GINSENG, une infrastructure de grille au service de l'e-santé et de l'épidémiologie

Sébastien Cipière[1,2,3], Paul De Vlieger[1],Sébastien Gaspard[5], David Manset[5], Jérôme Revillard[5], David Sarramia[1,2], David R.C. Hill[1,2,4], Lydia Maigne[1,2]
cipiere@clermont.in2p3.fr, vlieger@moniut.univ-bpclermont.fr, sgaspard@maatg.fr, dmanset@maatg.fr, jrevillard@maatg.fr, sarramia@clermont.in2p3.fr, david.hill@univ-bpclermont.fr, maigne@clermont.in2p3.fr.
[1] Clermont Université, Université Blaise Pascal, LPC, BP 10448, F-63000
[2] CNRS/IN2P3, UMR 6533, LPC, F-63177
[3] CNRS, UMR 6158, Université Blaise Pascal, LIMOS, F-63173
[4] ISIMA, Institut Supérieur d'Informatique, de Modélisation et de leurs Applications, BP 10125, F-63177
[5] MAAT France, GNUBILA Group, Argonay, F-74370

Résumé. Le projet GINSENG a pour but de mettre en place une infrastructure de grille pour l'e-santé et l'épidémiologie en Auvergne. Il s'agit de mettre en réseau des bases de données médicales distribuées de manière sécurisée afin de pouvoir les interroger et obtenir des informations statistiques pour des études épidémiologiques. L'objectif étant de décentraliser au maximum les données médicales et bénéficier des technologies de grille informatique pour la mise en place de l'infrastructure. Les sites médicaux participant au projet se rassemblent autour de deux thématiques médicales : le suivi des cancers (plus particulièrement les cancers suivis dans le cadre d'un dépistage organisé) et les soins périnataux (11 maternités). Dans chaque site médical est déployé un serveur disposant des services de grille sur lequel est dupliquée la base de données médicale. La solution adoptée permet un haut niveau de sécurité, de confidentialité, de disponibilité, et de tolérance aux fautes. Les requêtes effectuées sur les bases de données médicales distribuées sont réalisées via un portail web sécurisé. Les requêtes utilisent pour effectuer leurs calculs le logiciel R en mode distribué via des services webs. Le service de santé publique bénéficie de cette infrastructure informatique pour la veille sanitaire, des études épidémiologiques plus spécifiques et l'évaluation des pratiques médicales.

Mots clés: grille, bases de données, e-santé, épidémiologie, réseau de surveillance.

Introduction

Le projet français GINSENG (Global Initiative for Sentinel E-health Network on Grid), financé depuis 2011 par l'Agence Nationale pour la Recherche (ANR) et pour une période de 3 ans, vise à mettre en place un réseau sentinelle sur grille informatique pour l'e-santé et l'épidémiologie en France, la région Auvergne étant la région pilote. Ce projet associe plusieurs partenaires publics et privés pour mettre en réseau les bases de données médicales de plusieurs sites hospitaliers et les laboratoires de pathologie en région).
Ce projet bénéficie pour l'instant à deux applications médicales majeures : il participe à l'amélioration de la qualité du dépistage des cancers grâce à un partage sécurisé, rapide et sans erreurs des données d'intérêt majeur dans le suivi des patients atteints de la maladie (comme les données anatomo-cytopathologiques). Les cancers du sein, du col de l'utérus et du colon sont ceux actuellement concernés par le dépistage organisé. De la même façon, le projet GINSENG bénéficie au suivi des parturientes et de leurs nouveau-nés afin d'améliorer leur prise en charge pendant toute la grossesse et les premières périodes de la vie du bébé.
D'autres projets européens ont déjà posé les jalons de l'utilisation d'une grille informatique pour l'e-santé : les projets MAMMOGRID (Amendolia et al. 2004), MEDIGRID (Montagnat et al. 2005) et Health-e-Child (Freund et al. 2006) ont d'ailleurs permis le développement de middleware de grille et d'outils webs pour la gestion des données et des images médicales. Le problème posé par la gestion centralisée des données du patient à travers un seul serveur n'est pas facilement solvable à l'échelle d'un pays, comme nous avons pu le constater pour le Dossier Médical Partagé (DMP) en France (Manaouil 2009). Elle est encore plus problématique à

