THÈSE DE DOCTORAT DE

l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Biostatistique

École Doctorale Pierre Louis de Santé Publique à Paris : Épidémiologie et Sciences de l'Information Biomédicale

Présentée par

Yohann MANSIAUX

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Analyse d'un grand jeu de données en épidémiologie : problématiques et perspectives méthodologiques

soutenue le 30 octobre 2014

devant le jury composé de :

M. Rodolphe THIÉBAUT	Rapporteur
M. Jean-Claude DESENCLOS	Rapporteur
M. Arnaud FONTANET	Examinateur
M. Didier GUILLEMOT	Examinateur
M. Charles-Edouard LUYT	Examinateur
M. Fabrice CARRAT	Directeur de thèse

Laboratoire de rattachement

UMR-S 1136, INSERM - Université Pierre et Marie Curie

Institut Pierre Louis d'Épidémiologie et de Santé Publique
Responsable : Dominique Costagliola

Équipe "Épidémiologie de la grippe et des hépatites virales : risque, pronostic et stratégies thérapeutiques"

Faculté de Médecine Pierre et Marie Curie, site Saint-Antoine
27 rue de Chaligny, 75571 Paris cedex 12, France

Remerciements

Arrivé au terme de cette aventure, c'est avec une certaine émotion que je tiens à remercier tous les acteurs qui ont permis, chacun à leur échelle, que cette thèse ait pu être menée à son terme dans la sérénité. Résumer trois années en quelques lignes est un exercice délicat, j'espère que celles-ci seront à la hauteur du plaisir que j'ai eu à travailler auprès de toutes ces personnes, de mon premier jour dans l'unité jusqu'à ma soutenance.

Je commencerais par remercier Fabrice Carrat de m'avoir accueilli dans son équipe durant mon stage de Master 2, de m'avoir fait découvrir l'épidémiologie et bien sur de m'avoir encadré pour cette thèse. Merci pour ta rigueur scientifique et ta disponibilité, ainsi que pour m'avoir accordé une grande liberté et une grande autonomie dans mon travail au quotidien. Tout ceci m'a permis de m'accomplir, non seulement en tant que scientifique, mais aussi en tant que personne.

Mes remerciements vont ensuite vers Nathanaël Lapidus, qui m'a encadré lors de mes premiers pas dans l'unité. Ses précieux conseils (toujours pertinents) et sa patience pour faire face à mes multiples questionnements, ont été d'une grande aide dans l'élaboration des travaux présentés dans ce document et m'ont permis d'aborder certaines périodes difficiles sereinement.

Je tiens ensuite à remercier Rosemary Delabre. Nos échanges, probablement plus récréatifs que riches en grandes réflexions scientifiques, m'ont permis de travailler au quotidien dans une atmosphère joyeuse et ont une part prépondérante dans le plaisir que j'ai eu à passer ces trois années dans l'unité ; ils auront dans tous les cas développé mon intérêt pour la cuisine américaine et pour l'œnologie en général.

Je remercie Magali Lemaitre, pour avoir favorisé mon intégration dans l'unité et pour nos conversations sur des questions professionnelles, des questionnements existentiels ou sur des sujets beaucoup plus légers. Isabelle Goderel, pour sa bonne humeur légendaire, et Frédéric Chau, pour sa connaissance sans faille des sportifs réunionnais, ne peuvent être oubliés. Merci également à tous mes collègues et ex-collègues de l'unité pour le plaisir que j'ai eu à travailler à vos côtés.

Je remercie bien évidemment mes amis pour leur soutien. En dépit du retour récurrent de la fameuse question "alors elle en est où ta thèse ?", leur présence à mes côtés a été une aide plus que précieuse.

Enfin, je souhaite remercier tout particulièrement Hélène, pour sa présence, son soutien sans faille et ses encouragements durant ces trois ans. Je conclurais en remerciant mes parents, pour m'avoir accompagné et conforté dans les choix que j'ai fait ces dernières années.

En conclusion et au risque de me répéter, merci à tous pour votre soutien, vos conseils, votre intérêt et votre bienveillance. Une thèse est loin d'être un travail solitaire, c'est la somme de ces petites attentions qui m'a permis de mener cette aventure à bien, et je ne peux que me souhaiter d'être aussi bien entouré pour la suite !

Résumé

L'augmentation de la taille des jeux de données est une problématique croissante en épidémiologie. La cohorte CoPanFlu-France (1450 sujets), proposant d'étudier le risque d'infection par la grippe pandémique H1N1pdm comme une combinaison de facteurs biologiques, environnementaux, sociodémographiques et comportementaux, et dans laquelle plusieurs centaines de variables sont récoltées pour chaque patient, en est un exemple. Les méthodes statistiques usuelles pour explorer des associations ont de nombreuses limites dans un tel contexte. Nous comparons l'apport de méthodes exploratoires *data-driven*, supposant l'absence d'hypothèses *a priori*, à celui de méthodes *hypothesis-driven*, nécessitant l'élaboration d'hypothèses préalables.

Une première approche *data-driven* est présentée, évaluant la capacité à détecter des déterminants de l'infection grippale de deux méthodes de *data mining*, les forêts aléatoires (FA) et les arbres de régression *boostés* (ARB), de la méthodologie d'analyse conventionnelle en régression logistique (régressions univariées suivie d'une régression multivariée – RURM) et de la régression logistique pénalisée par le *Least Absolute Shrinkage and Selection Operator* (LASSO), introduisant une pénalité dans le modèle logistique multivarié permettant d'effectuer une sélection des covariables importantes. Une approche par simulation a permis d'évaluer les taux de vrais et de faux positifs (TVP et TFP) de ces méthodes. Entre trois et vingt-quatre déterminants de l'infection ont été identifiés, le titre d'anticorps pré-épidémique étant la seule variable unanimement détectée. Les TVP moyens étaient les plus élevés pour FA (85%) et ARB (80%), suivis du LASSO (jusqu'à 78%), tandis que la méthodologie RURM était très peu performante (moins de 50%). Une légère augmentation du risque de première espèce (TFP moyen jusqu'à 9%) a été observée pour les modèles basés sur la régression logistique, LASSO y compris, tandis que les méthodes de *data mining* ont montré un TFP moyen de 4%.

Nous proposons ensuite une étude causale *hypothesis-driven* du risque d'infection, par le biais d'un modèle d'équations structurelles (SEM). Nous avons exploité la spécificité des SEM de modéliser des variables latentes pour étudier des facteurs très divers, leur impact relatif sur l'infection ainsi que leurs relations éventuelles. Seules les variables latentes décrivant la susceptibilité de l'hôte (modélisée par le titre d'anticorps pré-épidémique) et l'adoption de comportements préventifs étaient directement associées avec le risque d'infection. Les facteurs comportementaux décrivant la perception du risque et des mesures préventives ont montré une influence positive sur l'adoption de comportements préventifs. Aucun effet indépendant de l'intensité (durée et nombre) des contacts sociaux n'a pu être montré.

Cette thèse montre la nécessité de considérer de nouvelles approches statistiques pour l'analyse des grands jeux de données en épidémiologie. Le *data mining* et le LASSO sont des alternatives crédibles

aux outils conventionnels pour la recherche d'associations parmi un grand nombre de variables. Les SEM permettent l'intégration de variables décrivant différentes dimensions et la modélisation explicite de leurs relations ; ces modèles sont dès lors d'un intérêt majeur dans une étude multidisciplinaire comme CoPanFlu.

Summary

The increasing size of datasets is a growing issue in epidemiology. The CoPanFlu-France cohort (1450 subjects), intended to study H1N1 pandemic influenza infection risk as a combination of biological, environmental, socio-demographic and behavioral factors, and in which hundreds of covariates are collected for each patient, is a good example. The statistical methods usually employed to explore associations have many limits in this context. We compare the contribution of data-driven exploratory methods, assuming the absence of *a priori* hypotheses, to hypothesis-driven methods, requiring the development of preliminary hypotheses.

Firstly a data-driven study is presented, assessing the ability to detect influenza infection determinants of two data mining methods, the random forests (RF) and the boosted regression trees (BRT), of the conventional logistic regression framework (Univariate Followed by Multivariate Logistic Regression - UFMLR) and of the Least Absolute Shrinkage and Selection Operator (LASSO), with penalty in multivariate logistic regression to achieve a sparse selection of covariates. A simulation approach was used to estimate the True (TPR) and False (FPR) Positive Rates associated with these methods. Between three and twenty-four determinants of infection were identified, the pre-epidemic antibody titer being the unique covariate selected with all methods. The mean TPR were the highest for RF (85%) and BRT (80%), followed by the LASSO (up to 78%), while the UFMLR methodology was inefficient (below 50%). A slight increase of alpha risk (mean FPR up to 9%) was observed for logistic regression-based models, LASSO included, while the mean FPR was 4% for the data-mining methods.

Secondly, we propose a hypothesis-driven causal analysis of the infection risk, with a structural-equation model (SEM). We exploited the SEM specificity of modeling latent variables to study very diverse factors, their relative impact on the infection, as well as their eventual relationships. Only the latent variables describing host susceptibility (modeled by the pre-epidemic antibody titer) and compliance with preventive behaviors were directly associated with infection. The behavioral factors describing risk perception and preventive measures perception positively influenced compliance with preventive behaviors. The intensity (number and duration) of social contacts was not associated with the infection.

This thesis shows the necessity of considering novel statistical approaches for the analysis of large datasets in epidemiology. Data mining and LASSO are credible alternatives to the tools generally used to explore associations with a high number of variables. SEM allows the integration of variables describing diverse dimensions and the explicit modeling of their relationships ; these models are therefore of major interest in a multidisciplinary study as CoPanFlu.

Mots-clés

- Grippe H1N1 pandémique
- Big data
- Data-driven
- Hypothesis-driven
- Data mining
- Régression pénalisée LASSO
- Régression logistique
- Équations structurelles

Keywords

- H1N1 pandemic influenza
- Big data
- Data-driven
- Hypothesis-driven
- Data mining
- LASSO penalized regression
- Logistic regression
- Structural equations

Production scientifique au cours de la thèse

1. Production liée au sujet de thèse

Articles publiés

- **Contribution of genome-wide association studies to scientific research : a bibliometric survey of the citation impacts of GWAS and candidate gene studies published during the same period and in the same journals.** Mansiaux Y, Carrat F. PLoS One. 2012 ;7(12) :e51408. doi : 10.1371/journal.pone.0051408. Epub 2012 Dec 11.
- **Detection of independent associations in a large epidemiologic dataset : a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections.** Mansiaux Y, Carrat F. BMC Medical Research Methodology 2014, 14 :99.

Articles soumis

- **Causal analysis of H1N1pdm09 influenza infection risk in a household cohort** Mansiaux Y, Salez N, Lapidus N, Setbon M, Andreoletti L, Leruez-Ville M, Cauchemez S, Gougeon ML, Vély F, Schwarzinger M, Abel L, Markovic Delabre R, Flahault A, de Lamballerie X, Carrat F. En révision dans le Journal of Epidemiology and Community Health.
- **Risk factors of pandemic influenza A/H1N1 in a prospective household cohort in the general population : results from the CoPanFlu-France cohort.** Markovic Delabre R, Lapidus N, Salez N, Mansiaux Y, de Lamballerie X, Carrat F. En révision dans le journal Influenza and Other Respiratory Viruses.

Communication orale

- **Détection d'associations dans un grand jeu de données en épidémiologie - Comparaison des performances du data mining, de la régression logistique conventionnelle et de la régression logistique pénalisée pour détecter des associations indépendantes.** Mansiaux Y, Carrat F. Conférence "Epiclin 8, 21èmes Journées des Statisticiens du CLCC" (Bordeaux, France, 05/2014).

Communications affichées

- **Quantifying the respective roles of exposure and susceptibility to H1N1pdm infection : a structural equation modeling approach.** Mansiaux Y, Carrat F. Conférence "Options for the Control of Influenza VIII" (Le Cap, Afrique du Sud, 09/2013).
- **Causal analysis of H1N1pdm09 influenza infection risk in a household cohort.** Mansiaux Y, Lapidus N, Markovic Delabre R, Carrat F. Conférence "Epiclin 8, 21èmes Journées des Statisticiens du CLCC" (Bordeaux, France, 05/2014).

2. Production hors du sujet de thèse

Article publié

- **Clusterin Expression in Gastrointestinal NeuroEndocrine Tumours is Highly Correlated with the Location and it is Helpful to Determine the Origin of Liver Metastases.** Mourra N, Scriva A, Mansiaux Y, Gozlan S, Bennis M, Balaton A. Histopathology. 2014 May 7. doi : 10.1111/his.12450.

