessaire face à cette explosion de l'information. En 1990, Denning, dans sa publication « *Saving All the Bits* », expose les défis du stockage des données liés aux expérimentations scientifiques. Il prédit qu'il sera nécessaire de construire des machines capables d'analyser et de résumer de vastes quantités de données. Si les problématiques liées à la croissance de la quantité d'informations étaient déjà connus, les enjeux concernant le volume des données numériques ont commencé à être évoqués à la fin des années 90 par Fayyad et al. (1996). En 1997, Michael Cox et David Ellsworth dans leur publication « Application-controlled demand paging for out-of-core visualization » présentent la visualisation comme un défi à relever pour les systèmes informatiques, notamment lorsque le nombre de données à représenter est très important. Ils appellent ce problème « big data ». Il s'agit du premier article de la bibliothèque numérique d'ACM à utiliser le terme "données volumineuses". Le domaine de la visualisation est étroitement lié aux défis du Big Data puisqu'en 1999 Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes publient « Visually exploring gigabyte data sets in real time », l'article décrit l'augmentation des données comme une conséquence des progrès accomplis en termes de puissance de calculs des ordinateurs. En 2000, Peter Lyman et Hal Varian ont tenté de quantifier l'information crées dans le monde chaque année sur 4 supports : papier, film, optique et magnétique. Les auteurs ont estimé que pour l'année 1999 le monde a produit 1,5 exaoctects d'informations non redondantes. Ils ajoutent que la production de données digitales n'est pas seulement la plus importante mais aussi celle qui croit le plus rapidement. L'émergence du web,

ou chaque utilisateur devient à la fois producteur et consommateur d'information, a ensuite amplifié ce phénomène. Pour preuve, en 2005, Tim O'Reilly, fondateur de la maison d'édition O'Reilly Media, publie « What is Web 2.0 » (31), ou il affirme que les données sont le nouveau « Intel Inside », c'est-à-dire qu'à l'instar des microprocesseurs elles constituent le composant essentiel des systèmes dont l'architecture est le plus souvent issue du monde open source. L'apparition d'outil comme le « *cloud computing* » a drastiquement permis de réduire le coût du stockage des données. A titre d'exemple, le prix d'un gigaoctet pour un disque dur est passé d'environ 12,30 euros en février 2000 0,07 euros en août 2010.

L'International Data Corporation a défini le Big Data en 2011 comme : « *une nouvelle génération de technologies et architectures, conçues pour extraire économiquement à partir de volumes très importants d'une grande variété de données, en permettant la capture, la découverte et/ou l'analyse à grande vitesse* ». »

Aujourd'hui, on estime que 2,5 téraoctets sont générés dans le monde chaque jour. Dans le domaine de la santé, on estime que leur volume devrait atteindre 2,3 milliards de giga-octets d'ici à 2020 dans le monde (32). L'origine et le format de ces données sont très disparates : elles peuvent être produites par l'hôpital lors des soins, être collectées lors des essais cliniques ou encore provenir de bases de données médico-administratives (par exemple le SNIIRAM et ses 20 milliards de lignes).

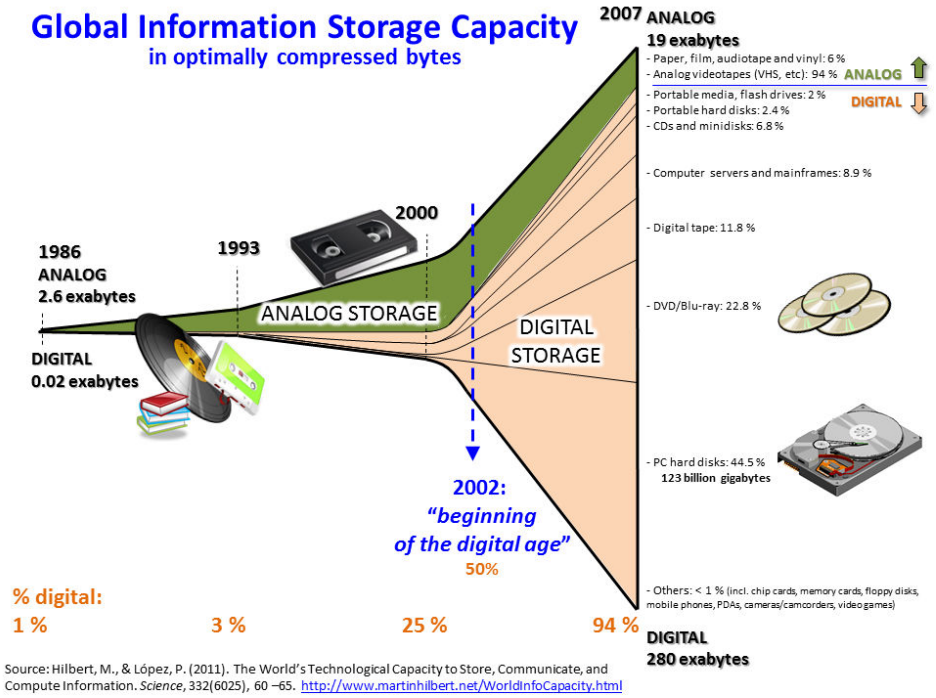


Figure 1 Evolution du volume des données stockées au cours des 30 dernières années (Source : Hilbert M, Lopez P, The World's Technological Capacity to Store, Communicate, and Compute Information, Science, 332(6025), 60 -65, 2011

b. Les enjeux

Le Big Data peut être vu comme une évolution de la Business Intelligence (BI). L'analyse de ces importants volumes de données numériques, appelée aussi « *Data*

Mining », a pour objectif de détecter les informations pertinentes et d'établir des corrélations entre elles. Aujourd'hui le métier de *Data Scientist* est de plus en plus demandé par les différents acteurs économiques (services publics et grandes entreprises), il a pour mission de permettre le traitement, la visualisation et la catégorisation de ces énormes flux de données. Ces analyses permettront d'assister les structures dans leurs prises de décision, de comprendre et d'anticiper le comportement des individus (clients ou patients) ou encore de faciliter les démarches qualité des organisations. L'évolution du Big Data rentre aussi en convergence avec celui du domaine de *l'intelligence artificielle* (IA). En effet, au travers de différents algorithmes, notamment des mécanismes d'apprentissage automatique, les différents systèmes d'intelligence artificielle vont être capables de déterminer des caractéristiques spécifiques dans les données pour comprendre et interpréter différents phénomènes. Aujourd'hui on trouve des exemples d'application des technologies de l'IA dans les chatbots ou encore la conduite assistée dans l'automobile.

c. Les dimensions du Big Data

Le Big Data, ou données massives, se définit selon 5 dimensions :

- Volumétrie
- Variabilité
- Vélocité
- Véracité
- Valeur

La volumétrie est la quantité de données à stocker, sa croissance est exponentielle (l'ordre de grandeur est le Tera/Peta octet de données). La problématique liée au stockage des données est plutôt bien maîtrisée et est toujours en constante évolution, notamment avec le phénomène du « cloud computing ».

La vélocité (ou vitesse) fait référence à deux composantes : la vitesse de production des données et la vitesse de traitement de ces données (extraction de connaissance, indicateurs statistiques). Le passage de la verticalisation (supercalculateur) à l'horizontalisation (architecture distribuée) a permis de répartir les charges de calcul et de traitement sur plusieurs microprocesseurs. Ce besoin a émergé avec la fin de validité de la loi de Moore. En effet, la fouille de données nécessite parfois des traitements devant résister au passage à l'échelle, ce que permet plus facilement le traitement distribué.

La variabilité (ou variété) renvoie à la diversité de formats de données devant être gérés. Les données peuvent en effet être structurées ou semi structurées dans un format JSON ou XML, ou être non structurées lorsqu'elles sont contenues dans des textes ou des images. De plus, la connaissance du schéma de production de ces données est susceptible de ne pas être disponible à l'avance. Aujourd'hui, on estime que la part des données non structurées dans le volume des données produites au total est comprise entre 80 et 85%.

La véracité est le niveau de confiance accordée aux données. Plusieurs problématiques sont associées à cette dimension du Big Data :

- La transformation des données pouvant altérer leur contenu
- La consolidation des données à l'aide de source de données
- La temporalité d'une donnée : tel le vocabulaire d'une langue vivante, le sens et la structure sémantique d'une donnée peuvent évoluer avec le temps, par exemple avec les évolutions des terminologies
- La gestion des données manquantes
- Les données bruitées (bruit dans un signal, doublons,…)

La cinquième dimension, la valeur, désigne la plus-value qu'il est possible pour la structure ou l'entreprise de tirer de ces données. Il faut également souligner le caractère très sensible des données de santé.

Les dimensions du Big Data que nous avons présenté ont abouti, entre autres, aux développements de nouvelles technologies de bases de données que nous allons décrire ci-après. La mise en œuvre de telles technologies fut réalisée dans la perspective d'une mise en l'échelle de nos outils de visualisation et de traitement de données, notamment sur un jeu de données plus grand que le contenu de notre entrepôt de données.

2. Les bases de données NoSQL

a. Le rationnel

Les premières bases de données étaient fondées sur des systèmes de gestion de fichiers très sophistiqués. Dans les années 70, E.F. Codd a posé la théorie des relations et des fondements des bases de données relationnelles avec l'algèbre relationnel. (33) Depuis, si d'autres modèles de bases de données ont émergés tels que les systèmes orientés objets, les bases de données relationnelles ont largement dominé le marché. Cependant, ces systèmes peuvent être mis à mal devant les défis posés par les différentes dimensions du Big Data. En effet, si les propriétés ACID (atomicité, cohérence, isolation, durabilité) décrites par Haerder et Reurer (34) de ces systèmes garantissent que les transactions informatiques se réalisent de façon fiable, elles sont aussi très contraignantes pour leur mise à l'échelle.

Dans les années 2000, les géants du web confrontés à cette problématique de volumes de données ont commencé à développer leurs propres systèmes de gestion de bases de données. Ces systèmes ont pour objectif de pouvoir maintenir de bonnes performances avec la montée en charge tout en ayant un coût financier raisonnable. Le terme « NoSQL » pour « *Not Only SQL* » a été inventé par Oskarsson en 2009 pour désigner ces nouveaux systèmes « *open-source, distribués et non-relationnels* ». Ils ont pour caractéristiques de permettre la manipulation de très larges volumes de données tout en ayant la possibilité d'être mis à l'échelle horizontalement.

Les bases de données NoSQL ont une structuration relationnelle faible et reposent sur le théorème de CAP d'Eric Brewer (35) :

- Cohérence (*Consistency*) : Tous les clients voient la même donnée au même moment
- Disponibilité (*Availability*) : Toutes les requêtes doivent recevoir une réponse
- Tolérance à la partition (*Partition Tolerance*)

En pratique, seulement deux de ces trois principes peuvent être respectés à la fois. En général, on privilégie les deux derniers principes. Nous verrons dans la prochaine partie comment s'organise l'information dans ces bases de données à travers différents exemples de systèmes.

b. Les différentes catégories

i. Clé-valeur et orientée document

Le modèle de stockage de l'information sous forme de tableaux associatif, ou modèle-valeur est un modèle typique du NoSQL. Un excellent exemple de l'implémentation de ce modèle est le système REDIS (de REmote DIctionary Server) qui conserve l'intégralité des données en RAM pour éviter les accès disques coûteux. (36)

Les bases orientées document reposent sur le paradigme clé-valeur mais où la valeur est remplacé par un document de type JSON ou XML. Ceci permet de retrouver l'ensemble des informations de manière hiérarchique et d'effectuer des requêtes plus complexes. On peut prendre comme exemple les bases CouchDB (37) et MongoDB (38).

Figure 2 Schéma clé-valeur

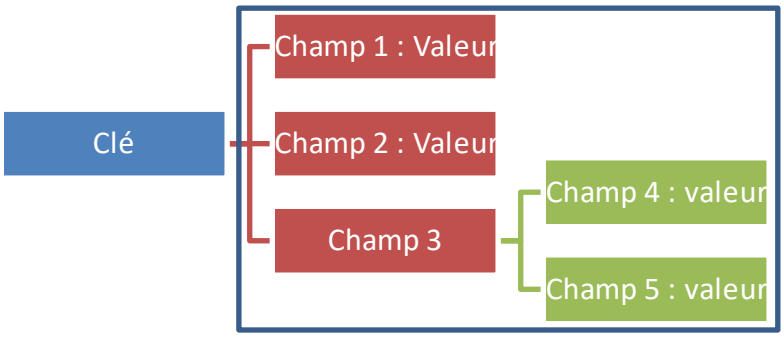


Figure 3 Schéma orientée document

ii. Orientée colonne

Une base de données orientée colonnes est une base de données qui stocke les données par colonne et non par ligne. Ceci permet d'ajouter de la flexibilité au schéma en permettant d'ajouter des colonnes beaucoup plus facilement aux tables. Les données peuvent aussi être fortement compressées, rendant beaucoup plus

rapides les opérations en colonnes telles que MIN, MAX, SUM et COUNT. Sybase IQ et Vertica sont de bons exemples d'implémentation de bases de données orientées colonnes.

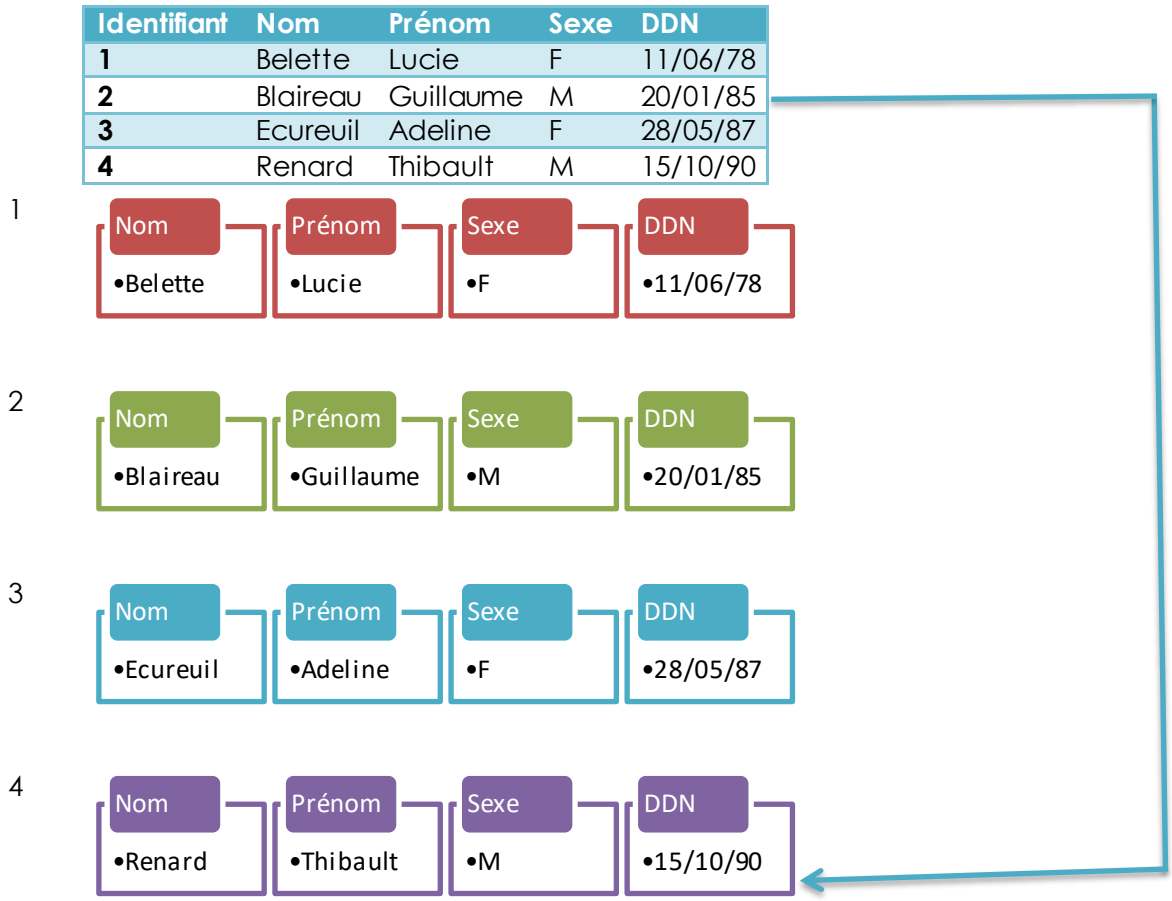


Figure 4 Schéma orientée colonne

iii. *Orientée graphe*

Il s'agit d'un type de base de données orientée objet fondé sur la théorie des graphes. Pour décrire les données, elles utilisent des arcs et des nœuds. Elles ont pour avantage d'être adapté à la gestion de données relationnelles. L'exemple le plus connu de base de données orientée graph est Neo4J (39). On peut aussi citer l'exemple d'OrientDB qui est également multi-modèle (40).

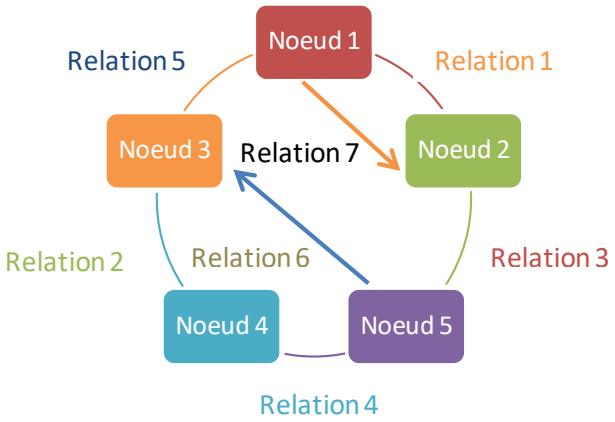


Figure 5 Schéma orientée graph

Dans les prototypes que nous avons réalisés, nous avons utilisé des bases de données orientés colonnes (MongoDB). Ces bases nous ont permis de stocker l'information sans contraintes rigides de schéma. Nous avons également utilisé les bases de données orientées graph pour stocker les différentes terminologies, leur structure sous forme d'arborescence étant particulièrement bien adaptée à ce type de base de données.

3. Le passage de la verticalisation à l'horizontalisation

a. La fin de la loi de Moore

En 1965, Gordon Moore, docteur en chimie et en physique et futur co-fondateur d'Intel, fait le constat que le nombre de transistors regroupés sur un même circuit intégré croit de manière exponentielle depuis plusieurs années. Il annonce en 1975 la loi de Moore qui énonce que le nombre de transistors intégrés dans les circuits en silicium double tous les deux ans jusqu'à atteindre la taille d'un atome vers 2015. Une interprétation erronée de cette loi est que la puissance de calcul, la capacité de stockage, la fréquence d'horloge doublent tous les 18 mois. Si cette interprétation s'est révélée à peu près exacte depuis 1973, depuis le milieu des années 2000 la fréquence d'horloge des microprocesseurs tend à stagner voire même à régresser. En effet, les efforts des fabricants de micro-processeurs portent plus sur les architectures multi-cœurs pour contourner les problèmes techniques liés aux très hautes fréquences d'horloge (dissipation de chaleur par exemple). Ce changement d'architecture a eu des très forts impacts sur le logiciel, qui doit être réécrit et spécialement adapté en parallélisant au maximum les calculs (applications multi-thread) pour tirer parti des avancées matérielles.

b. Traitement parallélisé et distribué

La parallélisation est une technique d'accroissement des performances des ordinateurs en utilisant plusieurs processeurs fonctionnant simultanément. Le parallélisme peut se faire à différents niveaux, soit à l'intérieur du processeur (multi cœurs ou instructions en parallèle), soit répartis sur plusieurs processeurs. L'exécution d'instructions en parallèle permet de gagner du temps en découpant un gros problème en plusieurs petits problèmes. Cependant, la loi d'Amdahl (41) montre qu'une limite théorique de l'accélération en latence de l'exécution d'une tâche

existe, et elle est proportionnelle à la proportion d'instructions ne pouvant pas être parallélisée (la partie sérielle).

Une architecture informatique est dite distribuée lorsque tous ses composants ne se situent pas sur la même machine. Ces composants communiquent par l'intermédiaire du réseau et ont la possibilité d'utiliser des objets situés sur différentes machines. Le calcul distribué est la répartition de traitement sur plusieurs unités centrales. Un exemple de projet d'utilisation du calcul distribué est SETI@home, développé par l'université de Berkeley, en Californie. (42) Il s'agissait d'utiliser la puissance de calcul des ordinateurs reliés à internet pour faire les calculs nécessaires pour détecter la vie intelligente non terrestre.

c. Exemple de Map Reduce & Hadoop

Pour répondre aux besoins de parallélisations et de distribution des calculs, Google a inventé un patron d'architecture de développement informatique, nommé *MapReduce*. Son intérêt réside dans sa capacité à manipuler et de traiter un nombre importants de données sur une architecture distribuée. Ce modèle de programmation peut être découpé en 2 fonctions : *map* et *reduce*. La fonction *map* va permettre de répartir la charge de travail sur les différents nœuds du système. La fonction *reduce* va permettre de rassembler les résultats.

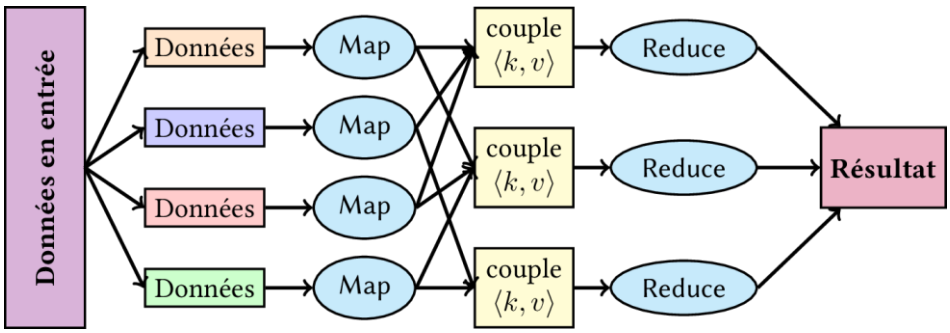


Figure 4 Illustration de l'algorithme Map Reduce Par Clém IAGL — Travail personnel, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=22688163

Le traitement des données pour l'extraction des évènements d'intêret qui seront utilisés dans les cas d'usage de pharmacovigilance ont nécessité de parallèliser les traitements pour réduite les temps de calcul. Il s'agissait de rendre ces informations disponibles à l'utilisateurs pour lui permettre d'interagir avec ces données de manière fluide.

4. Réutilisation des données de santé d'intêret et systèmes disponibles

a. Les données produites à l'hôpital

Avec l'informatisation des systèmes d'informations hospitaliers (SIH), l'hôpital est devenu un producteur de données. En effet, avec la mise en place des plans successifs Hôpital 2012 et Hôpital numérique, les hôpitaux français ont dématérailisé

les données qu'ils produisent pour un objectif sans papier. Ainsi les hôpitaux produisent actuellement des gisements de données patient qui font l'objet d'une réutilisation secondaire Les données produites par les SIH sont de natures variées et hétérogènes puisqu'elles proviennent de différentes sources. On retrouve des données administratives (identité, âge, lieu de résidence du patient), des données de prescriptions et d'administrations médicamenteuses (dose, nom du médicament administré), des données d'analyses biologiques. On retrouve aussi des informations sous forme de texte libre dans les comptes rendus d'hospitalisation ou les lettres de sortie. Dans le cadre du PMSI (Programme de médicalisation des systèmes d'information), les actes et les diagnostics sont recueillis obligatoirement pour chaque hospitalisation. Ces données sont ensuite collectées pour constituer les bases nationales du PMSI. Ces bases sont mises à disposition des structures (établissements de santé, sociétés, associations...) par l'ATIH (Agence Technique de l'Information Hospitalière).

Le tableau ci-dessous présente les types de données et des exemples de données produites à l'hôpital.

Tableau 1 Types et sources des données dans les systèmes d'information hospitalier

Type de données	Exemples de sources	Exemples de données
Symbolique	Diagnostics CIM-10	C900 Myélome
	Actes médicaux CCAM	DBLF001 Pose d'une bioprothèse de la valve aortique, par voie artérielle transcutanée
	Prescriptions médicamenteuses ATC	N05BA12 Alprazolam
Numérique	Résultats d'analyses biologiques	INR 2,5
	Dosage d'administrations médicamenteuses	Alprazolam 0,5 mg
Textuelle	Courrier de sortie	
	Compte-rendu d'imagerie	
Temporelle	Dates de prescriptions médicamenteuses	Prescription de Seroplex du 05/02/2016 au 05/03/2016
	Dates de séjour	Séjour en réanimation médicale du 09/09/2016 au 16/09/2016

b. La réutilisation des données de santé

La réutilisation des données, souvent désignée par le terme anglais « data reuse » ou « secondary use of data », à une utilisation des données à d'autres fins que celles de la mission pour les besoins de laquelle elles ont été produites ou conçues. En santé, elles concernent l'usage de données administratives et/ou recueillies au cours des soins pour une autre finalité comme la recherche. Une grande majorité des données médicales sont non structurées et contiennent de l'information de forte valeur et les sources de données sont très variées (par exemple, les systèmes d'imagerie, notes et correspondances, résumés d'internes, les données de l'assurance maladie. La réutilisation des données en vie réelle ouvre la voie à de très nombreux travaux

épidémiologiques (rapport de Bégaud, Polton et Von Lennep (43)) encouragés par l'état. D'autre part, l'exploitation de ces données par ces méthodes d'exploitation des données (comme exemple par exemple l'extraction de connaissances, la recherche d'information, ou le raisonnement automatique via des algorithmes d'IA) permet d'entrevoir l'émergence de systèmes d'aide à la décision à la fois à l'échelle individuelle ou populationnelle (rapport Villani (44)).

c. Etat de l'art des entrepôts de données biomédicaux

Wade (45) définit les entrepôts de données cliniques comme des dépôts d'information provenant de dossiers cliniques, et parfois de recherche, d'un seul organisme, comme un fournisseur de soins ou un payeur. Un entrepôt possède également un haut niveau d'intégration permettant d'interroger de façon flexible son contenu.

MacKenzie et al ont mené en 2010 (46) une enquête auprès de centres de santé universitaires américains. Cette enquête a révélé que 22 entrepôts de données institutionnels se trouvaient dans 35 établissements. Le nombre de patients dans un entrepôt variait de 43 000 à 10 millions, avec une médiane de 1,6 million. Ces chiffres reflètent les populations généralement importantes de patients dans de nombreux hôpitaux universitaux.

L'un des objectifs de ces entrepôts est de permettre la découverte et la constitution de cohortes pour la recherche clinique ou épidémiologique mais également pour de la détection de signaux utiles en pharmacovigilance.

Ces entrepôts possèdent le plus souvent des moyens de standardiser les données en fonction de leur chargement par des processus dits ETL (pour « Extract Transform Load »). Ces processus seront en charge de transformer les données quel que soit leur source (flux de la pharmacie, dossier patient informatisé) au « format pivot » utilisé par l'entrepôt.

Inmon (47) décrit les différences fondamentales entre un entrepôt de données dans un hôpital et un entrepôt dans d'autres domaines opérationnels :

- En santé, une transaction est relativement unique alors qu'elle est répétitive dans d'autres domaines
- En santé, les données sont hétérogènes (format textuel, nombre, code), alors que l'information est essentiellement sous forme de nombre dans les autres domaines
- En santé, il n'existe pas de vocabulaires communs entre chaque hôpital (on peut prendre l'exemple des terminologies utilisées dans les laboratoires d'analyse qui différent en fonction de l'établissement). Une normalisation est donc requise.
- En santé, la valeur temporelle de l'information est primordiale
- En santé, l'entrepôt de données devra être en mesure d'intégrer plusieurs sources de données externes

Aujourd'hui un des systèmes d'entrepôt de données cliniques le plus connu est la plateforme i2b2 (Informatics for Integration Biology & the Bedside) (48). Il s'agit d'un framework *open-source* permettant l'utilisation des données cliniques avec des données génomiques. STRIDE (Stanford Translational Research Integrated Database Environment) (49) est un entrepôt de données cliniques intégrant les données d'un réseau d'entrepôts et propose un portail d'accès aux données biomédicales avec procédures adaptées en gérant les aspects de sécurité et confidentialité des données.

d. eHOP

Au Centre Hospitalier Universitaire de Rennes, nous avons développé notre propre solution d'entrepôt de données biomédicales, appelé eHOP (anciennement Roogle). (50) Cet entrepôt intègre tous les types de documents produits par le système d'information hospitalier et liés aux soins de santé:

- Des données structurées utilisant des terminologies de référence (par exemples des diagnostics de la CIM-10 à partir du PMSI, codes terminologiques locaux pour les tests de laboratoire, codes de *l'Association for the Development of Informatics in Cytology and Pathology* (ADICAP) pour les diagnostics de pathologie, terminologie anatomique thérapeutique chimique (ATC) correspondant aux prescriptions et à l'administration des médicaments)
- Des données non structurées, telles que notes narratives cliniques, protocoles chirurgicaux, radiographies ou rapports pathologiques.

Par conséquent, eHOP permet aux utilisateurs de rechercher des informations à partir de données structurées et non structurées. De plus, il est possible de combiner deux façons différentes d'interroger les données. Les utilisateurs peuvent construire des requêtes basées sur des terminologies de référence, ou simplement soumettre des mots-clés pour récupérer des documents structurés ou non qui contiennent ces termes ou mots-clés. Les utilisateurs peuvent ensuite accéder aux documents au moyen d'une interface dédiée qui intègre des fonctionnalités permettant de naviguer dans l'ensemble du dossier électronique du patient. eHOP est couramment utilisé au CHU de Rennes pour soutenir la recherche clinique dans les études de faisabilité, ou pour le repérage de patients sur des critères d'éligibilité.

L'entrepôt de données eHOP offre actuellement la possibilité de rechercher parmi 25 millions de données non structurées et 170 millions d'éléments structurés. Certaines données non structurées, comme les résultats de laboratoire ou les diagnostics, sont également enregistrées sous une forme structurée grâce aux codes terminologiques correspondants. Toutes ces données proviennent du système d'information hospitalier et couvrent plus de 1,6 million de patients.

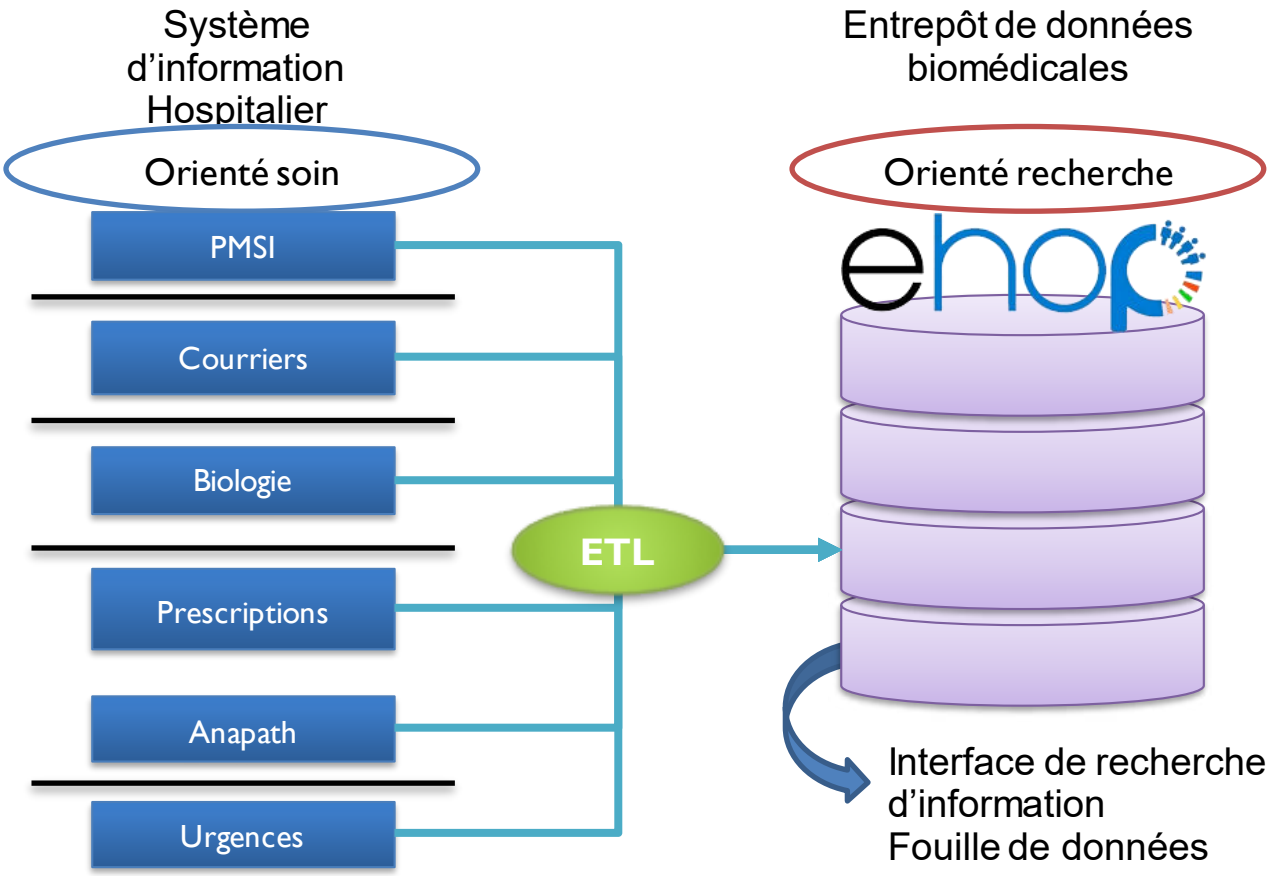


Figure 5 Intégration des données dans eHOP

e. Exemples d'application

Les entrepôts de données biomédicaux ont un rôle important à jouer pour la recherche. En effet, la réutilisation automatisée des données de santé trouve son intérêt dans plusieurs domaines :

- La recherche clinique pour le préscreening de patients correspondant aux critères d'inclusion d'une étude (39).
- La recherche épidémiologique pour la constitution de cohorte.
- Les études médico-administratives (organisation raisonnée des soins, pilotage des activités, analyse de trajectoires de santé).
- La médecine personnalisée (médecine 4P).
- Pour les collectivités :
    - Détection de signal pour les vigilances (ex : Détection d'effets indésirables médicamenteux par recherche d'information dans les textes) (51)
    - Prédiction d'épidémie (ex : Surveillance syndromique par l'utilisation des données du réseau Sentinelles et de eHOP pour produire des indicateurs temps réels d'activité grippale (52)).