l'échelle de l'Europe. Le développement de la technologie des grilles informatiques ouvre la perspective d'un stockage distribué des données médicales dans des dépôts sécurisés. Dans le cadre du projet GINSENG, l'infrastructure déployée permet de relever le défi de l'information épidémiologique en temps réel par une réactivité du système bien plus accrue que s'il s'agissait d'un système centralisé. L'accès sécurisé au réseau est effectué par le biais du portail web www.e-ginseng.org. Il permet entre autre aux utilisateurs de s'authentifier et d'avoir une vue d'ensemble de l'activité de santé pour le dépistage et de diagnostic des cancers et la périnatalité en région Auvergne. Dans cette publication, nous présentons tout d'abord le contexte scientifique et médical qui a favorisé la mise en place du projet GINSENG. Une deuxième partie explique l'infrastructure informatique et les différents services proposés. Une dernière partie présente les cas d'utilisation du réseau pour la santé publique.

1.1 Contexte

Ce projet permet d'envisager des innovations importantes dans deux champs majeurs de l'épidémiologie : la veille sanitaire (et l'analyse des facteurs de risque) et l'évaluation tant des politiques de santé que des pratiques médicales. La veille sanitaire est l'un des enjeux primordiaux de la santé publique: l'accroissement de la circulation des hommes sur le globe entraîne un risque accru de propagation de maladies émergentes(Rotureau et al. 2007). Le développement des politiques de santé, et parmi celles-ci les politiques de prévention secondaires, rend nécessaire de disposer d'outils adaptés à leur évaluation. Dans les deux cas, les mesures de risque ou d'efficacité se font à partir de recueils créés ad hoc avec toutes leurs limites : perte d'information, biais de sous déclaration, absence de données pour un risque non connu, biais de mesure (par exemple pour les données de nature médico-économiques). Or, le projet GINSENG se base sur des informations produites dans le cadre courant des soins, sans nouvelles modalités de recueil. Ces données existent indépendamment de toute centralisation et sont utilisées quotidiennement par les professions de santé. L'accès via des techniques de grille permet à la fois une vitesse d'accès à l'information et une exhaustivité accrue. Ce recueil se fait par ailleurs avec de meilleures garanties d'anonymat, seules des données agrégées étant exportées. Le partage sécurisé de données médicales entre différentes structures médicales publiques et/ou privées est à ce jour en pleine mutation technologique. Les technologies proposées doivent rendre possible un partage électronique et sécurisé de ces données de manière à les rendre disponible à tout instant dans le cadre notamment d'une veille sanitaire ou d'analyses épidémiologiques (allant de l'observation sanitaire à l'évaluation de prises en charge ou de politiques de santé). Le projet GINSENG permet de répondre à ce problème en respectant ces contraintes sans nécessité de déclaration médicale mais en interconnectant des bases de données existantes et distribuées. Les avantages majeurs d'un réseau de grille distribué sont :

- Le coût de l'infrastructure.
- Le travail sur des données qui reposent sur le maintien de leur distribution et non leur centralisation.
- L'interopérabilité
- Les alarmes sanitaires efficaces, nécessitant moins d'étapes intermédiaires de déclaration et de traitement des données

Le projet GINSENG s'appuie sur des solutions déjà éprouvées dans le cadre du projet de grille européen European Grid Infrastructure (EGI) (Kranzlmüller et al. 2010). Ces solutions telles qu'AMGA et VOMS (que nous expliciterons en 2.2 et 2.3) permettent la gestion des droits d'accès des utilisateurs et des bases de données distribuées. Le projet européen de grille de calculs a été initié au CERN par les physiciens des particules, qui produisent des quantités de données très importantes au travers des différents capteurs qui constitue le LHC. Ainsi les solutions utilisées au CERN, répondent très bien à nos problématiques où nos quantités de données sont bien moins importantes. Le fait d'utiliser les technologies de grille de calcul permet une décentralisation du système d'information. Ce qui se traduit en un des avantages majeur du projet, la tolérance aux fautes. Contrairement aux systèmes centralisés de type Dossier Médical Partagé (DMP)(Catherine Quantin et al. 2009), où toutes les informations se retrouvent agrégées dans un seul et unique

point géographique, le réseau sentinelle du projet GINSENG est réparti dans tous les sites médicaux appartenant au projet.
L'infrastructure proposée dans la partie suivante fait état des avantages technologiques de l'information médicale répartie sur plusieurs sites et aux mécanismes d'authentification, de sécurité et de tolérance aux fautes pour l'identification du patient.