Table des matières

1 Introduction **1**

2 Article : L'apport des grands jeux de données en épidémiologie, l'exemple des GWAS **6**

2.1 Résumé . . . 6

2.2 Article . . . 7

2.3 Discussion . . . 16

3 L'étude CoPanFlu-France : une approche intégrative du risque d'infection grippale **18**

4 La recherche d'associations dans un grand jeu de données en épidémiologie : problématiques et approches exploratoires **21**

4.1 La recherche d'associations en épidémiologie hypothesis-driven . . . 21

4.2 Problématiques de l'analyse d'un grand jeu de données et limites de la régression . . . 23

4.3 Analyse statistique des genome-wide association studies . . . 25

4.4 Méthodes de data mining . . . 25

4.4.1 Réseaux de neurones artificiels . . . 28

4.4.2 Machines à vecteurs de support . . . 29

4.4.3 Arbres de classification et de régression . . . 30

4.4.4 Bagging . . . 33

4.4.5 Forêts aléatoires . . . 33

4.4.6 Boosting . . . 35

4.4.7 Performances prédictives des méthodes d'ensemble . . . 38

4.4.8 Comment extraire de la connaissance des modèles data mining ? . . . 39

4.5 Méthodes de régression pénalisée . . . 44

4.5.1 Régression ridge . . . 44

4.5.2 Régression LASSO . . . 45

4.5.3 Régression elastic net . . . 46

4.5.4 Estimation des paramètres de pénalisation dans la régression pénalisée . . . 47

5 Article : le risque d'infection grippale abordé sous l'angle data-driven **48**

5.1 Résumé . . . 48

5.2 Article . . . 49

5.3 Discussion . . . 69

5.3.1 Détection d'associations indépendantes . . . 69

5.3.2 Utilisation des mesures d'importance des variables . . . 70
5.3.3 Tests de permutation pour la détection d'associations . . . 71
5.3.4 Limites du plan d'analyse classique en régression logistique . . . 72

6 Les modèles d'équations structurelles, une alternative hypothesis-driven aux méthodes de régression 73

6.1 Introduction . . . 73
6.2 Concepts généraux . . . 73
6.2.1 Variables observées et variables latentes . . . 73
6.2.2 Variables exogènes et variables endogènes . . . 74
6.2.3 L'analyse factorielle . . . 74
6.2.4 Le modèle d'équations structurelles complet . . . 76
6.3 Le diagramme de chemin et les équations structurelles . . . 76
6.4 Construction d'un modèle d'équations structurelles . . . 79
6.4.1 Spécification du modèle . . . 79
6.4.2 Identifiabilité du modèle . . . 79
6.4.3 Estimation du modèle . . . 80
6.4.4 Evaluation globale du modèle . . . 81
6.4.5 Évaluation des paramètres estimés . . . 82
6.4.6 Manipulation du modèle . . . 83
6.4.7 Modèle final . . . 83
6.5 Avantages et limites des modèles d'équations structurelles . . . 84

7 Article : Le risque d'infection grippale abordé sous l'angle hypothesis-driven 86

7.1 Résumé . . . 86
7.2 Article . . . 87
7.3 Discussion . . . 131
7.3.1 L'absence d'association entre l'infection et les contacts sociaux . . . 131
7.3.2 Modélisation de la susceptibilité individuelle . . . 132
7.3.3 L'infection vue comme une variable latente . . . 132

8 Conclusions et perspectives 135

Références 139

Introduction

Analyste des données (*data scientist*) est devenu le métier “le plus sexy du XXIème siècle”, voici l’opinion de la *Harvard Business Review* [53]. La “science des données” (*data science*), un terme inventé il y a à peine plus de cinq ans par deux analystes des sociétés *LinkedIn* et *Facebook*, serait la discipline scientifique amenée à prendre une place prépondérante dans les prochaines années [107]. Cette tendance trouve sa source dans l’explosion du *“big data”*. En sciences de l’information, ce terme décrit le recueil et la gestion de bases de données se distinguant par un volume important, une variété des types de données et une grande vitesse de génération [194]. L’observation de l’évolution du nombre de requêtes contenant le terme *“big data”* dans le moteur de recherche *Google* depuis 10 ans (Figure 1.1) montre qu’il s’agit à la fois d’un phénomène non seulement récent mais également en pleine expansion.

Le “déluge de données” touche des domaines très variés. Dans le commerce et le marketing, les entreprises cherchent à récolter un maximum d’informations sur les consommateurs (notamment via Internet) afin d’élaborer des profils clients aussi complets que possible. Les technologies modernes, smartphones et réseaux sociaux notamment, génèrent également une énorme quantité de données ; à titre d’exemple si le réseau social *Facebook* était un pays, celui-ci serait la troisième nation la plus peuplée du monde avec plus d’un milliard d’habitants, offrant des perspectives inégalées en terme d’analyses sociales ou comportementales [131]. La politique est également concernée ; il a en effet été observé qu’une analyse approfondie *big data*, notamment par l’étude de facteurs géographiques et socio-démographiques, de résultats d’élections antérieures et de sondages d’opinion visant à connaitre les thèmes majeurs susceptibles d’attirer un maximum d’électeurs, aurait joué un rôle prépondérant dans la victoire de Barack Obama lors des élections présidentielles américaines de 2012 [88, 159].

Ces courts exemples ne représentent qu’une petite part de la place prise par le *big data* tant les secteurs d’activités concernés sont divers. D’une manière générale, on estimait début 2013 que plus de 2.5 trillions d’octets étaient générés chaque jour dans le monde [117]. De même, une publication récente affirmait début 2014 que 90% des données mondiales avaient été produites ces deux dernières années [189]. Le besoin de *data scientists*, capables de traiter ces données massives, les analyser et en extraire de l’information prend ici tout son sens.

Les sciences biomédicales sont bien sur également concernées par cette problématique de l’accroissement du nombre de données. Le succès du projet de séquençage du génome humain en 2001 [103] a entrainé l’émergence d’études "*genome-wide association studies*" (GWAS) proposant d’analyser des centaines de milliers, voire plusieurs millions, de variants génétiques et leurs liens potentiels avec des maladies ou des traits phénotypiques [86]. Les technologies d’analyse “haut-débit” [172] se sont de-

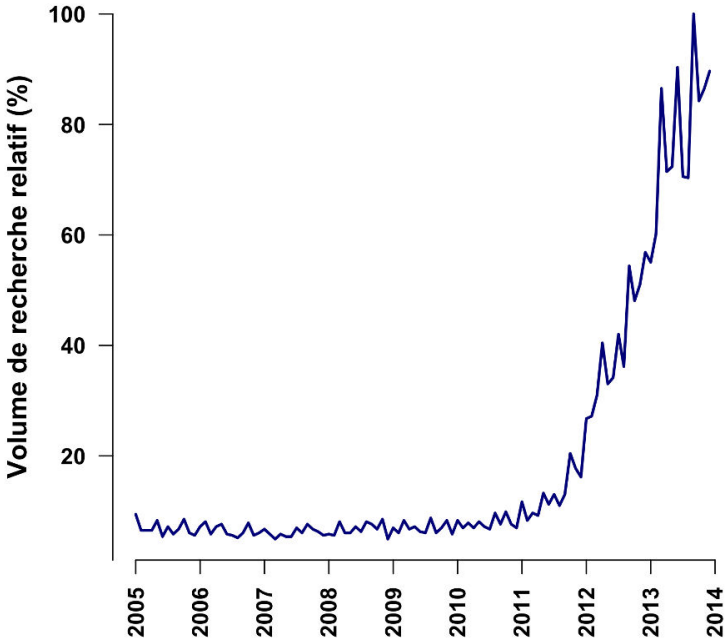


FIGURE 1.1 : Évolution du nombre de requêtes contenant le terme *"big data"* dans le moteur de recherche *Google* (données fournies par l'outil Google Trends ®) entre 2005 et 2014. Les données représentées correspondent au volume de recherche mensuel après comparaison avec le volume le plus élevé indiqué sur le graphique.

puis développées et la philosophie *wide association study* s'applique maintenant à des domaines très divers, proposant d'étudier non seulement le génome (le matériel génétique d'un individu) mais aussi le transcriptome (l'ensemble des ARNs issus de la transcription), le métabolome (les métabolites présents dans un échantillon biologique) ou encore le protéome (l'ensemble des protéines exprimées dans une cellule ou un tissu à un instant donné).

Ces différentes études peuvent être englobées sous un terme plus général de sciences "*omiques*". Le nombre de thématiques *omiques* se chiffre aujourd'hui en plusieurs centaines, celles pertinentes d'un point de vue pratique étant probablement bien moins nombreuses. Outre les dimensions citées précédemment, un intérêt croissant est montré pour des domaines tels que le variome (l'ensemble des variations génétiques au sein d'une population), l'épigénome (les éléments non codés dans l'ADN contrôlant l'expression des gènes) ou encore l'interactome (les interactions moléculaires dans un système) [14]. La biologie systémique, décrivant l'intégration de ces divers niveaux d'information [48,93], offre de nouvelles possibilités d'exploration plus globale (et non concentrée sur un pan particulier de la biologie) des systèmes biologiques [47, 65]. Les bénéfices attendus en terme de compréhension des causes et des mécanismes des maladies, notamment en terme de personnalisation du suivi médical, sont très importants [44]. La médecine conventionnelle avec un diagnostic orienté sur les symptômes a certaines limitations ; l'utilisation de profils *omiques* détaillés pourrait outrepasser ces limitations

en permettant un diagnostic et un traitement individualisés [45, 113]. Enfin, l'intégration de la biologie systémique avec les données de dossiers médicaux électroniques (permettant de définir un profil médical exhaustif pour chaque individu : antécédents médicaux, résultats d'analyse, traitements ...) offre également de nombreuses promesses, par exemple en terme de compréhension des relations génotype/phénotype [90, 133].

L'explosion de l'usage d'Internet et des téléphones mobiles a donné naissance à un nouveau genre d'épidémiologie : l'épidémiologie digitale [162]. Les outils de communication modernes permettent d'accéder à une énorme quantité d'information utilisable en épidémiologie, par exemple sur le mode de vie ou le déplacement des individus, information qu'il aurait été impossible de recueillir par des approches traditionnelles, par exemple par des questionnaires. Le service *Google Flu Trends*, proposant de suivre les épidémies de grippe à partir des requêtes effectuées par les utilisateurs du moteur de recherche *Google*, est un des exemples les plus connus de l'application de l'épidémiologie digitale [75]. Si la fiabilité de ce service est encore perfectible [106, 142], d'autres exemples, notamment basés sur l'utilisation du réseau social *Twitter* ont permis de modéliser avec succès la propagation de maladies infectieuses [149,166,200], ou encore la détection d'intoxications alimentaires suite à un repas pris dans un restaurant [160]. L'épidémiologie digitale reste une discipline jeune et ses applications sont encore relativement peu nombreuses. Ces médias modernes, de part la quantité de données qu'ils peuvent permettre de récolter, sont des outils avec un grand potentiel. De plus, contrairement à des données fournies par un réseau sentinelle de médecins ou un centre de surveillance, les données sont ici obtenues quasi-instantanément, ceci étant d'un intérêt majeur par exemple dans la surveillance de la propagation d'une épidémie.

Un autre domaine en expansion en épidémiologie est l'exposome. Ce concept, introduit en 2005 par Christopher Wild [208], définit l'ensemble des expositions environnementales auxquelles un individu est confronté de sa naissance jusqu'à son décès, et ayant un impact sur l'environnement chimique interne [151]. Sous le terme expositions on peut, entre autres grandeurs, regrouper le régime alimentaire, la pollution sous toutes ses formes, le mode de vie, les infections potentielles, les traitements médicamenteux ou encore le stress.

La prise en compte des facteurs environnementaux est essentielle en épidémiologie, il est en effet estimé que les facteurs non-génétiques contribuent pour environ 90% au risque d'avoir une maladie chronique [150]. L'utilisation de ces nouvelles approches environnementales offre de nouvelles perspectives dans la découverte de facteurs de risque et dans l'optique d'une compréhension accrue de l'étiologie des maladies. Les applications des techniques *wide association study* à des données environnementales sont encore rares [109, 144], et ne concernent à ce jour qu'un nombre limité d'expositions.

La multiplication de nouvelles approches et le développement d'approches déjà existantes supposent, en plus de nouvelles perspectives de compréhension des maladies, de nouvelles problématiques d'un point de vue de l'analyse statistique. Les méthodes statistiques usuellement employées en épidémiologie (notamment les différents types de régressions) pour tester des associations sont limitées dans un contexte d'augmentation de la taille des jeux de données. L'utilisation de méthodes spécifiques, dites *data-driven*, devient dès lors indispensable. Le terme *data-driven* suppose l'absence d'hypothèses *a priori* au moment de l'analyse des données. Cette philosophie est à l'opposé de l'épidémiologie *hypothesis-driven* plus classique, pour laquelle un certain nombre d'hypothèses sont faites quant aux associations qu'on peut s'attendre à observer, par exemple entre un trait ou une maladie et un certain nombre de variables pré-sélectionnées, et sur la forme de ces associations (linéaire ou non). Les approches exploratoires *data-driven*, par leur côté "pêche à la ligne" [28], sont attendues comme étant de grands générateurs d'hypothèses, hypothèses à confirmer dans des études de validation approfondies [198].

L'objectif de ce travail est de présenter des méthodes statistiques alternatives à celles habituellement utilisées en épidémiologie pour l'analyse d'un jeu de données de grande dimension. Les problématiques propres aux grands jeux de données en épidémiologie sont abordées, ainsi que les limitations des méthodes usuelles dans un tel contexte. Différentes pistes sont proposées, notamment via des modèles s'inspirant de ceux utilisés dans l'analyse de données *omiques*, en particulier dans les GWAS.