Chapitre 2 : Méthodes de traitement des données dans une perspective de pharmacovigilance

Dans le chapitre précédent, nous avons présenté différentes techniques de stockage adaptées à de grand volume de données. Dans les chapitres suivants, nous présenterons les techniques que nous avons mises en œuvre pour stocker les données des trajectoires de soins, qui sont hétérogènes et volumineuses. Nous allons désormais aborder différentes méthodes d'exploitation de ces données permettant l'élaboration et/ou la vérification de différentes hypothèses en pharmacovigilance.

1. Méthodes de fouilles de données

a. Définition, objectifs et principes

L'augmentation du volume de données a rendu nécessaire le développement de méthodes adaptées pour ne pas être « submergé » par l'information. Nous avons vu dans le chapitre précédent que les entrepôts permettaient de stocker efficacement ces données, nous allons voir ici comment les exploiter par l'emploi de méthodes automatiques ou semi-automatiques permettant d'extraire la connaissance intéressante (règles d'association, corrélation, motifs récurrents) ou de la résumer (statistiquement ou visuellement). Cet ensemble d'outils informatiques et statistiques forme le domaine de l'exploration de données ou «*Data Mining* ».

b. Méthodes de datamining d'intêret en pharmacovigilance

De nombreux travaux se sont intéressés à l'utilisation de technique de fouille de données dans ce domaine, nous présentons ci-dessous une liste non exhaustive des principales méthodes d'intérêt. L'ensemble de ces méthodes de fouille de données constituent les étapes du Processus de Découverte des Connaissances (en anglais, *Knowledge Discovery Process* KDD).

Selon M Wu. (53), ces méthodes permettent d'effectuer 3 types d'analyses :

- Une analyse descriptive permettant de résumer l'information. Il s'agit du type d'analyse le plus connu (moyenne, dispersion)
- Une analyse prédictive basée sur des modèles utiles pour anticiper certains types d'événements. Les analyses prédictives permettent par exemple d'anticiper les épidémies de grippe (52). Ces analyses sont par nature probabilistes (séries temporelles)
- Une analyse prescriptive. Ici, il s'agit non seulement de prédire un événement mais de recommander une action pour aboutir à la meilleure solution.

On peut distinguer parmi les méthodes d'exploration de données non-supervisées qui sont par nature exploratoire et qui vont permettre d'organiser l'information. Parmi elles, on retrouve des méthodes de clusterisation (classification hiérarchique ascendante) permettant de répartir les individus dans un certain nombre de classes. Les analyses factorielles (analyse en composantes principales, analyses des correspondances multiples) mettent en évidence des liens entre variables. Dans les travaux présentés dans cette thèse, nous avons mis en application des méthodes non-

supervisées de recherche d'associations telles que les algorithmes Apriori et GSP. Ces méthodes nous ont permis de détecter les séquences les plus communes dans les trajectoires patients.

Les méthodes supervisées sont utilisées pour faire des prédictions à partir d'une base d'apprentissage. Parmi ces méthodes, on retrouve les réseaux de neurones, les réseaux bayésiens ou encore les arbres de décisions. Dans les travaux présentés dans cette thèse, nous avons mis en œuvre des techniques de classifications hiérarchiques (CAH) non-supervisées pour regrouper des trajectoires patients similaires entre elles.

c. Fouilles de données séquentielles

La fouille de données séquentielles est un domaine particulier du *Data Mining* faisant parti des approches d'apprentissage symbolique non supervisé. Un des exemples les plus connus de ce type d'algorithme est le panier de la ménagère décrivant un ensemble d'achat réalisé au supermarché.

Les règles d'associations, popularisée par un article d'Agrawal en 1993, permettent de découvrir des relations entre plusieurs événements dans une base de données en recherchant des motifs fréquents. Ces règles introduisent plusieurs notions définies par Agrawal et Srikant en 1995 pour décrire le domaine.

Les algorithmes de fouilles de données séquentielles sont appliqués sur une base de données transactionnelle constituée d'un ensemble fini d'items. Un item est un « atome », dans l'exemple du panier de la ménagère il s'agira d'un article dans le supermarché (ex : jambon, biscuit, bière). Une transaction est un ensemble d'items, par exemple <Jambon, Fromage, Œufs>. Le nombre d'item contenu dans la transaction constitue le volume. Le support d'un ensemble d'items (ou itemset) désigne le nombre de transactions en base de données qui le contiennent. Ce sont ces transactions qui vont permettre de déterminer les règles d'association. Une règle d'association est notée X → Y avec X et Y des ensembles d'items disjoint. Une séquence est une liste ordonnée d'un ou plusieurs ensembles d'items.

Tableau 2 Exemple de bases de données transactionnelles

Identifiant Séquence	Séquence
1	<A,C><B,D,E>
2	<A,B,C><E,G>
3	<A,B,D,E,F>
4	<C><A,C><B,A,C>
5	<A><C><B,F,G>

d. Exemples d'algorithmes de fouilles de données séquentielles

Dans cette partie nous présenterons succinctement et de façon non exhaustive quelques algorithmes spécifiques à la fouille de données séquentielles. Il s'agit d'un ensemble de méthodes utilisées pour découvrir des structures ou des modèles récurrents à partir d'un très grand nombre de données. On peut classer ces algorithmes en deux catégories : l'approche « *Pattern-Growth* » et l'approche basée

sur l'algorithme Apriori. En introduisant des contraintes telles qu'une valeur de support minimale spécifiée par l'utilisateur, l'écart minimum ou le temps, les algorithmes répondent à différentes questions de recherche.

i. Apriori

Apriori est le plus connu des algorithmes de recherche d'association. Il a été conçu en 1994, par Agrawal et Sikrant. (54) Il a été conçu pour retrouver les règles d'associations à partir des transactions présentes en base de données. Il ne garde que les itemsets avec une fréquence (support) supérieure à un seuil fixé par l'utilisateur. L'image ci-dessous montre le fonctionnement et les résultats renvoyés par l'algorithme Apriori.

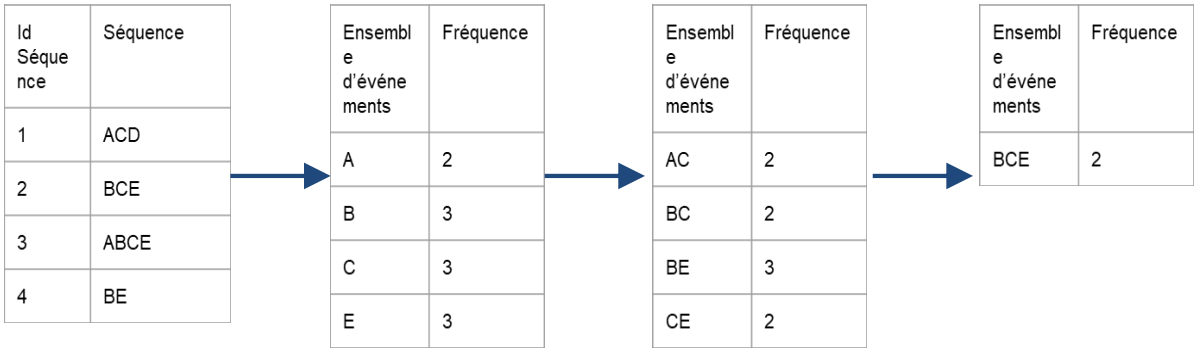


Figure 6 Illustration du fonctionnement de l'algorithme Apriori

L'algorithme Apriori permet donc de voir l'ensemble des événements co-occurrentss dans une cohorte de patients. Cependant il présente plusieurs limites :

- Plusieurs lectures de la base de données sont nécessaires, la parallélisation n'est pas possible
- L'algorithme ne tient pas compte du nombre d'occurrence des événements au sein d'un itemset
- Le calcul des supports et la génération des règles sont couteux en temps de calcul

ii. GSP

L'algorithme GSP (Generalized Sequential Pattern) est un des premiers algorithmes d'extraction de règles séquentielles également proposé par Agrawal et Srikant. (55) Contrairement à l'algorithme Apriori basé sur la théorie ensembliste à qui il fait suite, GSP permet d'extraire et de classer des ensembles d'items à partir de leur support dans la base de données en tenant compte de l'ordre d'apparition des événements. Une séquence est considérée comme fréquente si elle est supérieure à une valeur seuil définie par l'utilisateur. GSP est un dérivé de l'algorithme Apriori auquel les auteurs ont rajouté la notion de fenêtre temporelle (en anglais « *sliding-window* ») afin de prendre en compte la simultanéité des transactions. L'algorithme prend en entrée deux paramètres donnés par l'utilisateur : le min-gap et le max-gap. Le min-gap définit l'unité fondamentale de temps d'une transaction (par exemple une journée) et le max-gap le temps maximum pour considérer que deux ensembles d'items font partie de la même séquence.

Id Séquence	Séquence
1	ABCDE
2	BADC
3	DFEABC

Séquences détectées	Support
BC	3
ABC	2
DE	2
...	....

Figure 7 Illustration du fonctionnement de l'algorithme GSP

Le fonctionnement de l'algorithme GSP s'effectue en plusieurs étapes :

1. Le tri de la base de données pour récupérer une liste en ligne de paires d'identifiants de séquences et de séquences d'événements.
2. La génération des candidats a lieu en deux phases :
   2.1. La phase de jointure, qui permet de générer les possibles séquences candidates (par exemple, pour les séquences <AB> et <BC>, la phase de jointure va générer le candidat <ABC>)
   2.2. La phase d'élagage qui servira à garder les séquences candidates ne pouvant être fréquentes (il s'agit de la même logique que l'algorithme Apriori)
3. Calcul du support des candidats en fonction des paramètres min-gap et max-gap

L'algorithme GSP présente plusieurs limites. Si la base de données comporte un grand nombre de types d'événements différents, on aura un grand nombre de séquences candidates de longueur 2 (pour illustrer, si on a 500 types d'événements, on aura 124750 séquences candidates à rechercher en bases de données). Comme pour l'algorithme Apriori, un grand nombre de lectures de la base est nécessaire (une lecture après chaque phase de jointure), et il n'est pas possible de paralléliser les calculs.

iii. SPAM

Il s'agit d'un algorithme pour compter les itemsets fréquents présenté par Ayres et al. (56) La particularité de l'algorithme SPAM est de représenter la base de données sous forme de bitmap (ou tableau de bits, voir Fig. 8). Cette collection ordonnée de bits permet d'accélérer les opérations logiques en considérant les bits comme des opérateurs booléens. Dans le cas de l'algorithme SPAM, une colonne du bitmap correspondra à un itemset tandis qu'une ligne correspondra à une séquence (voir Fig.ou 8) afin de réduire les temps de calcul du support de chaque itemset.

Numéro de séquence	<A>	<B>	<C>	<D>
1	1	1	0	0
2	1	1	1	0
3	0	0	0	1
4	1	0	1	0

Figure 8 Représentation de séquences sous formes de bitmap

iv. SPADE

SPADE est un algorithme proposé par Zaki et al. permettant de réduire l'espace des solutions possibles en regroupant les motifs séquentiels par catégorie par une méthode de « préfixage » (57). Autre particularité, SPADE représente les séquences sous forme de bases de données verticales, c'est-à-dire en inversant la méthode d'indexation (voir Fig. 9). Cependant cette approche a l'inconvénient d'être très gourmande en mémoire, même si le temps de comptage du support des candidats générés est assez court.

<A>	
Numéro de séquence	Nombre de transactions
1	1
2	2
2	1
3	3
4	1

<C>	
Numéro de séquence	Nombre de transactions
1	1
2	2
2	4
4	1
5	1

<AC>	
Numéro de séquence	Nombre de transactions
1	1
2	2
4	1

Figure 9 Représentation des séquences sous forme de bases de données verticales

v. PSP

L'algorithme PSP (Prefix Tree for Sequential Pattern) utilise une structure arborescente de préfixe pour améliorer la génération des candidats. (58) Il S'agit d'une amélioration de l'algorithme GSP visant à améliorer ses performances. La structure en forme d'arbre permet de retrouver les séquences candidates en les « factorisant » par leur préfixe. La figure 10 illustre la structure arborescente utilisée par PSP. Les branches symbolisées par des traits pleins représentent l'ajout d'un événement dans la séquence dans un nouvel itemset. Les traits en pointillés représentent quant à eux l'ajout d'un événement dans le dernier itemset de la séquence. Parcourir l'arbre depuis sa racine permet donc de retrouver toutes les séquences candidates dans la base de données.

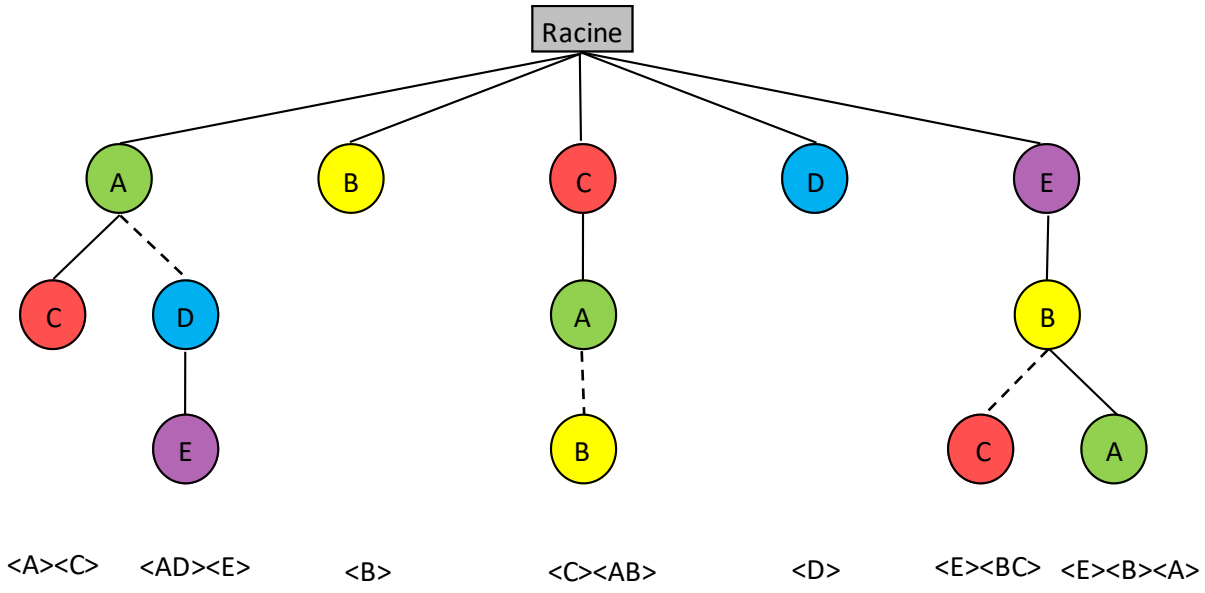


Figure 10 Illustration de fonctionnement de l'algorithme PSP

La liste ci-dessus n'est pas exhaustive, et nous aurions également pu citer les algorithmes PrefixSpan (59), ou encore LAPIN (60) pour l'extraction des longues séquences. Dans les travaux réalisés dans le cadre de cette thèse, nous utiliserons les algorithmes Apriori et GSP pour la découverte de motifs d'événements récurrents dans notre base de données. Si ce ne sont pas les plus performants en termes de rapidité de calculs, ils ont l'avantage d'être relativement faciles à mettre en place et leur « simplicité » algorithmique peut permettre de leur apporter des modifications pour d'adapter aux cas d'usage de pharmacovigilance.