2 Les services de grille

L'architecture matérielle du projet GINSENG est résumée dans la figure 1. Cette architecture est le support de différents services de grille que nous détaillerons dans les parties suivantes.

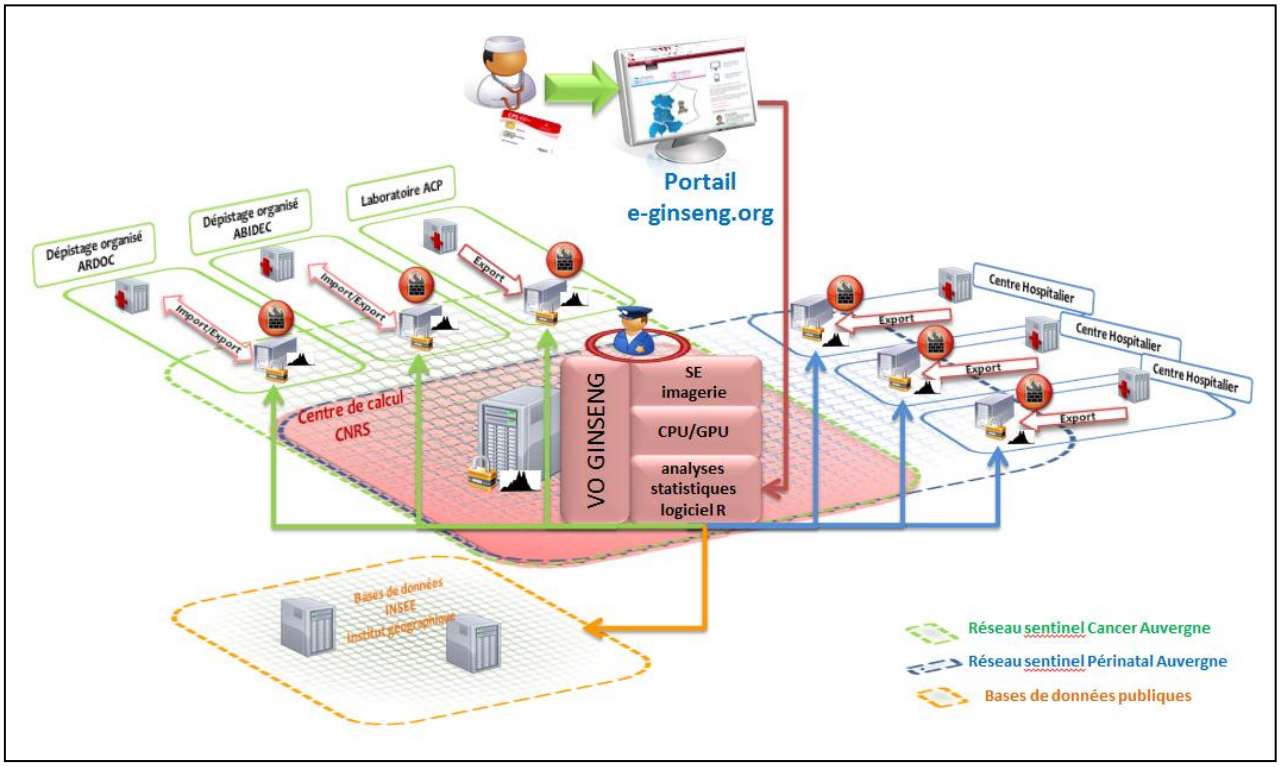

Figure 1 : Architecture du réseau GINSENG

2.1 L'accès au réseau par un serveur Web :www.e-ginseng.org

Le serveur Web héberge le site Internet du projet, qui est le point d'accès unique à l'infrastructure. Le site Internet poursuit trois objectifs principaux, l'information, l'identification, et la gestion des requêtes épidémiologiques. Au travers d'une interface ergonomique et attractive, des informations statistiques sont rendues disponibles aux différentes catégories d'utilisateurs. Quatre groupes d'utilisateurs sont identifiés : les biostatisticiens, épidémiologistes les chercheurs en santé publique, les praticiens hospitaliers, les structures de santé publique de l'État (Institut National de Veille Sanitaire, Agences régionales de santé), et la population française. Pour chacun de ces groupes d'utilisateurs, nous avons prévu une interface différente, qui est adaptée à leurs besoins. Cette interface est en quelque sorte un tableau de bord qui peut intégrer différents outils permettant de retranscrire les informations calculées à l'intérieur du système. Les différents outils que nous avons à notre disposition pour communiquer l'information sont des tableaux, des graphiques (histogrammes, courbes,…), ou encore des informations géolocalisables de type graphiques sur représentation planisphère en 3D. Le citoyen lambda n'a pas besoin de s'identifier pour accéder aux données publiques publiées sur le site. Cet accès le connecte à une interface simple qui surveille en "temps réel" quelques informations de santé publique d'intérêt général. Une authentification est nécessaire pour les utilisateurs

autorisés sur le réseau. Les utilisateurs sont identifiés par un serveur Central Authentification Service (CAS) grâce à leur carte de professionnel de santé (CPS). Le serveur CAS permet une seule et unique authentification de l'utilisateur, tout en lui octroyant l'accès à tous les services, ainsi c'est le serveur CAS qui authentifie utilisateurs auprès du Virtual Organization Membership