L'ensemble des travaux présentés dans ce document porte sur les données récoltées dans l'étude CoPanFlu-France [105]. Cette grande cohorte nationale lancée en 2009 propose une approche intégrative pour étudier le risque d'infection par la grippe pandémique H1N1 (H1N1pdm). Le caractère "intégratif" de cette étude sous-entend la transdisciplinarité [199]. Dans CoPanFlu, le risque d'infection grippale est abordé comme une combinaison complexe de facteurs biologiques, environnementaux, socio-démographiques ou encore comportementaux ; ceci suppose la collection d'un ensemble large et varié de données pour chaque individu. Le déroulement de cette étude ainsi que les données récoltées sont détaillés dans le chapitre 3.

Une analyse exploratoire *data-driven* des facteurs de risque associés à l'infection grippale dans les données CoPanFlu par des approches *data mining* est ensuite présentée. Le *data mining* repose sur l'idée d'extraction de connaissances dans de grandes bases de données sans hypothèses *a priori* sur la structure des données et les relations éventuelles qui peuvent exister au sein des données.

Nous nous concentrons pour finir sur une analyse des données CoPanFlu, cette fois sous l'angle *hypothesis-driven*. Cette approche, à l'opposé de l'analyse précédente, propose une modélisation du

risque d'infection par des modèles d'équations structurelles. Ces modèles permettent d'exprimer et de tester des relations causales directes ou indirectes entre des variables observées et/ou des variables latentes (*i.e.* des variables non directement observables et mesurables).

Avant d'aborder l'étude CoPanFlu et les problématiques d'ordre statistique liées à l'analyse de ces données nous proposons une réflexion sur l'apport des études GWAS en épidémiologie génétique ; bien que débutées il y a moins de 10 ans, ces études sont l'exemple le plus "ancien" de l'utilisation de grands jeux de données en biomédecine. Cette réflexion est illustrée par une étude bibliographique comparant l'impact dans la communauté scientifique des GWAS par rapport à la méthodologie de référence pour la recherche d'association génétique jusqu'à l'explosion des GWAS : l'étude gène-candidat.

Article : L'apport des grands jeux de données en épidémiologie, l'exemple des GWAS

2.1 Résumé

Les *genome-wide association studies* (GWAS) sont des études se proposant de rechercher des associations avec une maladie ou un trait phénotypique à partir de centaines de milliers de variants génétiques sur un grand nombre d'individus [118]. Dans la majorité des cas, les facteurs génétiques étudiés sont des *single-nucleotide polymorphisms* (SNPs), correspondant à la variation d'une seule paire de bases du génome. Ces polymorphismes représentent 90% des variations du génome humain.

L'étude de variants génétiques à l'échelle d'un génome complet sans faire d'hypothèses *a priori* sur les associations les plus probables se démarque du paradigme d'analyse de facteurs génétiques en vigueur jusqu'alors, l'étude "gène-candidat". L'approche gène-candidat se propose de tester les effets d'un ou plusieurs gènes sélectionnés, par le biais d'une étude d'association.

L'analyse de facteurs génétiques par le biais des GWAS a débuté en 2005 et le nombre de ces études est depuis en augmentation constante. La science de l'analyse du génome est ainsi parmi les thématiques *omiques* celle pour laquelle le recul est aujourd'hui le plus important, permettant ainsi une évaluation de la plus value apportée par cette approche *data-driven* comparativement aux études gène-candidat *hypothesis-driven*.

Nous avons choisi d'évaluer l'impact des GWAS dans la communauté scientifique par le biais d'une étude bibliographique et de deux métriques de citation : le nombre de citations et l'indice de crown. L'indice de crown correspond au rapport du nombre de citations reçues par un article sur le nombre de citations attendues sachant l'année de publication, le type de publication et le journal dans lequel l'article a été publié. Ces deux métriques de citation font partie des outils de référence employés par l'*Institute of Scientific Information* (ISI) pour évaluer l'impact des publications scientifiques [147].

Quatre-vingt-dix-sept articles de GWAS publiés entre 2005 et 2009 dans 22 revues à fort *impact factor* ont été appariés avec des études gène-candidat publiées la même année et traitant de la même thématique (62 études différentes). Nous avons observé que les GWAS avaient des valeurs significativement supérieures à celles des études gène-candidat pour les deux métriques de citations, un et deux ans après la publication de ces études. Une analyse approfondie des 62 études gène-candidat sélectionnées a révélé que 40% d'entre elles dérivaient d'hypothèses issues de résultats de GWAS. Quoique l'évaluation de l'impact des GWAS par des métriques de citations puisse être critiqué, il est néanmoins apparent que ces études permettent la génération d'hypothèses de recherche. Il est possible que ces hypothèses

aient été difficiles à élaborer sans l'apport des GWAS, proposant dès lors de puissants compléments de leurs pendants *hypothesis-driven*, les études gène-candidat.

2.2 Article

OPEN ACCESS Freely available online

PLOS ONE

Contribution of Genome-Wide Association Studies to Scientific Research: A Bibliometric Survey of the Citation Impacts of GWAS and Candidate Gene Studies Published during the Same Period and in the Same Journals

Yohann Mansiaux[1,2]*, Fabrice Carrat[1,2,3]

1 Epidemiology, Information System, Modeling, UMR-S 707, University Paris 6-UPMC, Paris, France, **2** Inserm U707, Paris, France, **3** Public Health Unit, Saint-Antoine Hospital, Paris, France

Abstract

In genetic epidemiology, genome-wide association studies (GWAS) are used to rapidly scan a large set of genetic variants and thus to identify associations with a particular trait or disease. The GWAS philosophy is different to that of conventional candidate-gene-based approaches, which directly test the effects of genetic variants of potentially contributory genes in an association study. One controversial question is whether GWAS provide relevant scientific outcomes by comparison with candidate-gene studies. We thus performed a bibliometric study using two citation metrics to assess whether the GWAS have contributed a capital gain in knowledge discovery by comparison with candidate-gene approaches. We selected GWAS published between 2005 and 2009 and matched them with candidate-gene studies on the same topic and published in the same period of time. We observed that the GWAS papers have received, on average, 30±55 citations more than the candidate gene papers, 1 year after their publication date, and 39±58 citations more 2 years after their publication date. The GWAS papers were, on average, 2.8±2.4 and 2.9±2.4 times more cited than expected, 1 and 2 years after their publication date; whereas the candidate gene papers were 1.5±1.2 and 1.5±1.4 times more cited than expected. While the evaluation of the contribution to scientific research through citation metrics may be challenged, it cannot be denied that GWAS are great hypothesis generators, and are a powerful complement to candidate gene studies.

Citation: Mansiaux Y, Carrat F (2012) Contribution of Genome-Wide Association Studies to Scientific Research: A Bibliometric Survey of the Citation Impacts of GWAS and Candidate Gene Studies Published during the Same Period and in the Same Journals. PLoS ONE 7(12): e51408. doi:10.1371/journal.pone.0051408

Editor: Struan Frederick Airth Grant, The Children's Hospital of Philadelphia, United States of America

Received August 22, 2012; **Accepted** November 2, 2012; **Published** December 11, 2012

Copyright: © 2012 Mansiaux, Carrat. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Ministry of Higher Education and Research of France. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mansiaux@u707.jussieu.fr

Introduction

The first genome-wide association study (GWAS) was published in 2005 and identified a genetic variant associated with a higher risk of age-related macular degeneration [1]. The completion of the Human Genome Project [2] and the HapMap Project [3] yielded tools to identify common genetic variations, mainly single-nucleotide polymorphisms (SNPs), associated with many traits or diseases. In 2008, the National Human Genome Research Institute (NHGRI) published an online catalog of Genome-Wide Association Studies [4], listing genetic marker-trait associations published in GWAS. As of 01 November 2012, this database included 1416 publications and 7688 disease- or trait-associated genetic variants.

These "wide association" approaches, initially devoted to genetic associations, have since been extended to transcriptomic [5,6], metabolomic [7] and environmental factors (chemical toxicants, pollutants or nutrients) [8]. While examination of many common genetics variants [9] or environmental exposures [10] is relevant because of their known role in the development of diseases, one may wonder if simultaneous analysis of such an overwhelming amount of data using data-driven methods leads to a significant value in the generation of knowledge, compared with the classical hypothesis-driven candidate-gene studies.

To answer this question, we conducted a bibliometric survey using two citation metrics, the citation count and the crown index, as a surrogate marker for the contribution to the scientific research of GWAS and candidate-gene studies focusing on the same trait or disease and published during the same time period in a predefined set of journals.

Materials and Methods

Data

We searched for GWAS published between 2005 and 2009 in *Nature*, *Science* and journals ranked among the 10 leading journals according to the 2010 Journal Impact Factors (JIFs), in the "Genetics & Heredity" or "Medicine, General & Internal" categories (see Table S1). Meta-analyses, reviews and other publication types were excluded. GWAS had to be performed in humans.

The first search was conducted in the MEDLINE database with the search terms "GWAS", "Genome-Wide Association Study", "Genome-Wide Association Studies", "Genome Wide Association

Study", "Genome Wide Association Studies", "Genomewide Association Study" and "Genomewide Association Studies", "Whole Genome Association Study", "Whole Genome Association Studies", "WGAS", "WGA study" and "WGA studies". To ensure an exhaustive selection process, we retrieved from the NHGRI Catalog of Genome-Wide Association Studies those studies not identified by the MEDLINE search.

The search for candidate-gene studies was conducted using BioPython [11], a set of freely available tools for biological computation, written in Python. For each paper in the GWAS group, we searched for papers focusing on the same disease or trait and published the same year in the same set of journals. We selected the paper that allowed us to reduce both the difference between the 2010 JIFs of the journals in which the GWAS and candidate-gene study were published, and to reduce the interval (in months) between the 2 publication dates. The studied trait or disease was, if available, retrieved from the NHGRI Catalog of Genome-Wide Association Studies, or was otherwise added manually.

Citation Metrics

The "quantity" of knowledge provided by the GWAS and candidate-gene studies was assessed with 2 metrics, namely the citation count and the crown index, computed 1 and 2 years after the publication dates. The citation count was recorded in the Web of Science database. The crown index depends on the expected citation rate, which indicates how frequently cited a paper is expected to be according to its year of publication, its publication type, and the journal in which it was published. The crown index is the ratio of the citation count to the expected citation rate.

Our main objective was to explore the differences of citation metrics between the GWAS and candidate-gene studies. To help interpretation of findings, we also compared candidate-gene studies derived from GWAS with those not derived from GWAS; we investigated the association between the reported P-value and citation metrics and we evaluated the impact of residual confounding, i.e. whether residual differences in impact factor or publication dates despite matching may have influenced the results.

Statistical Analysis

The statistical analysis was conducted using R 2.13.0. The alpha level for all tests was 0.05.

Wilcoxon signed-rank tests for paired samples were used to test differences of quantitative variables between GWAS and candidate gene studies. Linear mixed-effects models were used to assess the effect of the approach used, "Genome-Wide" or "candidate gene", on the crown index. A random effect was included to account for the correlation between the citation metrics in each pair of papers. The models were fitted with and without the effect of the approach and likelihood ratio tests were performed to compare the two fits.

GWAS and candidate-gene studies statistical tests for association do not usually use the same threshold to assess the statistical significance of an association. For example, the significance threshold in GWAS is often defined between 10^{-8} and 10^{-6} while it is usually close to 10^{-2} in candidate-gene studies. To investigate the association of P-values with citation metrics, we therefore calculated in each study the ratio between the reported P-value and the corresponding probability threshold used to conclude statistical significance (hereafter P-ratio). We explored the association between the log-transformed P-ratio, the type of study and the crown indexes using mixed models as described above.

We finally used linear regression models to assess the impact of the difference between the 2010 JIFs and of the interval between the publication dates on the difference in the crown indexes within each pair of papers.

Note: Summary statistics are reported as mean ($\pm$ sd). The estimates of the linear mixed-effects models are reported as estimate ($\pm$ se).

Results

GWAS and Candidate-gene Study Selection

The initial search in the MEDLINE database identified 511 articles, and 68 additional articles were identified in the NHGRI Catalog. A total of 211 GWAS were selected, of which a candidate-gene study was identified in 97 cases (62 different papers) (see Tables S2, S3 and S4 for summary statistics of the 97 GWAS/candidate-gene study pairs and Tables S5, S6, S7 and S8 for the PMID PUBMED identifiers of the articles used in the survey). The full article selection process is summarized in Figure 1. Of the 62 candidate-gene selected, 23 of them (37%) were derived from former GWAS (not necessarily from the GWAS with which the candidate-gene study was paired).

Citation Metrics Analysis

We observed that the GWAS papers were more frequently cited than the candidate-gene studies, whatever the metric used (see Table 1 and Figure 2). This was confirmed by comparing the citation counts and the crown indexes, 1 and 2 years after the publication dates. The two-sided Wilcoxon signed-rank tests were all significant (P value $=1.8\times10^{-8}$ and 7.8×10^{-11} for the citation counts; P value $=2.6\times10^{-7}$ and 4.7×10^{-12} for the crown indexes). The distribution of the crown indexes were not significantly different between the candidate-gene studies derived from GWAS, compared with those not derived from GWAS, both 1 year after the publication date (Wilcoxon test P value $=0.67$) and 2 years after the publication date (Wilcoxon test P value $=0.46$).