e. Algorithme d'alignement de séquences

Les problèmes de l'alignement de séquences sont largement étudiés dans le domaine de la bio-informatique. Il s'agit de représenter deux ou plusieurs séquences de macromolécules de façon à mettre en évidence les régions semblables entre elles (voir l'exemple de la Figure 3). Aligner des séquences permet d'identifier des sites fonctionnels d'une protéine, ou de retrouver des mutations ayant survenu lors de l'évolution d'une même espèce. De nombreux algorithmes ont été proposés pour résoudre ce problème d'alignements. Il s'agit d'un problème complexe d'un point de vue optimisation étant donné le volume très important de données à traiter (pour illustrer la taille du génome humain est de 3,4 milliards de paires de bases). Nous allons vous présenter dans cette partie deux algorithmes d'alignement de séquences utilisés en bio-informatique.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :*****.:* ********..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:************:***:**.: .****************    :  *.: :
```

Figure 11 Alignement des séquences produites avec le programme libre ClustalW entre deux séquences de protéines, publiquement disponibles dans GenBank, Opabinia regalis

i. Algorithme de Needleman-Wunsch

L'algorithme de Needleman-Wunsch (NW) (61) est un algorithme d'alignement global de deux chaines de caractères. Il est très proche de l'algorithme de Levenshtein proposée en 1965, dont l'objectif est de produire une mesure de la différence entre deux chaines de caractères. Il s'agit d'une application de la programmation dynamique dont le principe est de résoudre un problème en le découpant en plusieurs sous-problèmes, le tout en mémorisant les résultats de calculs intermédiaires qui seront la plupart du temps répétés un grand nombre de fois. Cependant, l'alignement global est un problème coûteux à calculer en terme de temps de calcul, et les algorithmes peuvent retourner un grand nombre de solutions différentes. L'algorithme NW s'effectue en plusieurs étapes :

1. Initialisation d'une matrice, avec en têtes de ligne et de colonnes les deux séquences à comparer
2. Calcul des scores et remplissage de la matrice
3. Parcours de la matrice pour calculer l'alignement

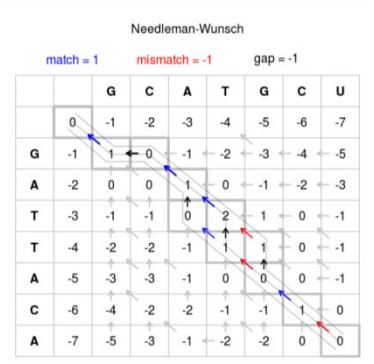


Figure 12 Illustration du fonctionnement de l'algorithme Needleman-Wunsch, Kamil Slowikowski, CC0 (https://commons.wikimedia.org/wiki/File:Needleman-Wunsch_pairwise_sequence_alignment.png#filelinks)

ii. Algorithme de Smith-Waterman

L'algorithme de Smith-Waterman est un algorithme optimal d'alignement de séquences permettant la recherche du meilleur alignement possible entre deux séquences de nucléotides ou d'acides aminés. (62) L'algorithme traite les séquences comme des chaînes de caractère. Dans ces chaînes, on peut insérer des « caractères vides » pour maximiser le nombre de coïncidence entre caractères (dans le cas d'insertion ou de délétion). Contrairement à l'algorithme de Needleman-Wunsch, l'alignement peut commencer à partir de n'importe quel caractère de la séquence.

C'est un algorithme utilisant également la programmation dynamique en calculant une matrice de similarité et en la parcourant pour retrouver l'alignement opitmal (Figure 13 et 14) Il s'agit d'un algorithme considéré comme plus sensible que l'algorithme de Needleman-Wunsch pour comparer des séquences de longueurs

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	2
T	0	3	0	4	9	7	5	3	1
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

différentes.

Figure 13 Parcours de la matrice pour calculer l'alignement (Exemple tiré de la page wikipedia :https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm)

T	G	T	T	-	A	C	G	G
G	G	T	T	G	A	C	T	A

Figure 14 Alignement obtenu

Dans le cadre de cette thèse, nous avons modifié cet algorithme pour qu'il puisse prendre en compte l'aspect temporel des données médicales. L'alphabet utilisé sera généré automatiquement en fonction de la question de recherche posée.

2. Recherche d'information

L'information en santé est également contenue dans des documents textuels tels que les comptes-rendus d'imagerie ou les questionnaires. Il est donc essentiel de pouvoir extraire l'information de ces documents. De nombreuses techniques d'indexation et de fouilles de textes ont déjà été développées dans le domaine de la santé ((63), (64), (65)). Cependant, extraire l'information des documents médicaux est un vrai défi étant donné les nombreuses nuances apportées dans le langage naturel utilisé dans les textes : référence à des antécédents familiaux, datation des événements, notion d'incertitude d'un diagnostic, extraction de valeurs numériques. De plus, la

signification de l'information extraite peut varier fortement en fonction du domaine médical, par exemple un acronyme peut avoir plusieurs sens différents en fonction du champ disciplinaire dans lequel il est utilisé. Plusieurs « moteurs d'indexation » ont été réalisés pour extraire des textes médicaux des concepts appartenant à un thésaurus médical tel que le MeSH (*Medical Subject Headings*), qui est une hiérarchie de termes normalisés pour l'indexation d'articles par la NLM, et qui peut servir à l'analyse de documents biomédicaux. Une autre terminologie est le MedDRA (66) (Medical Dictionary for Regulatory Activities), utilisée le partage d'informations réglementaires concernant les médicament, notamment la notification d'effets indésirables lors d'un essai clinique. Les ontologies ont également été utilisées ces dernières années dans un contexte médical (67) (68). Une ontologie est une représentation formelle de connaissance basée sur une un ensemble structuré de concepts. Ces concepts sont reliés entre eux sous forme de graphe par des relations sémantiques ou hierarchiques.

Dans les travaux présentés dans cette thèse, nous nous développerons notre méthode pour afficher les termes et le résumé des informations contenues dans les documents médicaux.

3. Visualisation de données

La visualisation de données est un domaine dont le développement s'est accéléré en en même temps que l'émergence du *Big Data*. Cet ensemble de techniques permet de prendre connaissance des dominantes principales d'un jeu de données assez large ou encore de repérer rapidement les caractéristiques remarquables. L'analyse visuelle est souvent un préambule à une exploration plus formelle des données par des méthodes statistiques. L'exemple du quartet d'Anscombe (69) est très instructif sur l'intérêt d'utiliser des méthodes de visualisations en complément des méthodes statistiques classiques. Le quartet d'Ascombe est constitué de quatre ensembles de données qui ont les mêmes propriétés statistiques (moyenne, écart-type, coefficient de corrélation) mais qui apparaissent très différent lorsqu'on les représente sous forme de graphiques.

I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

Figure 15 Données du quartet d'Anscombe

Tableau 3 Résumé des données du quartet d'Ascombe

Propriété	Valeur
Moyenne des x	9,0
Variance des x	10,0
Moyenne des y	7,5
Variance des y	3,75
Corrélation entre les x et les y	0,816
quation de la droite de régression linéaire	Y = 3 + 0,5x
Somme des carrés des erreurs relativement à la moyenne	110,0

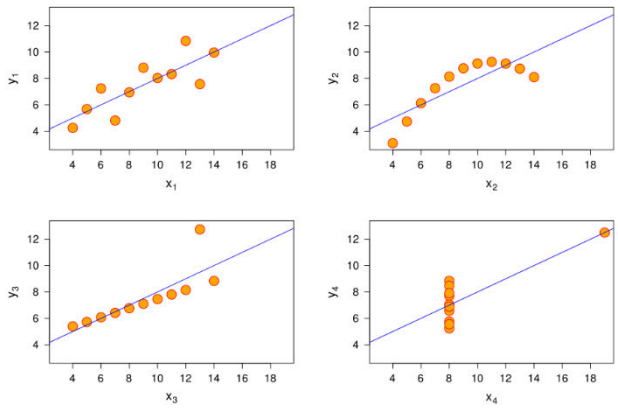


Figure 16 Illustration du quartet d'Ascombe

a. Processus de conception d'un outil de visualisation de données

Ben Fry décrit dans son ouvrage *Visualizing Data* (70) les 7 étapes indispensables de la visualisation de données. Ces étapes sont une marche à suivre pour pouvoir répondre à une question posée.

1. **Acquisition** : Il s'agit tout simplement d'obtenir les données nécessaires pour répondre à la question posée. C'est une étape pouvant être soit très complexe (c.-à-d. essayer de glaner des données utiles à partir d'un grand système) ou très simple (lecture d'un fichier texte facilement disponible).
2. **Analyse et transformation** : Une fois les données acquises, il faut les analyser et les transformer en un format exploitable en fonction de l'usage auquel elles sont destinées.
3. **Filtrage** : Cette étape consiste à filtrer les données pour supprimer les parties qui ne sont pas pertinentes pour répondre à la question posée
4. **Fouille** : Cette étape fait intervenir les mathématiques, les statistiques et l'exploration de données. Les données peuvent recevoir un traitement simple :

(rechercher du minimum du maximum) ou beaucoup plus complexe (mécanismes d'agrégation, de mise à l'échelle…)

5. **Représentation** : Cette étape détermine la forme de base qu'un ensemble de données prendra. Par exemple, certains ensembles de données sont présentés sous forme de listes, d'autres sont structurés comme des arbres. Il s'agit d'une des étapes des plus importantes dans un projet de visualisation et peut amener à repenser les étapes précédentes.
6. **Affinage** : Dans cette étape, les méthodes de conception graphique sont utilisées pour clarifier davantage la représentation en attirant davantage l'attention sur des données particulières (établissement d'une hiérarchie) ou en changeant les attributs (comme la couleur) qui contribuent à la lisibilité.
7. **Interagir** : La dernière étape du processus ajoute l'interaction, permettant à l'utilisateur de contrôler ou d'explorer les données. L'interaction peut couvrir des actions comme la sélection d'un sous-ensemble de données ou la modification du point de vue. Cette étape peut également affecter les traitements antérieurs des données, car un changement de point de vue peut nécessiter une conception différente des données.

Dans la partie suivante nous allons présenter quelques exemples de méthodes de représentation des données.

b. Exemple de méthodes de visualisation de données

i. Diagramme de Sankey

Les diagrammes de Sankey permettent de représenter des flux sous forme de flèches. La largeur de chaque flèche est fonction de la proportion du flux représenté. Ainsi, les transferts les plus importants sont directement interprétables visuellement. Les nœuds connectés par les flèches représentent un état. Les diagrammes de Sankey ont été initialement utilisés dans le domaine de l'énergie par Matthew Sankey qui a utilisé ce type de diagramme dès 1898. Dans l'exemple ci-dessous, les transitions de l'énergie du pays du Canada sont représentées sous forme de diagramme de Sankey. Les couleurs des flux représentent les sources d'énergie (uranium, pétrole, gaz naturel).

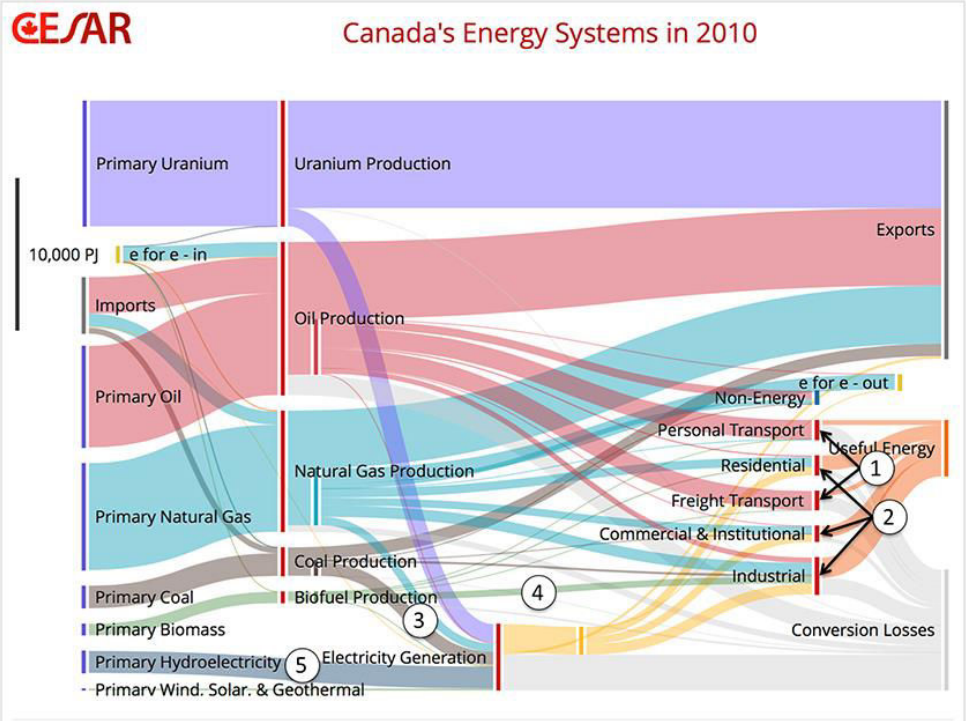


Figure 17 Diagramme de Sankey représentant les flux d'energie au Canada au cours de l'année 2010

ii. Diagramme de Venn

Les diagrammes de Venn (ou diagramme logique) représentent toutes les combinaisons possibles d'une combinaison finie d'ensembles. Ils ont été inventés en 1880 par John Venn. Ils contiennent 2^n zones correspondant à toutes les combinaisons possibles d'inclusion et d'exclusion des éléments avec n le nombre d'ensembles. Ils permettent de représenter facilement des relations de probabilité et de logique. Ils ne doivent pas être confondus avec les diagrammes d'Euler où l'appartenance à l'ensemble est indiquée par le chevauchement ainsi que la couleur. Les diagrammes de Venn sont difficilement applicables avec un nombre de variable supérieur à 3, et les tables de Karnaugh peuvent leur être préférées (Hill et Peterson, 1968, 1964).

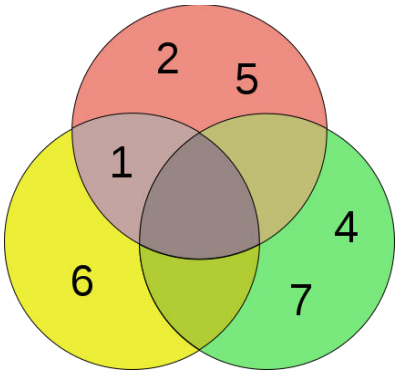


Figure 18 Exemple d'un diagramme de Venn avec trois ensembles

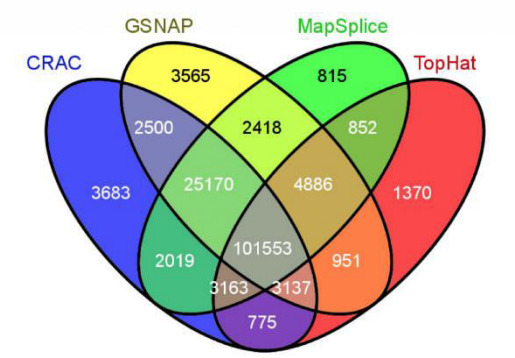


Figure 19 Exemple d'un diagramme de Venn à quatre ensembles. Reproduction de la figure 4b de Genome Biology, 14(3):R30, 2013

iii. Carte de chaleur (Heatmap) et carte proportionnelle (Treemap)

Une carte de chaleur (ou heatmap) est une représentation graphique de données où les valeurs individuelles contenues dans une matrice sont représentées en couleurs. Le terme «*heatmap*» a été inventé par le concepteur de logiciels Cormac Kinney en 1991, pour décrire un affichage 2D décrivant l'information sur les marchés financiers. Elles sont également utilisées dans le domaine du webmarketing pour visualiser le parcours des visiteurs sur un site marchand. Dans l'exemple ci-dessous, la carte de chaleur représente l'évolution des températures au cours des heures de la journée et des mois pour l'année 2013.

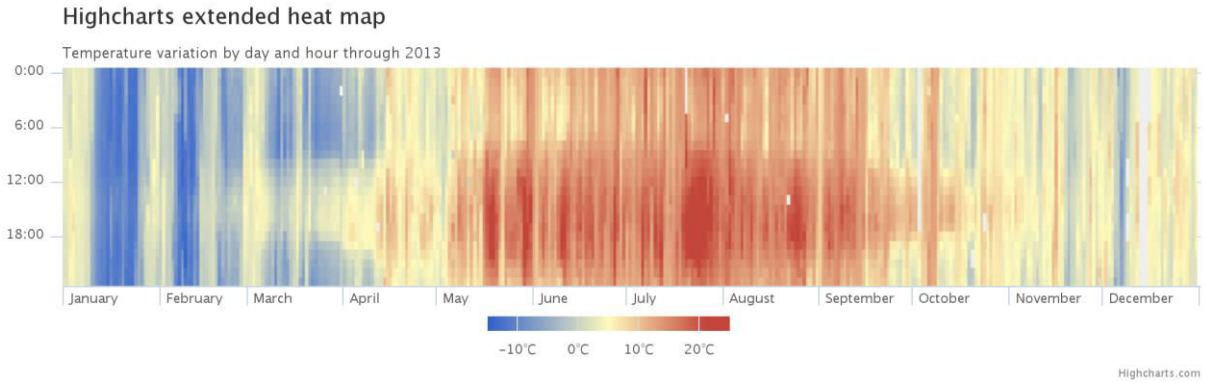


Figure 20 Carte de chaleur représentant l'évolution des températures au cours des heures de la journée et des mois pour l'année 2013

Une représentation proche des cartes de chaleurs est la Treemap qui est une représentation hiérarchique des données. Elle a été inventée en 1990 par le Pr Ben Shneiderman pour visualiser les proportions d'espace occupées par les différents dossiers et fichiers sur le disque dur du serveur de son laboratoire. L'arborescence des fichiers est représentée dans un plan par plusieurs rectangles dont la surface et la couleur dépendent de la taille et des attributs de ces fichiers. Ben Schneiderman a implémenté cette représentation dans le logiciel TreeSize dont une capture d'écran est disponible ci-dessous.

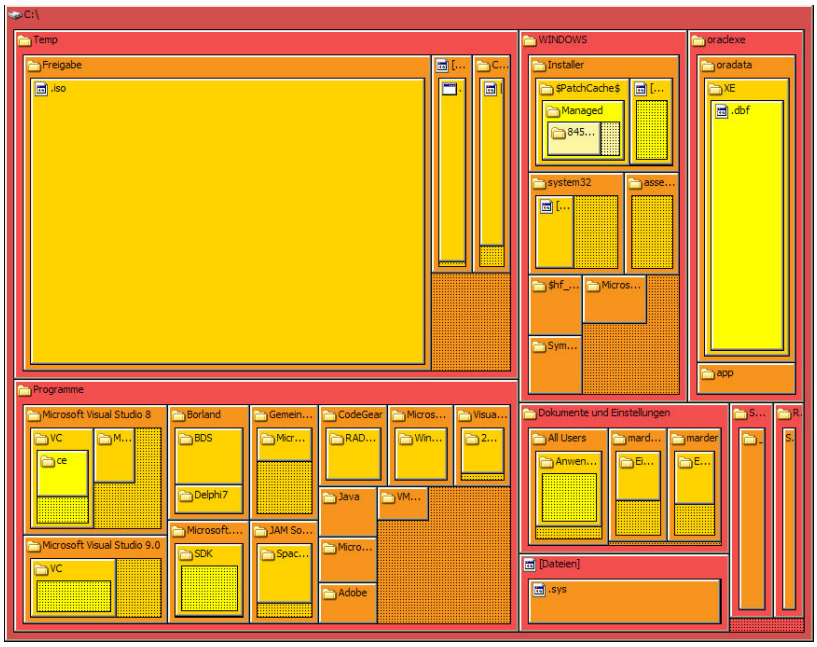


Figure 21 Visualisation de l'occupation de l'espace disque dans le logiciel TreeSize

iv. Nuages de mots

Les techniques de TAL (Traitement Automatique du Langage) peuvent être utilisées conjointement avec des techniques de visualisation de données comme les nuages de mots. Le nuage de mots permet de mettre en évidence les concepts ou expression jugées comme les plus importants dans un document pour pouvoir résumer visuellement ce dernier. L'importance peut être traduite comme le nombre d'occurrence d'un terme dans un corpus ou par son score *tf-idf* (term frequency – inverse document frequency).
Un des premiers exemples de nuage de mot-clés anglais fut les « *subconscious files* » de Douglas Coupland (1995).

Figure 22 Nuage de mots représentant la population de chaque pays proportionnellement à la police de caractère

v. *Langage iconique de représentation des connaissances médicales*

Ce langage a été développé par Jean-Baptiste Lamy et Catherine Duclos. (71)

VCM représente par des icônes les principaux concepts médicaux (symptômes, traitements, antécédent de maladies, test cliniques ou de laboratoire, état physiologique). Un exemple d'icone représentant des troubles du rythme cardiaque est proposé ci-dessous.

Figure 23 Exemple d'icône VCM

La construction d'une icône VCM va obéir à plusieurs types de règles. La forme générale de l'icône indique le type d'état : un cercle pour les états physiologiques et un carré pour les états pathologiques. Un pictogramme blanc peut être ajouté au centre de l'icône pour spécifier l'emplacement anatomique du problème : par exemple, un pictogramme en forme de cœur ou de poumon pour représenter les problèmes cardiaques ou pulmonaires. La couleur de l'icône indique la temporalité de l'état : rouge pour les états actuels, orange pour les risques menaçant le patient dans le futur, marron pour les antécédents, vert pour la surveillance, et bleu pour le type de traitement (médical, chirurgical, par exemple) ou de surveillance (clinique, laboratoire, imagerie, ou autre)

Un exemple d'application de ces règles est proposé sur la figure ci-dessous.

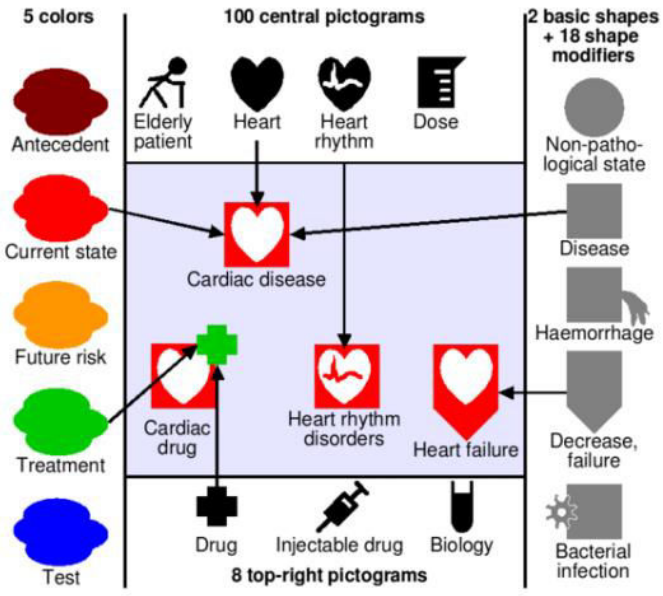


Figure 24 Illustration du fonctionnement du langage VCM

Des tests d'utilisabilité démontrent que les médecins s'approprient deux fois plus vite le contenu du dossier patient après une formation au langage VCM. Un alignement sur les codes de la classification CIM-10 est proposé avec VCM. Il a été choisi d'utiliser cet alignement pour afficher les données CIM-10 issues du PMSI sur les outils de visualisation développés.

Chapitre 3 : Application aux trajectoires de soins mono-patients

1. Contexte et problématique

La pharmacovigilance nécessite la collecte rétrospective d'information à partir des dossiers médicaux des patients afin de trouver des indices sur la responsabilité des médicaments dans la survenue d'effets indésirables médicamenteux (EIM). Cela signifie qu'il faut trouver de l'information pertinente dans les dossiers des patients qui peuvent constituer une preuve d'un effet indésirable associé à un médicament. Comme expliqué dans la partie sur les entrepôts de données biomédicaux, les données cliniques sont hétérogènes en termes de structure (données structurées ou textuelles) et de domaine (p. ex. résultats cliniques/laboratoires ou données de prescription). La recherche et l'organisation de ces informations nécessitent une expertise spécifique et sont souvent des tâches chronophages en raison du manque d'outils adaptés à cette fin.

Parmi les différentes méthodes d'extraction de données, l'analyse visuelle des données pourrait être utile pour explorer des données complexes. En effet, la visualisation des données sur l'histoire de patients individuels ou de groupes de patients pourrait permettre de mettre en évidence des indices chronologiques et des relations possibles entre l'exposition au médicament et les EIM. Une ligne du temps est une représentation chronologique et interactive d'une liste d'événements qui peut inclure différents types de données et échelles de temps. Dans la littérature, plusieurs outils de visualisation des données des dossiers de santé électroniques (DSE) ont été décrits pour explorer les données cliniques individuelles ou à l'échelle de la population (p. ex. Lifeline (72), Outflow (73), VisuExplore (74), Eventflow (75)).CareCruiser (76) fournit simultanément une vue des données du patient et des protocoles thérapeutiques pour l'analyse des réponses au traitement (par exemple, pour un patient sous oxygène, les valeurs de saturation en O2 varient en fonction du traitement). VisuExplore (74) propose différents types de graphiques (courbes de distribution, histogrammes) en fonction du type de données (par exemple, pour un patient diabétique, l'administration d'insuline est affichée sous forme de points, tandis que les paramètres biologiques, tels que la concentration de glucose, sont affichés sous forme de courbes de distribution). KNAVE-II (77) est un outil de visualisation interactif et sémantique de données cliniques basé sur l'utilisation d'ontologies de domaine. Cependant, aucun de ces outils n'a été évalué ou développé pour la pharmacovigilance.

Nous présentons dans ce chapitre un outil de visualisation sous forme de timeline que nous avons développé dans le cadre des projets de recherche RAVEL (78) et Pharmaco-Epidémiologie des Produits de Santé (PEPS) dont la mission principale est le développement de nouvelles méthodes de recherche d'information et la visualisation de données cliniques hétérogènes. Dans ce chapitre, nous présentons également l'évaluation de l'utilisabilité de ce prototype et la quantification de sa contribution à

la pratique courante de la pharmacovigilance en termes de gain de temps et de qualité des données collectées.

2. Etat de l'art

a. Vizualization of changes in neuro oncology patient (Visualisation des changements chez les patients en neuro-oncologie)

Le projet présenté intègre un outil de visualisation sous forme de timeline (79). Il intègre un outil de traitement naturel du langage (TAL) permettant de décrire les changements d'état d'un problème médical. Les dossiers étudiés concernent des patients admis dans un service de neuro-oncologie.

Les problèmes médicaux sont identifiés par le processus TAL et sont ensuite assignés dans 9 classes différentes décrivant les changements d'état de ces problèmes (voir Figure 1). Ces 9 classes sont ensuite visualisées sur une interface en forme de Timeline permettant à l'utilisateur de zoomer et de filtrer les données. Cet outil avait donc pour objectif de relever 2 défis :

1. Caractériser automatiquement les changements dans un problème médical pour chaque observation
2. Pouvoir visualiser graphiquement cette caractérisation

Pour un affichage pertinent sur la timeline, l'outil de TAL permet d'attribuer aux données les attributs suivants :

- La localisation anatomique du signe ou du symptôme (ex : cerveau) et la relation dans l'espace avec d'autres éléments (« à côté de »)
- L'existence du problème (à partir de mots clés tels que « certainement », etc.)
- Les données relatives à un problème (ex : « changement de taille ») (voir Figure 26)

Class	Description
Existing	Assigned to problems that are known to have existed prior to the start of available documentation.
Improving	Given to a problem that has changed for the better based on context. # of arrows denote greater improvement.
New	Assigned to the first mention of a problem that has never previously been observed.
Not stated	Given to problems that are not described in a report.
Recurrent	Given to the first mention of a problem that has already been observed and resolved in the past.
Resolved	Assigned to problems that have been mentioned in the past but explicitly observed to be not present or resolved in the current report.
Unassigned	Given to a problem that is mentioned but does not contain sufficient context to make a class assignment.
Unchanged	Assigned to a problem that has remained in the same state as the last observation.
Worsening	Given to a problem that has changed for the worse based on context. # of arrows denote greater decline.

Figure 26 Classes de changement d'état

Existence Attributes	
Certainty	Definite, appears to be, less likely, unlikely, does not exist
How determined	Observation, inference
Multiplicity	Single, a few, multiple, not stated
Newness	Previously seen, newly diagnosed, recurrent, resolved, not stated
Relevancy	Significant, incidental, not stated
Study quality	Okay, poor
Visibility	Clearly seen, appears, difficult
Finding Attributes	
Presence	Yes, no, not stated
Severity	None, mild, moderate, severe, not stated
Change	Improved, worsened, increased, decreased, stable, not stated
Degree	Slightly, significantly, not stated

Figure 25 Attributs de changement d'état

b. Using Timeline Displays to Improve Medication Reconciliation (Utilisation de timeline pour améliorer le bilan comparatif des médicaments)

Cet outil propose une méthode de visualisation de données par Timeline permettant l'intégration et la visualisation de données temporelles portant sur le médicament (80). Ces données proviennent à la fois de données structurées mais aussi de données textuelles (non structurées). L'originalité de la méthode de visualisation présentée est de tenir compte à la fois de la granularité (niveau de détail) et de l'incertitude des données. L'aboutissement d'une telle méthode de visualisation est de parvenir à améliorer les processus de comparaison du traitement que le médecin pourrait prescrire au patient et les traitements que le patient reçoit déjà, enfin d'éviter les erreurs de prescriptions.

Une classification des évènements a été conçue afin de représenter les différentes combinaisons d'incertitude portant à la fois sur la partie « clinique » de la donnée mais aussi de la partie « temporelle ». Une étape d'alignement a ensuite lieu entre les données récupérées et la taxonomie. Chaque catégorie issue de l'étape d'alignement est assignée à un symbole graphique représentant l'évènement sur la timeline. Ces symboles ont différentes formes et couleurs afin de représenter :

1. La longueur de l'évènement
2. Le degré d'incertitude de la donnée (clinique et temporelle)
3. La nature de l'évènement (début de la prise du médicament, arrêt, etc.)

Les différentes combinaisons de la taxonomie sont présentées dans le tableau ci-dessous.

Event Category 3: Medication event happened over a period of time.

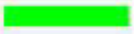
Taxonomy Group	Start Time	Stop Time	Active Usage	Examples from patient records	Illustration
1	Known	Known	Active	Hospital course 04/03/1999 - 05/11/1999: Her diabetes was brought under control with Glyburide 20 mg q day	
2	Known	Known	Uncertain	Ambulatory order 01/01/1997 - 01/24/1997: Tolinase 500 mg, p.o. (Specimen received)	
3	Known	Unknown	Active	She was diagnosed with chronic HTN after her last delivery which was complicated by PEC. She has been maintained on a variety of medications, and started regimen of Enalapril (10mg BID), Spironolactone (25mg qd) and lasix (40mg qd) on 10/04.	
4	Known	Unknown	Uncertain	Hospital course 01/31/2005 - 02/03/2005: patient was discharged home on Calcium, Magnesium, Vicodin and probably Ceftin for antibiosis.	
5	Unknown	Known	Active	In ER, 160/78 89 24 96% RA. She was given Prednisone 60 mg po qd, Albuterol, Atrovent and admitted.	
6	Unknown	Known	Uncertain	Pre-admission medications: Dilantin, Neurontin, Prednisone, multivitamin and Ibuprofen.	