Service (VOMS)(Alfieri et al. 2005). En fonction de leur profession et de leur rôle au sein de l'organisation, les personnes authentifiées n'ont pas accès aux mêmes tableaux de bord. Les médecins ont accès à une interface de consultation de données de références au niveau régional. Pour les épidémiologistes ils sont capables de tester et valider de nouvelles hypothèses sous forme de requêtes épidémiologiques. Une fois validée, les requêtes peuvent être mises à disposition sur le portail comme des « sentinelles sanitaires. Les utilisateurs authentifiés peuvent s'il le désire souscrire à une ou plusieurs alertes de veille sanitaire.
Une Virtual Organisation (VO) ou Organisation Virtuelle a été créée spécifiquement pour les besoins du projet. Elle a pour nom VO Sentinelle. D'après (Cummings et al. 2008) une VO regroupe une communauté d'utilisateurs partageant des ressources de la cyber-infrastructure suivant une politique d'usage commune. Autrement dit, c'est la structure sociale résultant d'une cyber-infrastructure. Les VO sont souvent multidisciplinaires. Elles sont disséminées dans le temps et l'espace. Leur existence est permise par la puissance informatique (construites sur des grilles) et leur production est facilitée par l'accès à distance à de larges banques de données. Au sein d'une VO, les utilisateurs bénéficient de droits spécifiques. Il est ainsi possible de partager les droits d'accès aux fichiers des utilisateurs d'une même VO. Il existe aussi la notion de groupes, qui permet d'obtenir des droits spécifiques la VO (administration ou simple utilisateur par exemple). Des rôles peuvent aussi être attribués à des utilisateurs afin d'ajuster encore plus finement les droits. Les groupes permettent aussi, en cas d'extension du réseau sentinelle, de définir des droits propres à une application sans devoir créer une nouvelle VO. Par exemple un groupe « cancer-depistage » au sein de la VO Sentinelle rassemble tous les acteurs ayant besoin de récupérer les comptes rendus pathologiques nominatifs tandis qu'un groupe « cancer-epidemio » agrège seulement les utilisateurs ayant besoin de lancer des requêtes statistiques. L'accès à cette VO sera conditionné par une authentification auprès d'un serveur VOMS qui sera relié à un serveur CAS.