Assessment of the effect of the "genome-wide" or "candidate-gene" approach on the crown indexes based on linear mixed-effects models confirmed that this metric increased in GWAS papers compared with candidate-gene papers. The estimates of the "genome-wide" effect on the crown indexes were 1.31 ($\pm$0.23, likelihood ratio test P value $=3.3\times10^{-6}$) and 1.45 ($\pm$0.19, likelihood ratio test P value $=4.3\times10^{-7}$), 1 and 2 years after the publication dates.

The average log transformed P-ratio was lower in the GWAS group than in the candidate-gene group : -25.8 ($\pm$54.7) for the GWAS group, -12.6 ($\pm$24.9) for the candidate-gene group (Wilcoxon test P value $=6.9\times10^{-9}$). The estimates of the mixed models assessing the effect of the log transformed P-ratio on the crown index 1 and 2 years after the publication date were 3.0×10^{-13} ($\pm8.6\times10^{-8}$, P value $=0.97$) and -4.4×10^{-13} ($\pm1.0\times10^{-7}$, P value $=0.98$).

There was no evidence for residual confounding: the average difference between the 2010 JIFs of the journals in which the paired papers were published was 1.5 ($\pm$15.9, Wilcoxon test P value $=0.24$) and the average interval between the publication dates was 2.7 ($\pm$2.9) months (Wilcoxon test P value $=0.71$). The effects of these differences on the difference in the crown indexes within each pair of papers in linear models were not significant, either one year after publication (P value $=0.56$ for the difference in 2010 JIFs, and P value $=0.37$ for the interval between the publication dates) or 2 years after publication (P value $=0.90$ for

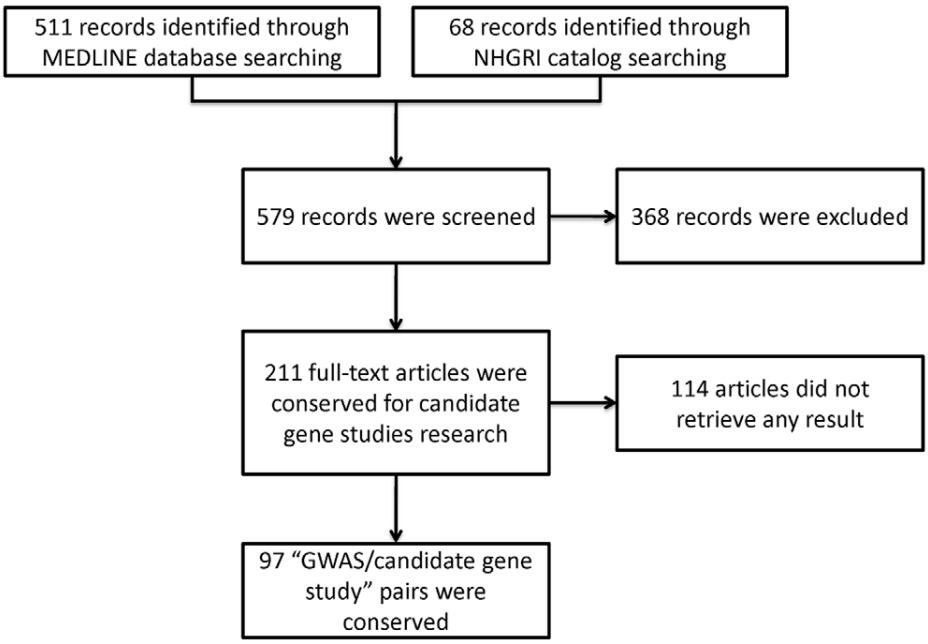


Figure 1. Study flow diagram.
doi:10.1371/journal.pone.0051408.g001

the difference in 2010 JIFs and P value = 0.36 for the interval between the publication dates).

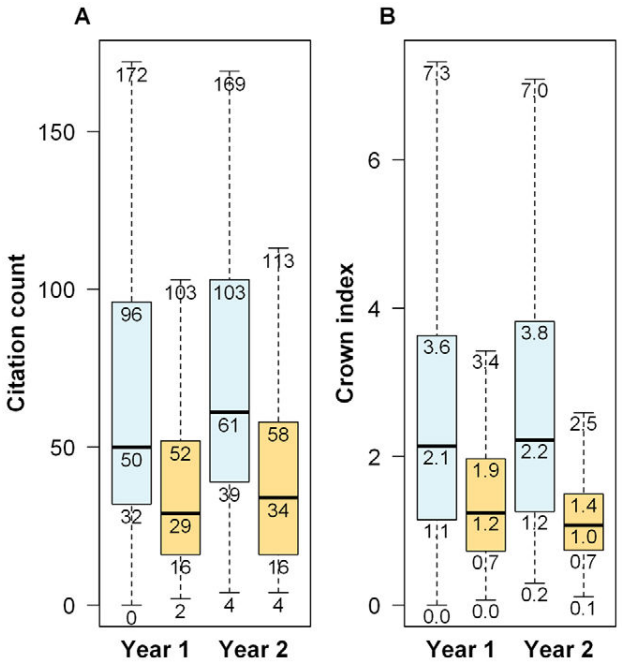


Figure 2. Boxplots of the citation count and of the crown index.
Distributions of the citation count and of the crown index are depicted on panel A and panel B, respectively, 1 year (Year 1) and 2 years (Year 2) after the publication dates. Blue boxes refer to GWAS and orange boxes to candidate-gene studies. The boxplots depict five statistics: the sample minimum, the lower quartile, the median, the upper quartile, and the sample maximum.
doi:10.1371/journal.pone.0051408.g002

Discussion

We showed that GWAS had a higher impact in the scientific literature than candidate gene studies. We assume that our study selection process, although not systematic, produced robust results by focusing on the genetic analyses published in the papers with the most important impact factors, and that our matching process did not alter the quality of the results provided.

The use of citations metrics to assess the ability of GWAS and candidate-gene studies to generate knowledge can be questioned. Indeed it was not possible to directly link the citation metrics to the relevance or the applicability of the scientific outcomes provided by these studies, or to their contribution to understanding the disease's etiology. An in-depth analysis of the papers citing the GWAS and the candidate-gene studies would have been required to observe how the findings provided by those papers were used. However, we can consider that the citation metrics used are good markers for "research" generation, and therefore good, albeit indirect, markers for knowledge generation. Indeed, we observed that nearly 40% of the candidate-gene studies selected in our survey were inspired by hypotheses derived from GWAS.

The investigation of the potential association of the reported P values in the GWAS and the candidate-gene studies and of the crown index did not permit to show a significant relationship. Although the statistical threshold to assess significance is stricter in GWAS because of the multiplicity of tests, their findings were not associated with a higher scientific impact, and not associated with a higher credibility for follow-up.

Conclusions

We have explored the impact of GWAS and candidate-gene studies through 2 citation metrics and observed that the GWAS papers were more cited than the candidate-gene papers, whatever the metrics used. If it cannot formally be claimed that the knowledge generated by GWAS outweighs that provided by classical genetic association studies, we can say that this

Table 1. Summary statistics of the citation metrics of the "GWAS" and "candidate-gene study" papers.

Group	GWAS	Candidate-gene studies
citation count 1 year after the publication date	69.1±54.9	38.7±35.5
citation count 2 years after the publication date	85.3±69.5	45.6±47.8
crown index 1 year after the publication date	2.8±2.4	1.5±1.2
crown index 2 years after the publication date	2.9±2.4	1.5±1.4

Mean±sd.
doi:10.1371/journal.pone.0051408.t001

manuscript provides evidence, at least indirect evidence, that GWAS contribute to knowledge production by allowing candidate-gene studies to focus on credible candidates.

Supporting Information

Table S1 2010 Journal Impact Factors of the Journals used for the MEDLINE database and NHGRI Catalog research.
(PDF)

Table S2 Years of publication of the 97 pairs of papers in the "GWAS" and "candidate-gene studies" groups.
(PDF)

Table S3 Journals of publication of the "GWAS" papers.
(PDF)

Table S4 Journal of publication of the "candidate-gene studies" papers.
(PDF)

Table S5 PMID PUBMED identifiers of the 511 articles identified through MEDLINE research.
(XLS)

Table S6 PMID PUBMED identifiers of the 68 articles identified through NHGRI catalog research.
(XLS)

Table S7 PMID PUBMED identifiers of the 211 GWAS articles conserved for candidate-gene study research.
(XLS)

Table S8 PMID PUBMED identifiers of the 97 GWAS/candidate-gene study pairs.
(XLS)

Acknowledgments

We thank Rosemary M. Delabre and David Young for critical review of the manuscript.

Author Contributions

Conceived and designed the experiments: YM FC. Performed the experiments: YM. Analyzed the data: YM. Wrote the paper: YM FC.

References

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
3. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
4. Hindorff L, MacArthur JEBI, Wise A, Junkins H, Hall P, et al. (2009) A Catalog of Published Genome-Wide Association Studies. National Human Genome Research Institute.
5. Tang B, Di Lena P, Schaffer L, Head SR, Baldi P, et al. (2011) Genome-wide identification of Bcl11b gene targets reveals role in brain-derived neurotrophic factor signaling. PLoS One 6: e23691.
6. Marques FZ, Campain AE, Tomaszewski M, Zukowska-Szczechowska E, Yang YH, et al. (2011) Gene expression profiling reveals renin mRNA overexpression in human hypertensive kidneys and a role for microRNAs. Hypertension 58: 1093–1098.
7. Yap IK, Brown IJ, Chan Q, Wijeyesekera A, Garcia-Perez I, et al. (2010) Metabolome-wide association study identifies multiple biomarkers that discriminate north and south Chinese populations at differing risks of cardiovascular disease: INTERMAP study. J Proteome Res 9: 6647–6654.
8. Patel CJ, Bhattacharya J, Butte AJ (2010) An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One 5: e10746.
9. Rappaport SM (2011) Implications of the exposome for exposure science. J Expo Sci Environ Epidemiol 21: 5–9.
10. Rappaport SM, Smith MT (2010) Epidemiology. Environment and disease risks. Science 330: 460–461.
11. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

Table S1 2010 Journal Impact Factors of the Journals used for the MEDLINE database and NHGRI Catalog research

Journal	2010 Journal Impact Factor
Medicine, General & Internal	
New Engl J Med	53.486
Lancet	33.633
Jama-J Am Med Assoc	30.011
Ann Intern Med	16.729
Plos Med	15.617
Brit Med J	13.471
Annu Rev Med	12.457
Arch Intern Med	10.639
Can Med Assoc J	9.015
Cochrane Db Syst Rev	6.186
Genetics & Heredity	
Nat Genet	36.377
Nat Rev Genet	32.745
Annu Rev Genet	21.774
Annu Rev Genom Hum G	17.182
Trends Ecol Evol	14.448
Genome Res	13.588
Gene Dev	12.889
Am J Hum Genet	11.680
Trends Genet	11.364
Plos Genet	9.543
Multidisciplinary Sciences	
Nature	36.104
Science	31.377

Table S2 Years of publication of the 97 pairs of papers in the "GWAS" and "candidate-gene studies" groups

Year of publication	**2005**	**2006**	**2007**	**2008**	**2009**
Number of papers	2	2	25	23	45

Table S3 Journals of publication of the "GWAS" papers

Journal	**Number of papers**
Am J Hum Genet	6
Gene Dev	1
Genome Res	1
Jama-J Am Med Assoc	1
Nature	7
Nat Genet	59
New Engl J Med	5
Plos Genet	9
Science	8

Table S4 Journal of publication of the "candidate-gene studies" papers

Journal	**Number of papers**
Am J Hum Genet	16
Nature	3
Nat Genet	47
New Engl J Med	11
Plos Genet	11
Science	9

2.3 Discussion

Ce travail a porté sur l'analyse de l'impact du premier type d'étude *big data* en biomédecine. La question de la plus-value apportée par une analyse "à l'aveugle" d'une très grande quantité de données par rapport à la méthodologie habituelle de construction/test d'hypothèses méritait d'être posée.

Il apparait délicat de lier l'apport de connaissance dans la communauté scientifique ou la pertinence des résultats d'un article avec le nombre de citations récoltées, au moins d'une manière directe. D'une manière plus générale, le scepticisme est d'ailleurs toujours de rigueur concernant les GWAS. En effet la taille des effets (*effect size*) montrés dans ces études est souvent modeste, et les implications fonctionnelles des variants identifiés ne sont pas toujours évidentes [101]. Bien que certaines applications en clinique de résultats de GWAS existent maintenant [119], et qu'un grand nombre de tests génétiques soient disponibles pour de nombreuses maladies [8], le débat reste ouvert sur l'impact réel des résultats de GWAS dans la pratique médicale courante [49, 99, 118].

Comme observé sur la figure 2.1, le nombre de GWAS est en augmentation continue. A la date du 1er août 2014, le catalogue d'associations issues de GWAS publié par le *National Human Genome Research Institute* (NHGRI) faisait état de 1961 publications rapportant 14012 SNPs associés avec plus de 300 maladies ou traits phénotypiques [86, 137]. Bien que parfois critiquées, ces études à très grande échelle sont devenues des études de routine en épidémiologie génétique. On peut penser, comme observé dans les résultats de l'article, que le côté "générateur d'hypothèses" des GWAS est une des explications de leur popularité [50, 78]. La génération de pistes de recherche qu'il aurait probablement été difficile à élaborer *a priori* fait des GWAS un complément puissant des études gène-candidat. Cet aspect est moins apparent du point de vue de l'application clinique des variants détectés, expliquant probablement en partie le scepticisme entourant ces études.