7	Unknown	Unknown	Active	The patient has a history of hypertension in the past, previously on Lisinopril and Hydrochlorothiazide, recently diet-controlled, history of increased cholesterol on Mevacor.	
8	Unknown	Unknown	Uncertain	Patient previous medications may include Methotrexate and Fosamax.	

Figure 27 Tableau des iconographies des événements

c. CareCruiser

CareCruiser est un outil de visualisation de données dont l'approche est de fournir plusieurs vues simultanées des données (76). Ces multiples vues permettent de couvrir les différents aspects de la complexité des données patients. Cette approche porte à la fois sur les données des protocoles de soins et les données patient. Les vues calculées sont basées sur l'outil LifeLine et le langage Asbru, langage de représentation temporelle pour les plans thérapeutiques et les bonnes pratiques de soins (format XML). L'objectif de CareCruiser est d'intégrer et combiner différents types de données et de les présenter d'une façon cohérente pour être analysées par un humain. L'analyse de l'information ne se limite pas à l'analyse isolée des données brutes mais peut également prendre en compte le contexte lié aux différentes étapes du traitement suivi par le patient.

Pour visualiser l'exécution d'un plan thérapeutique, CareVis utilise 2 jeux de données :

- L'avancement sur la planification thérapeutique et les actes cliniques effectués
- L'état du patient à chaque étape du traitement

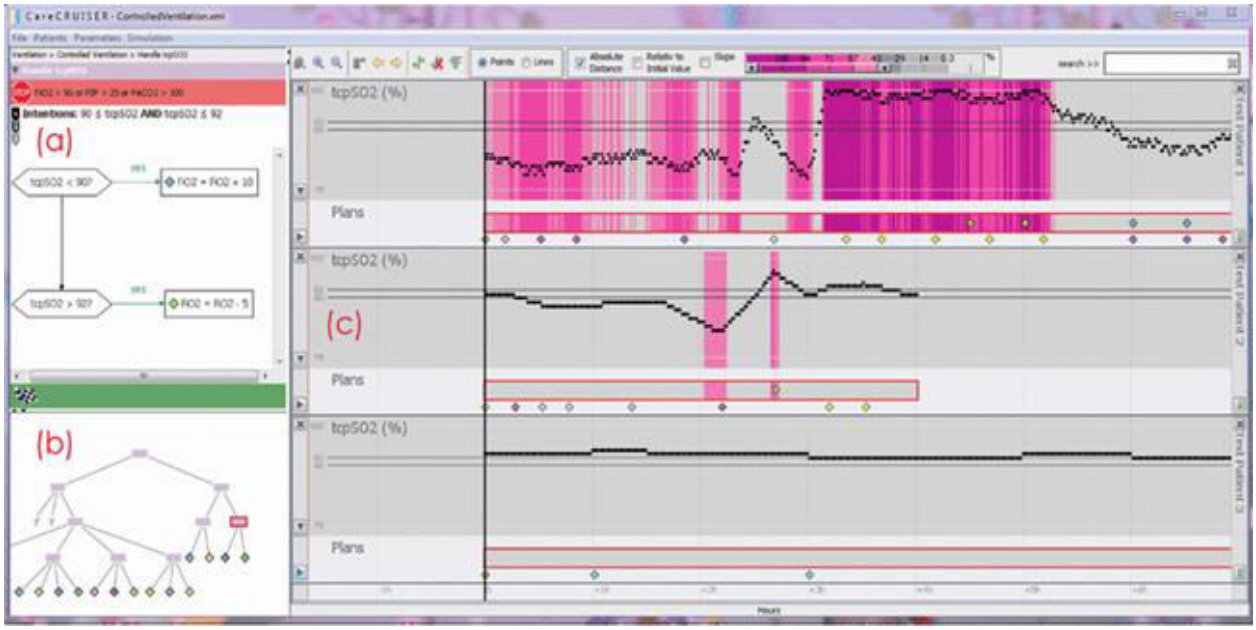


Figure 28 Interface de CareCruiser

Sur l'image ci-dessus est présentée l'interface utilisateur de CareCruiser :

- En (a) la vue logique représente la structure logique du plan thérapeutique avec une représentation en diagramme
- En (b) un graphique en forme d'arbre permet de visualiser la structure hiérarchique du plan thérapeutique (plan et sous plan)
- En (c) la vue temporelle représente à la fois les paramètres du patient, l'avancement dans l'exécution du plan thérapeutique et des actes cliniques

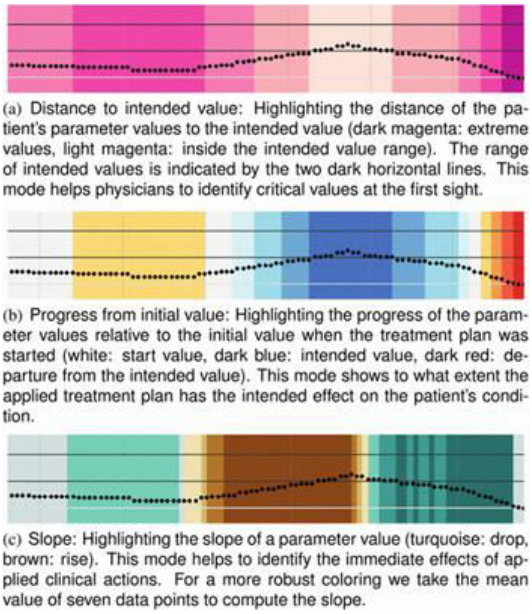
effectués sur le patient. Cette partie est agrandissable et l'utilisateur peut naviguer le long de la Timeline

En (c), on peut voir les données de 3 patients différents sur un même plan de traitement alignées verticalement. Les graphiques et les plans de traitement sont coloriés suivant la distance à la valeur attendue.

Les actes cliniques sont représentés sous forme de diamants sous les graphiques.

L'image ci-dessous permet de comprendre les différents modes de mise en valeur des effets des traitements.

- Sur (a), la mise en évidence des écarts de valeurs des paramètres du patient aux valeurs attendues permet d'identifier rapidement les valeurs critiques. L'intervalle des valeurs attendues est indiqué par les 2 lignes noires horizontales. La distance par rapport à cet intervalle est représentée par une échelle de saturation des couleurs (du magenta foncé pour les plus grands écarts au magenta pâle pour les plus petits)
- Sur (b), la coloration a pour référence la valeur initiale au début du traitement (en blanc). Les couleurs du jaune au rouge mettent en évidence une baisse par rapport à la valeur initiale tandis que les couleurs dans les tons bleus montrent une hausse. Ce mode de coloration permet de voir si le traitement a bien les effets escomptés.
- Sur (c), la coloration se fait en fonction du sens de variation des valeurs pour identifier les effets immédiats des actes cliniques.



(a) Distance to intended value: Highlighting the distance of the patient's parameter values to the intended value (dark magenta: extreme values, light magenta: inside the intended value range). The range of intended values is indicated by the two dark horizontal lines. This mode helps physicians to identify critical values at the first sight.

(b) Progress from initial value: Highlighting the progress of the parameter values relative to the initial value when the treatment plan was started (white: start value, dark blue: intended value, dark red: departure from the intended value). This mode shows to what extent the applied treatment plan has the intended effect on the patient's condition.

(c) Slope: Highlighting the slope of a parameter value (turquoise: drop, brown: rise). This mode helps to identify the immediate effects of applied clinical actions. For a more robust coloring we take the mean value of seven data points to compute the slope.

Figure 29 Illustration des modes de mises en valeur des effets des traitements sur CareCruiser

d. VisuExplore

VisuExplore est un système d'assistance aux médecins pour la fouille de données temporelles chez des patients atteints de maladies chroniques (74).

Les auteurs ont utilisé une approche centrée sur l'utilisateur impliquant les potentiels utilisateurs dès le début de la conception pour remplir ces 5 prérequis :

- Interface utilisateur simple (prise d'information rapide et non ambiguë)
- Flexible pour les différents types de variables
- Prise en compte de l'évolution des variables au cours du temps
- Interface interactive (zoom sémantique, mise à l'échelle, ouverture du document depuis la visualisation, inscrire des annotations)
- Pouvoir afficher plusieurs patients en même temps (étude de cohorte)

Les auteurs ont pris le parti de s'appuyer sur plusieurs techniques de visualisation connues. En effet, une seule technique ne peut pas être efficace pour tous les types de données utiles dans le domaine du diabète. VisuExplore utilise 4 techniques de visualisations :

- Les graphiques en ligne (voir v1 sur la capture d'écran ci-dessous) et en barre (voir v2) classiques (l'échelle l'axe des ordonnées se définit soit automatiquement grâce aux métadonnées soit manuellement par l'utilisateur aux valeurs extrêmes du jeu de données)
- Un graphique présentant des événements pour les données nominales (voir v3 : les diamants vides correspondent aux médicaments non prescrits et les diamants remplis aux prescriptions, les noms et dosage sont visibles en i2)
- Une timeline pour les données indiquant un intervalle de temps à la manière de LifeLines (voir v4)

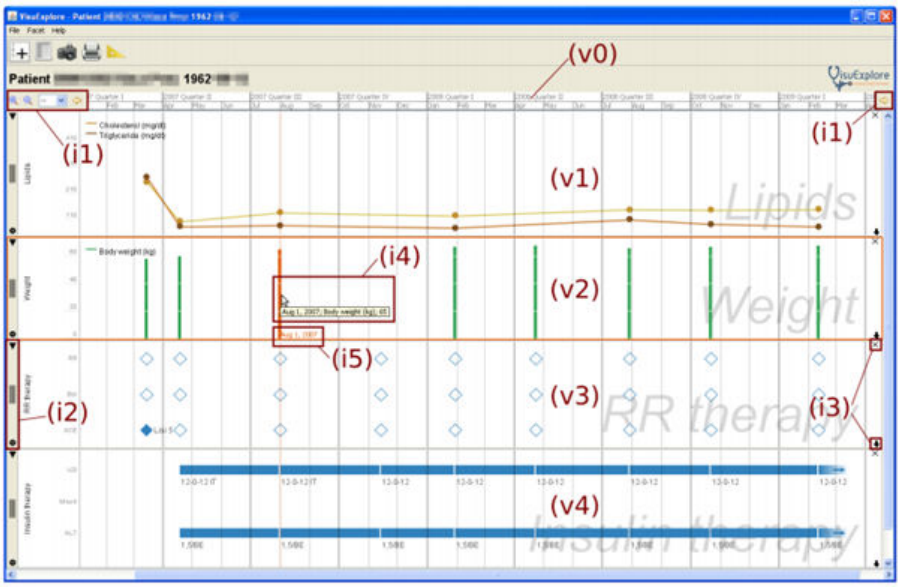


Figure 30 Interface de VisuExplore

L'état de l'art a permis d'identifier plusieurs exemples d'outils prenant en compte les données structurées comme textuelles. On peut noter les apports suivants de ces outils:

- Plusieurs modes de représentation des données en fonction de leur nature avec plusieurs modalités de visualisation des données
- Processus d'abstraction des données (numériques surtout)
- Navigation temporelle

Cependant, ces outils ont également présenté plusieurs limites :

- Ces outils ne sont pas développés spécifiquement pour la pharmacovigilance
- Leur Interface est souvent jugée complexe
- Les possibilités de navigation sont inexistantes ou se sont révélées peu intuitives lors de l'évaluation

3. Méthodes utilisées

a. Développement de l'interface

La conception de la timeline s'est basée sur des règles de sémiologies graphiques simples. Les variables visuelles et les types de données mettant en évidence des différences qualitatives ou quantitatives sont basés sur les travaux de Bertin (1967). Le tableau ci-contre présente les règles de visualisation des données en fonction des variables visuelles (formes, couleurs, surface, etc…). Ces règles permettent de mieux mettre en évidence les différences d'ordre qualitatives et quantitatives des données. Le développement de l'interface a débuté pour le projet ANR RAVEL (78) dont un des objectifs était de développer des méthodes de recherche et de représentation multimodales appliquées aux données médicales. Le développement du projet s'appuyait sur des cas d'usage de rhumatologie, de cancérologie, et de pharmacovigilance. Le prototype de timeline a donc été réalisé en recueillant les besoins des utilisateurs, aux travers de différents entretiens. Il a également été guidé par les conseils d'une spécialiste en conception d'interface homme-machine, Catherine Plaisant. Ces entretiens ont permis d'éviter des problèmes de conception et d'accélérer le développement du prototype.

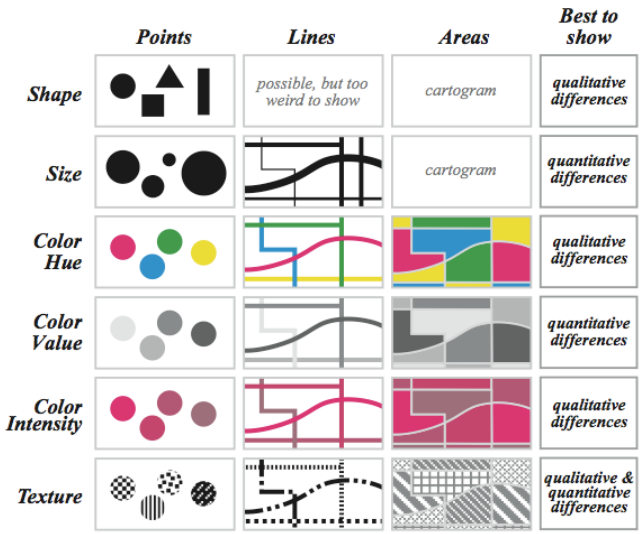


Figure 31 Illustration des règles de sémiologie graphique

Pour représenter efficacement les éléments de données sur la timeline, il est apparu pertinent d'utiliser le langage VCM.

Un alignement sur les codes de la classification CIM-10 est proposé avec VCM. Il a été choisi d'utiliser cet alignement pour afficher les données CIM-10 issues du PMSI sur la timeline.

Un exemple d'iconique représentant des troubles du rythme cardiaque est proposé ci-dessous.

Figure 32 Exemple d'icône VCM

b. Modèle de données

La structure de données nécessaires pour l'affichage dans la timeline est illustrée sur le schéma ci-dessous. Tout événement est associé à une information temporelle (au moins une date de début et une date de fin).

Type d'évènement	Classe	Liste des événements {e}
PARACETAMOL	ATC	{date : 15/10/2014, datefin: 31/12/2012,...},{...}
Hémoglobine	Labo	{date : 05/05/2014, value: 45,...},{...}
CR Hospitalisation	Texte	{date : 14/06/2014, text: La patiente admise…,...},{...}

Figure 33 Modèle de données

Un type d'évènement correspond à une ligne sur la timeline. Un type d'évènement peut correspondre à un code PMSI, un type de document, ou encore à une catégorie de prescription. La classe permet de caractériser le type d'évènement : un code CIM-10 correspond à un diagnostic, un document à du texte, etc… Les éléments de même classe sont ensuite regroupés entre eux. Pour chaque type d'évènement correspond une liste d'évènement avec les dates et d'autres données permettant de les placer sur la timeline.

c. Structure des terminologies

Les données structurées sont nativement codées ou secondairement alignées sur une terminologie hiérarchisée. Tous les codes appartiennent ainsi à des terminologies qui prennent la forme d'arborescence. Chaque code correspond à une place (c'est-à-dire un nœud) dans l'arborescence. Il est possible à partir d'un nœud fils de remonter sur le nœud parent. Ceci permet d'effectuer des regroupements, d'agréger, et ainsi de diminuer le nombre d'objets visuels à l'écran.

Les terminologies utilisées pour l'affichage des données sont au nombre de quatre dans la timeline :

- Le système de classification Anatomique, Thérapeutique et Chimique (ATC)
- La Classification statistique internationale des maladies et des problèmes de santé connexes, 10e révision (CIM-10)

- La Classification Commune des Actes Médicaux (CCAM)
- La terminologique locale du laboratoire

L'arborescence des terminologies est extraite à partir de l'entrepôt de données eHOP. eHOP sert de squelette pour la formation de l'arborescence. Pour obtenir l'arborescence finie des terminologies à ce patient, une étape de correspondance est réalisée pour ne retenir que les parties utiles de l'arborescence initiale.

4. Choix architecturaux

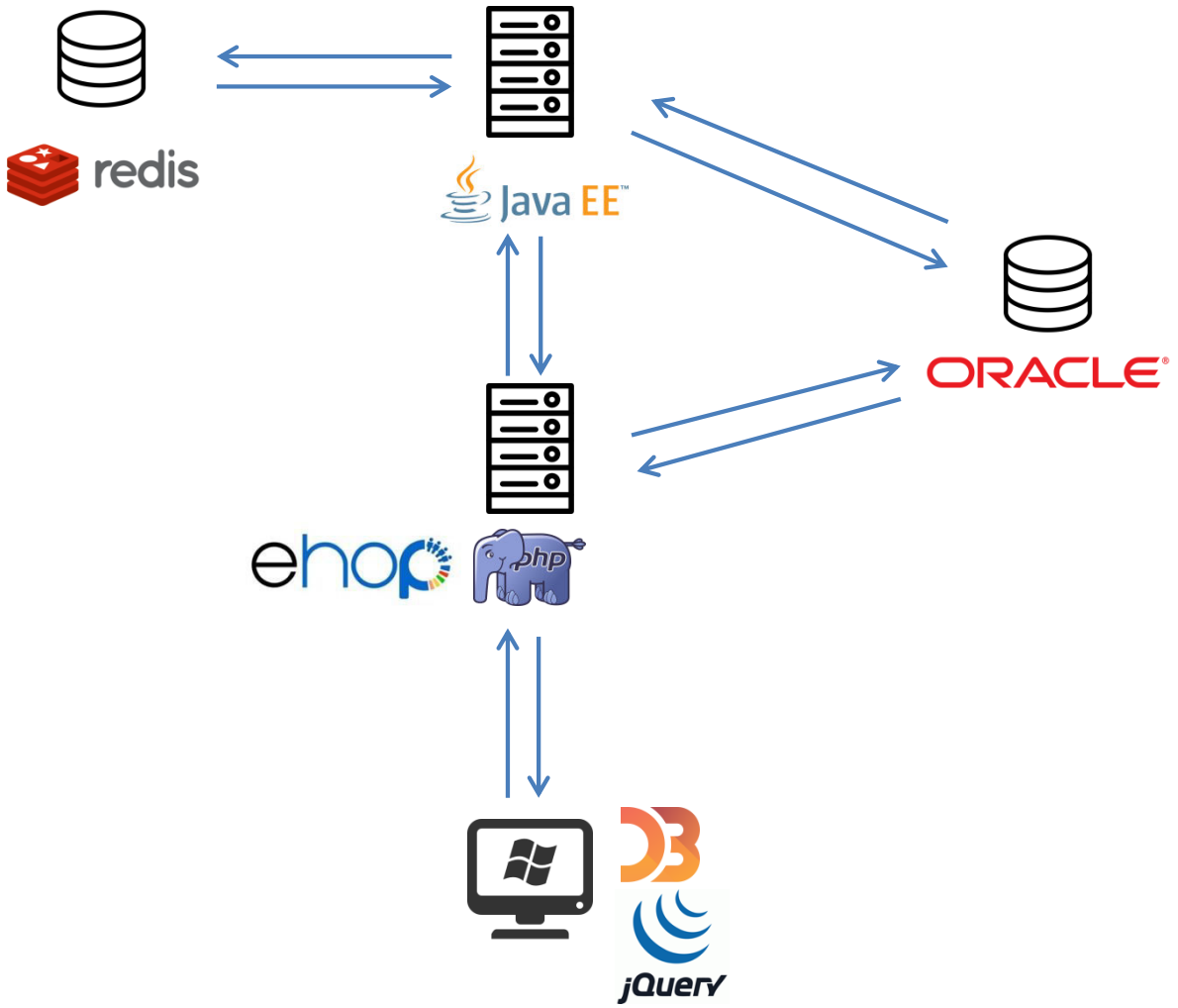


Figure 34 Archirecture de la timeline

a. Front-end

Le client de l'application a été implémenté en JavaScript. La libraire jQuery a été utilisée pour structurer le code tandis que la librairie d3js a été utilisée pour les aspects interaction et visualisation de l'application. Le client communique directement avec le serveur d'eHOP.

b. Back-end

La partie serveur a été implémentée sous la forme d'un web service Java. Celui-ci communique via le protocole SOAP avec le serveur d'eHOP écrit en PHP. Ces

interactions permettent de vérifier si l'utilisateur a bien les droits d'accès aux données du patient demandé. Ces droits d'accès sont établis en fonction du rôle de l'utilisateur (administrateur, pharmacovigilant). Le service Java charge les données du patient à partir de la base de données eHOP, les transforme dans le format de données adapté et stocke les résultats sous la forme d'un objet JSON (JavaScript Object Notation) dans une base de données REDIS. Ce système de cache permet d'éviter le retraitement systématique des données d'un patient ayant déjà été demandé par l'utilisateur.

5. Description de l'application

a. Description de l'interface

Le développement de cette timeline est inspiré de la timeline proposée par l'équipe de Catherine Plaisant, dans l'ebook InspiredEHR. (81) Une capture d'écran est présentée ci-dessous avec les différents composants annotés.

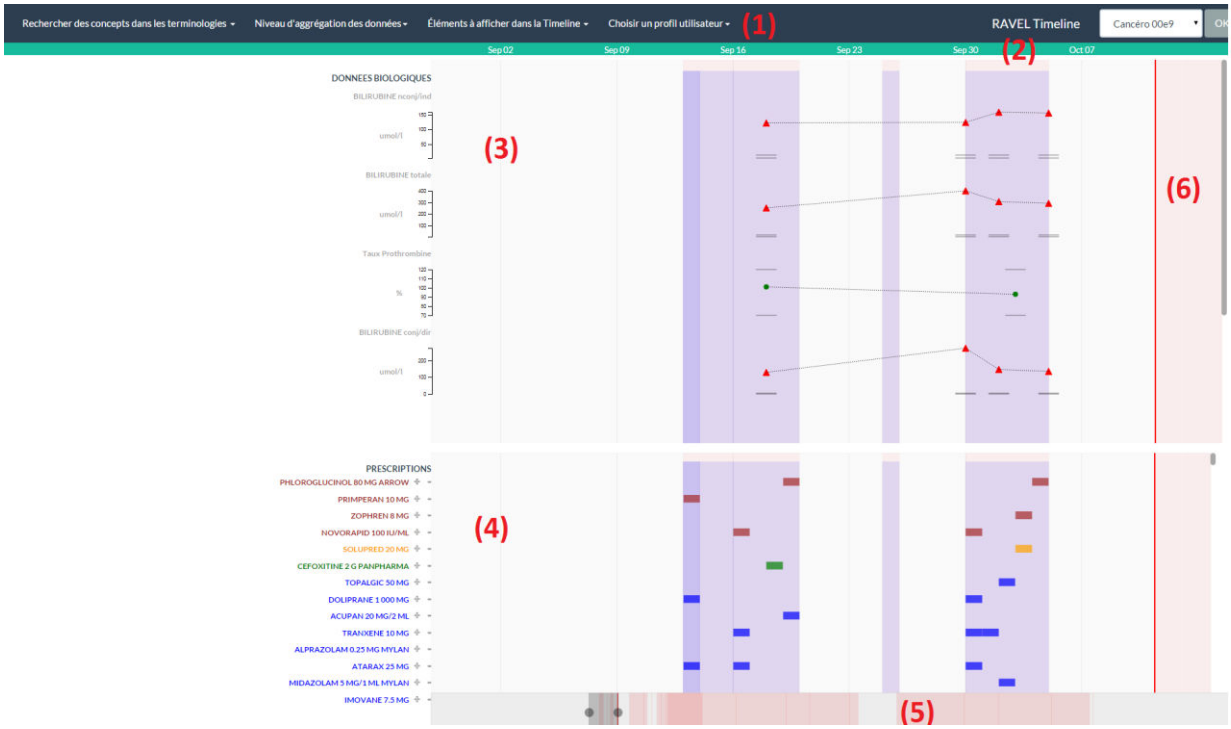


Figure 35 Interface de la timeline

Les différents composants de l'interface sont les suivants :

1) Le bandeau de navigation. Celui-ci permet de choisir le patient à afficher, sélectionner les données, ou encore rechercher des concepts
2) La bande verte est une frise chronologique permettant de repérer les dates sur la timeline
3) Un bloc pour l'affichage des données biologiques du patient
4) Un bloc pour l'affichage des prescriptions

5) Une mini-timeline permettant de naviguer dans le temps
6) La ligne verticale permet de choisir d'afficher les détails pour une date précise

Pour répondre aux problèmes de l'hétérogénéité des données, chaque type de données a son propre type de représentation :

- Les prescriptions sont représentées par un rectangle. Le côté gauche du rectangle représente le début de la prescription et le côté droit la fin. Une couleur arbitraire a été assignée à chaque classe ATC de premier niveau pour faciliter la visualisation. Au survol d'un rectangle, une pancarte de prescription s'affiche avec des informations complémentaires (dosage, U.F. prescriptrice,...)

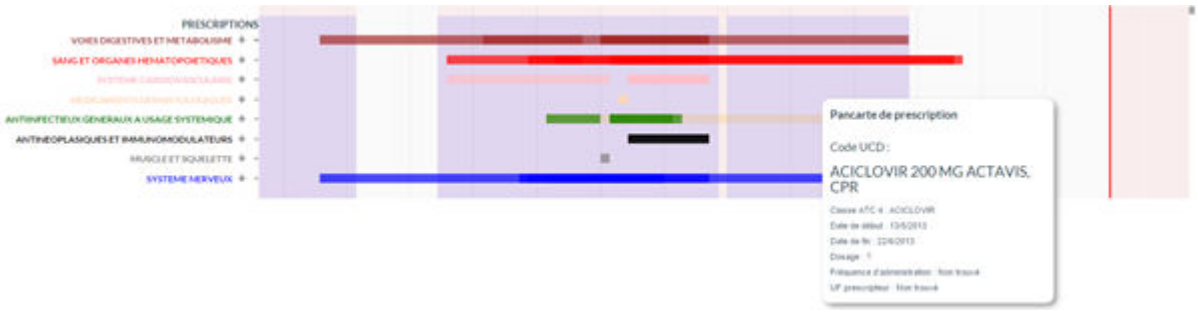


Figure 36 Blocs de prescriptions

- Comme montré dans la partie concernant les méthodes de visualisation, les diagnostics sont représentés par les icônes du langage VCM

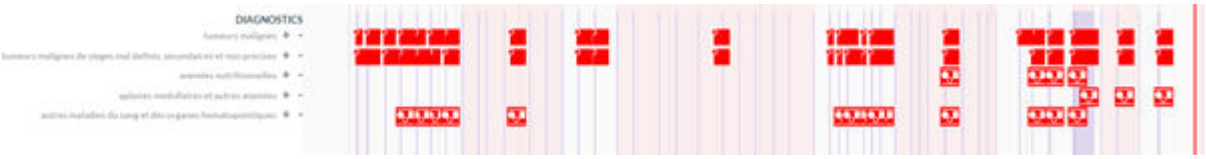


Figure 37 Bloc des diagnostics

- Les actes médicaux sont représentés par des cercles verts

Figure 38 Bloc des actes

- Les documents sont représentés par des icônes de document habituellement utilisés dans les représentations des interfaces homme-machine. L'utilisateur

peut cliquer sur le document pour afficher le contenu complet du document. Au survol de l'icône du document, un nuage de mots s'affiche. Ces mots correspondent aux concepts indexés par le processus de TAL. Ils s'affichent en deux couleurs différentes : vert pour les concepts indexés "positivement" et rouge pour les termes indexés « négativement » (exemple : « le patient ne présente pas d'ascite »). La taille des mots est proportionnelle à leur tf*idf pour le document.

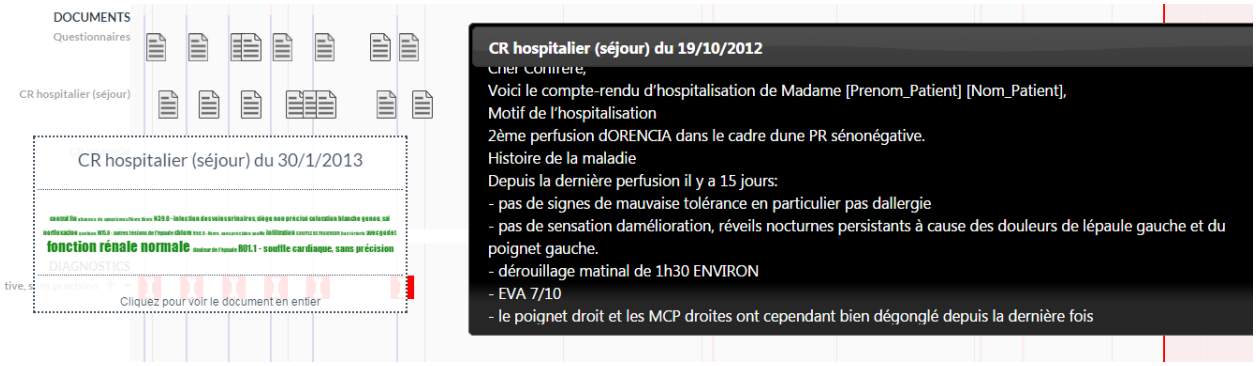


Figure 39 Bloc des documents textuels

- Les analyses biologiques sont des données numériques et sont représentés par des courbes. Les unités sont affichées sur la gauche du graphique. Pour permettre au clinicien de distinguer rapidement l'évolution des variables biologiques du patient, les valeurs normales apparaissent sous forme de cercles verts tandis que les valeurs en dehors des bornes sont affichées sous forme de rectangle rouge. Les bornes biologiques, qui peuvent évoluer dans le temps, prennent la forme de traits fins gris pour permettre de quantifier l'écart par rapport aux bornes de la normale. Sur la capture d'écran ci-dessous, on peut voir que le patient a un taux de prothrombine dans la normale alors que son taux de bilirubine conjuguée est largement au-dessus de la normale.

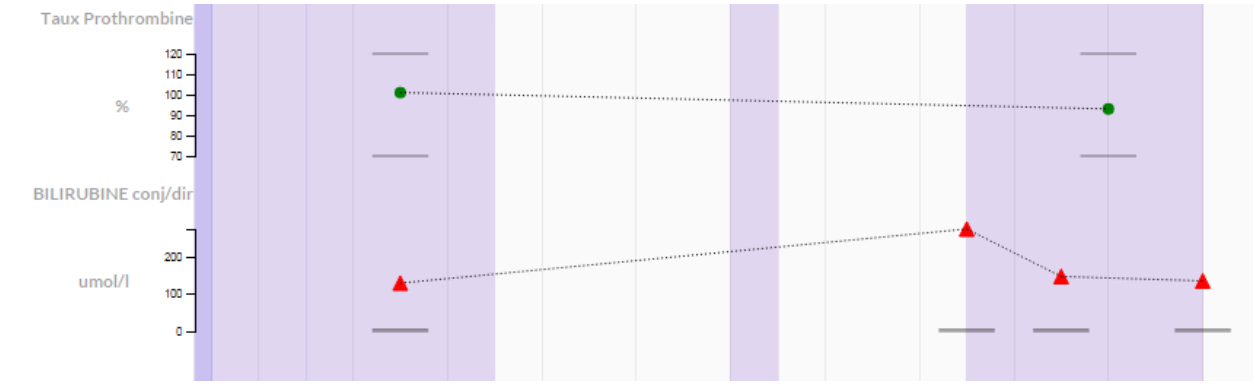


Figure 40 Bloc des analyses biologiques

- Les passages dans les séjours et les unités médicales sont représentés respectivement par des rectangles oranges et violets prenant toutes la hauteur de la timeline.

b. Navigation

De nombreux éléments peuvent être affichés à l'écran simultanément. Comme dit précédemment, il est nécessaire de naviguer à travers cette grande quantité d'informations pour rendre l'affichage pertinent pour un clinicien.

i. Navigation temporelle

L'aspect temporel dans une timeline est bien évidemment primordial. La navigation temporelle peut se faire de deux façons :

- En élargissant/rétrécissant la fenêtre d'intérêt, c'est à dire en approchant ou en éloignant les dates de début et de fin de la timeline.
- En changeant les dates de début et de fin de la timeline sans modifier l'intervalle.

Pour que l'utilisateur puisse réaliser cette tâche de navigation de façon intuitive, une "mini timeline" est implémentée dans l'interface comme illustrée sur l'image ci-dessous.

Figure 41 Mini-timeline pour la navigation temporelle

La "zone de contexte", autrement dit la fenêtre temporelle affichée sur la timeline, est ici sous forme du rectangle translucide gris. L'utilisateur peut faire glisser ce rectangle pour changer les dates de début et de fin de la timeline. Il peut également utiliser les ronds gris présents sur les côtés du rectangle pour agrandir ou rétrécir cette fenêtreou bien utiliser la molette de la souris. Les séjours hospitaliers du patient sont représentés sous forme de rectangles roses pour faciliter le repérage des zones d'intérêt et leur succession.

ii. Filtrage par type de données

L'utilisateur a la possibilité de choisir le type de données qu'il veut afficher dans la timeline. Un menu lui est présenté dans le menu de navigation sous la forme de case à cocher comme illustrée sur l'image ci-contre. Si la case est cochée et que le dossier du patient contient le type de données associé à la spécialité, le bloc correspondant s'affichera sur la timeline. Cette fonctionnalité est utile pour ne pas afficher de bloc prenant de la place sur l'écran, comme c'est le cas pour les données biologiques par exemple. De même, l'affichage du bloc de rhumatologie ne sera pas utile pour un cancérologue.

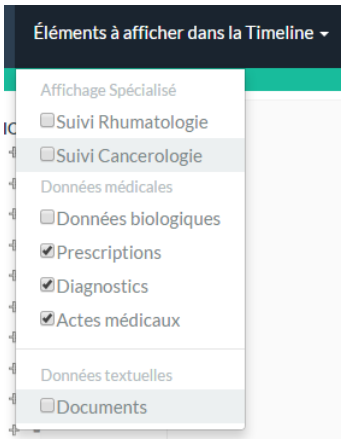


Figure 42 Filtrage par type de données

iii. Recherche et sélection de concept dans les terminologies

Il est possible de rechercher des concepts dans les trois terminologies utilisées. Le système de recherche permet de choisir précisément les données structurées à afficher dans la timeline. Dans l'exemple proposé ci-contre, l'utilisateur recherche le terme "pancrea" et retrouve des concepts à la fois de la CIM-10 et de la CCAM.

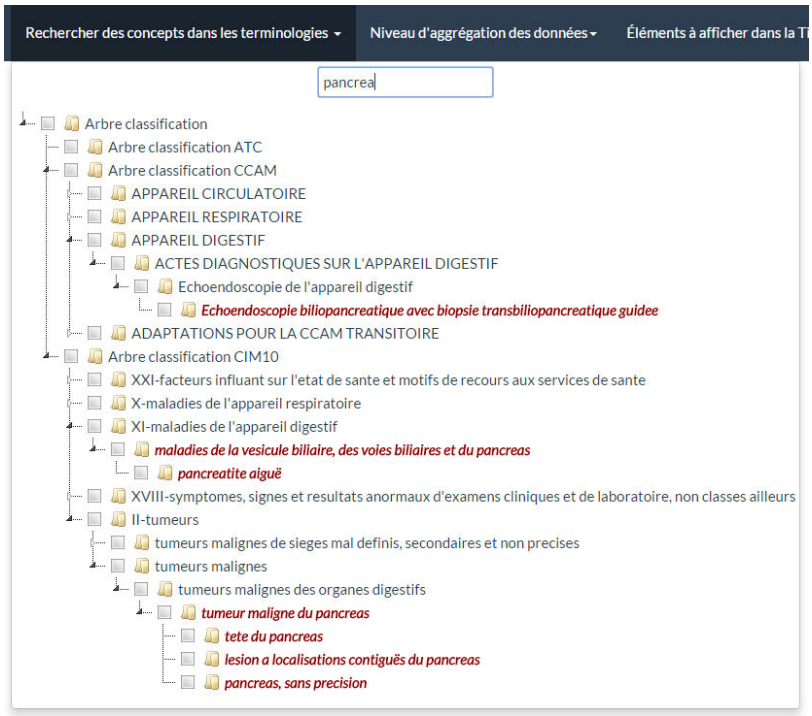


Figure 43 Menu de recherche de concepts dans les terminologies

iv. Filtrage par spécialité

Le filtrage par spécialité combine les deux types de filtrages vu précédemment. Il est prédéfini pour chaque type de spécialité un ensemble de codes dans les terminologies :

- Pour la classification ATC, la sélection des codes se fait sur les groupes anatomiques concernés par la spécialité.
- Pour les codes CIM-10, la sélection est faite sur les têtes de chapitres de la classification.

- Pour la classification CCAM, la sélection se fait selon la topographie anatomique. Les actes les plus pratiqués dans la spécialité (information extraite du site institutionnel ameli.fr) sont ajoutés dans cette sélection.

Les blocs spécifiques à la spécialité (rhumatologie ou cancérologie) sont aussi affichés ou non selon la sélection.

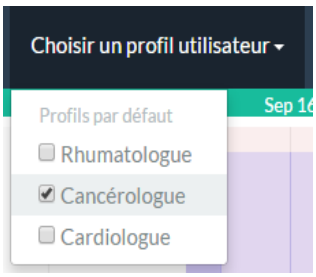


Figure 44 Menu de choix du profil utilisateur

v. *Agrégation des données structurées*

Dans les différentes classifications, l'information peut paraître redondante d'un point de vue clinique. Par exemple, au niveau UCD, Paracétamol et Doliprane ont deux codes différents et seront donc sur deux lignes différentes sur la timeline (médicament générique et médicament princeps). Il s'agit pourtant de la même molécule. L'utilisateur a la possibilité d'agréger ces 2 données pour pouvoir l'afficher sur une seule ligne en cliquant sur le (-) à côté du nom de la ligne (type d'événement). Au survol, il sera affiché avec quelle autre donnée la donnée sera agrégée. Il est possible d'inverser l'opération en cliquant sur le (+) sur une ligne contenant différents types de données. Cette agrégation se fait en remontant dans l'arbre de la classification.

L'exemple ci-dessous illustre cette fonctionnalité. La donnée source reste à disposition au survol de l'item.

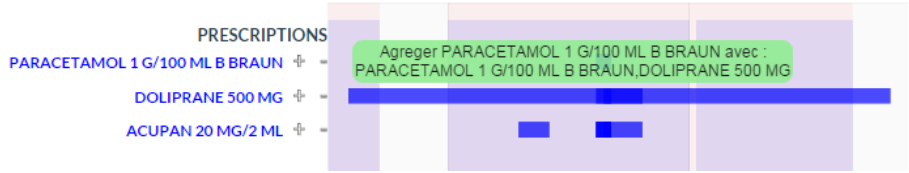


Figure 45 Prescriptions avant agrégation

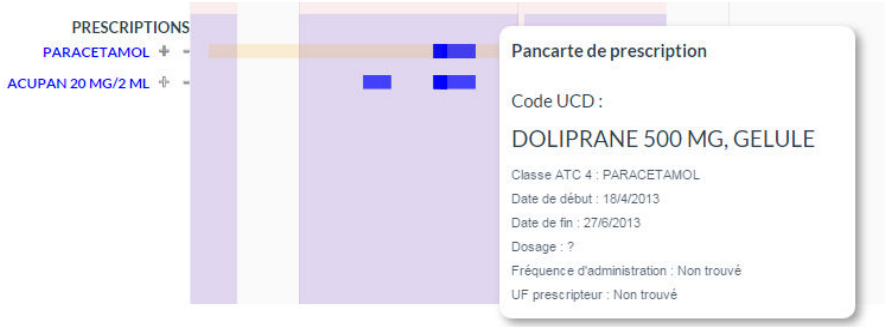


Figure 46 Prescriptions après agrégation

L'utilisateur peut aussi choisir le niveau de granularité des données grâce au menu "Niveau d'agrégation des données". Il peut choisir le niveau d'agrégation globale des données.

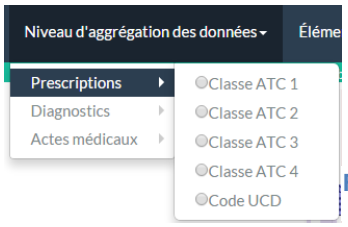


Figure 47 Menu d'agrégation globale

Cette fonction d'agrégation est possible sur les prescriptions, sur les diagnostics, et sur les actes médicaux.

vi. Recherche dans les documents textuels

Pour faciliter la navigation dans les documents ainsi que la recherche d'information, trois champs de recherche sont proposés à l'utilisateur :

- Un champ de recherche en texte libre. Si un document contient le terme recherché, son icône sur la timeline sera coloriée en bleu. Si l'utilisateur clique sur le document complet, le terme dans le champ de recherche sera surligné dans le texte comme montré dans l'image ci-dessous.
- Les champs de recherche « Index + » « et Index – » permettent de rechercher des termes indexés

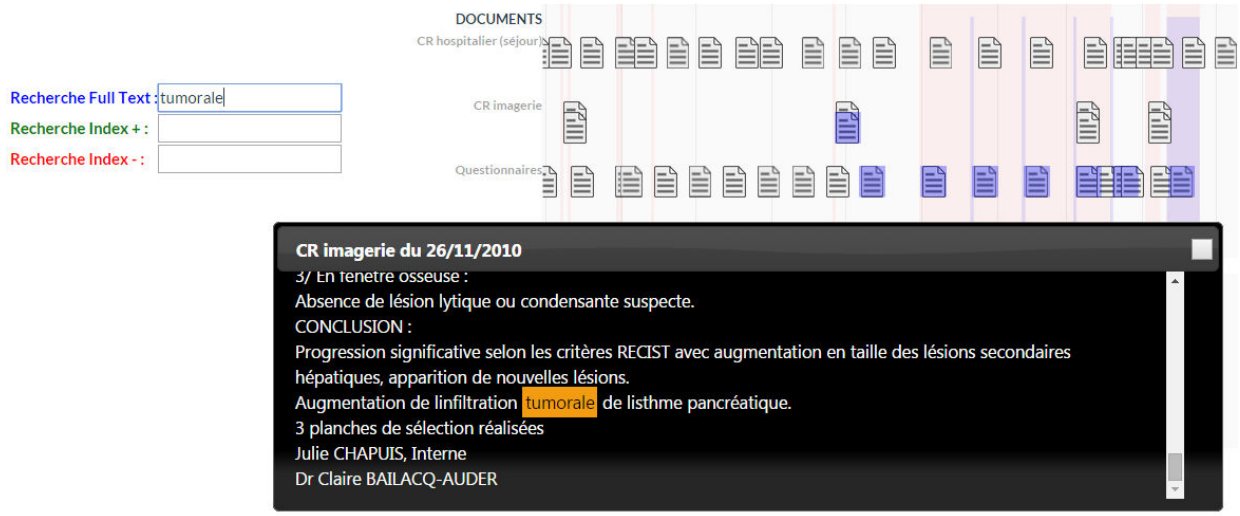


Figure 48 Illustration de la recherche dans les documents textuels

6. Evaluation de l'application

a. Evaluation de l'utilisabilité

i. Matériel et méthode

1) Participants

Les utilisateurs experts ont été recrutés parmi une équipe de pharmaciens et de médecins qui travaillent au Centre Régional de Pharmacovigilance de Rennes et qui font partie du public cible de notre outil. Dans la littérature (82), il a été déterminé que cinq utilisateurs étaient suffisants pour identifier 80% des problèmes liés à la facilité d'utilisation de l'application. (83) Enfin, six personnes (cinq femmes et un homme, dont cinq pharmaciens et un médecin) ont pu être recrutées pour l'étude d'utilisabilité, et trois d'entre elles ont également participé à l'étude de quantification de l'impact. Ces participants effectuent régulièrement des activités pratiques de pharmacovigilance. Les six personnes n'ont reçu aucune compensation financière en échange de leur participation. L'âge moyen était de 31,3 ± 4,5 ans (de 25 à 40 ans). Tous utilisaient régulièrement des ordinateurs dans le cadre de leurs activités de pharmacovigilance, en particulier pour étudier les dossiers des patients.

2) Méthode d'évaluation de l'utilisabilité

L'Organisation internationale de normalisation (ISO) définit l'utilisabilité comme "la mesure dans laquelle un produit peut être utilisé par des utilisateurs spécifiés pour atteindre des objectifs spécifiés avec efficacité, efficience et satisfaction dans un contexte d'utilisation spécifié". (84) Il existe de nombreuses méthodes d'évaluation de l'utilisabilité (17) et plusieurs sont utilisées dans le développement et l'évaluation de logiciels liés à la santé (85) (86) (87). Nous avons choisi l'approche « *think-aloud protocol* » qui a été décrite par Nielsen comme "la méthode d'évaluation la plus fiable pour obtenir une estimation d'utilisabilité". (88) Elle est basée sur l'enregistrement de l'opinion des participants pendant qu'ils exécutent une série de tâches conçues spécifiquement pour l'évaluation de l'utilisabilité. (89) On demande aux participants d'exprimer oralement leurs pensées, leurs perceptions et leurs opinions tout en interagissant avec l'application (méthode de réflexion simultanée à haute voix). La performance des participants pendant l'exécution de ces tâches est ensuite analysée par l'évaluateur. Cette méthode peut être appliquée à toutes les étapes du développement de l'application Web et fournit des données qualitatives qui aideront les évaluateurs à identifier les problèmes d'utilisabilité. Il peut également déclencher le développement de nouvelles fonctionnalités pour l'interface en réponse aux commentaires des utilisateurs. La gravité des problèmes d'utilisabilité rencontrés par les utilisateurs a été évaluée à l'aide de l'échelle de Nielsen (de 0=pas de problème d'utilisabilité à 4=catastrophe pour l'utilisabilité). (90) Peute et al (91) ont montré que par rapport à la méthode rétrospective « *think aloud* », la méthode concomitante augmente la verbalisation des participants et conduit à un taux plus élevé de détection des problèmes d'utilisabilité. De plus, cette méthode semble plus fiable que l'analyse rétrospective de l'enregistrement du test, réalisée en présence des utilisateurs.

3) Cas d'usage