2.2 VOMS

VOMS est un système qui gère l'ensemble des autorisations des composants de grille. Il permet surtout aux utilisateurs de s'authentifier sur la VO et bien entendu faire que cette authentification ait validité sur l'ensemble de l'infrastructure de grille que nous utilisons. La collaboration entre VOMS et les Autorités de Certification (AC) est très étroite et constitue la clé de voûte de toute la sécurité du système. Toute demande d'authentification utilisant un certificat passe par une vérification qui dans notre cas s'appuie sur l'utilisation des cartes CPS. VOMS est aussi le service qui, lors de l'authentification d'un utilisateur sur une VO, crée un proxy grille, issue de son certificat, lui permettant d'utiliser les ressources de la VO pour une durée comprise entre quelques heures et quelques jours (24h le plus souvent). Le proxy est alors valide sur l'ensemble de la VO pour cette durée, que ce soit pour la gestion de données ou l'exécution de tâches. Cette durée limitée est un élément supplémentaire de sécurité, car si une personne arrive à intercepter le proxy, ou arrive à s'authentifier sur une machine où un utilisateur dispose d'un proxy, son champ d'action reste limité dans le temps.

2.3 AMGA

AMGA, ARDA Metadata Catalog Project (Koblitz et al. 2007) est un catalogue de métadonnées pour les environnements de grille. Le besoin de ce type de logiciel pour la grille s'est vite rendu primordial car les systèmes de gestion des données comme le Logical File Catalog (LFC), ne permettaient pas d'annoter suffisamment les fichiers stockés sur la grille. Le LFC se chargeant uniquement d'associer à un nom logique, un (ou plusieurs) emplacement(s) physique(s) sur la grille. Le plus souvent, la mise à disposition de fichiers sur la grille ne permettait pas à un autre utilisateur de les exploiter correctement, souvent par manque d'information sur le contenu du fichier. Un catalogue de métadonnées permet d'associer à un élément un certain nombre d'informations supplémentaires caractérisant cet élément. AMGA diffère aussi d'une base de données

classique en proposant une interface arborescente aux données. La base dispose ainsi d'une racine et d'un ensemble de répertoires qui contiennent un ensemble d'entrées et d'attributs caractérisant ces entrées. C'est grâce à AMGA que seront agrégées les informations qui serviront de base aux requêtes R des épidémiologistes, que nous présenterons dans la partie 2.4.
Le support matériel de ces services est une machine que nous appelons gateway. La gateway est la partie visible de chaque site, toutes les gateway sont interconnectés pour former un réseau. Le réseau ainsi créé constitue notre base de données distribuée. Dans un souci de sécurité maximale et de disponibilité des données patients, nous mettons en œuvre différents protocoles tel que la fragmentation de la base de données et une duplication de ces fragments. Ce qui nous permet de continuer d'avoir accès à la totalité des informations contenues dans la base de données bien que certains sites puissent être inaccessible.

2.4 Le logiciel d'analyses statistiques

Le logiciel R (Bates et al. 2010) est issu d'un projet open source très largement répandu dans la communauté des statisticiens. R est disponible sous licence GNU GPL sur le site http://www.r-project.org/. L'un de ses principaux avantages outre le fait qu'il soit distribué librement est de permettre d'effectuer des requêtes distribuées. Dans la figure 1 les machines ayant vocation à effectuer des requêtes R sont symbolisées par une icône représentant une courbe. Il sera interrogé par les épidémiologistes au travers d'une console située à l'intérieur du site Internet. Nous présenterons dans la partie suivante des cas d'utilisations auxquels les épidémiologistes peuvent répondre au moyen de l'outil que nous mettons à leur disposition.

3 La gestion des requêtes épidémiologiques

Pour qu'une enquête épidémiologique ait des résultats fiables il faut absolument que les données en entrée soient les plus propres possible. Le nettoyage des jeux de données est une étape préalable fastidieuse et souvent plus coûteuse que l'analyse en elle-même des données. C'est pourquoi les analyses épidémiologiques sur le cancer ne sont souvent disponibles qu'après 3 années de travail. Une étude menée par (Friedman & Sideli 1992) a montré jusqu'à 27% d'erreurs d'identification de patients au sein de trois bases de données hospitalières de 100.000 patients sur le même site. Le but du rapprochement des identités médicales distribuées est alors de réduire drastiquement le temps de disponibilité des données pour l'épidémiologie. L'objectif final étant de les mettre à disposition en temps quasi réel, c'est-à-dire peu de temps après l'intégration des données au réseau.