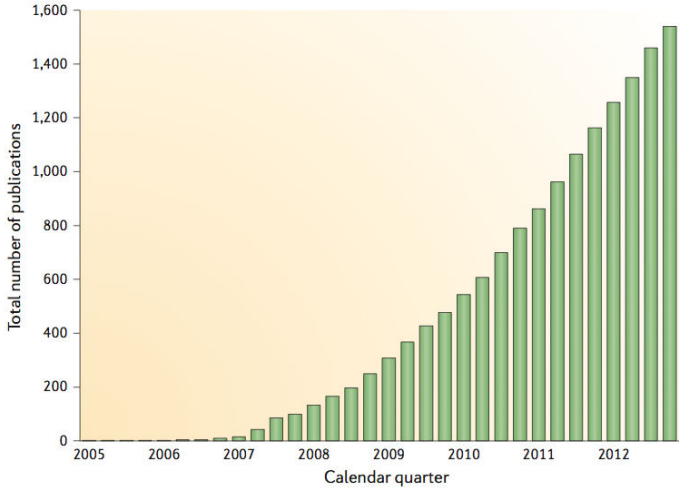


FIGURE 2.1 : Nombre total de GWAS publiées depuis 2005 (issu de Manolio 2013 [119])

Cette dernière réflexion pose la question des attentes liées à ces études, il convient de garder à l'esprit que ces approches ont moins d'une dizaine d'années et que la plupart des découvertes issues de GWAS n'ont été faites qu'après 2010 [119]. Il est donc probablement prématuré de vouloir tirer des conclusions cliniques fermes des études *omiques* à grande échelle ; l'impact à court terme du *big data* en épidémiologie génétique a probablement été surestimé tandis que son apport à long terme est quant à lui potentiellement sous-estimé [49].

Nous présentons maintenant les données récoltées lors de l'étude CoPanFlu France, cas pratique d'un grand jeu de données en épidémiologie. Nous verrons par la suite comment nous pouvons nous inspirer de la méthodologie d'analyse *data-driven* des GWAS pour explorer ces données et en extraire de l'information relative au risque d'infection par la grippe pandémique H1N1pdm.

L'étude CoPanFlu-France : une approche intégrative du risque d'infection grippale

Les travaux présentés dans la suite de ce document portent sur l'analyse des données issues de l'étude CoPanFlu-France (voir Lapidus *et al.* 2012 [105], pour de plus amples informations concernant le protocole de l'étude). Cette cohorte fait partie du consortium international CoPanFlu (Cohorts for Pandemic Influenza) formé en 2009 suite à l'émergence du nouveau virus grippal H1N1 pandémique (H1N1pdm) [54], afin d'étudier les déterminants individuels et collectifs de l'infection par ce virus. Un ensemble de cohortes de foyers ont été constituées à travers le monde, la cohorte CoPanFlu-France étant une étude en population générale en France métropolitaine.

Cette cohorte de 601 foyers (1450 individus), inclus entre décembre 2009 et juillet 2010 s'est proposée d'étudier durant 2 saisons grippales consécutives (2010-2011 et 2011-2012) les facteurs de risque et les mécanismes de l'infection par le virus grippal H1N1pdm. Un suivi actif des sujets a été utilisé pour permettre une détection efficace de la survenue de syndromes grippaux. En cas de symptômes de syndrome grippal rapportés par un sujet, des prélèvements nasaux étaient effectués chez l'ensemble des individus du foyer du sujet en question, permettant ainsi l'identification de divers virus respiratoires. Des titrations sérologiques ont également été réalisées chez l'ensemble des sujets entre chaque saison grippale, autorisant une comparaison des sérologies pré/post saison épidémique.

Cette étude intégrative propose d'aborder l'infection grippale comme une combinaison complexe de caractéristiques biologiques, de facteurs comportementaux individuels et collectifs et de facteurs environnementaux. En plus des prélèvements biologiques réalisés, l'utilisation de questionnaires et de bases de données environnementales a permis la récolte d'un grand nombre de variables explorant des dimensions aussi diverses que les caractéristiques socio-démographiques et médicales des sujets, l'environnement au sein des foyers et l'environnement extérieur, les comportements sociaux des sujets ou encore leur perception du risque d'infection grippale [104, 105]. Au total plus de 300 variables ont été récoltées chez l'ensemble des sujets au moment de leur inclusion dans l'étude (voir la figure 3.1).

Depuis l'émergence du virus H1N1pdm en 2009, de nombreuses études de facteurs de risque d'infection ont été publiées, basées principalement sur des études de ménages, explorant la transmission à des cas secondaires lorsqu'un cas index était identifié chez un sujet du ménage, mais également sur des études de cohortes. Un large panel de facteurs ont été rapportés, relatifs aux caractéristiques individuelles des sujets (*e.g.* être un sujet jeune [35, 39, 43, 67, 97, 120, 156, 167, 183], souffrir d'une maladie respiratoire [35] ou d'une maladie chronique [167, 183], avoir un titre d'anticorps pré-épidémique faible [41, 43], ou encore être de sexe féminin [197]), à l'environnement des foyers des sujets (*e.g.*

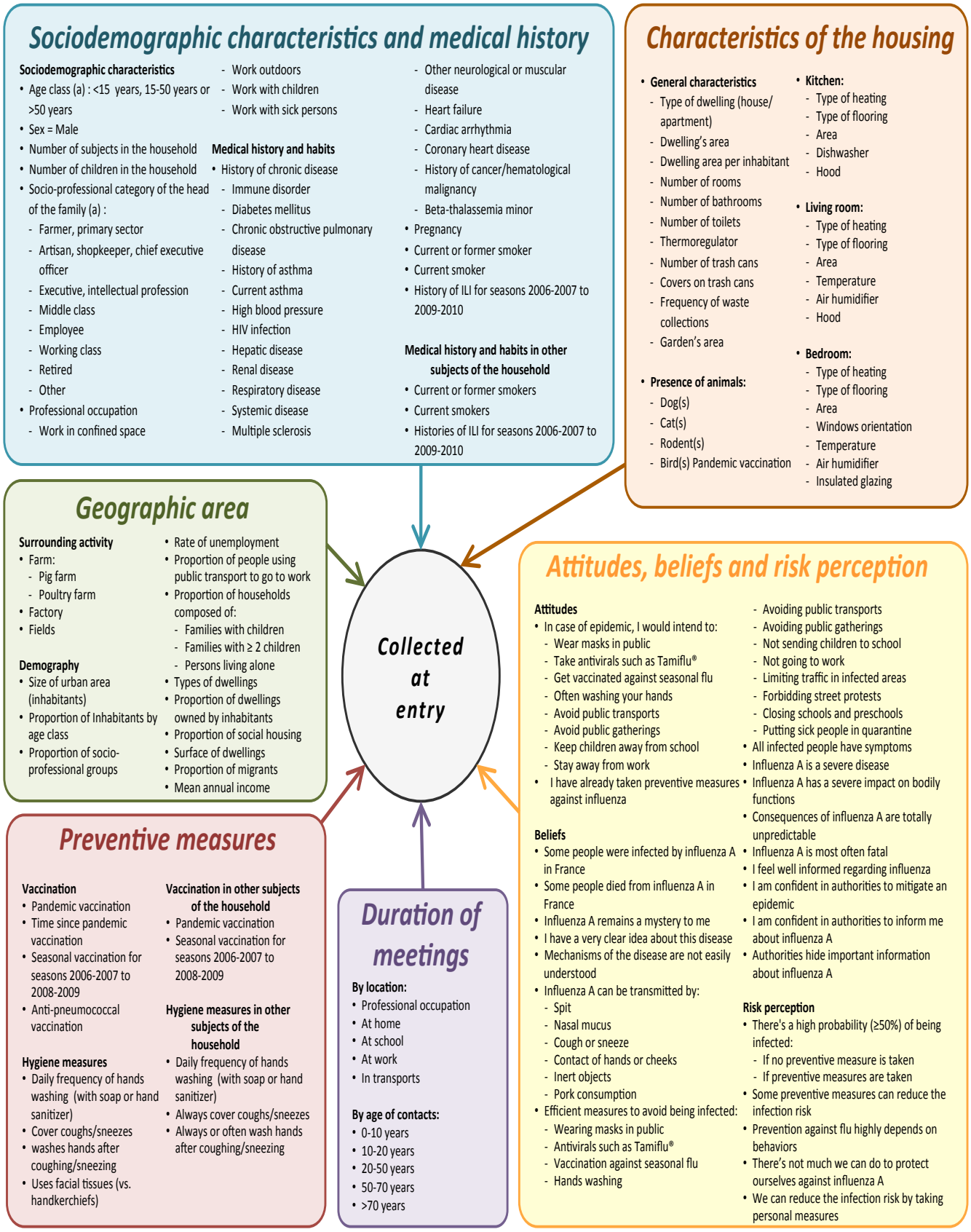


FIGURE 3.1 : Données collectées par questionnaires à l'inclusion des sujets dans l'étude CoPanFlu-France (issu de Lapidus *et al.* 2012 [105])

le nombre d'individus dans le foyer [39]), à l'environnement extérieur (*e.g.* vivre dans une zone urbaine [41]), aux contacts avec des individus à risque (*e.g.* être en contact avec des enfants [120, 156] ou des individus infectés [43, 67, 112]), à l'usage de mesures préventives pharmaceutiques (*e.g.* prophylaxie antivirale du sujet [97, 141, 145, 197] ou traitement antiviral du cas index [139, 145]) ou non pharmaceutiques (*e.g.* lavage des mains ou utilisation d'un masque facial [169]), ou encore à la perception du risque d'infection grippale (*e.g.* avoir des discussions à propos de la grippe dans le ménage [67]).

L'ensemble des facteurs de risque rapportés dans ces études peuvent être reliés à deux conditions indispensables pour être infecté : être exposé au virus et être susceptible (d'un point de vue biologique) de contracter une infection. Parmi les facteurs identifiés dans ces études on peut par exemple considérer que les contacts avec des individus à risque ou l'utilisation de mesures préventives sont des facteurs modulant l'exposition au virus tandis que le titre d'anticorps pré-épidémique peut être lié à la susceptibilité individuelle. Des facteurs comme la perception du risque peuvent agir conjointement sur l'exposition et la susceptibilité, en modifiant par exemple les comportements sociaux ou la décision de se vacciner.

La diversité des facteurs rapportés souligne la nécessité d'utiliser une approche intégrative pour explorer le risque d'infection par le virus grippal. Aucune des études citées n'a analysé autant de facteurs simultanément. Une approche multidisciplinaire telle que celle menée dans CoPanFlu autorise l'étude des rôles respectifs de l'exposition au virus et de la susceptibilité de l'hôte, ainsi que des relations complexes éventuelles entre des facteurs liés à l'une ou l'autre de ces dimensions.

L'analyse statistique de telles données, à la fois massives et diversifiées, pose un certain nombre de questions d'un point de vue méthodologique et requiert la remise en question de principes couramment admis en épidémiologie. Nous présentons maintenant un rappel des méthodes de référence employées en épidémiologie pour détecter des associations, en soulignons les limites, et suggérons des outils alternatifs plus adaptés à l'exploration d'un jeu de données tel que celui de l'étude CoPanFlu-France.

La recherche d'associations dans un grand jeu de données en épidémiologie : problématiques et approches exploratoires

4.1 La recherche d'associations en épidémiologie hypothesis-driven

En épidémiologie *hypothesis-driven*, la recherche d'associations implique généralement une modélisation statistique testant les relations entre une ou plusieurs covariables (ou prédicteurs) et une variable réponse cible. Les modèles de régression sont très populaires en épidémiologie. Ils sont simples à implémenter et permettent une interprétation aisée de la relation entre la variable réponse et les covariables testées. Différents modèles de régression existent selon le type de réponse que l'on souhaite étudier : modèle linéaire (pour une variable réponse continue), modèle de Poisson (utilisé lorsque la variable réponse est un nombre d'événements discret), modèle de survie ou encore modèle logistique. Ce dernier est utilisé lorsque la variable réponse suit une distribution binomiale. Les différentes modélisations présentées dans ce document étudient les facteurs associés au statut infectieux des individus de l'étude CoPanFlu, s'agissant d'une grandeur dichotomique "infecté/non-infecté" nous nous concentrons uniquement sur les propriétés du modèle de régression logistique binomial.

Dans les multiples modèles présentés dans la suite de ce document nous noterons Y la variable réponse à prédire, X le vecteur de p covariables prédictives et N le nombre de sujets.

Le modèle logistique

Le modèle logistique appartient à la famille des modèles linéaires généralisés [123], il est utilisé pour modéliser une relation linéaire entre des covariables X et le *logit* de la probabilité que la variable réponse étudiée Y soit présente.