Un cas d'effet indésirable médicamenteux (EIM) choisi en collaboration avec le responsable du Centre de Pharmacovigilance, a été utilisé pour tester toutes les caractéristiques du prototype en situation réelle. Il a été demandé aux participants de déterminer si l'effet indésirable subi par un patient admis aux soins intensifs résultait d'une exposition à un antibiotique. Pour ce faire, ils devaient accomplir quatre tâches: i) trouver le nom et l'adresse du patient, ii) déterminer la nature exacte de l'effet indésirable et recueillir tous les renseignements pertinents, iii) trouver des renseignements sur le médicament soupçonné (nom et dose, par exemple) et iv) recueillir des renseignements sur le résultat de l'EIM.

4) Déroulement de l'évaluation

Avant le test d'utilisabilité, tous les utilisateurs ont suivi, en même temps, une formation pour se familiariser avec les fonctionnalités de la timeline pendant environ une heure. Pour cette session de formation, des données fictives sur les patients ont été créées dans le même format que les données réelles afin de pouvoir montrer/tester les caractéristiques de la chronologie sans avoir à accéder aux données réelles sur les patients. Au cours de cette séance de formation, les participants pouvaient visualiser les données d'un patient fictif dans la chronologie (par exemple les ordonnances de médicaments, les données de laboratoire, les rapports) et utiliser chaque fonctionnalité au moins une fois. Si un utilisateur rencontrait un problème, le formateur l'aidait à surmonter ce problème. Cette formation et le test d'utilisabilité du prototype ont eu lieu au Laboratoire d'Informatique Médicale de l'Université de Rennes. L'environnement de travail (système d'exploitation, ordinateurs, etc.) était le même que celui du Centre de pharmacovigilance. Des tests d'utilisabilité (six séances avec un participant par séance) ont été effectués en présence de l'évaluateur, qui a observé les interactions entre le participant et l'interface et noté ses observations sur papier. Chaque participant a reçu un carnet de notes pour enregistrer toutes les données pertinentes récupérées sur le cas d'utilisation. Les séances de tests d'utilisabilité ont été enregistrées sur vidéo, ainsi que toutes les interactions avec l'interface (saisie au clavier et mouvements de souris au sein de l'application). L'enregistrement et l'analyse subséquente ont été réalisés à l'aide de MORAE (TechSmith, version logicielle 3.3.3.4). Les enregistrements vidéo ont été analysés après que tous les participants aient accompli toutes les tâches. Le logiciel MORAE peut capturer des séquences vidéo à l'écran, tout en enregistrant simultanément les interactions entre le participant et l'interface graphique. De plus, le visage des participants a également été enregistré, ce qui a permis aux évaluateurs d'évaluer, rétrospectivement, leur expression et leurs réactions tout en interagissant avec le prototype.

A la fin du test d'utilisabilité, chaque participant a rempli un questionnaire System and Usability Scale (SUS), traduit en français. Ce questionnaire permet de mesurer l'utilisabilité de l'application de manière rapide et fiable (92). Enfin, chaque participant a été interrogé à l'aide de questions ouvertes afin de déterminer les caractéristiques de la timeline qu'il appréciait le plus et s'il avait rencontré des problèmes particuliers.

ii. Résultats

1) Score SUS et autres métriques

La note SUS moyenne des participants était de 82,5 sur 100 points. Après avoir répondu au questionnaire, tous les participants ont exprimé le souhait d'utiliser la timeline dans leur pratique quotidienne. Tous les participants ont pu exécuter avec succès les tâches requises pour le cas d'utilisation. Le temps de traitement des cas d'utilisation variait grandement d'un participant à l'autre (temps de traitement moyen = 24,44 ± 9,97 minutes). Le nombre de clics était également variable parmi les participants (nombre moyen de clics : 174 ± 92,41, P = 0,06).

La distribution des erreurs dans le traitement de l'interface ne variait pas beaucoup d'un participant à l'autre (nombre moyen d'erreurs : 1,33 ± 1,5). La tâche de recueillir de l'information sur les résultats de l'EIM comportait le plus grand nombre d'erreurs (P = 0,16).

2) Problèmes rencontrés

Le tableau 1 énumère les problèmes d'utilisabilité rencontrés par les participants pendant le test. La gravité de chaque problème d'utilisabilité a été évaluée à l'aide de l'échelle de Nielsen. (90)

Tableau 1 – Liste des problèmes d'utilisabilité rencontrés

Problème	**Source du problème**	**Fréquence du problème**	**Score sur l'échelle de Nielsen**	**Observations**
Fonctionnalité d'augmentation/diminution du zoom jugée insuffisamment précise par les participants.	Interface utilisateur (IU)	6/6 (100%)	3/4	L'aperçu de la timeline ne permet pas un zoom précis pour sélectionner la plage de vue détaillée.
Absence des données	Source des données	6/6 (100%)	3/4	Les prescriptions n'ont pas été informatisées dans le service des soins intensifs.
La mise en surbrillance du texte n'a pas fonctionné pendant la recherche.	IU	1/6 (17%)	3/4	Problèmes avec la détection des accents

Les administrations médicamenteuses n'ont pas été affichées.	IU	6/6 (100%)	3/4	Les données sur les administrations médicamenteuses n'étaient pas disponibles
Arbre hiérarchique des concepts biologiques : problèmes de navigation	Complexité de la terminologie dans l'UI	3/6 (50%)	2/4	L'arborescence hiérarchique des résultats de laboratoire était trop grande pour permettre aux utilisateurs de trouver facilement l'information sans filtrage.
Repositionnement des fenêtres de documents	IU bug	2/6 (34%)	2/4	La fenêtre a glissé sous la barre de menu.

3) Corrections effectuées

L'évaluation de l'utilisabilité a permis d'améliorer l'interface utilisateur et d'ajouter de nouvelles fonctions conçues spécifiquement pour la pharmacovigilance, comme l'extraction automatique des noms des médicaments des documents textuels et la possibilité d'accéder, directement à partir de la ligne de temps, à un module de visualisation de la monographie du médicament.

iii. Discussion

L'évaluation de l'utilisabilité de la Timeline s'est montrée globalement positive avec un score SUS élevé (82.5). Les problèmes rencontrés lors de la séance ou mentionnés lors des entrevus par les utilisateurs ne remettent pas en cause la conception même de la Timeline mais ont révélé des problèmes d'ergonomie ou plus simplement des bogues logiciels pour la plupart liés au « jeune âge » de l'application. Ces bogues ont étés corrigés juste après cette évaluation et nous avons développé de nouvelles fonctionnalités pour parfaire les problèmes d'ergonomie : fonction de zoom à la molette de la souris, arborescence des terminologies scrollable. De plus, à la demande des utilisateurs nous avons rajouté quelques fonctionnalités spécifiques à leur métier, à savoir la recherche de noms de médicaments dans les différents textes ou encore l'affichage des administrations médicamenteuses. Cependant, le design de notre évaluation de l'utilisabilité rend difficile la comparaison du temps passés pour chaque tâches entre utilisateurs. En effet, certains participants ont accompli leurs tâches en parallèle, tandis que d'autres les ont exécutées l'une après l'autre. Cette

explication s'applique également à la différence dans le nombre de clics, car certains participants ont dû rechercher les mêmes informations plus d'une fois.

b. Evaluation de l'impact sur les pratiques

i. Matériel et méthode

1) Sélection des dossiers patients et randomisation

Pour étudier l'impact sur les pratiques, 743 cas potentiels issus des données d'eHOP sur l'année 2015 ont été sélectionnés en fonction d'une liste de codes CIM-10 potentiellement liés à I4EIM en fonction la méthode décrite par Osmont et al. (93) Cette liste a été définie dans une étude interne antérieure du Centre de pharmacovigilance basée sur les données de 2014. Seuls les codes CIM-10 ayant la plus forte probabilité de faire correspondre un EIM ont été sélectionnés (Figure 2). Ensuite, 85 cas ont été sélectionnés au hasard, afin d'avoir à peu près la même proportion des différents codes de la CIM-10 pour les trois participants et pour chaque méthode d'analyse (méthode habituelle par rapport à la timeline, voir ci-dessous). Un script Python a été utilisé pour randomiser les cas entre les trois participants.

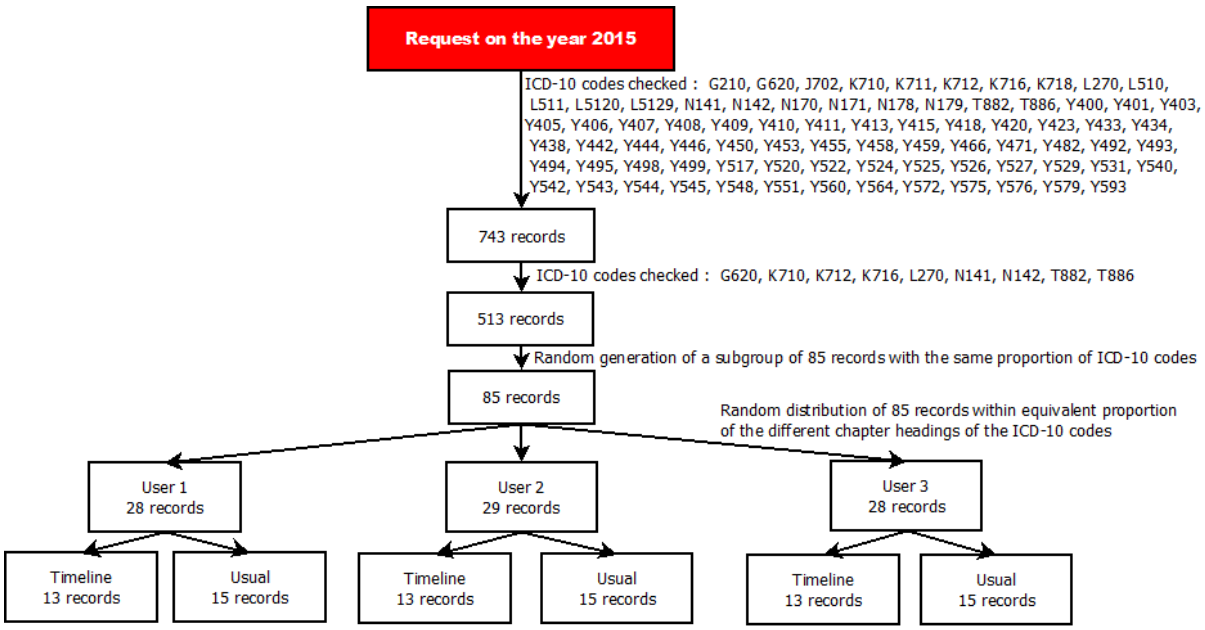


Figure 49 Schéma de randomisation des dossiers de l'évaluation

2) Design de l'étude

Ces 85 cas ont été examinés par trois membres du Centre Régional de Pharmacovigilance de Rennes selon deux méthodes : la "méthode habituelle" (DSE et logiciel DxCare®, voir Dossier complémentaire 2), et la méthode "timeline" (prototype de timeline). Les participants ont reçu une liste des dossiers des patients et des instructions sur la méthode à utiliser pour chaque dossier (Figure 49). L'environnement de travail (système d'exploitation, ordinateurs, etc.) était le même que celui du Centre de pharmacovigilance. Comme dans la pratique courante de pharmacovigilance, ils devaient recueillir de l'information pour établir un lien de causalité possible entre le diagnostic et l'administration d'un médicament. Ils ont également noté l'heure de

début et l'heure de fin pour le traitement de chaque fichier. Le processus était considéré comme terminé lorsque le participant avait récupéré toute l'information requise ou lorsqu'il estimait que cette information n'était pas présente ou ne pouvait être trouvée avec la méthode utilisée.

Après avoir examiné un cas, les participants ont rempli un questionnaire fait maison détaillant les informations suivantes : complexité du cas, informations sur l'ADR, quelles fonctions de la chronologie ont été utilisées, et commentaires généraux.

La qualité du rapport a été mesurée à l'aide d'une méthode basée sur le score de complétude vigiGrade de l'OMS qui inclut des indicateurs de qualité, tels que le nom du médicament ou la description de l'EIM. De plus, l'absence d'information a affecté le score obtenu. Un membre du Centre de pharmacovigilance qui n'a pas participé aux tests de la timeline a procédé à la notation.

Toutes les analyses ont été effectuées sur les 85 cas, mais pour la comparaison du score d'exhaustivité vigiGrade qui ne concernait que les cas considérés comme étant des ADR par les participants. Les différences temporelles entre les méthodes ont également été comparées selon les participants et selon le chapitre des codes de la CIM-10.

3) Analyses statistiques

Les variables quantitatives ont été exprimées sous forme de moyenne et d'écart-type. Les variables continues ont été comparées à l'aide du test t de Student ou du test de Wilcoxon, selon le cas, pour deux groupes, et avec le test Kruskal-Wallis pour plus de deux groupes. Les valeurs de $P < 0,05$ ont été jugées significatives. Les analyses ont été effectuées avec le logiciel statistique R, version 3.4 (94)

ii. Résultats

1) Comparaison du temps passé pour effectuer les tâches

Trois participants (deux pharmaciens et un médecin) ont été recrutés pour cette évaluation. Tous ont également participé à l'étude d'utilisabilité préalable.

Dans l'ensemble, le temps nécessaire à la réalisation de la tâche était comparable entre les méthodes (16' 29" et 15' 51" pour la méthode habituelle et la ligne du temps, respectivement ; $P = 0,38$) (Figure 50).

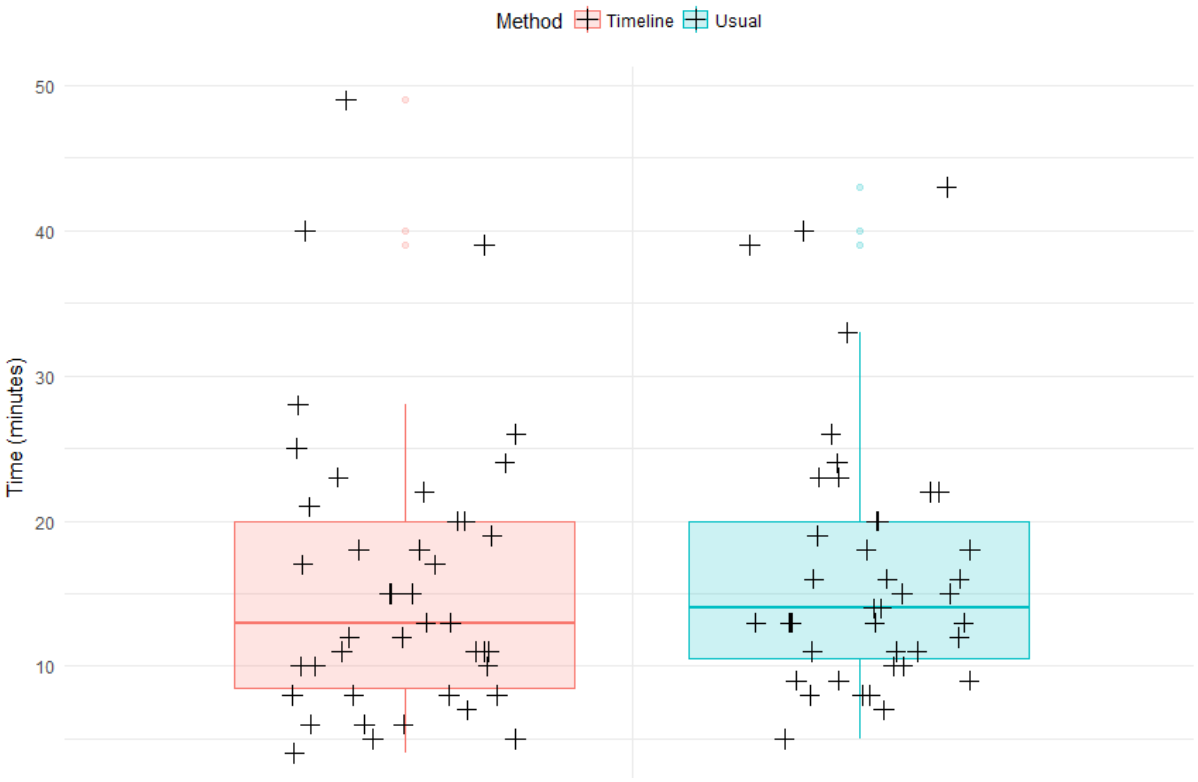


Figure 50 Box-plot affichant le temps passé par cas selon la méthode

La comparaison entre les trois participants a montré que pour l'utilisateur 1, le temps de réalisation était significativement plus court avec la ligne du temps qu'avec la méthode habituelle (P = 0,02). Aucune différence entre les méthodes n'a été observée pour les deux autres participants (Tableau 4 et Figure 50).

Tableau 4 Temps moyen passé par cas en fonction de la méthode et du participant

Method / Participant	Classique	Timeline	Difference	P-value
Participant 1	17' 26"	11' 38"	- 5' 48"	0.02
Participant 2	15' 55"	18' 11"	+ 2' 16"	0.71
Participant 3	16' 8"	17' 40"	+1' 32"	0.59

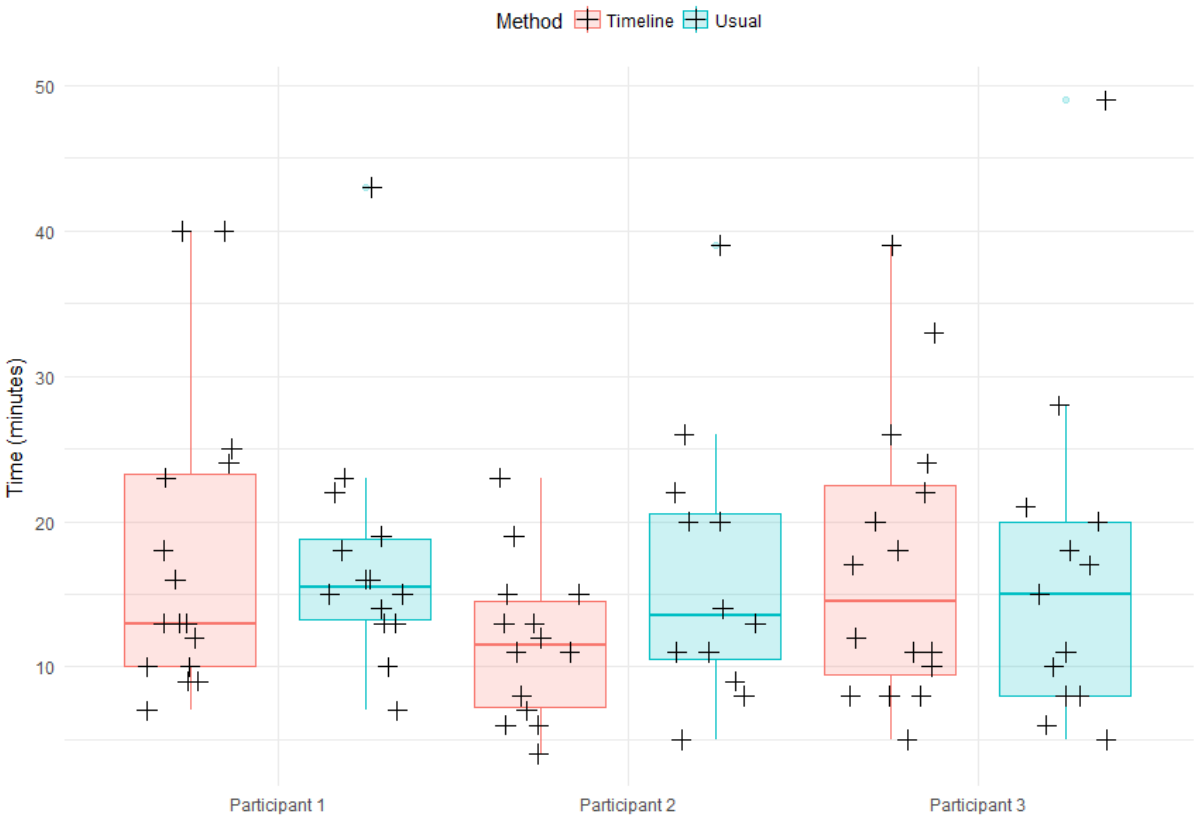


Figure 51 Box-plot montrant le temps passé par cas en fonction de la méthode et du participant

Enfin, le temps de réalisation a été comparé entre les méthodes et le chapitre de la CIM-10 (Tableau 4). Bien qu'il n'y ait pas de différence significative, le temps de réalisation a eu tendance à être légèrement plus rapide avec la méthode avec la timeline qu'avec la méthode habituelle seulement pour les cas concernant les maladies de l'appareil génito-urinaire (P=0,08). Cette différence pourrait s'expliquer par la nécessité systématique de consulter les résultats de laboratoire du patient pour juger de l'implication du médicament dans ce type d'effet indésirable.

Tableau 5 Temps moyen passé par dossier en fonction du chapitre de la CIM-10 et de la méthode utilisée

Chapitre CIM-10	Habituelle (n)	Timeline (n)	Différence
XIV - Maladies de l'appareil génito-urinaire	17' 36" (20)	13' 10" (18)	- 4' 26"
XII - Maladies de la peau et du tissu cellulaire sous-cutané	15' 50" (12)	15' 20" (9)	- 30"

XX - Causes externes de morbidité et de mortalité	15' 17" (7)	15' 15" (4)	- 2"
VI - Maladies du système nerveux	17' (5)	20' (9)	+ 3'
XI - Maladies de l'appareil digestif	13'	17' 45"	+ 4' 45"
XIX - Lésions traumatiques, empoisonnements et certaines autres conséquences de causes externes	12' 40"	14' 12"	+ 1' 32"

2) Comparaison de la qualité des données recueillies

Parmi les 85 cas sélectionnés, 68 correspondaient à un EIM réel. Par conséquent, la qualité des données n'a été évaluée et comparée que pour ces enregistrements à l'aide du score de complétude vigiGrade. Dans l'ensemble, la qualité moyenne des notifications produites avec les deux méthodes n'était pas significativement différente (P=0,49). Des résultats similaires ont également été obtenus en comparant la qualité des notifications entre les trois participants (Tableau 4).

Tableau 6 Qualité moyenne des notifications en fonction de la méthode et de l'utilisateur

Participant	Qualité moyenne avec la méthode habituelle	Qualité moyenne avec la timeline	Qualité moyenne globale
Participant n°1	0,87	0,81	0,84
Participant n° 2	0,81	0,85	0,84
Participant n° 3	0,81	0,74	0,78
Total	0,83	0,80	0,81

3) Résultats des interviews

Au cours des interviews, les participants ont exprimé le souhait d'utiliser la timeline dans leur pratique quotidienne. Les caractéristiques les plus appréciées étaient la fonctionnalité de recherche dans l'affichage du texte et des ordonnances. Les entrevues ont également mis en lumière certains points négatifs : des problèmes dans le choix de l'intervalle de temps de la ligne du temps et des difficultés dans le choix des éléments à afficher dans la ligne du temps. De plus, les participants souhaitaient voir les administrations médicamenteuses en plus des prescriptions.

Après l'évaluation de l'impact de la timeline sur la pratique de la pharmacovigilance, les participants ont réitéré leur intérêt à obtenir la demande d'utilisation dans leur pratique quotidienne. Ils ont déclaré que la timeline leur permettait de gagner du temps dans tous les cas qui nécessitaient des résultats de laboratoire de consultation. De plus, ils ont ajouté que la possibilité d'afficher les données de prescription et les informations sur l'administration des médicaments dans un format chronologique était un avantage considérable.

iii. Discussion

Les résultats du test d'utilisabilité de la timeline ont été très bons, comme l'indiquent le score SUS (82,5) et les commentaires des utilisateurs. Le score du questionnaire SUS indique un bon niveau d'utilisabilité. Ce score place la timeline dans les 10% d'applications avec des scores d'utilisabilité supérieurs à 80% (30) et est, à titre d'exemple, supérieur à celui de l'outil KNAVE-II (69.1) (95).Les problèmes d'utilisabilité signalés étaient, en général, des défauts mineurs qui pouvaient être corrigés rapidement. En outre, sur la base des réactions des utilisateurs, de nouvelles fonctionnalités ont été introduites afin de mieux répondre aux besoins des utilisateurs de la pharmacovigilance. Le nombre significativement plus élevé d'erreurs lors de la recherche du résultat de l'ADR peut s'expliquer par la grande complexité de cette tâche (recherche impliquant des résultats de laboratoire, lecture simultanée de plusieurs documents, etc. Les différences dans les délais d'exécution peuvent s'expliquer par des différences dans la méthode de travail de chaque participant. Certains participants ont accompli leurs tâches en parallèle, tandis que d'autres les ont exécutées l'une après l'autre. Cette explication s'applique également à la différence dans le nombre de clics, car certains participants ont dû rechercher les mêmes informations plus d'une fois.

L'évaluation de l'impact quantitatif de la timeline sur la pratique courante a montré un gain de temps significatif pour l'un des participants. Cet utilisateur est celui qui a le plus participé au développement de l'application et qui a le moins d'expérience avec la méthode habituelle, par rapport aux autres. Cette observation suggère que le gain de temps aurait pu être plus général avec une séance de formation préliminaire plus longue pour mieux familiariser tous les participants sur les fonctionnalités de la timeline. De plus, étant donné que le chronogramme a été comparé à la méthode utilisée au Centre de pharmacovigilance depuis plus de 10 ans, l'absence globale de gain de temps pourrait être interprétée comme un résultat encourageant, d'autant plus que

les utilisateurs voulaient pouvoir continuer à utiliser le chronogramme après l'évaluation.

Dans l'ensemble, pour les cas simples pour lesquels la récupération du dossier d'hospitalisation est suffisante, la valeur ajoutée temporelle est moins apparente.

Enfin, l'absence de toute différence significative dans la qualité des rapports montre que l'utilisation exclusive du calendrier n'est pas un facteur limitant pour la collecte de données. De plus, dans certains cas traités selon le calendrier, la piètre qualité des rapports pourrait s'expliquer par la présence de documents textuels seulement dans la demande. Ce manque de données n'était pas lié à l'application en elle-même, mais aux données disponibles dans l'entrepôt de données.

Sur la base des résultats et des retours d'expérience obtenus dans le cadre de cette étude, nous prévoyons d'introduire de nouvelles fonctionnalités dans la prochaine version de l'application. En particulier, nous nous concentrerons sur une meilleure présentation des documents textuels, qui est le point de départ pour localiser l'information dans les dossiers des patients. De plus, il sera possible de comparer plus facilement le contenu d'un document avec les données structurées affichées dans la ligne de temps.

Pendant le test d'utilisabilité, la décomposition du cas d'utilisation en tâches causa des problèmes dans l'analyse des enregistrements MORAE. En effet, ces tâches ne se sont pas déroulées de manière séquentielle, mais en parallèle car, en pharmacovigilance, le processus d'investigation ne suit pas un seul chemin critique. De plus, pour la même tâche, les participants n'ont pas recherché le même niveau de détails. Certains participants voulaient des renseignements très précis et détaillés (p. ex. la demi-vie de la molécule administrée au patient) et, par conséquent, ils avaient besoin de plus de temps pour accomplir leurs tâches. Inversement, les autres participants ont simplement utilisé l'information contenue dans le document texte récupéré par la requête. Il est donc difficile d'attribuer les différences de temps observées lors de l'exécution de différentes tâches à un problème d'utilisabilité de l'application.

Pour l'étude d'impact sur les pratiques, seuls trois participants étaient disponibles, en raison des ressources humaines et matérielles limitées dont nous disposions, et il n'a donc pas été possible de traiter deux fois le même enregistrement en utilisant les deux méthodes. Une comparaison plus approfondie entre les méthodes nécessiterait une plus grande puissance d'essai.

c. Conclusion de l'évaluation

L'évaluation s'est porté à la fois sur l'utilisabilité de la timeline et sur les impacts sur pratiques. Si, les résultats du test d'utilisabilité indiquent un bon niveau général d'utilisabilité, le gain en termes de temps est moins net. Cependant, l'application ayant été mise en production, et toujours utilisée dans le service de pharmacovigilance, nous pouvons estimer que la timeline nécessitait un temps d'adaptation avant de pouvoir prouver un gain sensible en terme de temps.

7. Discussion

Les points forts méthodologiques de la timeline sont les suivants :

- Intégration des données hétérogènes avec chacune leur représentation visuelle correspondante (graphique pour des données biologiques,...)
- Méthode d'agrégation des données basée sur les terminologies médicales
- Méthode de recherche et de visualisation des données textuelles

Les évaluations de l'application ont montré une bonne utilisabilité générale. Même la deuxième évaluation ne montre pas que la timeline permet un gain de temps significatif pour tous les utilisateurs, elle est encore utilisée quotidiennement au sein du service de pharmacovigilance du CHU de Rennes. Elle a donc répondu à un besoin spécifique de visualisation dans le dossier patient pour la pharmacovigilance.

Nous avons développé un prototype permettant d'étendre l'utilisation de la timeline avec l'intégration des données issues du SNDS (Système National des Données de Santé). L'ensemble des consommations de soins donnant à lieu à remboursement par l'Assurance maladie en ville ou à l'hôpital sont collectés par ce système. Nous avons donc, avec l'aide de l'équipe REPERES (Recherche en Pharmaco-épidémiologie et recours aux soins, UPRES EA-7449), spécialiste dans le traitement des données du SNDS, extraits les dispensations médicamenteuses de cette base de données pour pouvoir l'intégrer à la timeline.

Un ingénieur de l'équipe a ensuite procédé au chaînage des données des données du SNDS (Système National des Données de Santé) et des données d'eHOP. Cette étape consiste à mettre en correspondance les données pour retrouver à quel patient d'eHOP appartient les données de dispensation médicamenteuses. Le chaînage des données a été réalisé grâce aux variables suivantes : année de naissance, sexe, date de début de séjour, date de fin de séjour, diagnostic principal.

L'intérêt de ce prototype réside dans le fait de pouvoir visualiser à la fois les dispensations médicamenteuses faites en ville. Ceci permet de faire des hypothèses sur un effet indésirable ayant eu lieu en dehors de l'hôpital et ayant abouti à une hospitalisation et un passage aux urgences. Il est cependant important de préciser qu'une donnée de dispensation est moins fiable qu'une donnée d'administration à l'hôpital, car elle nécessite de faire des hypothèses sur l'observance thérapeutique du patient.

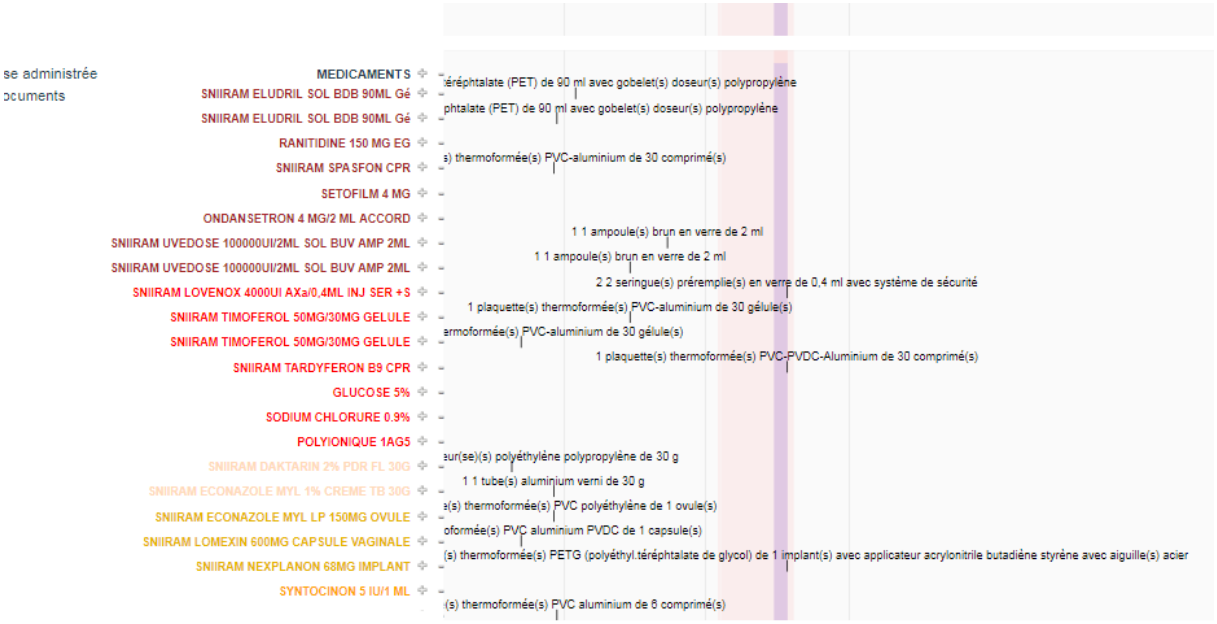


Figure 52 Intégration des données du SNIIRAM

8. Conclusion

La complexité et la richesse des données contenues dans les dossiers médicaux nécessitent le développement d'outils spécifiques pour leur exploitation efficace en pharmacovigilance. Parmi les caractéristiques requises, l'affichage et l'interrogation de l'information au sein d'un fichier demeurent des éléments critiques. Dans le cadre du cette thèse, nous avons développé une méthode intégrée de recherche et de visualisation de l'information, sous forme de timeline. Nous avons dans cette partie décrit l'interface développé ainsi que détailler les résultats du test d'utilisabilité indiquent un bon niveau général d'utilisabilité, et de nouvelles fonctionnalités adaptées à la pratique courante de la pharmacovigilance pourraient être ajoutées à la demande des participants.

Chapitre 4 : Application aux trajectoires de soins multi-patients

1. Contexte et problématique

Les séquences d'événements temporelles sont actuellement générées dans presque tous les domaines de l'analyse des données. Un exemple typique est un site de commerce électronique suivant chacun de ses utilisateurs par le biais d'une des pages de produits qu'il a consulté jusqu'à ce qu'un achat soit effectué. Les données de santé à la disposition des chercheurs et des institutions (entrepôt de données hospitaliers, données de l'assurance maladie,..) ont suivi la même tendance ce qui a eu pour conséquence de faire émerger de nouveaux besoins en termes d'outils d'analyses, notamment pour l'analyse de séquences de différents événements. Ces séquences peuvent révéler les modalités d'observance d'un traitement, le lien possible entre un traitement et de potentiels effets secondaires ou encore l'efficacité d'un traitement. (96) Les séquences d'événements temporels se composent de milliers ou de millions d'événements, et comportent l'identifiant d'un patient, un horodatage (peut être par année ou par jour ou à la seconde, et une catégorie d'événements (diagnostic, donnée de laboratoire,…). Cette information sur des événements ponctuels peut être rassemblée en séquences de plusieurs milliers d'événements.