3.1 L'identification des patients

Le système a pour but de mettre en commun des bases de données provenant de divers sites médicaux. Il est alors primordial de s'assurer de l'identité des patients lors de l'utilisation de leurs données médicales pour des études épidémiologiques. Il est possible par exemple, qu'un patient consulte dans plusieurs sites médicaux, d'autre part la probabilité qu'un homonyme fréquente le même établissement de santé n'est pas négligeable. Il est aussi envisageable que lors de l'enregistrement d'un dossier dans le système d'information une erreur se soit produite. Ces trois différents cas doivent être correctement pris en charge par l'infrastructure. En effet en aucun cas le système ne devra associer des patients distincts à une même identité. De la même façon qu'il est important que le système puisse associer deux dossiers provenant de deux services d'information différents au bon patient. Comme notre solution est développée sur le territoire français elle respecte les critères de la Commission Nationale de l'Informatique et des Libertés (CNIL) qui garantit notamment le respect des conditions relatives à la gestion des bases de données intégrant des informations personnelles.

3.1.1 Les algorithmes d'identification

Pour répondre aux différentes problématiques liées à l'identification des patients nous avons retenu deux algorithmes : Jaro Winkler (Jaro 1995)et Soundex (C Quantin et al. 2004) . L'algorithme Jaro Winkler est basé sur l'analyse de chaînes de caractères. L'algorithme Soundex est quant à lui un algorithme qui compare les données en se basant sur les phonèmes, ce qui permet de considérer « Philippe » et une version mal orthographiée de ce prénom telle que « filipe » comme étant similaire. Ces deux algorithmes auront à comparer

les données des patients ; les informations que nous avons retenues sont le prénom, le nom, la date de naissance, et l'adresse. Pour améliorer la vitesse de traitement des algorithmes nous étudions actuellement des solutions basées sur la puissance des cartes graphiques de type GP GPU(Owens et al. 2008), qui seront déployées sur le serveur ayant la charge de l'identification des patients.

3.2 Les cas d'utilisation de l'infrastructure

Dans cette partie nous présenterons les avantages que peut procurer le projet GINSENG à ses utilisateurs, au travers d'exemples concrets.

3.2.1 La veille sanitaire pour les cancers du dépistage organisé

Le principe fondamental de la veille sanitaire est la surveillance continue et rapprochée d'un processus temporel ou spatio-temporel (par exemple le nombre de cas d'une maladie) afin d'en évaluer les paramètres (la vitesse moyenne d'apparition des nouveaux cas, saisonnalité, tendance séculaire) et de déterminer si et quand ces paramètres évoluent de façon anormale. Une évolution anormale peut signifier l'apparition d'un problème de santé publique (sous forme épidémique ou endémique) justifiant le déclenchement d'une alerte sanitaire (facilement propagée grâce au site Internet). Cette surveillance n'a d'intérêt que lorsque elle est capable d'aboutir au déclenchement rapide de l'alerte dès que le processus est dit « hors contrôle » c'est-à-dire dés le début d'une épidémie afin que de nouvelles mesures de régulation plus adaptée puissent être prises dans un délai suffisamment court pour être efficaces. La qualité d'un système de surveillance est en partie dépendante de la rapidité avec laquelle il a accès aux données relatives au processus étudié. Ces données doivent décrire le processus selon les trois domaines « temps-lieux-personnes », ce qui est le cas des données manipulées à l'intérieur de GINSENG. Une fois la définition d'un cas posée, chacun doit pouvoir être enregistré par le système avec le repère chronologique (e.g. date de début de la maladie) et spatial pertinents (e.g. lieu d'habitation), ainsi que tout autre facteur potentiellement source de confusion (e.g. l'âge pour les cancers, la centralisation des moyens de diagnostic,…). En particulier, l'évolution d'une maladie étant toujours liée à l'évolution de la population à risque de cette maladie, la taille de cette population est toujours prise en compte dans la surveillance. De par la nature des données sous-jacentes, la construction du signal relatif au processus est presque toujours issue du croisement de plusieurs sources de données. Dans le cadre du cancer, les bases de données des laboratoires d'anatomo-cytopathologie doivent être croisées avec celles des centres de dépistage afin d'assurer la meilleure exhaustivité dans le repérage des cas mais aussi dans le recueil des variables qui leurs sont associées (repères chronologiques et temporels, autres variables d'intérêt). Ces données concernant les cas sont ensuite croisés avec une source de données de population de l'Institut National de la Statistique et des Études Économiques (INSEE) comme le montre la figure 1. L'infrastructure de GINSENG permet d'utiliser plusieurs sources de données médicales exhaustives, pouvant être entrecroisé pour permettre des analyses de veille sanitaire efficace.