La formule 4.1 permet de représenter cette probabilité. Le coefficient β_0 correspond à l'*intercept* ou risque de base en l'absence des covariables étudiées (indépendant d'un sujet donné), les termes β représentent les coefficients de régression estimés, relatifs à la force d'association entre les covariables X et la probabilité p que la variable réponse Y vaille 1.

$$f(X) = p(Y = 1|X) = \frac{1}{1 + \exp(-(\beta_0 + \sum_{j=1}^{p} \beta_j X_j))} \tag{4.1}$$

Le *logit* de la probabilité p s'exprime ainsi :

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{4.2}$$

Les coefficients β du modèle sont estimés par la méthode du maximum de vraisemblance, qui permet de déterminer les valeurs des paramètres du modèle qui rendent le plus vraisemblable les observations recueillies. Les coefficients sont estimés de telle façon à maximiser une fonction de vraisemblance (*likelihood*, L) (voir la formule 4.3). En pratique pour des raisons de simplification des calculs c'est généralement le logarithme de la vraisemblance, appelé log-vraisemblance (*log-likelihood*, logL) qui est la fonction à maximiser (voir la formule 4.4).

$$L(\hat{\beta}) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \tag{4.3}$$

$$logL(\hat{\beta}) = \sum_{i=1}^{N} y_i log(p_i) + (1-y_i) log(1-p_i) \tag{4.4}$$

avec y_i la valeur de la variable réponse pour l'individu i et p_i la probabilité de présence de la variable réponse prédite par le modèle pour ce même sujet.

La popularité du modèle de régression logistique en épidémiologie tient au fait que les coefficients estimés sont facilement interprétables en terme de rapport de cotes (généralement connu sous le terme d'*odds ratio*, OR). Un *odd* est un rapport de probabilité, en épidémiologie ceci pourrait correspondre au rapport entre la probabilité d'être malade et la probabilité de ne pas être malade. L'*odds ratio* correspond quant à lui au rapport de deux *odds*. L'*odds ratio* de la variable réponse Y en fonction d'un facteur de risque à deux classes X_1 peut s'exprimer comme :

$$OR_{Y|X_1} = \frac{p(Y=1|X_1=1)/p(Y=0|X_1=1)}{p(Y=1|X_1=0)/p(Y=0|X_1=0)} \tag{4.5}$$

Si la prévalence de la maladie étudiée est faible cette mesure est assimilable à un risque relatif, qui peut s'interpréter aisément comme le facteur par lequel le risque de la maladie étudiée est multiplié en cas de présence du facteur de risque considéré. L'OR est la mesure d'association obtenue dans les modèles de régression logistique, il s'obtient très facilement à partir des coefficients estimés par le modèle (formule 4.6)

$$OR_{Y|X_1} = \exp(\beta_1) \tag{4.6}$$

L'utilisation de ce modèle sous-entend la vérification d'un certain nombre d'hypothèses ; en plus de l'hypothèse fondamentale d'une relation linéaire entre le *logit* de la probabilité de présence de la variable réponse et les valeurs des covariables incluses dans le modèle, le modèle de régression logistique suppose l'indépendance des observations.

4.2 Problématiques de l'analyse d'un grand jeu de données et limites de la régression

Le développement de grands jeux de données en épidémiologie a provoqué l'apparition de nouvelles problématiques d'un point de vue de l'analyse statistique. Sinha *et al.* ont proposé une revue des problèmes analytiques rencontrés avec un grand jeu de données [170]. Certains aspects tels que la gestion du risque de première espèce suite à des tests statistiques multiples (discuté plus en détails par la suite), également rencontrés dans des bases de données de taille relativement modeste, se trouvent exacerbés dans un tel contexte. La grande dimensionnalité des données induit également des problèmes aussi divers que la multicolinéarité (*i.e* l'existence de relations linéaires entre des variables), le sur-ajustement des modèles aux données (ou *overfitting*) ou encore une complexité de calculs informatiques rédhibitoire.

La multicolinéarité des données est associée avec une estimation erronée des paramètres du modèle, tout comme l'est l'utilisation simultanée d'un grand nombre de covariables [168]. Dans un modèle de régression logistique il est généralement admis qu'une covariable peut être ajoutée dans le modèle pour 10 évènements (*e.g.* 10 individus infectés) afin d'obtenir une estimation robuste des paramètres [146]. Enfin, l'hypothèse fondamentale d'une liaison linéaire entre les covariables et le *logit* de la probabilité d'observer la variable réponse apparait discutable dans un contexte de grande dimensionnalité du jeu de données, où l'on peut notamment s'attendre à des interactions non linéaires entre différents prédicteurs [126].

Tests multiples et inflation du risque de première espèce

Dans un test statistique, le risque de première espèce α correspond à la probabilité de rejeter l'hypothèse nulle du test alors que cette hypothèse est vraie. Ce risque, également appelé *comparisonwise error rate* [21], est la plupart du temps fixé à 5%. Analyser un grand jeu de données par des tests statistiques usuels (*e.g.* test exact de Fisher ou test de Mann-Whitney) ou des modèles de régression (un test effectué par paramètre estimé) entraine une augmentation du nombre de tests à réaliser simultanément et par conséquent augmente fortement la probabilité de détecter des associations à tort. La probabilité de rejeter à tort au moins une fois l'hypothèse nulle dans une série de tests indépendants est appelée

l'*experimentwise error rate* (EER, ou encore *familywise error rate* - FWER). Pour t tests indépendants réalisés le EER peut être estimé ainsi : $1-(1-\alpha)^t$, avec α le risque de première espèce pour un test. En considérant l'exemple d'une étude GWAS où des centaines de milliers de variants génétiques sont testés simultanément, le nombre de faux positifs avec un tel risque de première espèce est alors très important, le EER atteint en effet 99% dès 90 tests simultanés.

La méthode de Bonferroni est l'une des techniques les plus courantes pour contrôler l'augmentation du EER causée par des tests multiples [58]. Il s'agit simplement d'ajuster le critère permettant d'évaluer si la *p-value* d'un test statistique est significative ou non. Si le EER souhaité sur l'ensemble des t tests est de 0.05, la significativité de chacun des tests sera évaluée par rapport à un risque de première espèce α ajusté : $\alpha' = \frac{EER \text{ souhaité}}{t} = \frac{0.05}{t}$.

La principale limite de cette méthode est son impact sur la puissance, un contrôle trop strict du risque de première espèce va avoir pour effet d'augmenter le risque de deuxième espèce β, correspondant à la probabilité de ne pas rejeter l'hypothèse nulle alors que celle-ci est fausse. En d'autres termes, à trop vouloir contrôler le nombre de "faux positifs", on risque d'augmenter le nombre de "faux négatifs".

Le développement des approches *omiques* a entrainé l'émergence d'une nouvelle littérature dédiée au problème des tests multiples et de nombreuses méthodes permettant un ajustement du risque global de première espèce tout en limitant l'impact sur la puissance sont maintenant disponibles [122, 178, 212, 215]. L'idée générale de ces méthodes est de contrôler la proportion de "fausses découvertes" (*false discovery rate*, FDR) (*i.e.* le nombre d'hypothèses nulles rejetées à tort) parmi l'ensemble des tests indépendants réalisés [22]. Pour chaque test on cherche à estimer le risque de première espèce permettant de maintenir le FDR sous un seuil choisi au préalable (*e.g.* 0.05). Une méthode populaire est la méthode des *q-values*. Il s'agit d'une variante ajustée de la *p-value* correspondant au FDR minimal pour lequel le test peut être considéré significatif [177]. Plutôt que de comparer des *p-values* à un seuil de significativité ajusté, il est possible de comparer les *q-values* au seuil de référence 0.05.

Les difficultés pour analyser un grand jeu de données sont multiples, et les méthodes usuelles, particulièrement les modèles de régression, sont peu adaptées dans de tels cas. Nous présentons maintenant des solutions alternatives aux méthodes conventionnelles utilisées en épidémiologie pour détecter des associations, attendues comme permettant d'outrepasser les limitations du modèle logistique. Ces méthodes s'inspirent de celles utilisées dans l'analyse de données *omiques*, en particulier dans le cas des GWAS.

4.3 Analyse statistique des genome-wide association studies

Bien que les GWAS soient maintenant devenues des études très courantes, il est difficile d'observer un consensus quant à la méthodologie statistique à employer pour en extraire les résultats. L'analyse des données de GWAS n'implique pas nécessairement l'utilisation de méthodes sophistiquées complexes à implémenter. En effet des tests statistiques communs tels que les tests de Pearson, de Cochran-Armitage ou encore le test exact de Fisher sont couramment utilisés pour détecter des associations univariées [15], les régressions linéaire ou logistique font également partie des méthodes parfois employées [34].

Si les tests statistiques standards utilisés sur des variants génétiques pris séparément ont permis de montrer des associations avec des traits phénotypiques et des maladies, des approches alternatives sont nécessaires pour aborder les données dans leur ensemble et ainsi capturer des associations indépendantes ou des marqueurs influençant conjointement la variable réponse par le biais d'interactions. Ces dernières années ont ainsi vu le développement de méthodes alternatives pour l'analyse de grands jeux de données. Les méthodes dites de *data mining* se sont ainsi multipliées depuis le début des années 1990 [171]. Le terme *data mining* décrit l'analyse de données sans faire d'hypothèses *a priori* sur la structure ou les relations potentielles qui peuvent exister dans les données. Parallèlement à l'émergence des méthodes de *data mining* sont apparues des méthodes de régression pénalisée, reposant sur les mêmes concepts théoriques que les régressions usuellement employées, tout en proposant des adaptations permettant la gestion d'un grand nombre de covariables. Les méthodes de *data mining* et la régression pénalisée font maintenant partie des outils statistiques fréquemment employés dans des problématiques de type GWAS [13, 77, 100, 130, 187, 214].

4.4 Méthodes de data mining

Le *data mining* est un champ multidisciplinaire basé sur les statistiques, les mathématiques, l'intelligence artificielle ou encore la gestion de bases de données. Le *data mining* s'inscrit dans une tâche plus globale de "*Knowledge Discovery in Databases*" (KDD) [64], c'est à dire l'extraction de connaissance dans les bases de données. Le processus de KDD est résumé sur la figure 4.1, sur laquelle le *data mining* consiste uniquement en l'analyse de données préalablement traitées. En pratique l'ensemble de ces étapes sont maintenant confondues et l'expression *data mining* n'est plus limitée à la simple analyse statistique des données. Les applications du *data mining* sont larges et loin d'être limitées à des problématiques de recherche scientifique. En effet, ces méthodes sont couramment utilisées dans des domaines comme la finance, l'assurance ou le marketing pour la détection de fraude ou l'analyse de consommation [19, 116]. L'utilisation du *data mining* en sciences de la vie est en pleine expan-

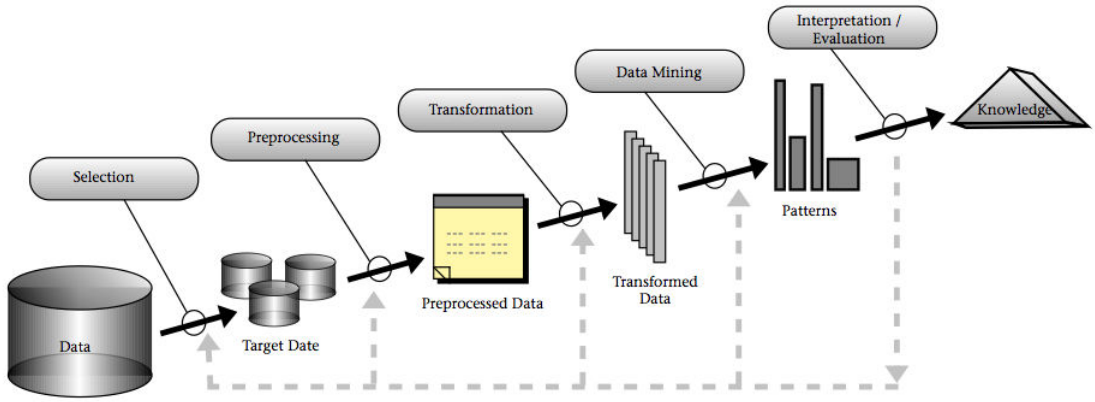


FIGURE 4.1 : Résumé du processus de "Knowledge Discovery in Databases" (KDD) (issu de Fayyad *et al.* 1996 [64])

sion, comme le montre notamment la création en 2008 du journal *BioData Mining* [3], centré sur l'analyse des grandes bases de données biologiques et biomédicales. On peut également observer un intérêt croissant pour ces approches dans des revues consacrées à l'utilisation d'approches informatiques en biomédecine telles que *Artificial Intelligence in Medicine*, *Journal of the American Medical Informatics Association* ou encore des journaux comme *PLoS Computational Biology* ou *BMC Bioinformatics* [18, 164].

Deux philosophies d'apprentissage

L'analyse et l'exploration de données reposent sur deux principaux types d'algorithmes, les algorithmes d'apprentissage non supervisé et les algorithmes d'apprentissage supervisé [82]. Les premiers sont à visée exploratoire et descriptive tandis que les seconds sont utilisés dans un but prédictif.