La visualisation de ces séquences peut apporter une aide précieuse pour permettre de révéler des informations différentes des tests statistiques classiques. En effet, un mode de visualisation bien adapté permet à l'utilisateur de mieux apprécier le contexte des données et de repérer immédiatement des séquences remarquables. De plus, contrairement aux outils statistiques qui répondent à des questions connues, les outils de visualisation permettent une analyse visuelle exploratoire flexible des données. En ce sens, elle peut être une étape préliminaire à l'analyse statistique des séquences, pour faire émerger des hypothèses qui seront secondairement analysées sur le plan statistique. L'outil de visualisation CoCo (97) propose une approche à l'interface de la fouille visuelle des données et des outils statistiques. Un autre avantage des outils de visualisation est de permettre de résumer l'information contenue dans les séquences d'une cohorte de patients. Eventflow (75), Outflow (73), et Peekquence (98) sont des exemples d'outils de visualisations permettant de résumer l'information contenue dans les séquences mais avec des approches différentes : Eventflow regroupe les séquences entre elles, Outflow est basé sur les diagrammes de Sankey pour présenter les différentes séquences, et Peekquence se base sur les résultats de différents algorithmes de fouilles de données. Peerfinder (99) propose lui de retrouver des séquences similaires à une séquence de référence.

Nous présentons dans ce chapitre un outil de visualisation et de fouille de motifs séquentiels développé dans le cadre de la plateforme Pharmaco-Epidémiologie des Produits de Santé (PEPS) dont la mission principale est le développement de nouvelles méthodes de recherche d'information et la visualisation de données cliniques hétérogènes. Cet outil permet la création et visualisation des séquences des patients à partir de l'entrepôt de données eHOP. Nous avons également implémenté une

version modifié de l'algorithme Smith-Waterman (62) pour la recherche de séquences similaires, ainsi que les algorithmes Apriori (54) et GSP (55) pour la recherche de motifs fréquents. Dans ce chapitre, nous présentons également l'évaluation du prototype développé sur un cas d'usage courant de pharmacovigilance en termes de gain de temps et de performances.

2. Etat de l'art

a. Coco

CoCo (pour « *Cohort Comparison* ») (97) est un outil d'analyse visuelle développé au sein du Human Computer Interaction Lab (HCIL) qui permet à un utilisateur de comparer deux ensembles de données de séquences temporelles. Il combine des tests statistiques automatisés avec une interface interactive pour permettre des aperçus ou la génération d'hypothèses. Les utilisateurs voient des statistiques sur leur ensemble de données, des statistiques au niveau de l'événement et un menu de métriques. CoCo affiche les tests de significativité (test de Student et du Khi²) sous une forme standardisé pour des mesures telles que la prévalence et la durée des écarts.

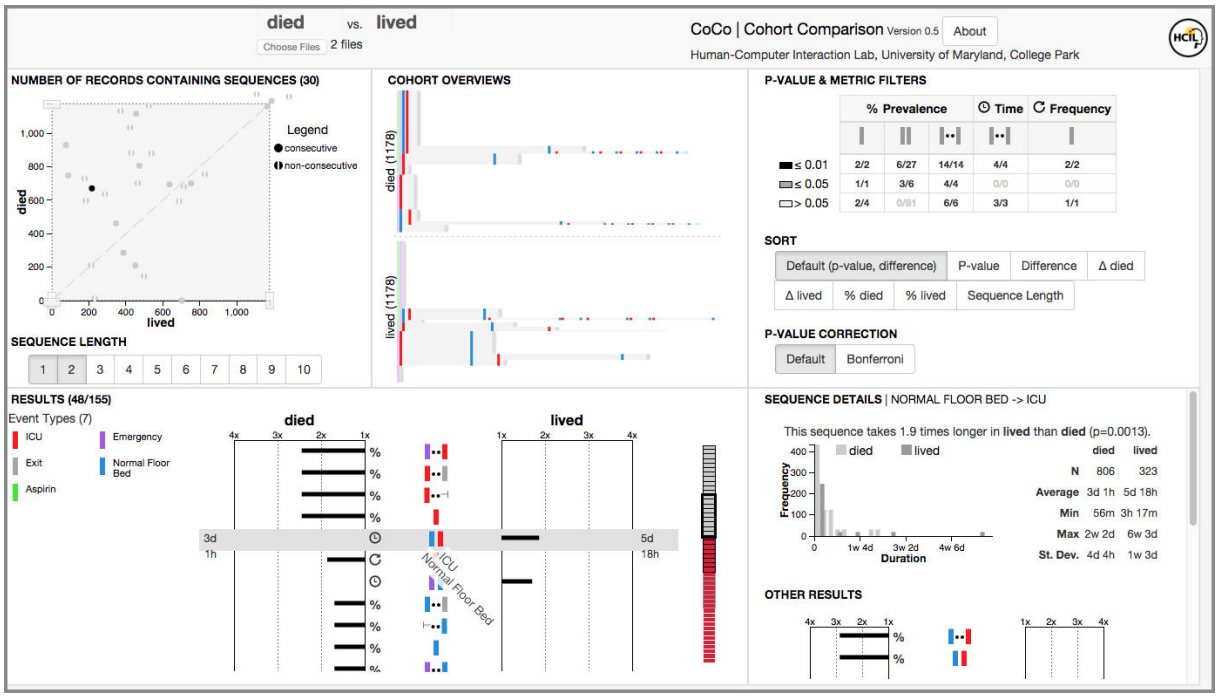


Figure 53 Capture d'écran du logiciel CoCo

Dans l'exemple de la capture d'écran ci-dessous, on compare les séquences d'événements dans deux cohortes de patients, une cohorte avec des patients décédés et une cohorte avec des patients vivants. Ici, la séquence « *Intensive Care Unit* » suivi par « *Normal Floor Bed* » apparait 2,5 fois plus souvent dans les enregistrements dans la cohorte de patients vivant. (avec une P-value < à 1%)

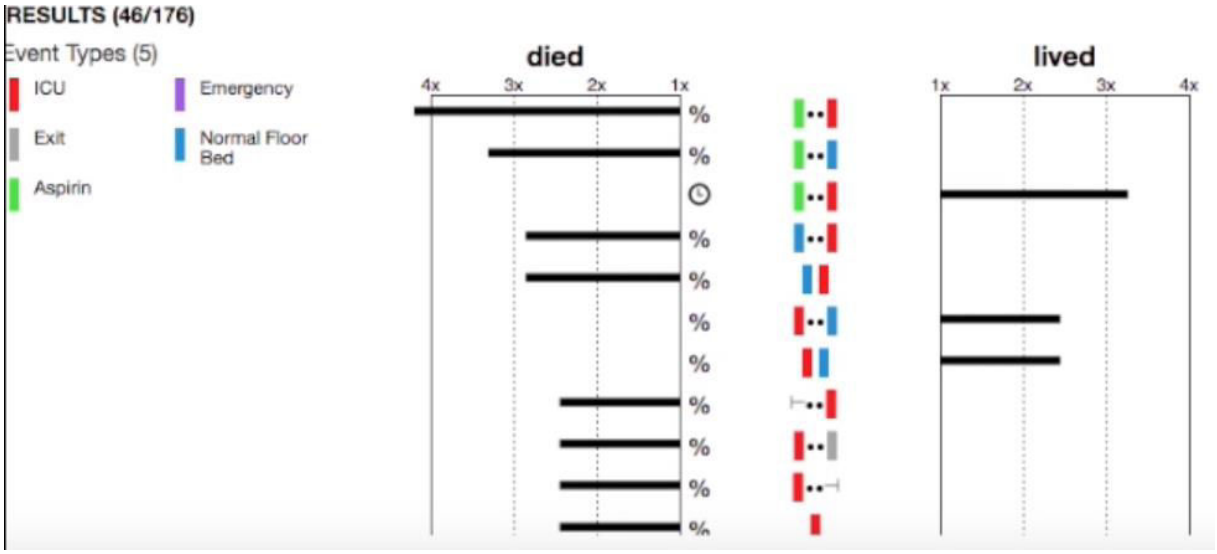


Figure 54 Capture d'écran du logiciel CoCo

b. Eventflow

Eventflow (75) est un outil interactif de visualisation de données temporelles également développé au sein du HCIL. Un des objectifs principaux de cet outil est de permettre aux chercheurs ou aux analystes de caractériser la nature et la fréquence de certains motifs redondants dans les bases de données et de pouvoir construire et tester des hypothèses sur les causes de ces changements.

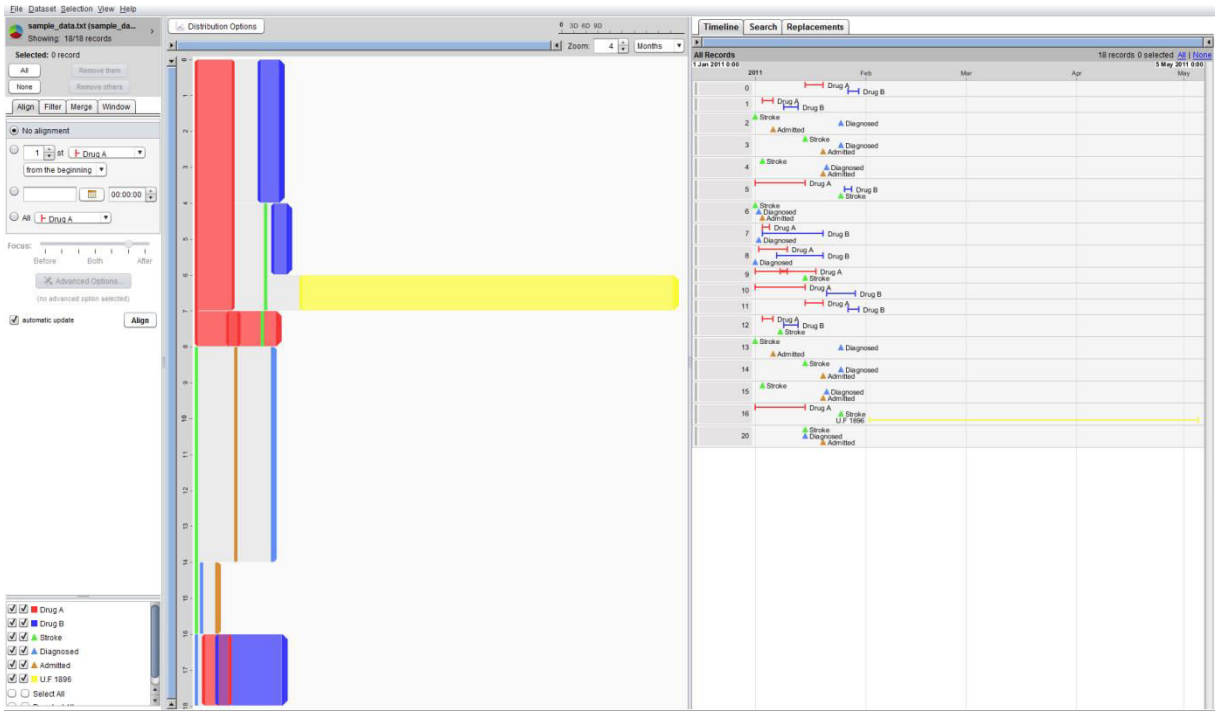


Figure 55 Interface principale d'Eventflow

L'interface principale d'Eventflow est construite en 3 parties, de gauche à droite de l'écran : les contrôles interactifs, une vue synthétique du jeu de données, et les Timelines individuelles.

1. L'interface de contrôle permet de régler l'affichage :
    - Aligner les éléments de la vue synthétique sur un élément particulier
    - Choisir les éléments à afficher ou filtrer ceux qui ne répondent qu'à certaines conditions (recouvrement par exemple)
    - Ajuster l'échelle de la Timeline (jour, mois, année,…)
2. La vue synthétique de l'enregistrement donne une vue d'ensemble sur le jeu de données. Eventflow repère les enregistrements ayant les mêmes motifs et les agrège dans des clusters. Ces clusters sont représentés par des rectangles dont la longueur correspond à la moyenne de la durée de chaque élément. La hauteur représente le nombre d'enregistrements compris dans le cluster.
3. A droite de l'écran sont disposés les Timelines individuelles. 3 onglets sont disponibles :
    - L'onglet « Timeline » : Les évènements ponctuels sont affichés sous forme de triangles colorés tandis que les intervalles sont représentés sous formes de lignes.
    - L'onglet « Search » permet de construire une requête basée sur les motifs temporels. Une timeline vide peut être remplie par l'utilisateur avec les évènements recherchés dans les enregistrements. Dans l'exemple ci-dessous, l'utilisateur recherche les enregistrements contenant le motif « évènement rouge puis évènement bleu sans recouvrement et sans survenue de l'évènement vert entre l'événement rouge et l'évènement bleu ». Les résultats de la requête sont ensuite affichés sous la TimeLine.
    - Le bouton « replace as » permet de créer des évènements à partir des événements existants. La capture d'écran ci-dessous illustre cette fonctionnalité : ici, l'utilisateur crée un nouvel évènement nommé « Nouvelle séquence » correspondant à l'enchainement « événement rouge puis évènement bleu sans évènements verts entre chaque ».

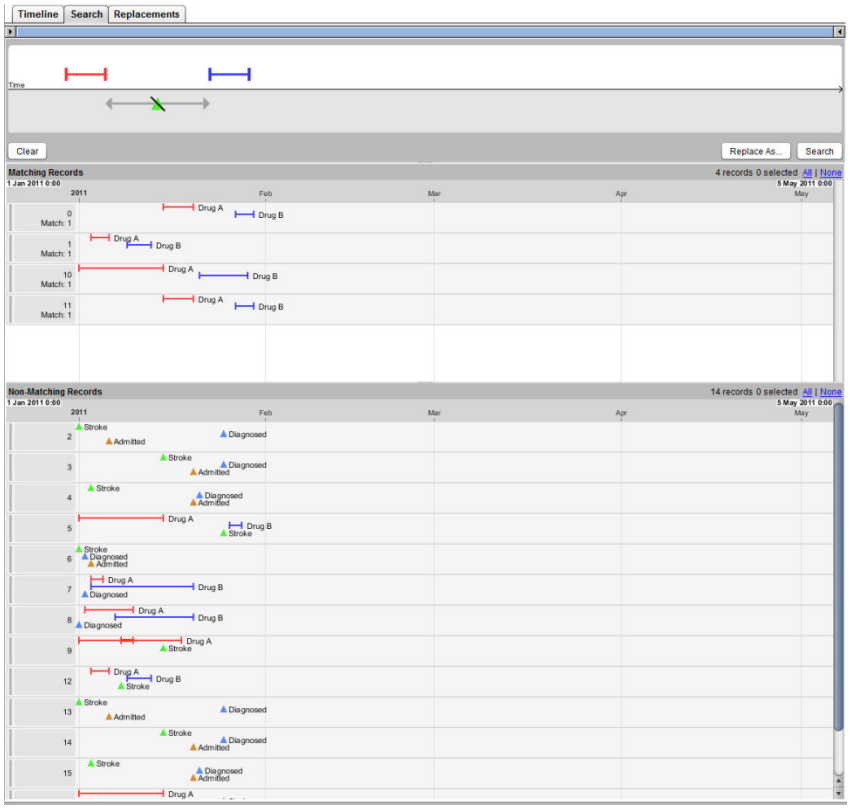


Figure 56 Recherche et visualisation de séquences individuelles dans Eventflow

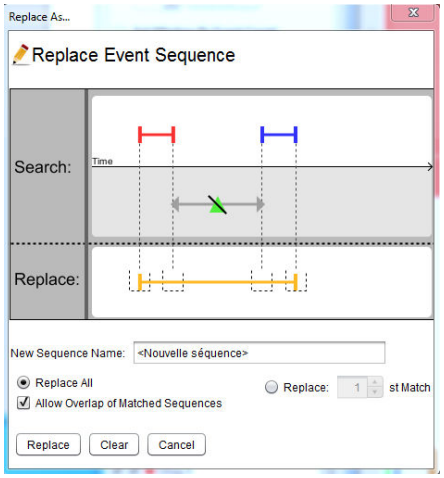


Figure 57 Interface de requêtage d'Eventflow

Eventflow a été testé en situation pour une étude de pharmacovigilance concernant les traitements reçus par 100 patients atteints d'asthme. Le but de l'étude était de comparer les traitements reçus par les patients aux bonnes pratiques de prescriptions, notamment grâce aux possibilités de détection des recouvrements d'Eventflow. L'expérience s'est avérée concluante quant à la satisfaction des utilisateurs notamment par rapport à des approches traditionnelles telles que les queries SAS ou SQL classiques. Eventflow a déjà été utilisé pour répondre à des cas d'usage de pharmacoépidémiologie. (100)

c. Peerfinder

PeerFinder (99) est un autre outil de visualisation développé au sein du HCIL. Il s'agit d'une interface visuelle qui permet aux utilisateurs de trouver et d'explorer des enregistrements similaires à un enregistrement de référence. PeerFinder utilise à la fois des attributs sur les individus (comme l'âge) et des informations sur les événements temporels. En termes d'interactivité, PeerFinder fournit différents niveaux de contrôle et de contexte qui permettent aux utilisateurs d'ajuster les critères de similarité. Il permet également aux utilisateurs de quantifier la similarité des enregistrements retrouvés par rapport à l'enregistrement de référence. Les résultats intermédiaires sont affichés et les utilisateurs peuvent affiner la recherche de façon itérative.

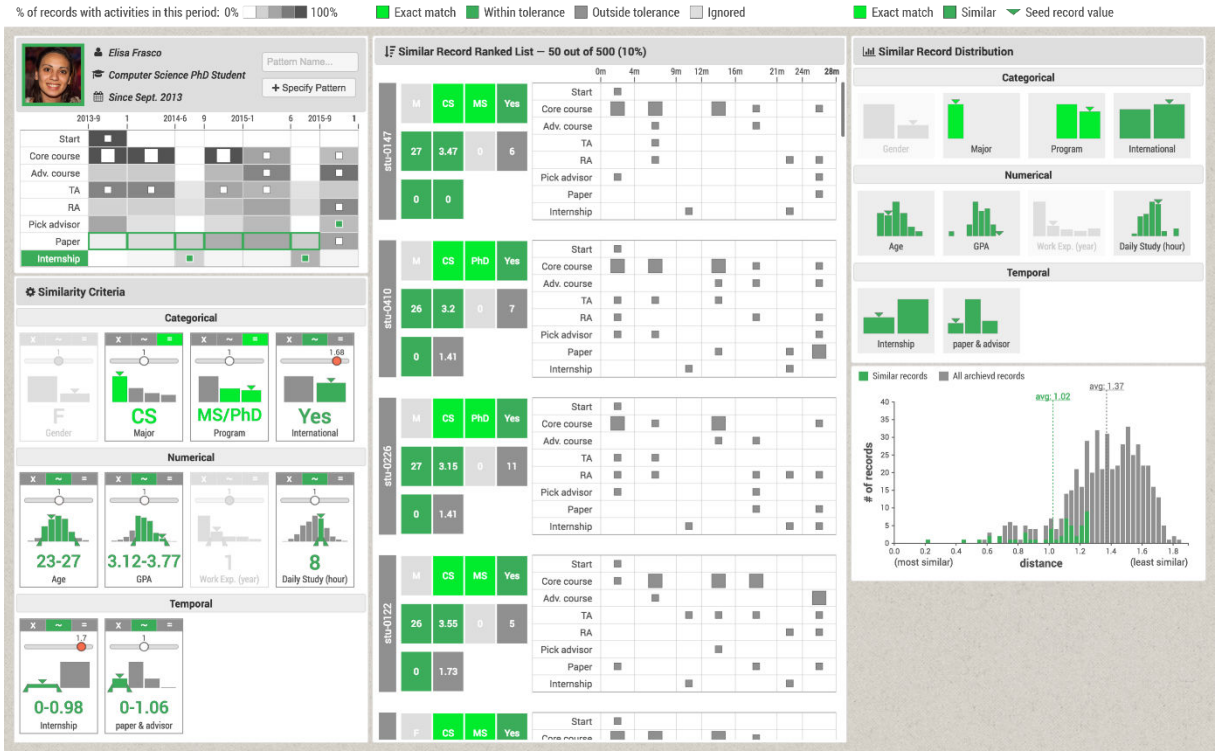


Figure 58 Interface de Peerfinder

d. Outflow

Outflow (73) est un outil de visualisation et d'analyse de séquences temporelles développé par IBM. Cet outil permet entre autre la visualisation des trajectoires des patients. Outflow est basé en partie sur les diagrammes de Sankey, mais les adapte pour prendre en compte la durée de chaque transition entre deux événements. Pour une trajectoire patient, chaque nœud représente un évènement survenu chez un patient. La largeur des flux correspond au nombre de patients pour lesquels sont survenus les deux mêmes événements à la suite. La couleur dépend de la probabilité d'aboutir à une issue définie de la séquence (par exemple « décés » ou « fin de séjour »). Outflow a pour objectifs :

- Agréger des séquences d'événements
- Afficher différents chemins représentant les changements d'états avec le temps et la cardinalité

- Permettre à l'utilisateur d'évaluer les facteurs externes possiblement corrélés aux différents changements d'état (par exemple l'administration de médicaments)

Figure 59 Interface d'Outflow

e. Peekquence

Peekquence (98) est un outre outil de visualisation développé par IBM pour l'analyse de séquence mettant en avant les motifs temporels. Son objectif principal est de rendre plus compréhensible et interprétables les résultats retournés par les algorithmes d'apprentissage automatique. Cet outil permet de choisir le modèle et les résultats (corrélation, variabilité), résumer l'information (événements les plus courants dans les trajectoires).

Peekquence se compose de quatre vues : la vue de la liste des motifs montrant les motifs extraits de l'algorithme SPAM, la vue en réseau des séquences montrant la fréquence des cooccurrences de type d'événement, l'histogramme des événements cooccurrents montrant la fréquence des événements cooccurrents pour un modèle sélectionné et l'affichage de la timeline du patient montrant les séquences d'événements des patients qui comprennent le ou les motifs sélectionnés.

Figure 60 Interface de Peekquence

Ces outils n'ont pas été développés pour les besoins spécifiques de la pharmacovigilance, et ils manquent de fonctions pour la recherche de séquences approximatives et de modules de traitement des données. À notre connaissance, l'alignement sur une séquence d'évènements n'a pas été réalisé. Pour contourner cette limitation, nous proposons d'adapter un algorithme d'alignement de séquence d'ADN au contexte de pharmacovigilance. L'originalité de notre approche est qu'elle permet la recherche et l'alignement d'une séquence de référence sur une séquence d'événements flous, tout en tenant compte de la temporalité des événements et en calculant un score de similarité pour classer les résultats. Les algorithmes d'alignement des séquences d'ADN (p. ex. les algorithmes Needleman-Wunsch et Smith-Waterman) ont été largement utilisés en bioinformatique pour comparer et aligner les séquences d'ADN.(62) L'algorithme Smith-Waterman (SW) est l'un des algorithmes de recherche de séquences les plus sensibles, quoique l'un des plus lents. (101) De plus, son utilité dans d'autres domaines que la bioinformatique, en particulier dans le traitement de l'image, a déjà été démontrée. (102) Nous sommes partis de l'hypothèse que la découverte de séquences temporelles spécifiques d'événements cliniques/exposition à un médicament dans une population est similaire à la découverte d'un modèle nucléotidique défini au sein d'une séquence d'ADN. De plus, une vision d'ensemble des séquences sera proposée grâce à l'utilisation des algorithmes Apriori et GSP permettant de visualiser les motifs et séquences fréquentes.

3. Extraction et traitement des trajectoires

a. Choix des informations pertinentes pour la constitution des séquences

Comme nous l'avons vu dans la partie concernant les trajectoires de soins, le contenu des séquences sera très dépendant de la question d'étude posée. Dans l'outil que nous avons développé, l'utilisateur pourra donc choisir le contenu des séquences qui seront extraites et générées à partir de l'entrepôt en fonction de la question scientifique posée. Cette recherche se base sur des « concepts » et des « critères ». Un

critère peut être constitué d'un ensemble d'un ou plusieurs codes issus des terminologies médicales (ATC, CIM-10, CCAM), ou un terme extrait dans un document. Un concept est une liste d'un ou plusieurs critères. Un évenement est un concept daté. A titre d'exemple, afin de définir l'événement indésirable médicamenteux d'intérêt, il est possible d'utiliser des requêtes SMQ (Standardised MeDRA Queries) pour sélectionner les codes CIM-10 qui correspondaient au diagnostic. Ces requêtes sont constituées de groupe de termes (symptômes, diagnostic,…) prédéfinis relatifs à un domaine médical précis. Les SMQ ont pour but d'aider à l'identification de cas potentiellement pertinents relatifs à la sécurité du médicament.

b. Recherche des séjours d'intérêt de l'entrepôt

La première étape du processus d'extraction des trajectoires sera de récupérer les séjours ayant un intérêt pour l'analyse. Cette étape consiste à filtrer les séjours ayant tous les concepts demandés par l'utilisateur. Le résultat de cette opération est affiché sous la forme d'un diagramme de Venn permettant à l'utilisateur d'assurer que le nombre de séjours retournés est suffisant et de vérifier qu'un concept n'est pas trop discriminant. .

c. Génération des séquences

Pour s'adapter au modèle de séquence proposé, nous avons traité les données du patient extraites de la base de données d'eHOP de différentes façons.

L'alphabet utilisé pour l'algorithme SW consiste en un ensemble de codes prédéfinis pour les événements présents dans la séquence du patient et dans la séquence de référence, et un code pour l'événement nul (c'est-à-dire une absence d'événement). Les événements ont été triés par ordre chronologique pour former une séquence. Pour avoir un alphabet cohérent avec des caractères limité, nous avons choisi d'offrir à l'utilisateur plusieurs possibilité de traitement des données numériques (à savoir les valeurs biologiques et les doses de médicaments administrés/prescrites).

i. Discrétisation

L'utilisateur peut choisir de discrétiser les valeurs numériques avec le nombre d'intervalles de son choix. L'utilisateur pourra nommer chaque intervalle. Cette transformation a intérêt lorsque l'on s'intéresse à la normalité d'une valeur biologique par exemple (exemple : un INR est considéré comme « idéal » chez un patient sain entre 2 et 3, alors qu'il est de 1 pour un patient sain).

ii. Suivi de l'évolution

L'évolution d'une valeur numérique peut être également un événement très important dans la séquence d'un patient. Pour les médicaments, calcul une évolution de dose peut permettre de mettre en évidence des événements ayant eu lieu avant ou après ce changement de dose (par exemple, l'utilisateur pourra rechercher une baisse de la dose administrée en anti-vitamine K après un INR trop élevé). Pour les valeurs biologiques, il est intéressant de rechercher une hausse ou une baisse d'une valeur biologique pouvant correspondre à une réalité clinique (par exemple, une

hausse brutale de la Protéine C réactive peut amener à considérer la survenue d'une infection chez le patient).

L'évolution d'une valeur numérique peut se faire de façon relative (calcul de l'évolution en pourcentage) ou absolue. Le calcul de l'évolution peut se faire entre deux événements consécutifs ou sur une durée définie par l'utilisateur (par exemple, pour suivre l'évolution d'une valeur biologique sur plusieurs jours). De plus, de la même façon que pour la discrétisation, l'utilisateur pourra spécifier des intervalles pour chaque calcul d'évolution. Ainsi, une évolution entre -5% et 5% pourra être considérée comme stable.

Les traitements présentés ci-dessus ne sont pas exclusifs et peuvent être combinés sur une même valeur numérique. La valeur numérique brute reste accessible à tout instant dans l'interface.

iii. Switch médicamenteux

Il est possible pour l'utilisateur d'approximer l'événement de switch médicamenteux. Par défaut, lorsque que deux événements d'administrations se suivent avec un code ATC identique mais des codes UCD différents, il est possible de créer un événement switch médicamenteux. Par exemple, si les codes ATC du Doliprane© et du Paracétamol générique sont identiques, leurs codes UCD sont différents et on peut considérer que l'on est en présence du switch médicament princeps-générique. Il est possible d'appliquer le même raisonnement avec des classes ATC plus hautes dans la hiérarchie (exemple : niveau 4 et 5 de l'ATC).

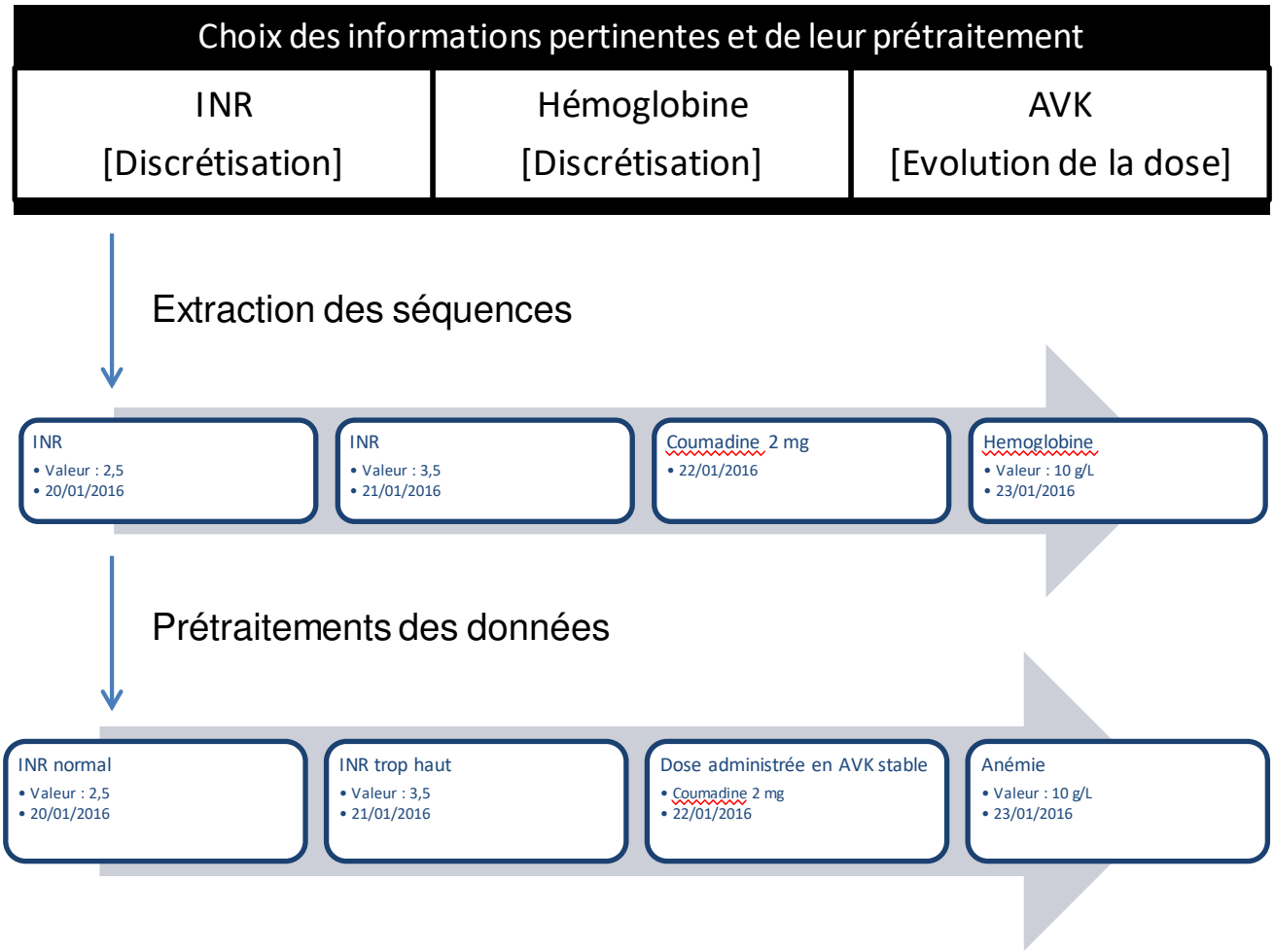


Figure 61 Workflow de création de séquences

4. Choix des méthodes de fouilles de données séquentielles

a. Adaptation de l'algorithme de Smith-Waterman aux données temporelles

Figure 62 Etapes d'application du Smith-Waterman adapté

i. Description

Nous avons utilisé l'algorithme Smith-Waterman (62), sans l'amélioration de Gotoh (103), pour calculer l'alignement local optimal des séquences cliniques. Les adaptations introduites dans l'algorithme pour répondre à nos besoins spécifiques sont décrites dans les sections suivantes.

Cet algorithme compare deux séquences : i) la séquence spécifiée par l'utilisateur (c'est-à-dire la séquence de référence) et ii) les séquences des patients. Chaque séquence est considérée comme une chaîne de caractères. Chaque caractère de

cette chaîne peut représenter un événement clinique, par exemple l'administration d'un médicament ou un résultat d'un test de laboratoire. Les types de caractères représentent l'alphabet utilisé par l'algorithme SW pour comparer chaque séquence. Une matrice de notation et un score de similarité sont calculés pour chaque comparaison. Les scores augmentent progressivement en fonction de la plus grande similitude entre la séquence de référence et les données du patient.

ii. Adaptation aux contraintes temporelles

Comme l'algorithme SW n'est pas conçu pour traiter des séquences temporelles, nous avons introduit divers changements dans l'algorithme. Tout d'abord, pour tenir compte des intervalles temporels entre les événements cliniques, nous avons créé un nouveau type de caractère correspondant aux événements nuls pour les jours où il n'y a pas d'événement dans l'historique du patient.

Deuxièmement, dans certains cas, l'utilisateur ne s'intéresse pas à trouver la correspondance exacte dans une séquence d'événements, mais seulement la présence d'un événement avant, pendant ou après un alignement de séquences (par exemple, "la recherche d'un diagnostic dans les jours suivant un schéma d'administration d'un médicament et une valeur biologique donnée"). Pour résoudre ce problème, nous avons mis en œuvre l'algèbre d'intervalle d'Allen21. Les critères d'Allen permettent de rechercher les liens temporels possibles entre les événements. Si la relation temporelle supposée par l'utilisateur n'est pas trouvée, le score de similarité de la séquence est plus faible.

iii. Calcul des matrices de substitution et de notation

La matrice de substitution a été utilisée pour calculer le score de chaque cellule de la matrice de notation en fonction des éléments de la séquence. La matrice de substitution a été calculée dynamiquement pour chaque comparaison entre une séquence de patients et la séquence d'interrogation (définie par l'utilisateur). Si les codes d'un événement dans les deux séquences correspond, le score retourné est le score correspondant m. Sinon, le coût de la différence était le contraire du score correspondant. La note S retournée a été calculée comme suit :

$$S(Eci, Ecj) = \begin{Bmatrix} +\, m\ (Eci = Ecj) \\ -\, m\ (Eci \neq Ecj) \end{Bmatrix}$$

où E est l'événement, c le code de l'événement, et i et j les identificateurs (référence et séquence du patient, respectivement). La matrice de notation permet de traiter les comparaisons individuelles entre tous les événements de la séquence du patient et la séquence de référence et d'enregistrer les résultats d'alignement optimal. Les événements nuls sont pris en compte lors du calcul de la matrice de notation.

La méthode de calcul de la matrice de notation M peut être résumée comme suit :

$$M(i,j) = max\begin{Bmatrix} 0 \\ M(i-1,j-1) + D(Ai,Bj) \\ M(i-1,j) + \Delta \\ M(i,j-1) + \Delta \\ M(i-1,j) + T(n) \end{Bmatrix}$$

où M est la matrice de notation, D est la fonction qui renvoie le score dans la matrice de substitution pour les événements A et B, Δ la pénalité pour les insertions ou suppressions, et T est la fonction mathématique choisie par l'utilisateur (logarithme naturel, exponentiel...) qui calcule la pénalité d'écart pour les jours sans événement.

iv. Processus de parcours de la matrice, alignement et vérification de la contrainte temporelle

Pour déterminer l'alignement de la séquence, nous avons utilisé la matrice de scoring pour identifier le chemin qui pourrait donner le score d'alignement maximal. Pour cela, nous avons supposé que la fonction D renvoie le score d'appariement entre l'événement A et l'événement B.

Pendant le processus de parcours de la matrice, pour chaque i, j (figure 2) :

Si M (i, j) = M (i-1, j-1) + D (Ai, Bj), alors Ai est comparé à Bj et l'algorithme retourne à M (i-1, j-1) ;
Si M (i, j) = M (i, j-1) + Δ (gap), alors Bj est associé à une cellule vide et l'algorithme retourne à M (i, j-1) ;
Si M (i, j) = M (i-1, j) + Δ, alors Ai est associé à une cellule vide et l'algorithme retourne à M (i-1, j) ;
Si M (i, j) = M (i-1, j) + T(n), alors Ai est associé à "jour sans événement" et l'algorithme revient à M (i-1, j) ;
Si M (i, j) = 0, l'alignement local est terminé.
Ceci est fait pour tous les maximums locaux M(x, y) afin d'obtenir tous les alignements locaux optimaux.

X		$INR\ normal$	$INR\ trop\ haut$	$Evenement\ nul$	$INR\ trop\ haut$	$AVK\ stable$	$AVK\ en\ hausse$	$INR\ trop\ haut$
	0	0	0	0	0	0	0	0
$INR\ too\ high$	0	0	3	3	3	1	0	3
$INR\ too\ high$	0	0	0	0	6	4	2	3
$VKA\ rising$	0	0	0	0	4	2	7	5

Figure 63. Exemple de processus de parcours de matrice avec des événements nuls. La séquence de référence est dans la première colonne et la séquence du patient dans la première rangée de la matrice. Les chiffres indiquent les scores correspondants. INR, rapport international normalisé ; AVK, antagoniste de la vitamine K.

Pour chaque contrainte temporelle, nous avons vérifié si l'événement sous contrainte temporelle était situé avant ou après le score d'alignement local maximum.

v. Calcul du score de similarité

Le score de similarité a été calculé à l'aide de l'équation suivante :

$$Ssim = \frac{Ml}{Lref * Ms} * \frac{Ntcc}{Ntct}$$

où Ml est le maximum local de la matrice de notation, Lref la longueur de la séquence de référence, Ms le score d'appariement, Ntcc le nombre de contraintes temporelles vérifiées et Ntct le nombre total de contraintes temporelles.

Pour l'alignement final de la figure 3, le score de similarité était de 77,8 %.

$INR\ normal$	$INR\ trop\ haut$	$Evenement\ Nul$	$INR\ trop\ haut$	$AVK\ stable$	$AVK\ en\ hausse$	$INR\ trop\ haut$
	$INR\ trop\ haut$	$-$	$INR\ trop\ haut$	$-$	$AVK\ en\ hausse$	

Figure 64. Alignement final

Pseudo Code

Input :
S: Sequence with n events such as S = {E(0), E(1),… , E(n)}
Sref: Reference Sequence with m events such as Sref = {E(0),… ,E(m)}
Ms: Matching Score
T: Mathematical function for computing time gap penalty
TempCons: List with k temporal constraints such as TempCons = { Tp(0), …, Tp(k)} with TpA is Allen Criterion and TpE is an event

Output :
ScoreSim: Similarity Score

Generate Sequence with Null Events from S called Sn
Let Se an empty Set of event.
for i ← 0 to n **do** :
    if Sn(i) **not in** Se :
        Se.push(Sn(i))
//Generation of Substitution Matrix
Let l the length of Se.
Let Msub an empty matrix of size l x l.
for i ← 0 to l **do** :
    for j ← 0 to l **do** :
        if Se(i) == Se(j) :
            Msub(j)(i) = Ms
        else :
            Msub(j)(i) = -Ms
//Computation of Scoring Matrix
Let Mscoring an empty matrix of size (n+1) x (m+1)
for i ← 0 to n **do** :
    Mscoring(0,i) = 0
for j ← 0 to m **do** :
    Mscoring(j,0) = 0
    for i ← 0 to n **do** :

$$Mscoring(i,j) = max\begin{Bmatrix} 0 \\ M(i-1,j-1) + D(Ai,Bj) \\ M(i-1,j) + \Delta ins \\ M(i,j-1) + \Delta del \\ M(i-1,j) + T(n) \end{Bmatrix}$$

Let Max the local maximum of Mscoring

//Checking Temporal Constraint
W = 0
for i ← 0 to k **do** :
    if TempCons(k)A == "after" **do** :