3.2.2 Intérêts dans les études épidémiologiques pour le cancer

Les questions élémentaires en épidémiologie relatives aux cancers concernent essentiellement l'étude des facteurs de risque de ces cancers. La portée des pathologies cancéreuses est classiquement quantifiée à l'aide de ce que les épidémiologistes appellent des « mesures de risque », dont certaines sont assimilables à des probabilités comme la prévalence (probabilité de cancer dans une population à un instant donnée) ou le risque proprement dit (probabilité de survenue du cancer dans une population « susceptible », i.e. initialement indemne de cancer, pendant une période de surveillance donnée) ou qui en dérivent comme les cotes (*odds* en anglais, rapport d'une probabilité et de son complément), et d'autres assimilables à une force ou vitesse de survenue du cancer comme l'incidence. En fonction du statut des individus vis-à-vis d'un facteur d'exposition, les épidémiologistes utilisent des « mesures d'associations de risque » sous la forme de différence ou de rapport de ces mesures de risque de façon à quantifier la force d'association entre cancer et facteur d'exposition. Les mesures d'association de risque les plus utilisées sont les rapports de cotes ou *odds ratio* (en anglais), notés OR et qui constituent une estimation ponctuelle de la force d'association entre cancer et exposition. Par exemple, un OR inférieur à 1 signe le caractère protecteur de l'exposition du facteur vis-à-vis du

risque. De façon analogue un facteur associé à un OR à 1 sera réputé neutre, et un facteur associé à un OR supérieur à 1 sera réputé être un « facteur de risque » de cancer. Pour donner une portée statistique à cette estimation ponctuelle de l'OR, il est nécessaire d'assortir cette mesure de ce que les biostatisticiens appellent un intervalle de confiance (en général calculé à 95%, noté IC95%) qui consiste en deux bornes à l'intérieur desquelles la vraie valeur de l'OR (qui n'est qu'estimée ponctuellement) a de grandes chances de se trouver (assimilable à une probabilité de 0.95). Enfin, lorsque les épidémiologistes constatent que la valeur 1 n'appartient pas à l'IC95%, ils concluent, si l'OR est supérieure à 1, à l'existence d'un facteur de risque significatif, qui doit dans l'idéal permettre la mise en œuvre de mesure de prévention dans le but d'en réduire le risque. Si le sujet ne peut être soustrait à ce facteur d'exposition, comme le sexe, par exemple, alors les épidémiologistes parlent de marqueur de risque, auquel cas, ils peuvent utiliser cette information pour cibler les sous-groupes à dépister ou à surveiller de façon particulièrement assidue. Le nouvel outil que constitue GINSENG est adapté au calcul du risque, de la prévalence et de l'OR en se basant sur les archives dont disposent les laboratoires participants au projet.

3.2.3 L'évaluation des pratiques médicales pour les soins périnataux

Une infrastructure de bases de données médicales distribuées peut également permettre de réaliser plus simplement une évaluation précise des pratiques médicales de par l'utilisation de données supplémentaires et nécessaires au bon établissement d'un diagnostic ou au meilleur suivi des patients. À titre d'exemple la pratique d'un frottis cervico-vaginal (FCV) avant le début de la grossesse est importante en ce qu'elle renseigne sur le suivi préalable de la femme et sur le traitement avant la grossesse de pathologies infectieuses (herpès génital, infection à Chlamydiae, gonocoque,…) qui peuvent avoir des conséquences péjoratives pour le bébé lors d'une naissance par voie basse. Or, il s'agit d'un item qui est mal renseigné dans le dossier de référence en obstétrique (dossier dit AUDIPOG – association des utilisateurs du dossier informatisé en pédiatrie obstétrique et gynécologie). Or, le suivi de la pratique régulière de FCV est l'un des points centraux de la politique régional de dépistage organisé des cancers du col utérin. Il est alors possible de croiser les données du dossier AUDIPOG avec celles du dépistage organisé des cancers du col utérin pour compléter les données manquantes. Il est également possible d'analyser la nature générale de ces données manquantes et notamment de savoir si l'absence d'informations concerne plutôt les femmes ayant eu des FCV normaux (ce qui peut justifier soit que le FCV soit oublié, soit qu'il ne soit pas noté dans le dossier puisque n'ayant pas de signification péjorative pour la grossesse) ou si une partie importante des FCV non recueillis dans l'AUDIPOG concerne des frottis pathologiques, ce qui serait une situation à risque de perte de chance pour les bébés.