Les méthodes d'apprentissage non supervisé reposent principalement sur des algorithmes de *clustering* cherchant à diviser un groupe de données en plusieurs sous-groupes (ou *clusters*) homogènes, en utilisant principalement des critères de distance ou de dispersion, l'idée étant que les individus regroupés au sein d'un même *cluster* soient plus "proches" les uns des autres que ceux assignés à un *cluster* différent. L'un des algorithmes de *clustering* les plus populaires est l'algorithme des *k-means*, où l'on cherche à séparer les individus en k groupes (k étant défini *a priori*), la mesure de distance utilisée ici étant simplement une distance euclidienne [82]. Une autre méthode populaire est la classification hiérarchique (*hierarchical clustering*) ; de manière similaire aux *k-means*, on cherche à diviser les individus en un certain nombre de groupes, l'une des différences étant que dans ce cadre le nombre de groupes n'est pas défini *a priori*. Les méthodes de *clustering* sont couramment appliquées pour l'étude de grandes bases de données en génomique, en particulier avec des données de puces à ADN [165, 182]. Ces modèles ne nécessitent pas qu'une valeur de variable réponse soit renseignée, se basant uniquement sur les valeurs des prédicteurs pour extraire les groupes éventuels.

La philosophie de l'apprentissage supervisé repose quant à elle sur l'idée de construire un modèle prédictif fiable et parcimonieux pour l'appliquer ensuite sur des données indépendantes à classifier [213]. Contrairement aux méthodes d'apprentissage non supervisé l'ensemble des covariables ainsi que la valeur de la variable réponse sont connues pour l'ensemble des individus. Le modèle est construit sur des données dites "d'apprentissage" et l'évaluation de ses performances prédictives (par exemple un taux de mauvaise classification dans le cas d'une variable réponse dichotomique) peut se faire soit par le biais d'un échantillon indépendant dit "de test" ou par une procédure d'évaluation interne appelée validation croisée (*cross-validation*).

Dans le premier cas, l'échantillon de départ est divisé en deux parties, l'une servant de jeu d'apprentissage et l'autre de jeu de test. Les valeurs de la variable réponse étant connues pour le jeu de test, l'évaluation de la qualité prédictive du modèle construit se fait aisément. L'utilisation de cette procédure sous-entend de travailler avec un jeu de données riche, pour lequel l'utilisation d'une proportion non négligeable de sujets (au moins 20%) uniquement à des fins d'évaluation du modèle n'est pas problématique. Dans le cas où l'on souhaite comparer différents algorithmes d'apprentissage (ou une même méthode paramétrisée de différente manière), il est même recommandé de diviser l'échantillon en trois parties : un échantillon "d'apprentissage" sur lequel les différents modèles seront construits, un échantillon de "validation" permettant de sélectionner le modèle ayant les meilleurs performances prédictives et enfin un échantillon de "test" permettant d'avoir une idée de la capacité du meilleur modèle à être performant sur des données indépendantes [82] (la répartition des individus dans les trois sous-échantillons étant généralement 50%-25%-25%).

Si le nombre d'individus est trop faible pour que le jeu de données puisse être divisé en de multiples sous-échantillons l'utilisation de la *cross-validation* est alors recommandée. La technique de *cross-validation* consiste en une division du jeu de données en k sous-échantillons (on emploie alors le terme *k-fold cross-validation*). De manière itérative, un des k sous-échantillon est défini comme jeu de test tandis que les autres sous-échantillons servent de jeu d'apprentissage. Un modèle est construit sur le jeu d'apprentissage et son erreur est estimée sur le jeu de test. Ce processus est répété jusqu'à ce que l'ensemble des k sous-échantillons ait servi de jeu de test [95]. La moyenne des k erreurs est alors calculée pour estimer l'erreur de prédiction globale du modèle.

Les méthodes *data mining* que nous présentons sont uniquement des méthodes d'apprentissage supervisé, celles-ci étant d'un intérêt supérieur pour rechercher des associations avec une variable réponse médicale (voir Hastie *et al.* 2009 [82], pour plus d'informations concernant l'apprentissage non supervisé).

4.4.1 Réseaux de neurones artificiels

Les réseaux de neurones artificiels sont des modèles utilisés dans des buts de classification ou de régression et s'inspirent du fonctionnement biologique des neurones dans le cerveau humain. Un réseau de neurones est constitué d'un ensemble de couches, chacune des couches prenant ses entrées sur les sorties de la couche précédente [98]. Le réseau présenté sur la figure 4.2 est constitué de trois couches, la couche d'entrée (les prédicteurs), la couche cachée (en pratique le nombre de couches cachées est variable) et la couche de sortie (la prédiction effectuée par le modèle). Un neurone particulier va recevoir des neurones qui lui sont connectés en amont un certain nombre d'entrées. Ce neurone va calculer une combinaison linéaire pondérée des entrées reçues par le biais d'une fonction de combinaison (cas le plus courant du perceptron multi-couches, *multi-layer perceptron*). Une fonction d'activation (généralement une transformation sigmoïde) est ensuite appliquée pour obtenir le signal fourni en sortie par le neurone [82, 188]. L'intérêt de la fonction d'activation est d'introduire de la non-linéarité dans le fonctionnement du réseau.

Les réseaux de neurones artificiels ont longtemps été considérés comme une méthode de référence pour des problématiques de classification en biomédecine [19]. Ces méthodes ont cependant un certain nombre d'inconvénients. Premièrement, l'obtention d'un réseau de neurones avec des performances prédictives optimales requiert l'optimisation d'un certain nombre de paramètres (notamment la valeur des poids utilisés pour la pondération des entrées issues des neurones situés en amont ou encore le nombre de couches cachées), ce qui peut conduire à des temps de calculs très importants [213]. Ensuite, même si les règles de décision des réseaux de neurones ont une forme mathématique explicite, celle-ci est trop complexe pour être interprétable, limitant fortement l'intérêt de cette méthode dans l'optique de détecter des associations [124].

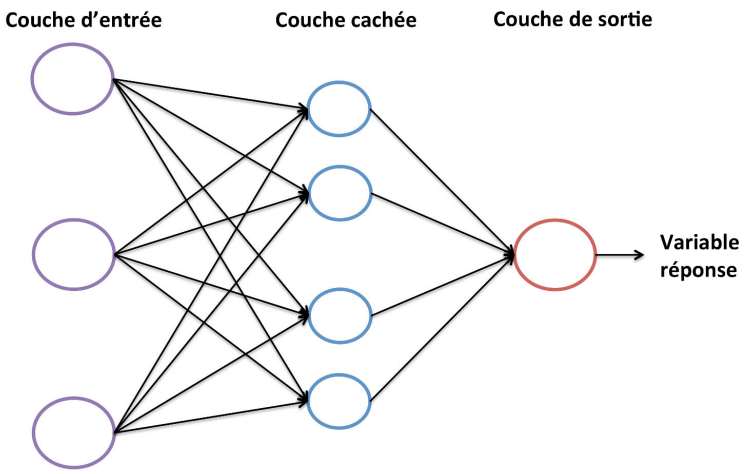


FIGURE 4.2 : Structure d'un réseau de neurones artificiels à trois couches

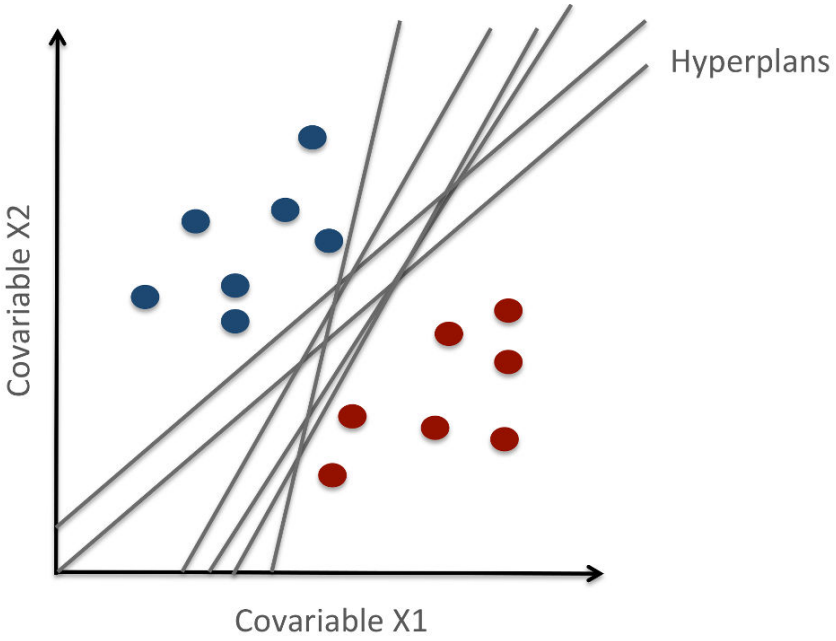


FIGURE 4.3 : Ensemble d'hyperplans possibles dans un espace à deux dimensions

4.4.2 Machines à vecteurs de support

Les machines à vecteurs de support (ou *support vector machine*, SVM) sont des modèles d'apprentissage supervisé généralement utilisés pour des problèmes de classification binaire (des extensions existent néanmoins pour des problèmes de régression ou de classification à plus de deux classes). L'idée fondamentale derrière l'algorithme des SVM est de projeter les données dans un nouvel espace où elles sont, plus ou moins, linéairement séparables et de déterminer une règle linéaire (ou hyperplan) permettant de séparer au mieux les groupes. Dans le cas de données linéairement séparables, il peut exister un grand nombre d'hyperplans permettant de séparer les données (voir la figure 4.3 pour un exemple simple de données à deux dimensions). L'hyperplan optimal est celui permettant de maximiser la distance entre l'hyperplan et les données d'apprentissage les plus proches (notion de marge maximale). Autrement dit, on cherche la droite (l'hyperplan) la plus éloignée des points qui lui sont le plus proche (les vecteurs supports) (Figure 4.4) [140, 213]. Cet hyperplan optimal minimise le risque d'*overfitting* et ainsi optimise la généralisation du modèle à de nouvelles données. La classification de nouvelles données se fait simplement en déterminant de quel côté de l'hyperplan optimal elles se situent [124].

Dans le cas où les données ne sont pas linéairement séparables, une transformation des données par l'utilisation d'une fonction non linéaire, appelée noyau (ou *kernel*), réalisant des produits scalaires des données dans un espace de plus grande dimension, permet de résoudre ce problème. Travailler dans un espace de plus grande dimension rend plus probable l'existence d'une séparation linéaire des données.

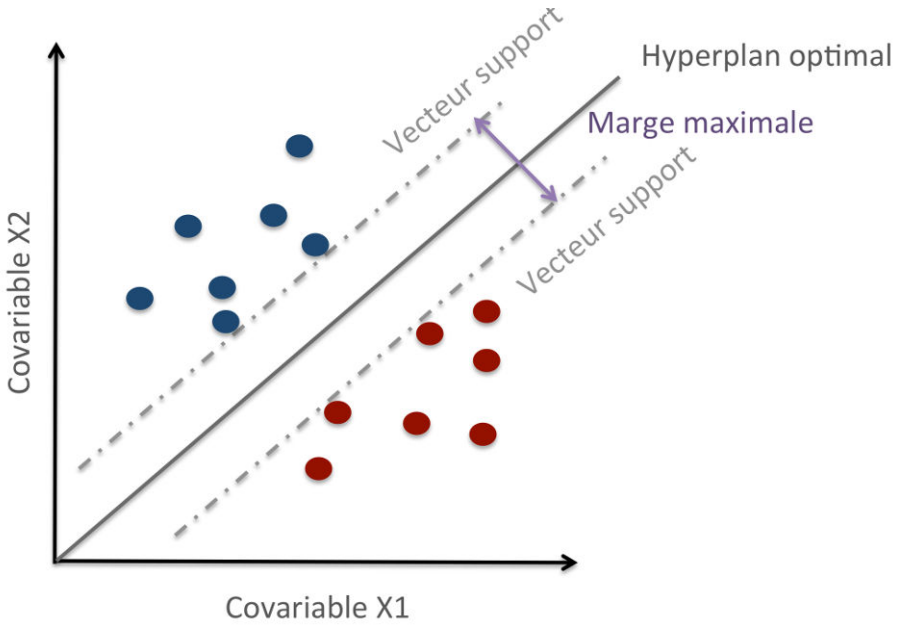


FIGURE 4.4 : Hyperplan optimal et vecteurs supports associés

Différents noyaux sont disponibles, noyau linéaire, gaussien ou encore polynomial [84]. Il est parfois impossible même après transformation des données de trouver une séparation linéaire, le recours à des marges souples (*soft margins*), introduisant une tolérance à l'erreur, est alors possible.

Les SVM sont des modèles très populaires, ils permettent de modéliser des phénomènes non linéaires et ont de bonnes performances prédictives [128]. En revanche, ils sont très sensibles au choix des paramètres, notamment au choix du noyau, et sont comme les réseaux de neurones des modèles très opaques, pour lesquels l'interprétation est complexe [213].

4.4.3 Arbres de classification et de régression

Les arbres de classification et de régression reposent sur l'idée d'un partitionnement récursif des individus et la construction d'un modèle simple au sein de chaque partition. Les partitions sont représentées sous la forme d'un arbre, les extrémités des branches correspondant aux différents résultats possibles en fonction des règles de décision construites par l'algorithme.