```
                for j ← Max to l do :
                        if TempCons(k)E == Se(k) do :
                                W++
                                break
        if TempCons(k)A == "before" do :
                for j ← 0 to Max do :
                        if TempCons(k)E == Se(k) do :
                                W++
                                break

//Compute Similarity Score
```

$$ScoreSsim = \frac{Max}{m * Ms} * \frac{W}{k}$$

b. Utilisation des algorithmes Apriori et GSP

Ces algorithmes sont utilisés pour identifier les motifs fréquents dans notre ensemble de séquences. Un ensemble d'événements cooccurrents sont appelés transactions. Une règle d'association peut être notée R : X => Y (A, B) avec X et Y deux événements, A la fréquence de cooccurrence des événements X et Y (il s'agit du support de la règle), et B la fréquence des individus qui ont l'événement Y sachant qu'ils ont également X (il s'agit de la confiance de la règle). Par exemple, l'association Diagnostic A -> Médicament C (10%, 70%) montre que 10% ont reçu le diagnostic A et le médicament C, et 70% des patients ayant reçu le diagnostic A se sont vu administrés le médicament C.

5. Choix architecturaux

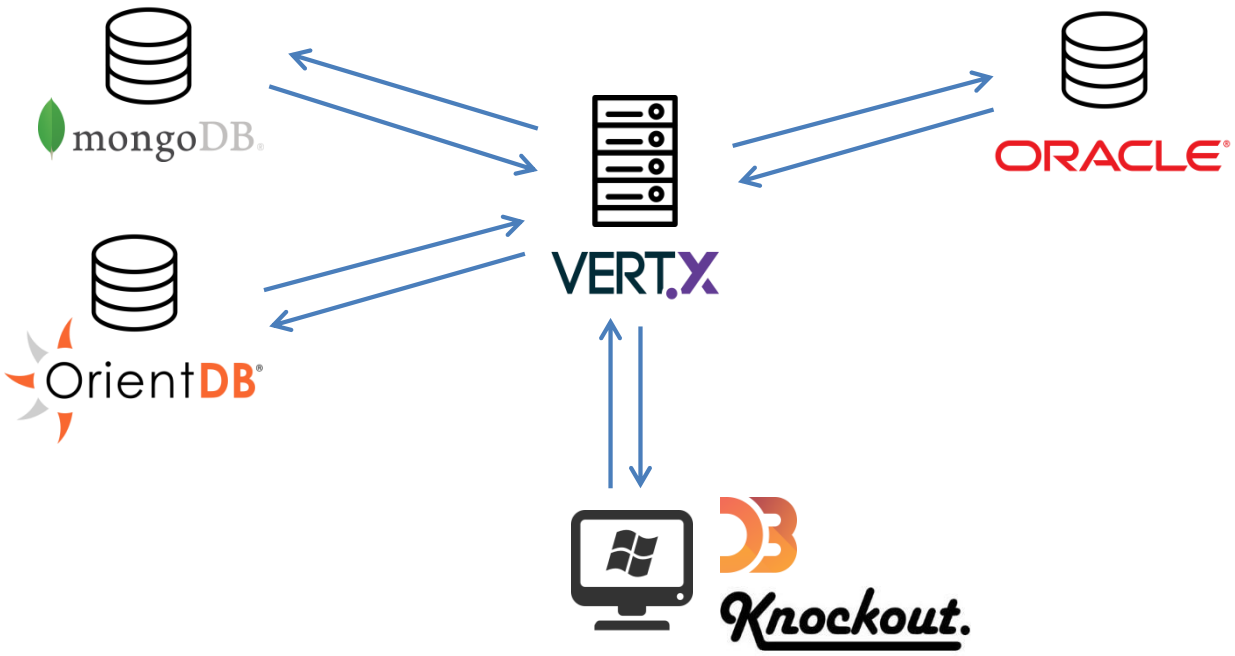


Figure 63 Schéma de l'architecture de l'application

a. Front-end

La partie client de l'application a été codée en JavaScript. Pour structurer notre code, nous avons utilisé la bibliothèque KnockoutJS (104) qui permet d'implémenter le pattern MVVM (Model-View-ViewModel) qui permet de séparer l'interface graphique de la couche logique de l'application web. La bibliothèque D3js (105) nous a permis de créer les éléments SVG (Scalable Vector Graphics) pour représenter graphiquement et dynamiquement les séquences d'événements.

b. Back-end

La partie serveur de l'application a été implémentée avec le framework Eclipse Vert.x. (106) Il s'agit d'un framework événementiel pour la Java Virtual Machine. Il est fortement inspiré de Node.js (107) mais se veut le plus agnostique possible concernant le choix du langage utilisé. Ce framework est une implémentation du design pattern reactor. Son fonctionnement est basé sur des composants nommés *verticles* communiquant via un *eventbus*. Vert.x permet une montée en charge très efficace et facilite la programmation concurrentielle. Si nous avons écrit notre propre implémentation de l'algorithme Smith-Waterman, les algorithmes Apriori et GSP proviennent de la librairie Java open-source SMPF dédiée à la fouille de données (108).

c. Bases de données

Nous avons utilisé plusieurs types de bases de données dans notre architecture. La base de données Oracle est la base de données d'eHOP, il s'agit de la source de données primaire. La base de données OrientDB est orientée graph et permet de récupérer rapidement l'arborescence des différentes terminologies. Enfin la base de données orientée document MongoDB permet de stocker et d'accéder aux différents résultats intermédiaires (alignement et recherche de motifs).

6. Description de l'application

Nous avons développé une application pour visualiser les séquences dans une cohorte de patients. Contrairement à d'autres outils de visualisation, tels que PeerFinder et OutFlow, notre outil gère l'ensemble du processus de visualisation des données, depuis le traitement des données CWD jusqu'à l'affichage à l'écran. L'interface web est divisée en deux onglets : l'un pour accéder à l'interface d'extraction de séquences et l'autre pour visualiser les séquences.

L'interface d'extraction de séquences est utilisée pour construire une cohorte de séquences de patients et pour choisir les méthodes de traitement des données (Figure 64). Pour construire une cohorte, l'utilisateur crée d'abord une liste de concepts en choisissant les codes relatifs à un concept dans une arborescence qui regroupe plusieurs terminologies médicales. ([1] sur la figure 64) Par conséquent, un code de laboratoire et un code de diagnostic peuvent faire partie du même concept. Une fois que la liste des concepts est établie (INR et VKA dans l'exemple de la figure 64), l'utilisateur clique sur le bouton "view sets" ([2] sur la figure 64). Cette action lance une recherche parmi les ensembles de séjours hospitaliers qui incluent les concepts. Le résultat de cette recherche est affiché sous forme de diagramme de Venn avec la

taille de chaque ensemble ([3] sur la figure 64). Ce mode de visualisation est utile pour vérifier si un ou plusieurs concepts ne sont pas trop discriminants par rapport aux autres. Si l'utilisateur est satisfait de la taille de tous les sets, il peut choisir les opérations de traitement des données. Dans l'exemple de la figure 64, ces traitements concernent les valeurs numériques des administrations de médicaments et des résultats d'analyses biologiques. Il est possible de dissocier les données des intervalles choisis par l'utilisateur, de calculer la variation relative ou brute d'une dose ou d'une constante entre les événements ou pendant un intervalle de temps. ([4] sur la Figure 64)

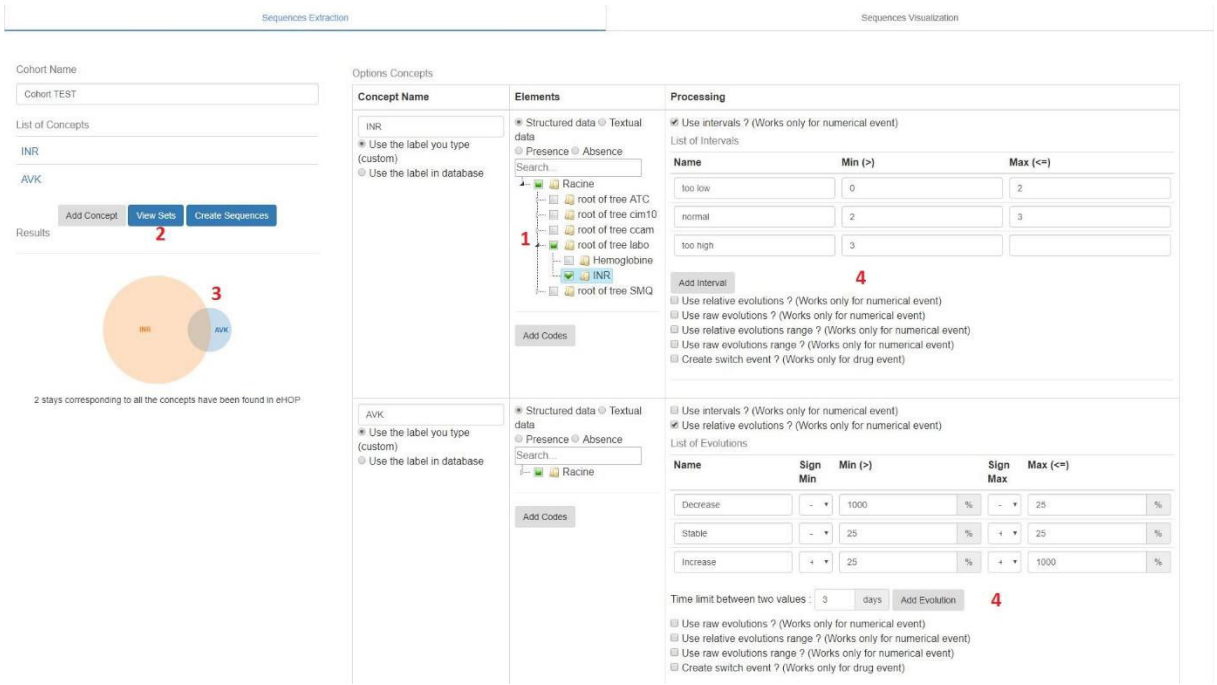


Figure 64 Interface de sélection de concepts et de méthodes de traitement des données

Le deuxième onglet est une interface visuelle interactive qui a été développée pour visualiser l'ensemble des séquences (Figure 65). Un dictionnaire de visualisation a été développé pour chaque type d'événement ([1] sur la Figure 65). Par exemple, les événements numériques discrets sont représentés par un carré. Les événements qui décrivent les changements d'une variable numérique sont représentés par des flèches. La direction de la flèche est basée sur la tendance du changement. Les événements occasionnels, comme un diagnostic, sont représentés par des croix. Chaque type d'événement a sa propre couleur. Les séquences d'événements sont représentées dans la partie principale de l'interface ([2] sur la Figure 65). Les résultats des algorithmes GSP et Apriori, qui sont respectivement les séquences et associations les plus courantes dans la cohorte ([3] et [4] sur la figure 65), sont représentés dans les colonnes de droite. Si l'utilisateur clique sur le numéro d'identification de la séquence, il accédera à la vue du dossier patient dans la timeline. Cette fonctionnalité permet de retourner facilement à une vue individuelle et plus fournie du dossier patient.

Figure 65 Interface d'affichage des séquences

L'utilisateur peut sélectionner la ou les séquences de référence à rechercher dans une interface d'interrogation (Figure 66). Tous les types d'événements sont énumérés dans la zone grise ([1] à la figure 66). En dessous de cette zone, chaque ligne représente une séquence de recherche ([2] sur la figure 66). L'utilisateur sélectionne les types d'événements présents dans la zone grise par un système de glisser-déposer pour constituer une séquence ([1] sur la Figure 66). Les petits espaces à gauche et à droite de la séquence de recherche sont utilisés pour spécifier les contraintes de temps ([3] sur la figure 66).

Figure 66 Interface de recherche de séquences

Les résultats de la recherche (Figure 67) sont ensuite classés en fonction de leur similitude avec la séquence de référence afin de faciliter leur visualisation ([1] dans la Figure 67). L'utilisateur peut choisir de filtrer les séquences en fonction de la présence ou de l'absence de types d'événements spécifiques ([2] dans la Figure 67), ou en fonction de la durée globale de la séquence ou du temps écoulé entre les événements alignés et l'événement temporellement contraint ([3] dans la Figure 67). Tous ces critères peuvent être combinés.

L'intervalle de temps entre l'alignement et l'événement à contrainte temporelle est indiqué par une ligne indiquant sa durée en jours ([4] sur la figure 67). Le calcul des

algorithmes GSP et Apriori est redémarré avec seulement les séquences filtrées, et les résultats sont mis à jour ([5] sur la Figure 67).

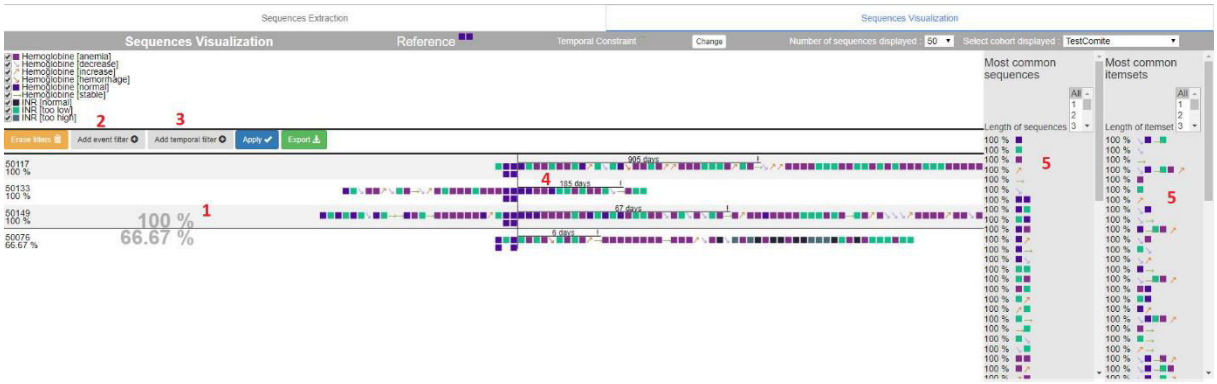


Figure 67 Affichage des résultats de la séquence similaire

7. Evaluation du Smith-Waterman adapté

a. Validation de l'algorithme

Pour valider l'algorithme, nous avons testé l'algorithme sur chaque patient d'un échantillon extrait au hasard de la base de données d'eHOP. Chaque patient a été recherché un par un dans toutes les bases de données des séquences. Si la première séquence retournée correspond exactement au patient recherché, le test passé avec succès pour ce patient.

b. Test de performance

Des tests de performance ont effectués pour mesurer les capacités de montée en charge de notre algorithme. Ces tests ont été effectués avec un nombre croissant de séquences à analyser, une longueur de séquence de référence croissante et un nombre croissant de séquences de référence. Les tests ont été effectués avec deux processeurs Intel® Xeon® E5-2603 v3 cadencé à 1,90GHz (2x6 cœurs).

c. Evaluation sur un cas d'usage de surveillance de bon usage du médicament

Pour évaluer la pertinence de l'algorithme du Smith-Waterman modifié pour l'identification des séquences d'événements cliniques pouvant correspondre à des cas de traitement inadéquat, nous avons utilisé les données cliniques des patients extraites du CDW eHOP.

Pour le cas d'utilisation, nous avons sélectionné tous les séjours à l'hôpital où il y avait à la fois des mesures de l'INR et des administrations d'antagonistes de la vitamine K (VKA) (10 882 séjours). L'objectif était de trouver des cas d'administration inappropriée de l'AVK. La dose d'AVK doit être réduite lorsque l'INR est supérieur à la valeur cible (>3 dans le cas présent, mais la cible dépend de l'objectif et de l'état du patient). Pour ce faire, nous nous sommes concentrés sur les schémas d'administration montrant des problèmes évidents (c.-à-d. une augmentation plutôt qu'une diminution de la dose administrée après deux valeurs consécutives du RIN au-dessus de la valeur cible).

Par conséquent, la séquence d'événements de référence était "INR trop élevé - INR trop élevé - INR trop élevé - augmentation de la dose de VKA".

Nous avons échantillonné au hasard l'ensemble de données d'évaluation (n=80 séquences) en fonction du score de similarité par rapport à la séquence de référence (requête) pour sélectionner des séquences qui étaient distribuées de façon équivalente dans quatre classes de score de similarité ([100;75[,[75;50[,[50,25[,[25,0]]]). Les données extraites pour cette évaluation ont été complètement anonymisées.

Dans la 1ère phase de l'évaluation, un expert en pharmacovigilance a passé en revue les 80 séquences et identifié tous les cas réels de traitement inadéquat, sur la base de la séquence d'événements affichés sur l'interface. Les séquences ont été triées au hasard, le score de similarité a été caché et aucun alignement n'a été fait pour éviter de biaiser le jugement de l'expert. Nous avons utilisé cet examen comme gold-standard pour évaluer la performance de l'algorithme. Nous avons estimé la sensibilité, la spécificité et la mesure F entre les résultats de l'examen du gold-standard et les scores de similarité calculés par l'algorithme SW. La taille de l'échantillon de 80 séquences nous permet d'avoir une puissance statistique de 90% avec une erreur de type I de 5% pour détecter une aire sous la courbe (AUC) d'au moins 0,70 entre notre gold-standard et la classification par l'algorithme.

Dans la deuxième phase de l'évaluation, nous avons comparé la performance des utilisateurs en examinant les 80 séquences avec et sans l'aide de l'algorithme,

Pour estimer le gain de temps que permet le système de classement par algorithme dans la recherche d'un traitement inadéquat, deux autres utilisateurs experts ont examiné indépendamment les séquences extraites pour les classer comme correspondant ou non à des cas de traitement inadéquat. Comme pour le premier expert, ils ont dû identifier les traitements inadéquats à partir des événements de ces 80 séquences. Ils ont fait ce travail deux fois avec deux méthodes de recherche différentes, avec un délai d'un jour entre les deux méthodes pour éviter un effet d'apprentissage :

- Première méthode, les séquences ont été triées au hasard et aucun score de similarité n'a été fourni. Aucun alignement n'a été effectué.
- Deuxième méthode, les séquences ont été alignées et affichées dans une liste triée en fonction de leur score de similarité par rapport à la séquence de référence.

Les deux utilisateurs ont passé en revue et analysé les séquences une par une, puis les ont rapportées dans un fichier : l'identification de la séquence, qu'elle corresponde ou non à un cas de traitement inadéquat, et le score de similarité (pour les séquences triées). Nous avons enregistré le temps nécessaire à l'analyse de la séquence pour chaque utilisateur et chaque méthode de recherche (affichage aléatoire suivi d'un tri en fonction de leur score de similarité).

Nous avons calculé le temps moyen nécessaire à l'examen d'une séquence avec chaque méthode et nous les avons comparés à l'aide du test de Wilcoxon.

Les analyses statistiques ont été effectuées avec le logiciel statistique R, version 3.4.1.1.22. (94).

8. Résultats

a. Résultats de l'évaluation du Smith-Waterman adapté

i. Validation de l'algorithme

Chaque patient de la base de données renvoie un score silencieux de 100% par rapport à sa propre séquence.

ii. Test de performance

Pour évaluer la mise à l'échelle de notre algorithme, nous avons redémarré le programme plusieurs fois avec un nombre croissant de séquences ainsi qu'une longueur de séquence de référence croissante. Le tableau et le graphique ci-dessous montrent les résultats de ces tests. Les temps de calcul apparaissent en millisecondes.

Tableau 7 Temps de calcul en fonction de la longueur de la séquence de référence et du nombre de séquences

	Longueur de la séquence de référence			
Nombre de séquences	3	6	8	12
50	1161	1220	1192	1210
100	2154	2195	2222	2301
500	9988	9970	10607	10268
1000	19358	20192	20081	21035
2000	38101	39207	40800	40695
4000	76690	78567	81401	80823
10000	188583	188449	203236	201381

Sur le graphique, on peut voir que le temps requis par l'algorithme évolue de manière linéaire en fonction du nombre de séquences à analyser. La longueur de la séquence de référence a très peu d'impact sur le temps requis pour l'algorithme.

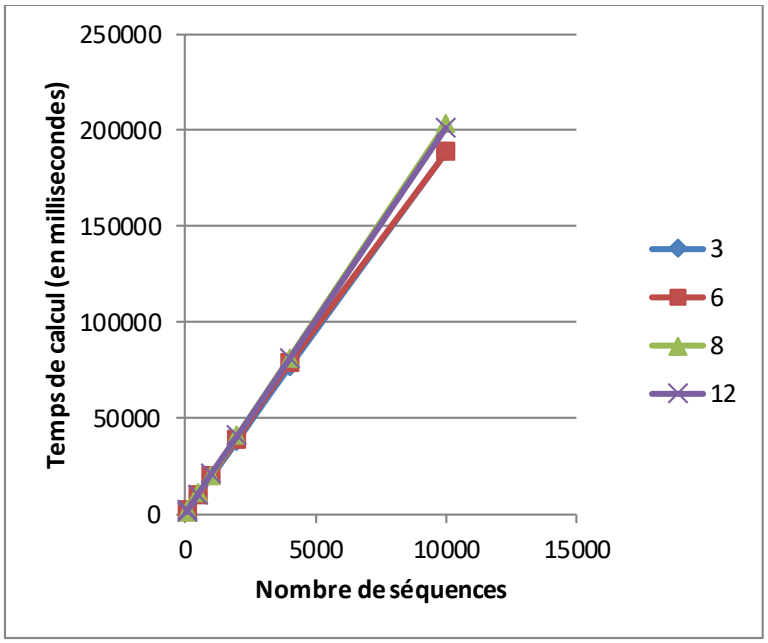


Figure 68 Evolution du temps de calcul en fonction de la longueur de la séquence de référence et du nombre de séquences

Nous avons également mesuré le temps pris par l'algorithme en fonction du nombre de séquences de référence à rechercher. Nous avons effectué la recherche avec 1000 séquences et une séquence de référence avec 3 événements.Il apparait que le temps de calcul de l'algorithme n'est pas très affecté.

Tableau 8 Temps de calcul en fonction du nombre de séquences de référence

Nombre de séquences de référence	Temps(ms)
1	19933
2	20562
5	20408
10	20891
20	21081

iii. Evaluation sur un cas d'usage de surveillance de bon usage du médicament

1) Résultats de l'évaluation de performance

Les résultats de l'expert en pharmacovigilance (gold standard review) ont été utilisés pour évaluer la performance de l'algorithme SW pour l'identification de séquences de cas d'administration de médicaments inappropriés. La précision de l'algorithme était de 1, le rappel était de 0,76, la mesure F était de 0,866 et l'AUC (Aire sous la courbe) était de 0,99 (voir courbe ROC)

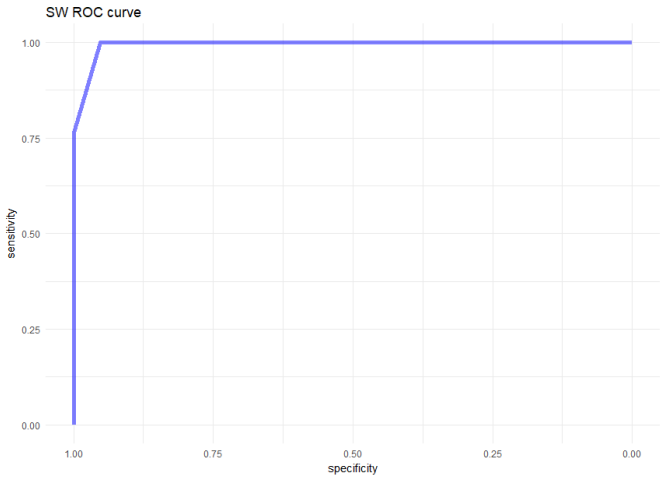


Figure 69 Courbe ROC montrant les performances de classification de l'algorithme Smith-Waterman adapté

2) Résultats de l'analyse temporelle

Le temps moyen d'analyse d'une séquence par l'utilisateur no 1 était de 2,86 ± 2,26 secondes pour les séquences alignées et triées et de 3,49 ± 3,54 secondes pour les séquences triées de façon aléatoire (p = 0,0003, test de Wilcoxon) (figure 7). Pour l'utilisateur 2, le temps moyen d'analyse d'une séquence était de 1,06 ± 1,06 secondes pour les séquences ordonnées et de 1,53 ± 0,85 secondes pour les séquences non ordonnées (p <0,0001, test de Wilcoxon).

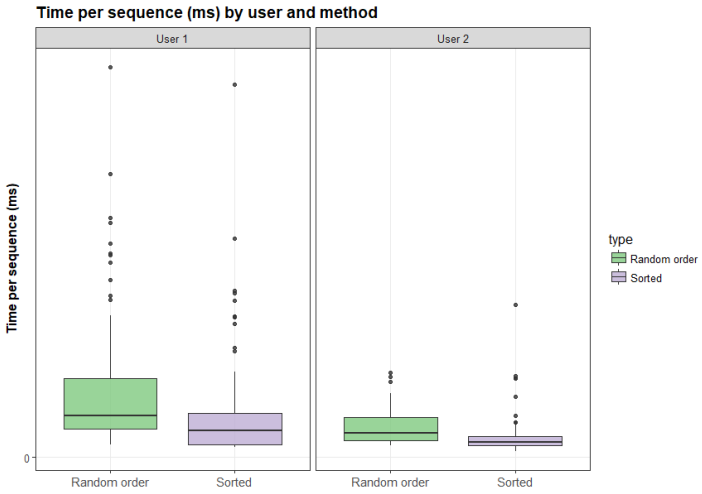


Figure 70 Boxplot montrant la distribution du temps passé sur une séquence par utilisateur et par méthode

9. Discussion

Notre algorithme Smith Waterman modifié présente plusieurs avantages : il permet de rechercher des motifs séquentiels similaires. Les modifications que nous avons apportées nous permettent de prendre en compte la dimension temporelle des données. D'autres outils (p. ex. EventFlow) peuvent rechercher des modèles avec des événements distants, mais seulement des correspondances exactes. Une autre originalité de l'outil est qu'il supporte toutes les étapes de visualisation : de la recherche de séquences dans la base de données, en passant par le traitement et la

transformation des données, jusqu'à l'affichage à l'écran. Toutes ces étapes sont sous le contrôle de l'utilisateur.

Cependant, dans notre étude de cas, nous avons constaté que l'étape de traitement des données était essentielle pour la recherche de séquences : les séquences, et donc les modèles fondés sur des algorithmes, varieront en fonction des choix faits par l'utilisateur. L'apport d'experts dans le domaine est essentiel pour générer une séquence adaptée à la question de recherche.

Les modifications que nous avons introduites dans l'algorithme SW prennent en compte la dimension temporelle dans le traitement des séquences, surmontant ainsi les limites des outils précédents, comme PeerFinder qui aligne seulement les séquences par rapport à la date du premier événement.

Les résultats de l'évaluation démontrent la pertinence de l'algorithme SW pour la détection des administrations de médicaments inappropriées. Plus précisément, le système de classement permet une identification rapide et précise des modèles d'intérêt dans les séquences de patients. En outre, les séquences qui ne présentent pas le modèle exact d'intérêt peuvent également être pertinentes. Par exemple, dans le cas d'utilisation, le schéma "INR trop élevé" - "INR trop élevé" - "INR trop élevé" - "AVK dose administrée augmentée", indique une administration de médicament inappropriée parce que la dose administrée n'a pas été réduite, selon les recommandations actuelles. C'est une illustration de la pertinence de la possibilité de recherche de séquences approchantes avec l'algorithme SW adapté. La valeur AUC (0,99) indique que l'algorithme SW était un très bon classificateur pour ce cas d'utilisation.

L'exploitation des temps n'incluait pas le temps nécessaire pour spécifier la requête ou le temps d'exécution de l'algorithme. Celles-ci varieront en fonction de la taille de l'ensemble de données et de la requête. Nous avons exécuté le programme avec un nombre croissant de séquences et plusieurs longueurs de séquences de référence. Pour une séquence de référence de longueur 3, le temps de traitement est de 2 sec. et 154 ms pour 100 séquences, 19 sec. et 358 ms pour 1000 séquences, 3 min, 8 sec. et 583 ms pour 10 000 séquences. Pour une séquence de référence de longueur 12, le temps de traitement est de 2 sec. et 301 ms pour 100 séquences, 21 sec. et 35 ms pour 1000 séquences, 3 min, 21 sec. et 381 ms pour 10 000 séquences. Le temps de calcul requis pour l'algorithme est linéaire en fonction du nombre de séquences, tandis que la longueur de la séquence a peu d'effet. En résumé, l'utilisation de l'algorithme peut retarder le début de l'examen humain des séquences, mais il est susceptible d'accélérer considérablement ce processus d'examen - surtout si un petit nombre de séquences similaires sont trouvées.

L'algorithme SW a des limites : il est inadéquat pour la recherche de modèles distants dans le temps. Par exemple, la médication pour l'hyperthyroïdie pendant plusieurs semaines suivie d'une hospitalisation pour des problèmes cardiaques est un modèle qui ne sera pas trouvé avec notre outil. De par sa conception, l'algorithme original de SW ne permet de récupérer que des modèles d'événements à court terme. D'autres

outils (p. ex. EventFlow) peuvent rechercher des modèles avec des événements distants, mais seulement des correspondances exactes.

Les modifications apportées à l'algorithme SW le rendent adapté pour traiter des modèles avec une absence d'événement, par exemple l'absence de prophylaxie avant l'intervention chirurgicale. La matrice de substitution de l'algorithme SW pourrait être également être paramétrée manuellement par l'utilisateur. Cela permettrait de pénaliser plus ou moins certaines substitutions d'événements. Par exemple, une substitution de Paracétamol et Doliprane pourrait ne pas être pénalisée dans le calcul de la matrice de scoring en le considérant comme une correspondance exacte.

Dans notre étude de cas, nous avons constaté que l'étape de traitement des données CDW était essentielle pour la recherche de séquences : les séquences, et donc la visibilité des motifs, varieront en fonction des choix faits par l'utilisateur. L'apport d'experts dans le domaine est essentiel pour générer une séquence adaptée à la question de recherche.

Par ailleurs, une des limites de la modélisation de la séquence que nous avons choisie et quelle ne prend pas compte des événements sous forme d'intervalles (avec une date de début et une date de fin), contrairement à ce qui est possible de faire avec un outil tel qu'EventFlow. Cette modélisation ne tient pas compte non plus des événements cooccurrents. Ces deux limitations pourraient être contournées en modifiant l'algorithme Smith-Waterman. Au moment de la création de la matrice de scoring, les événements seront sur plusieurs lignes comme illustré sur le schéma ci-dessous :

	X	INR normal	INR trop haut	Evenement nul	Prescription d'aspirine INR trop haut	Prescription d'asprine AVK stable	Prescription d'asprine AVK en hausse	INR trop haut
	0	0	0	0	0	0	0	0
Prescription d'asprine	INR trop haut	0	0	3	3	6	4	3	3
Prescription d'asprine	INR trop haut	0	0	0	0	9	7	5	3
Prescription d'asprine	AVK en hausse	0	0	0	0	4	2	10	5

Figure 71 Matrice de scoring avec événements cooccurrents

Dans cet exemple, l'utilisateur recherche des cas de non-respect des bonnes pratiques de prescription en AVK alors que le patient a reçu une prescription d'aspirine, ce qui est contre-indiqué. Le calcul des scores de la matrice de scoring est adapté en conséquence. Ici l'événement sous forme d'intervalles (la prescription d'aspirine) est représenté sous forme de plusieurs événements ponctuels consécutifs.

Une autre limite de l'application est de ne pas pouvoir générer des événements négatifs, c'est-à-dire des événements signifiant l'absence d'un événement (ex : « *Absence d'administration d'AVK* »). Une proposition pour contourner cette limitation serait de créer ces événements pour chaque jour ou l'événement n'est pas survenu avec cependant le risque de réduire les performances de l'algorithme en augmentant considérablement le nombre de séquences.

10. Perspectives

a. Comparaison de plusieurs cohortes de séquences

Lors de l'évaluation, les utilisateurs nous ont suggérés l'idée de pouvoir comparer plusieurs listes de séquences entre elles. Cette comparaison de séquence doit être réalisée sur des attributs tels que l'âge des patients, leur appartenance à telle service, ou encore leur antécédents médicaux (diagnostics ou actes médicaux). Cette fonctionnalité pourrait permettre de vérifier certaines hypothèses. Par exemple, les erreurs de doses administrées en anti-vitamines K sont plus nombreuses chez les patients porteurs de prothèses cardiaques valvulaires (risque de thrombose). Il serait donc facile de comparer la proportion de cas de non-respect des bonnes pratiques de prescription entre une cohorte de patient sains et une cohorte de patients porteurs de prothèses valvulaires. De plus, ces hypothèses pourraient être vérifiées avec une approche statistique de la même façon que l'application CoCo. (97)

b. Clustering de séquences

Afin d'améliorer l'identification de séquences d'événements similaires, nous prévoyons de regrouper les séquences. Plus précisément, en considérant les motifs extraits à l'aide d'un algorithme, tel que l'algorithme GSP, comme des termes et toutes les séquences comme un corpus, nous pouvons calculer un « *term frequency – inverse document frequency* » (tf-idf) pour chaque motif dans chaque séquence. La C-value (109) serait également une méthode de pondération pour la recherche d'information pertinente à utiliser. En effet, elle va pénaliser les séquences englobées dans des séquences plus grandes. Par exemple, si la séquence <AB> apparait 5 fois dans l'ensemble des transactions, et que la séquence <ABC> apparait 5 fois, alors la séquence <AB> serait pénalisée et seule la séquence <ABC> serait conservée dans l'analyse. Le calcul de ces mesures aboutirait sur une matrice de largeur m et de hauteur n, où n est le nombre de séquences dans le corpus, et m le nombre de sous-séquences dans le corpus. Cette matrice peut ensuite être analysée par différentes méthodes de classification comme la classification ascendante hiérarchique. La figure ci-dessous montre un exemple de cette analyse effectuée avec le logiciel R sur la cohorte utilisée pour l'évaluation (patients sous AVK avec mesure de l'INR). Pour maximiser l'inertie inter-classe nous avons choisi d'utiliser la distance de Ward. On obtient 6 clusters différentiés en grande partie par la longueur des différentes séquences.

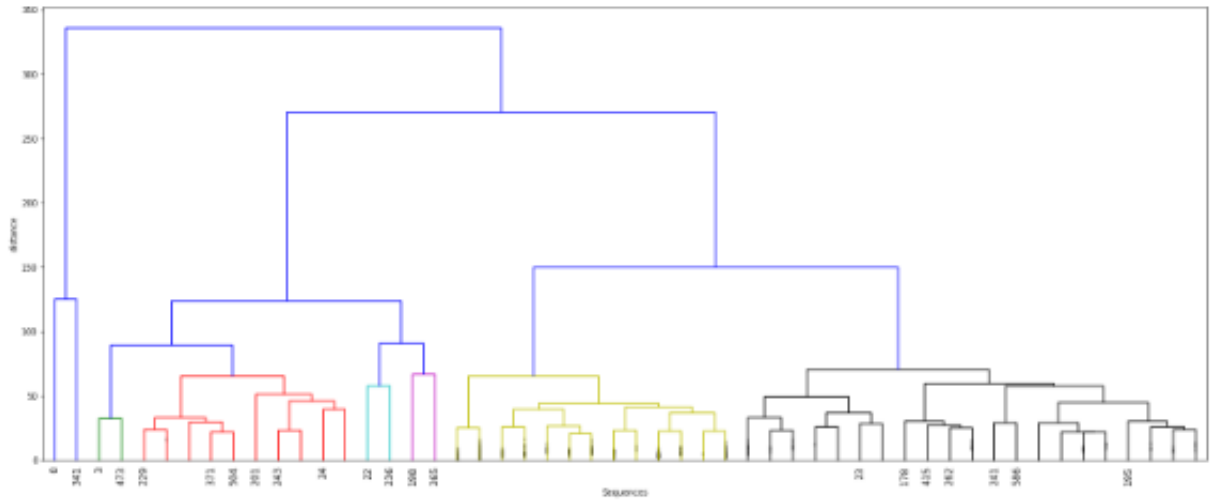


Figure 71 Exemple de classification hiérarchique des séquences réalisés avec R

11. Conclusion

Nous avons développé une application permettant d'extraire, de visualiser et de fouiller une cohorte de séquences de patients. Nous avons implémentés différent algorithmes. L'algorithme SW adapté permet de trouver des séquences d'intérêt dans le dossier électronique du patient représenté sous forme de séquences. Cette approche a l'avantage de permettre une recherche floue. De plus, les modifications introduites dans l'algorithme SW permettent de prendre en compte la dimension temporelle dans le traitement des séquences. L'utilisation des algorithmes Apriori et GSP permettent d'avoir une vue d'ensemble de la cohorte en se basant sur les séquences et les itemsets fréquents.

Discussion Générale

L'évolution des technologies de l'information a abouti à une explosion de la quantité de données produites et stockées. Le domaine de la santé n'échappe pas à ce phénomène de massification des données. La disponibilité de ces données constitue une véritable aubaine pour la recherche en santé, notamment dans le domaine de la surveillance épidémique, la prévention des maladies, ou encore la pharmacovigilance.

Nous avons développé des exemples d'application de technologies et de méthodes relatives aux données massives en santé. Durant ce travail de thèse, nous avons implémenté deux outils de visualisation et de fouille de données médicales et testés sur des cas d'usage de pharmacovigilance ou de suivi de pertinence des soins. En modélisant de façon simple une trajectoire patient, nous avons tenté de répondre à des questions cliniques de pharmacovigilance ou encore de surveillance de bon usage du médicament. Cependant, ces outils ne sont pas adaptés à tous les aspects de la pharmacovigilance, notamment la recherche d'effets secondaires non décrits.

La première application est un outil de visualisation sous forme de timeline. Cet outil permet d'afficher sur une frise chronologique toutes les informations contenues dans le dossier patient. Celle-ci résume l'information de façon originale en agrégeant l'information structurée et en intégrant des données hétérogènes avec chacune leur représentation visuelle. Elle permet aussi la recherche de termes dans les documents textuels. L'application a démontré son utilisabilité et son efficience lors des évaluations de l'utilisabilité et de mesure de l'impact sur les pratiques. Ces évaluations ont été réalisées avec des cas d'usage issus de la vie réelle sur des dossiers patients.