Conclusion

Dans cette publication nous avons présenté le projet GINSENG qui permet de partager des données médicales entre différents sites et de les interroger dans un but statistique. Les différents apports majeurs de GINSENG par rapport aux solutions existantes ou concurrentes sont un gain de temps dans le transfert, l'automatisation, la qualité de l'information, tout en assurant un niveau de sécurité très élevé. Pour l'instant les données gérées sont des données textuelles il est cependant envisageable à moyen terme que nous ajoutions le traitement de l'imagerie médicale. L'architecture matérielle et logicielle sur laquelle nous nous appuyons nous permettra le passage à l'échelle, lorsque nous déciderons d'étendre la solution en dehors de la région Auvergne ou à d'autres domaines médicaux en plus de la périnatalité et de la cancérologie.

Références

Alfieri, R. et al., 2005. From gridmap-file to VOMS: managing authorization in a Grid environment. *Future Generation Computer Systems*, 21(4), pp.549-558. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167739X04001682.

Amendolia, S.R. et al., 2004. MammoGrid: A Service Oriented Architecture based Medical Grid Application. *Grid and Cooperative*, p.12. Available at: http://arxiv.org/abs/cs/0405074.

Bates, D. et al., 2010. The R Project. *Power*, (Spring). Available at: http://www.r-project.org/.

Cummings, J. et al., 2008. *Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations*, National Science Foundation. Available at: http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf.

Freund, J. et al., 2006. Health-e-child: an integrated biomedical platform for grid-based paediatric applications. *Studies In Health Technology And Informatics*, 120, pp.259-270. Available at: http://arxiv.org/abs/cs/0603036.

Friedman, C. & Sideli, R., 1992. Tolerating spelling errors during patient validation. *Computers and biomedical research an international journal*, 25(5), pp.486-509.

Jaro, M.A., 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5-7), pp.491-498. Available at: http://doi.wiley.com/10.1002/sim.4780140510.

Koblitz, B., Santos, N. & Pose, V., 2007. The AMGA Metadata Service. *Journal of Grid Computing*, 6(1), pp.61-76. Available at: http://www.springerlink.com/index/10.1007/s10723-007-9084-6.

Kranzlmüller, D., Lucas, J.M. & Öster, P., 2010. The European Grid Initiative (EGI). In F. Davoli et al., eds. *Remote Instrumentation and Virtual Laboratories*. Springer US, pp. 61-66. Available at: http://www.springerlink.com/content/g02w441w533n4811.

Manaouil, C., 2009. Le dossier médical personnel (DMP) : « autopsie » d'un projet ambitieux ? *Médecine & Droit*, 2009(94), pp.24-41. Available at: http://www.sciencedirect.com/science/article/pii/S1246739109000037.

Montagnat, J., Breton, V & E Magnin, I., 2005. Partitioning medical image databases for content-based queries on a Grid. *Methods of Information in Medicine*, 44(2), pp.154-160. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15924166.

Owens, J.D. et al., 2008. GPU Computing. *Proceedings of the IEEE*, 96(5), pp.879-899. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4490127.

Quantin, C et al., 2004. Estimation de la valeur discriminante des traits d'identification utilisés pour le rapprochement des données d'un patient. *Revue d'Épidémiologie et de Santé Publique*, 52(5), pp.431-440. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0398762004990797.

Quantin, Catherine et al., 2009. Centralised versus decentralised management of patients' medical records. *Studies In Health Technology And Informatics*, 150, pp.700-704.

Rotureau, B. et al., 2007. International Epidemic Intelligence at the Institut de Veille Sanitaire, France. *Emerging Infectious Diseases*, 13(10), pp.1590-1592.