La deuxième application permet d'extraire, de visualiser, et d'explorer une cohorte de séquences de patients. Il est apparu que les choix faits en matière de traitement des données étaient primordiaux pour pouvoir vérifier correctement une hypothèse posée. L'outil a donc été conçu pour supporter toutes les étapes de visualisation : de la recherche de séquences dans la base de données, en passant par le traitement et la transformation des données, jusqu'à l'affichage à l'écran. Grâce à l'implémentation de différents algorithmes (Smith-Waterman, Apriori, et GSP), nous avons permis à l'utilisateur de rechercher des séquences similaires à une séquence de référence, ou de visualiser toutes les séquences d'événements récurrentes dans une cohorte. L'évaluation que nous avons réalisée montre que l'application aide à retrouver plus rapidement des cas de non-suivi des bonnes pratiques d'administration médicamenteuse. Le clustering des séquences est une fonctionnalité en cours d'étude qui permettrait de regrouper des séquences similaires automatiquement, ceci dans une approche plus exploratoire des données, afin de retrouver des variables discriminantes entre les groupes de patient (la prise d'un médicament ou un diagnostic spécifique à un groupe de patients par exemple).

Les travaux réalisés dans cette thèse montrent le potentiel des outils de visualisation et de fouilles de données pour la pharmacovigilance. D'autres outils pourraient être mis en place notamment pour la détection automatique du signal en

pharmacovigilance. Ces outils sont utiles pour une analyse avec des hypothèses déjà connues, une approche de type apprentissage automatique pourrait être pertinente pour détecter des séquences d'événements dans une cohorte par exemple.

D'autre part, l'impact de ces outils pourrait être beaucoup plus large que le domaine de la pharmacovigilance et les bonnes pratiques de prescriptions et il serait peut-être pertinent de les tester sur des cas d'usage d'autres domaines tels que la qualité des soins ou le phénotypage de patients.

Conclusion Générale

Le travail de cette thèse a porté sur l'analyse et visualisation de trajectoires de soins par l'exploitation de données massives hospitalières pour la pharmacovigilance. Nous avons retracé rapidement l'historique du développement des technologies de l'information ayant abouti au phénomène des données massives. L'essor du Big Data en santé est en effet corrélé à l'informatisation des systèmes d'information hospitaliers et à l'augmentation du périmètre couvert par ces systèmes. L'apparition des entrepôts de données biomédicaux a permis de stocker et réutiliser toutes les informations numérisées produites à l'hôpital par la capacité de traitement facilité (intégration et traitement en amont), et la possibilité de faire des outils facilement connectables à leurs données. Le Big Data en santé constitue aujourd'hui un marché en pleine explosion dirigé par le besoin de contrôler la hausse du coût des soins de santé et améliorer la prise en charge des patients.

Dans ce travail, nous nous sommes essentiellement concentrés sur le domaine de la pharmacovigilance, qui est une activité de surveillance du médicament après son autorisation de mise sur le marché, ou encore à des cas d'usage de suivi du bon usage du médicament. Le domaine d'utilisation de ces outils pourrait être élargi à des problématiques de groupage ou de phénotypage de patients. L'hypothèse posée dans cette thèse est qu'une approche visuelle interactive serait adaptée pour l'exploitation des données biomédicales hétérogènes et multi-domaines dans le champ de la pharmacovigilance.

Sur le plan méthodologique, les champs disciplinaires abordés durant cette thèse ont porté sur la visualisation et la fouille de données ainsi que les méthodes de conception et d'évaluation d'interface homme-machine.

Ce travail de thèse s'est effectué en plusieurs étapes. Nous avons tout d'abord réalisé un état de l'art sur le sujet des trajectoires de soins. Ce travail nous a permis de pouvoir définir formellement une trajectoire de soins. Nous avons ensuite commencé le développement de deux prototypes. Une fois le développement des applications suffisamment avancé, des tests d'utilisabilité ou d'impact sur les pratiques ont été fait pour valider l'intérêt de ces outils. Ces évaluations ont été réalisées par des praticiens sur des cas d'usages de la vie réelle.

La première application est un outil de visualisation du dossier patient sous forme de frise chronologique (timeline). Cette application permet de résumer l'information contenue dans le dossier ainsi que de l'explorer de façon plus rapide et plus adaptée aux besoins de la pharmacovigilance. Les évaluations de l'utilisabilité et de l'impact sur les pratiques ont montré des résultats satisfaisants, résultats qui ont été conforté par la mise en production de l'application pour le service de pharmacovigilance.

La deuxième application est un outil de visualisation et fouille d'une cohorte de séquences patient. Cette application permet de rechercher des séquences similaires entre elles ainsi que de retrouver les motifs récurrents dans une cohorte. Ces fonctionnalités sont rendues possibles grâce à l'implémentation des algorithmes

Smith-Waterman, Apriori, et GSP. L'évaluation de l'application nous a montré qu'elle permettait de gagner du temps sur l'activité de repérage des séquences pouvant correspondre à des cas de non-respect des bonnes pratiques d'administrations médicamenteuses.

Parmi les perspectives de cette thèse, pour la partie mono-patient, il serait pertinent de centrer les développements futurs sur la visualisation et l'exploration des documents textuels. Pour l'analyse et la visualisation d'une cohorte de séquences, il s'agira d'approfondir le travail réalisé sur la reconnaissance de séquences similaire entre elles. Ces futurs travaux pourront portés sur le clustering de ces séquences ou une détection automatique de séquences anormales par apprentissage automatique.

Publications

Ledieu T, Bouzillé G, Plaisant C, Thiessard F, Polard E, Cuggia M: Mining clinical big data for drug safety: Detecting inadequate treatment with a DNA sequence alignment algorithm. AMIA 2018, accepté le 12 juin 2018, choisi pour le Knowledge Discovery and Data Mining Student Innovation Award

Ledieu T, Bouzillé G, Polard E, Plaisant C, Thiessard F, Cuggia M: Mining clinical big data for drug safety : Clinical data analytics with time-related graphical user interfaces: application to pharmacovigilance. Frontiers Pharmacology, publié le 30 août 2018

Ledieu T, Bouzillé G, Thiessard F, Berquet K, Van Hille P, Renault E, Polard E, Cuggia M : Timeline representation of clinical data: usability and added value for pharmacovigilance. BMC Medical Informatics and Decision Making ; publiée le 19 octobre 2018

Bibliographie

1. Batzenschlager A, Dorner M, Weill-Bousson M. La pathologie tumorale du thorotrast chez l'homme. Oncology. 1963;16(1):28–63.

2. Vargesson N. Thalidomide-induced teratogenesis: History and mechanisms. Birth Defects Res. 2015 Jun;105(2):140–56.

3. française LD. Rapport sur la surveillance et la promotion du bon usage du médicament en France [Internet]. [cited 2018 May 4]. Available from: http://www.ladocumentationfrancaise.fr/rapports-publics/134000617/index.shtml

4. Dreyfus JC. Maladie de Creutzfeldt-Jakob et hormone de croissance. MS Médecine Sci Rev Pap ISSN 0767-0974 1986 Vol 2 N° 4 P220 [Internet]. 1986 [cited 2018 Aug 26]; Available from: http://www.ipubli.inserm.fr/handle/10608/3480

5. Tribouilloy C, Jeu A, Maréchaux S, Jobic Y, Rusinaru D, Andréjak M. Benfluorex (Mediator®) et atteintes valvulaires. /data/revues/07554982/v40i11/S0755498211004222/ [Internet]. 2011 Nov 2 [cited 2018 Aug 26]; Available from: http://www.em-consulte.com/en/article/668406

6. Delluc A, Moigne EL, Mottier D. Risque de maladie veineuse thromboembolique chez la femme en âge de procréer. Médecine Thérapeutique. 2011 Oct 1;17(3):213–33.

7. Malformations congénitales chez les enfants exposés in utero au valproate et aux autres traitements de l'épilepsie et des troubles bipolaires - Communiqué - ANSM : Agence nationale de sécurité du médicament et des produits de santé [Internet]. [cited 2018 Aug 26]. Available from: https://ansm.sante.fr/S-informer/Communiques-Communiques-Points-presse/Malformations-congenitales-chez-les-enfants-exposes-in-utero-au-valproate-et-aux-autres-traitements-de-l-epilepsie-et-des-troubles-bipolaires-Communique

8. WHO | Pharmacovigilance [Internet]. WHO. [cited 2017 Aug 21]. Available from: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/

9. Loi n° 93-5 du 4 janvier 1993 relative à la sécurité en matière de transfusion sanguine et de médicament.

10. LOI n° 2004-806 du 9 août 2004 relative à la politique de santé publique. 2004-806 Aug 9, 2004.

11. Pouyanne P, Haramburu F, Imbs JL, Bégaud B. Admissions to hospital caused by adverse drug reactions: cross sectional incidence study. French Pharmacovigilance Centres. BMJ. 2000 Apr 15;320(7241):1036.

12. La stratégie nationale de santé 2018-2022 [Internet]. Ministère des Solidarités et de la Santé. 2017 [cited 2018 May 4]. Available from: http://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/strategie-nationale-de-sante/article/la-strategie-nationale-de-sante-2018-2022

13. Lacoste-Roussillon C, Pouyanne P, Haramburu F, Miremont G, Bégaud B. Incidence of serious adverse drug reactions in general practice: a prospective study. Clin Pharmacol Ther. 2001 Jun;69(6):458–62.

14. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of Adverse Drug Events and Potential Adverse Drug Events: Implications for Prevention. JAMA. 1995 Jul 5;274(1):29–34.

15. Grall J-Y. Réorganisation des vigilances sanitaires [Internet]. Ministère des Affaires Sociales et de la Santé; 2013 Jul. Available from: http://solidarites-sante.gouv.fr/IMG/pdf/Rapport_JY_Grall_-_Reorganisation_des_vigilances_sanitaires.pdf

16. Moride Y, Haramburu F, Requejo AA, Bégaud B. Under-reporting of adverse drug reactions in general practice. Br J Clin Pharmacol. 1997 Feb;43(2):177–81.

17. L'ANSM soutient deux plateformes en épidémiologie des produits de santé - Point d'information - ANSM : Agence nationale de sécurité du médicament et des produits de santé [Internet]. [cited 2018 Jul 29]. Available from: https://www.ansm.sante.fr/S-informer/Points-d-information-Points-d-information/L-ANSM-soutient-deux-plateformes-en-epidemiologie-des-produits-de-sante-Point-d-information

18. École Polytechnique - Accueil site de l'Ecole Polytechnique [Internet]. [cited 2018 May 4]. Available from: https://www.polytechnique.edu/fr/content/lx-et-la-cnam-sunissent-pour-etudier-les-donnees-de-sante

19. Agnès Buzyn lance la mission de préfiguration du « Health Data Hub » un laboratoire d'exploitation des données de santé - Ministère des Solidarités et de la Santé [Internet]. [cited 2018 Jul 30]. Available from: http://solidarites-sante.gouv.fr/actualites/presse/communiques-de-presse/article/agnes-buzyn-lance-la-mission-de-prefiguration-du-health-data-hub-un-laboratoire

20. Glaser BG, Strauss AL. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine; 1967. 292 p.

21. Hornbrook MC, Hurtado AV, Johnson RE. Health care episodes: definition, measurement and use. Med Care Rev. 1985;42(2):163–218.

22. Perrier L. MODES DE FINANCEMENT DU SERVICE PUBLIC HOSPITALIER ET TRAJECTOIRE OPTIMALE DU PATIENT EN CANCÉROLOGIE PÉDIATRIQUE. [Lyon 2]; 2001.

23. Baszanger I. Les chantiers d'un intéractionniste américain. In: La trame de la négociation. A. Strauss; 1992. p. 11–63.

24. Duru G. P.M.S.I. Psychiatrie : l'approche « trajectoire de soins ». Ministère du Travail et des Affaires Sociales, Direction des Hôpitaux;

25. Vincent M, Lamure M, Pelc A, Dardennes J. Trajectoires patients par fouille de données. Document de travail du L.A.S.S.,; 1998.

26. GREMY F. Filières et réseaux. Vers l'organisation et la coordination du système de soins. Gest Hosp. 1997 07;(367):433–8.

27. Jay N. Découverte et représentation des trajectoires de soins par analyse formelle de concepts [Internet] [phdthesis]. Université Henri Poincaré - Nancy I; 2008 [cited 2018 Apr 12]. Available from: https://tel.archives-ouvertes.fr/tel-00585411/document

28. Riou F, Jarno P. Représentation et modélisation des trajectoires de soins. ITBM-RBM. 2000 Oct 1;21(5):313–7.

29. Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artificial Intelligence in Medicine [Internet]. 2012 Sep [cited 2015 Jun 10]; Available from: http://www.sciencedirect.com/science/article/pii/S0933365712000656

30. Press G. A Very Short History Of Big Data [Internet]. Forbes. [cited 2018 Feb 8]. Available from: https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

31. What Is Web 2.0 [Internet]. [cited 2018 Apr 17]. Available from: http://oreilly.com{file}

32. Big data : l'explosion de la production de données [Internet]. egora.fr. 2018 [cited 2018 Aug 11]. Available from: https://www.egora.fr/actus-medicales/sante-publique/39040-big-data-l-explosion-de-la-production-de-donnees

33. Codd EF. A Relational Model of Data for Large Shared Data Banks. Commun ACM. 1970 Jun;13(6):377–387.

34. Haerder T, Reuter A. Principles of Transaction-oriented Database Recovery. ACM Comput Surv. 1983 Dec;15(4):287–317.

35. Brewer EA. Towards Robust Distributed Systems (Abstract). In: Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing [Internet]. New York, NY, USA: ACM; 2000 [cited 2018 Apr 13]. p. 7–. (PODC '00). Available from: http://doi.acm.org/10.1145/343477.343502

36. Redis [Internet]. [cited 2018 Apr 13]. Available from: https://redis.io/

37. Apache CouchDB [Internet]. [cited 2018 Apr 13]. Available from: http://couchdb.apache.org/

38. MongoDB for GIANT Ideas [Internet]. MongoDB. [cited 2018 Apr 13]. Available from: https://www.mongodb.com/index

39. The Neo4j Graph Platform – The #1 Platform for Connected Data [Internet]. Neo4j Graph Database Platform. [cited 2018 Apr 13]. Available from: https://neo4j.com/

40. Graph Database | Multi-Model Database [Internet]. OrientDB. [cited 2018 Apr 13]. Available from: https://orientdb.com/

41. Amdahl GM. Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities, Reprinted from the AFIPS Conference Proceedings, Vol. 30 (Atlantic City, N.J., Apr. 18 #x2013;20), AFIPS Press, Reston, Va., 1967, pp. 483 #x2013;485, when Dr. Amdahl was at International Business Machines Corporation, Sunnyvale, California. IEEE Solid-State Circuits Soc Newsl. 2007 Summer;12(3):19–20.

42. Korpela E, Werthimer D, Anderson D, Cobb J, Lebofsky M. SETI@home—Massively Distributed Computing for SETI. Comput Sci Eng. 2001 Jan 1;3(1):78–83.

43. Bégaud B, Polton D, Von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé L'exemple du médicament [Internet]. Rapport réalisé à la demande de Madame la Ministre de la santé Marisol Touraine; [cited 2018 Jul 16]. Available from: http://solidarites-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf

44. Villani C. Donner un sens à l'intelligence artificielle. Mission confiée par le Premier Ministre Edouard Philippe; 2018 Mar.

45. Wade TD. Traits and types of health data repositories. Health Inf Sci Syst. 2014 Dec 1;2(1):4.

46. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc. 2012 Jun 1;19(e1):e119–24.

47. Inmon B. Data Warehousing in a Healthcare Environment [Internet]. TDAN.com. [cited 2018 Apr 16]. Available from: http://tdan.com/data-warehousing-in-a-healthcare-environment/4584

48. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar 1;17(2):124–30.

49. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. AMIA Annu Symp Proc. 2009;2009:391–5.

50. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform. 2011;169:584–8.

51. Bouzillé G, Osmont M-N, Triquet L, Grabar N, Rochefort-Morel C, Chazard E, et al. Drug safety and big clinical data: Detection of drug-induced anaphylactic shock events. J Eval Clin Pract. 2018;24(3):536–44.

52. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert M-L, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Comput Methods Programs Biomed. 2018 Feb;154:153–60.

53. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2014 Jan;26(1):97–107.

54. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Data Bases [Internet]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1994 [cited 2018 Apr 13]. p. 487–499. (VLDB '94). Available from: http://dl.acm.org/citation.cfm?id=645920.672836

55. Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In: Advances in Database Technology — EDBT '96 [Internet]. Springer, Berlin, Heidelberg; 1996 [cited 2018 Apr 13]. p. 1–17. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/BFb0014140

56. Ayres J, Flannick J, Gehrke J, Yiu T. Sequential PAttern mining using a bitmap representation. 2002. 429 p.

57. Zaki MJ. SPADE: An Efficient Algorithm for Mining Frequent Sequences. Mach Learn. 2001 Jan 1;42(1–2):31–60.

58. Chen Y, Guo J, Wang Y, Xiong Y, Zhu Y. Incremental Mining of Sequential Patterns Using Prefix Tree. In: Advances in Knowledge Discovery and Data Mining [Internet]. Springer, Berlin, Heidelberg; 2007 [cited 2018 Apr 13]. p. 433–40. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-540-71701-0_43

59. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In: Proceedings of the 17th International Conference on Data Engineering [Internet]. Washington, DC, USA: IEEE Computer Society; 2001 [cited 2018 Apr 13]. p. 215–. (ICDE '01). Available from: http://dl.acm.org/citation.cfm?id=876881.879716

60. Yang Z, Wang Y, Kitsuregawa M. LAPIN: Effective Sequential Pattern Mining Algorithms by Last Position Induction for Dense Databases. In: Advances in Databases: Concepts, Systems and Applications [Internet]. Springer, Berlin, Heidelberg; 2007 [cited 2018 Apr 17]. p. 1020–3. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-540-71703-4_95

61. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar 28;48(3):443–53.

62. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195–7.

63. Zweigenbaum P, Jacquemart P, Grabar N, Habert B. Building a text corpus for representing the variety of medical language. Stud Health Technol Inform. 2001;84(Pt 1):290–4.

64. Hamon T, Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. J Am Med Inform Assoc. 2010 Jan 9;17(5):549–54.

65. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform. 2007 Sep;8(5):358–75.

66. MedDRA | [Internet]. [cited 2018 Aug 5]. Available from: https://www.meddra.org/

67. Charlet J, Declerck G, Dhombres F, Gayet P, Miroux P, Vandenbussche P-Y. Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In: 23es journées francophones d'Ingénierie des connaissances [Internet]. Paris, France; 2012 [cited 2018 Aug 26]. p. 33–48. (IC 2012). Available from: https://hal.archives-ouvertes.fr/hal-00717807

68. Guefack PSVD. Modélisation des signes dans les ontologies biomédicales pour l'aide au diagnostic. [Internet] [phdthesis]. Université Rennes 1; 2013 [cited 2018 Aug 26]. Available from: https://tel.archives-ouvertes.fr/tel-01057310/document

69. Anscombe FJ. Graphs in Statistical Analysis. Am Stat. 1973;27(1):17–21.

70. Fry B. Visualizing Data. 1st ed. 2008.

71. Lamy J-B, Duclos C, Hamek S, Beuscart-Zéphir M-C, Kerdelhué G, Darmoni S, et al. Towards iconic language for patient records, drug monographs, guidelines and medical search engines. Stud Health Technol Inform. 2010;160(Pt 1):156–60.

72. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: using visualization to enhance navigation and analysis of patient records. Proc AMIA Symp. 1998;76–80.

73. Wongsuphasawat K, Gotz D. Outflow: Visualizing Patient Flow by Symptoms and Outcome. In.

74. Alexander Rind SM. VisuExplore: Gaining New Medical Insights from Visual Exploration. 2010;

75. Monroe M, Lan R, Lee H, Plaisant C, Shneiderman B. Temporal event sequence simplification. IEEE Trans Vis Comput Graph. 2013 Dec;19(12):2227–36.

76. Gschwandtner T, Aigner W, Kaiser K, Miksch S, Seyfang A. CareCruiser: Exploring and visualizing plans, events, and effects interactively. In: Visualization Symposium (PacificVis), 2011 IEEE Pacific. 2011. p. 43–50.

77. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artif Intell Med. 2006 Oct;38(2):115–35.

78. Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, et al. RAVEL: retrieval and visualization in ELectronic health records. Stud Health Technol Inform. 2012;180:194–8.

79. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AAT. Context-Based Electronic Health Record: Toward Patient Specific Healthcare. IEEE Trans Inf Technol Biomed. 2012 Mar;16(2):228–34.

80. Zhu X, Gold S, Lai A, Hripcsak G, Cimino JJ. Using Timeline Displays to Improve Medication Reconciliation. In: 2009 International Conference on eHealth, Telemedicine, and Social Medicine. 2009. p. 1–6.

81. Inspired EHRs | Timeline [Internet]. [cited 2018 Feb 20]. Available from: http://inspiredehrs.org/timeline/

82. Nielsen J. Why You Only Need to Test with 5 Users [Internet]. 2000 [cited 2017 Aug 21]. Available from: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

83. Rubin J, Chisnell D. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. John Wiley & Sons; 2011. 353 p.

84. ISO 9241-11:1998(en), Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability [Internet]. [cited 2017 Aug 21]. Available from: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en

85. Nair KM, Malaeekeh R, Schabort I, Taenzer P, Radhakrishnan A, Guenter D. A Clinical Decision Support System for Chronic Pain Management in Primary Care: Usability testing and its relevance. J Innov Health Inform. 2015 Aug 13;22(3):329–32.

86. Kamal J, Rogers P, Saltz J, Mekhjian H. Information warehouse as a tool to analyze Computerized Physician Order Entry order set utilization: opportunities for improvement. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2003;336–40.

87. Neri PM, Pollard SE, Volk LA, Newmark LP, Varugheese M, Baxter S, et al. Usability of a novel clinician interface for genetic results. J Biomed Inform. 2012 Oct;45(5):950–7.

88. Nielsen J. Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.

89. Olmsted-Hawala EL, Murphy ED, Hawala S, Ashenfelter KT. Think-aloud Protocols: A Comparison of Three Think-aloud Protocols for Use in Testing Data-dissemination Web Sites for Usability. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 2010 [cited 2016 Jan 6]. p. 2381–2390. (CHI '10). Available from: http://doi.acm.org/10.1145/1753326.1753685

90. Severity Ratings for Usability Problems: Article by Jakob Nielsen [Internet]. [cited 2017 Sep 18]. Available from: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

91. Peute LWP, de Keizer NF, Jaspers MWM. The value of Retrospective and Concurrent Think Aloud in formative usability testing of a physician data query tool. J Biomed Inform. 2015 Jun;55:1–10.

92. Brooke J. SUS: A quick and dirty usability scale. 1996.

93. Osmont M-N, Cuggia M, Polard E, Riou C, Balusson F, Oger E. Utilisation du PMSI pour la détection d'effets indésirables médicamenteux. Thérapie. 2013 Jul 1;68(4):285–95.

94. R Development Core Team (2008). R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing. Vienna, Austria; Available from: https://www.r-project.org/

95. Martins SB, Shahar Y, Galperin M, Kaizer H, Goren-Bar D, McNaughton D, et al. Evaluation of KNAVE-II: a tool for intelligent query and exploration of patient data. Stud Health Technol Inform. 2004;107(Pt 1):648–52.

96. Malik S, Shneiderman B, Du F, Plaisant C, Bjarnadottir M. High-Volume Hypothesis Testing: Systematic Exploration of Event Sequence Comparisons. ACM Trans Interact Intell Syst. 2016 Mar;6(1):9:1–9:23.

97. Malik S, Du F, Monroe M, Onukwugha E, Plaisant C, Shneiderman B. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. In: Proceedings of the 20th International Conference on Intelligent User Interfaces [Internet]. New York, NY, USA: ACM; 2015 [cited 2018 Apr 16]. p. 38–49. (IUI '15). Available from: http://doi.acm.org/10.1145/2678025.2701407

98. Kwon BC, Verma J, Perer A. Peekquence: Visual Analytics for Event Sequence Data. In 2016.

99. Du F, Plaisant C, Spring N, Shneiderman B. Finding Similar People to Guide Life Choices: Challenge, Design, and Evaluation. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 2017 [cited 2018 Jan 23]. p. 5498–5544. (CHI '17). Available from: http://doi.acm.org/10.1145/3025453.3025777

100. Monroe M, Meyer TE, Plaisant C, Lan R, Wongsuphasawat K, Coster TS, et al. Visualizing Patterns of Drug Prescriptions with EventFlow: A Pilot Study of Asthma Medications in the Military Health System. 2013 Jun.

101. Huang L-T, Wu C-C, Lai L-F, Li Y-J. Improving the Mapping of Smith-Waterman Sequence Database Searches onto CUDA-Enabled GPUs [Internet]. BioMed Research International. 2015 [cited 2018 Jul 29]. Available from: https://www.hindawi.com/journals/bmri/2015/185179/

102. Dumont E, Merialdo B. Rushes Video Parsing Using Video Sequence Alignment. In: 2009 Seventh International Workshop on Content-Based Multimedia Indexing. 2009. p. 44–9.

103. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol. 1982 Dec 15;162(3):705–8.

104. Knockout : Home [Internet]. [cited 2018 Apr 13]. Available from: http://knockoutjs.com/

105. Bostock M. D3.js - Data-Driven Documents [Internet]. [cited 2017 Aug 21]. Available from: https://d3js.org/

106. Eclipse Vert.x [Internet]. [cited 2018 Apr 13]. Available from: https://vertx.io/

107. Foundation N js. Node.js [Internet]. Node.js. [cited 2018 Apr 13]. Available from: https://nodejs.org/en/

108. Fournier Viger P, Lin C-W , Gomariz A, Gueniche T, Soltani A, Deng Z-H, et al. The SPMF Open-Source Data Mining Library Version 2. In 2016. p. 36–40.

109. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms:. the C-value/NC-value method. Int J Digit Libr. 2000 Aug 1;3(2):115–30.

Table des figures

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

Titre : Analyse et visualisation de trajectoires de soins par l'exploitation de données massives hospitalières pour la pharmacovigilance

Mots clés : Données Massives Hospitalières – Visualisation de données – Trajectoires de soins - Analyse de séquences - Pharmacovigilance

Résumé : Dans les travaux effectués au cours de cette thèse, nous présenterons des approches permettant d'exploiter la richesse et le volume des données intra hospitalières pour des cas d'usage de pharmacovigilance et de surveillance de bon usage du médicament. Cette approche reposera sur la modélisation de trajectoires de soins intra hospitalières adaptées aux besoins spécifiques de la pharmacovigilance. L'hypothèse posée dans cette thèse est qu'une approche visuelle interactive serait adaptée pour l'exploitation de ces données biomédicales hétérogènes et multi-domaines dans le champ de la pharmacovigilance.

Nous avons développé deux prototypes permettant la visualisation et l'analyse des trajectoires de soins. Le premier prototype est un outil de visualisation du dossier patient sous forme de frise chronologique. La deuxième application est un outil de visualisation et fouille d'une cohorte de séquences d'événements. Ce dernier outil repose sur la mise en œuvre d'algorithme d'analyse de séquences (Smith-Waterman, Apriori, GSP) pour la recherche de similarité ou de motifs d'événements récurrents. Ces interfaces homme-machine ont fait l'objet d'études d'utilisabilité sur des cas d'usage tirées de la pratique réelle qui ont prouvé leur potentiel pour un usage en routine.

Title : Analysis and visualization of care trajectories by using hospital big data for pharmacovigilance

Keywords: Hospital Big Data – Data Visualization – Care Trajectories – Sequence Analysis – Pharmacovigilance

Abstract: In this thesis work, we will present approaches to exploit the diversity and volume of intra-hospital data for pharmacovigilance use and monitoring the proper use of drugs. This approach will be based on the modelling of intra-hospital care trajectories adapted to the specific needs of pharmacovigilance. Using data from a hospital warehouse, it will be necessary to characterize events of interest and identify a link between the administration of these health products and the occurrence of adverse reactions, or to look for cases of misuse of the drug.

We have developed two prototypes allowing the visualization and analysis of care trajectories. The first prototype is a tool for visualizing the patient file in the form of a timeline. The second application is a tool for visualizing and searching a cohort of event sequences. The latter tool is based on the implementation of sequence analysis algorithms (Smith-Waterman, Apriori, GSP) for the search for similarity or patterns of recurring events. These human-machine interfaces have been the subject of usability studies on use cases from actual practice that have proven their potential for routine use.