Les arbres de classification et de régression diffèrent en terme du type de variable réponse à prédire. Les premiers sont utilisés pour prédire des variables catégorielles non ordonnées, tandis que les seconds sont utilisés pour des variables continues ou des variables catégorielles ordonnées. Ces méthodes se distinguent également par le critère utilisé pour évaluer la qualité de prédiction ; les arbres de classification cherchent à un minimiser un taux de mauvaise classification alors que les arbres de régression visent à réduire un terme d'erreur quadratique entre les valeurs prédites et les valeurs observées [111].

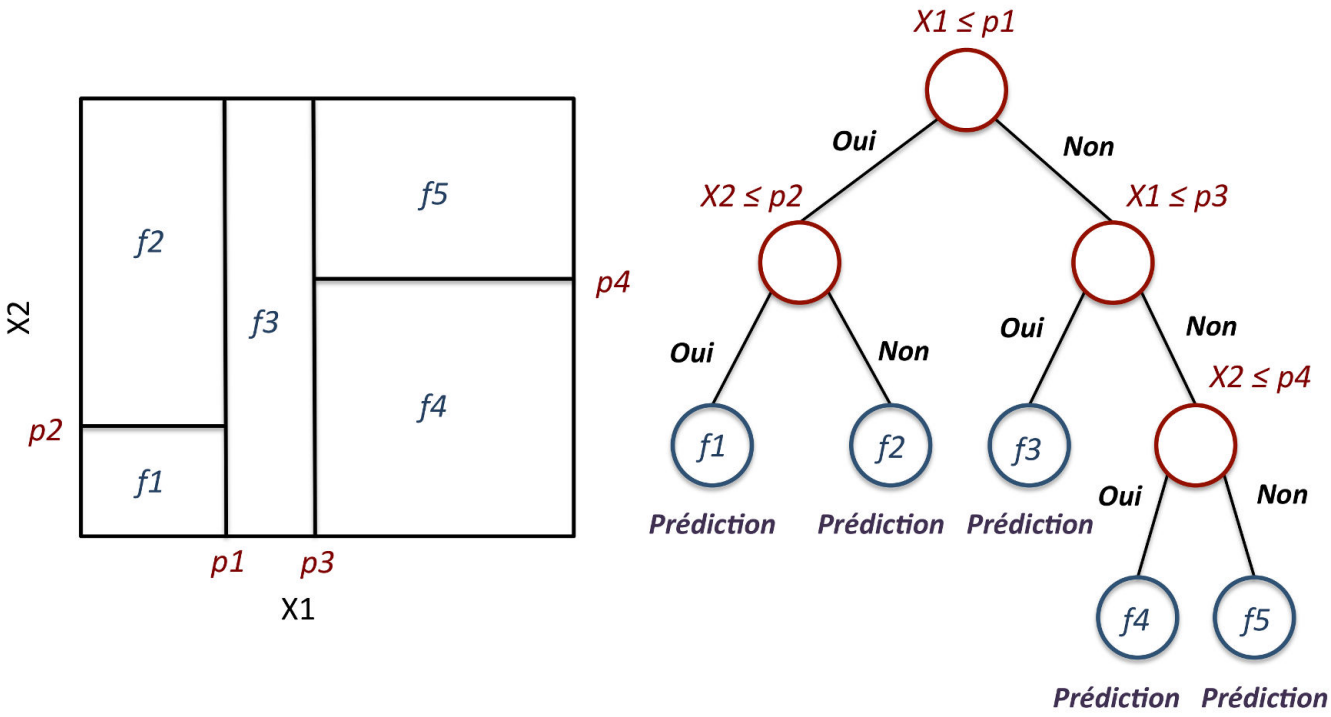


FIGURE 4.5 : Partitions d'un espace à deux dimensions et arbre associé (adapté de Hastie *et al.* 2009 [82])

Á chaque étape du processus l'ensemble des covariables sont évaluées en fonction de leur capacité à séparer les individus tout en minimisant un critère d'impureté nodale, différant en fonction du type de variable à prédire (les critères d'impureté sont détaillés par la suite). La figure 4.5 présente un cas simple de partitionnement récursif avec uniquement deux covariables et l'arbre associé. En partant du premier nœud de l'arbre, l'algorithme va de manière itérative déterminer pour l'ensemble des covariables (parmi *X1* et *X2* dans notre exemple), quelle covariable et quel partitionnement (p dans notre exemple) va permettre de minimiser le critère d'impureté. Le processus est itéré jusqu'à ce qu'une condition d'arrêt pré-spécifiée soit atteinte [82]. Le critère d'arrêt repose généralement sur l'effectif minimal dans chaque nœud terminal de l'arbre ou sur le nombre total de nœuds [46]. Le résultat du partitionnement présenté sur la figure 4.5 a permis de diviser les données en cinq régions, correspondant aux nœuds terminaux (ou feuilles) sur l'arbre résultant. Le modèle prédit la valeur de la variable réponse en suivant les différentes branches de l'arbre jusqu'au nœud terminal correspondant. La réponse modélisée par un arbre de régression sera la moyenne de la variable réponse dans la région correspondante, tandis que pour un arbre de classification il s'agira de la classe majoritaire.

Critères utilisés pour le partitionnement des nœuds

Dans le cas d'un arbre de classification, la sélection de la covariable utilisée pour diviser un nœud "père" en deux nœuds "fils" repose sur la minimisation d'un critère d'impureté nodale. L'algorithme

CART (pour *Classification and Regression Trees*), qui est l'algorithme généralement utilisé pour construire de tels modèles, utilise l'indice de Gini, qui s'exprime ainsi (formule 4.7) :

$$i(n) = \sum_{c=1}^{C} \hat{p}_{nc}(1 - \hat{p}_{nc}) \tag{4.7}$$

avec $\hat{p}_{nc}$ la proportion d'individus classifiés dans la classe c (parmi C classes) parmi les individus se retrouvant dans le nœud n. Un terme de déviance (ou d'entropie) est parfois également utilisé (formule 4.8) :

$$i(n) = -\sum_{c=1}^{C} \hat{p}_{nc} log(\hat{p}_{nc}) \tag{4.8}$$

Pour un arbre de régression l'indice précédent est remplacé par un terme d'erreur quadratique moyenne (formule 4.9)

$$i(n) = \frac{1}{N_n} \sum_{i \epsilon n} (y_i - \bar{y})^2 \tag{4.9}$$

avec N_n le nombre d'individus dans le nœud n, et $\bar{y}$ la valeur moyenne de la variable réponse dans le nœud n.

La simplicité de l'algorithme de partitionnement, l'interprétabilité des résultats (avec un nombre raisonnable de variables), notamment si l'on compare cette méthode aux réseaux de neurones artificiels et aux machines à vecteurs de support, et sa capacité à fonctionner avec un grand nombre de covariables sont à l'origine de la popularité des arbres [46, 92]. La complexité des arbres doit cependant être limitée pour éviter un sur-ajustement du modèle aux données et permettre une généralisation de l'arbre construit à de nouvelles données. Des techniques d'élagage des arbres existent, notamment basées sur des procédures de *cross-validation*, et permettent de retirer les nœuds pour lesquels la plus-value apportée au modèle est négligeable [46, 92].

Les arbres de classification et de régression sont des approches très flexibles ne nécessitant pas la spécification d'un modèle sous-jacent et sont capables de gérer un grand nombre de variables. Cependant la stabilité des solutions n'est pas garantie et le risque de sur-ajustement n'est pas négligeable [181]. Des extensions de ces modèles ont été proposées pour en améliorer la stabilité et les qualités prédictives, basées sur une agrégation des prédictions obtenues sur un ensemble d'arbres. Ces extensions peuvent être regroupées sous le nom de méthodes d'ensemble. L'idée des méthodes d'ensemble est de construire un modèle prédictif en combinant un grand nombre de modèles individuels, l'idée étant que les performances prédictives du modèle combiné soient meilleures que celles de chacun des modèles individuels. Les modèles basés sur des arbres s'adaptent parfaitement à cette philosophie. En combi-

nant les prédictions d'arbres construits à partir de ré-échantillonnages successifs des données, il est possible de compenser l'instabilité des arbres de classification et de régression, tout en profitant du fait que ces modèles produisent généralement de bonnes prédictions [16]. Les méthodes d'ensemble reposent sur trois principaux algorithmes : le *bagging* (ou *bootstrap aggregating*), les forêts aléatoires (*random forests*) et le *boosting*.

4.4.4 Bagging

L'algorithme du *bagging* est la plus ancienne des méthodes d'ensemble. Son principe est simple et intuitif : à partir d'un jeu de données composé de N individus, K sous-échantillons de taille N' (entre 75% et 100% de la taille de l'échantillon d'apprentissage [148]) sont créés par une procédure de *bootstrap* [60] ; ceci implique que certaines observations peuvent être dupliquées au sein de chaque sous-échantillon [143]. Les K modèles sont construits sur les K échantillons *bootstrap*. Ainsi l'ensemble des arbres peuvent être construits avec des prédicteurs différents et avoir un nombre de nœuds terminaux variable. Les K modèles sont combinés en moyennant leurs prédictions si la variable réponse est numérique, ou en choisissant la classe majoritairement prédite, si la variable réponse est catégorielle. Bien que le *bagging* soit très souvent utilisé avec des arbres de classification ou de régression, il peut être employé avec d'autres types de modèles, comme les réseaux de neurones artificiels [143]. En dépit de sa simplicité les performances des modèles construits par *bagging* sont satisfaisantes et supérieures à celles d'arbres construits individuellement [29].

4.4.5 Forêts aléatoires

Les forêts aléatoires, proposées en 2001, sont proches conceptuellement de l'algorithme de *bagging*. Cette méthode repose sur l'idée d'agréger des modèles (de type *CART*) construits à partir de ré-échantillonnages successifs des données d'apprentissage. L'algorithme des forêts aléatoires (pour un problème de classification) peut être résumé ainsi (voir également la figure 4.6) [76] :

Pour chaque arbre de classification T_b parmi B arbres :

1. Un échantillon *bootstrap* K_b de taille N est constitué à partir des données d'apprentissage ;
2. Un arbre T_b de profondeur maximale est construit avec l'échantillon K_b, en répétant récursivement les étapes suivantes à chaque nœud de l'arbre jusqu'à ce que les nœuds terminaux ne contiennent que des individus appartenant à la même classe ou ne contiennent qu'un seul individu ;
   (a) Choisir aléatoirement m variables parmi les p prédicteurs (m est généralement fixé tel que $m = \sqrt{p}$) ;

(b) Choisir parmi les m variables celle permettant de réduire au maximum le critère d'impureté de Gini (formule 4.7)

(c) Diviser le nœud en deux nœuds fils

Une fois la forêt aléatoire construite, la classification de nouveaux individus se fait simplement en gardant la classe majoritaire parmi les prédictions réalisées par les B arbres construits. Pour une forêt aléatoire appliquée à un problème de régression les différences algorithmiques sont minces. Les prédictions sur des nouveaux échantillons sont faites en moyennant les résultats obtenus sur chacun des arbres. Le nombre de covariables sélectionnées aléatoirement à chaque nœud correspond généralement au tiers du nombre total de prédicteurs, et le nombre minimal d'individus dans un nœud terminal est cinq (ce nombre est de un dans un modèle de classification). Enfin le critère de sélection de la meilleure covariable pour la division d'un nœud est le terme d'erreur quadratique moyenne présenté précédemment (voir la formule 4.9) [82].

Pour chaque arbre de la forêt aléatoire, environ $^2/_3$ des individus sont utilisés pour construire le modèle, le reste constituant un échantillon de test appelé échantillon *out-of-bag* (OBB). Les échantillons *OOB* permettent d'obtenir une estimation non biaisée de l'erreur de généralisation (une fonction mesurant à quel point un modèle est généralisable à des données indépendantes) de la forêt aléatoire, sans avoir besoin de recourir à l'utilisation d'un jeu de données indépendant ou d'utiliser une procédure de validation interne de type *cross-validation* [51, 52]. A la fin de l'apprentissage, l'erreur de test pour un individu i est évaluée à partir des arbres construits sans cet individu et correspond simplement à la proportion d'arbres prédisant la mauvaise classe pour l'individu i (pour un modèle de classification) ou à la moyenne des résidus (*i.e.* la différence entre les valeurs prédites et les valeurs observées) au carré (pour un modèle de régression).

Bien que proche conceptuellement du *bagging*, l'algorithme des forêts aléatoires se distingue par sa procédure de sélection aléatoire des variables utilisées pour construire les arbres. Ceci assure une faible corrélation entre les arbres, permet de prévenir des problèmes de sur-apprentissage [193], et offre également des performances prédictives supérieures au *bagging*. Il a été observé que l'erreur de généralisation d'un modèle d'ensemble composé d'arbres dépend de la corrélation entre les arbres, ainsi une faible corrélation est associée avec une faible erreur de généralisation [30].

Les forêts aléatoires sont devenues une méthode populaire notamment en raison de leur capacité à gérer un très grand nombre de variables [193], de leur robustesse au bruit et aux faux positifs [32], et de leurs très bonnes performances prédictives [37, 121, 128, 185, 201]. S'il a déjà été observé que les machines à vecteurs de support avec des paramètres optimisés sont capables de surpasser légèrement les

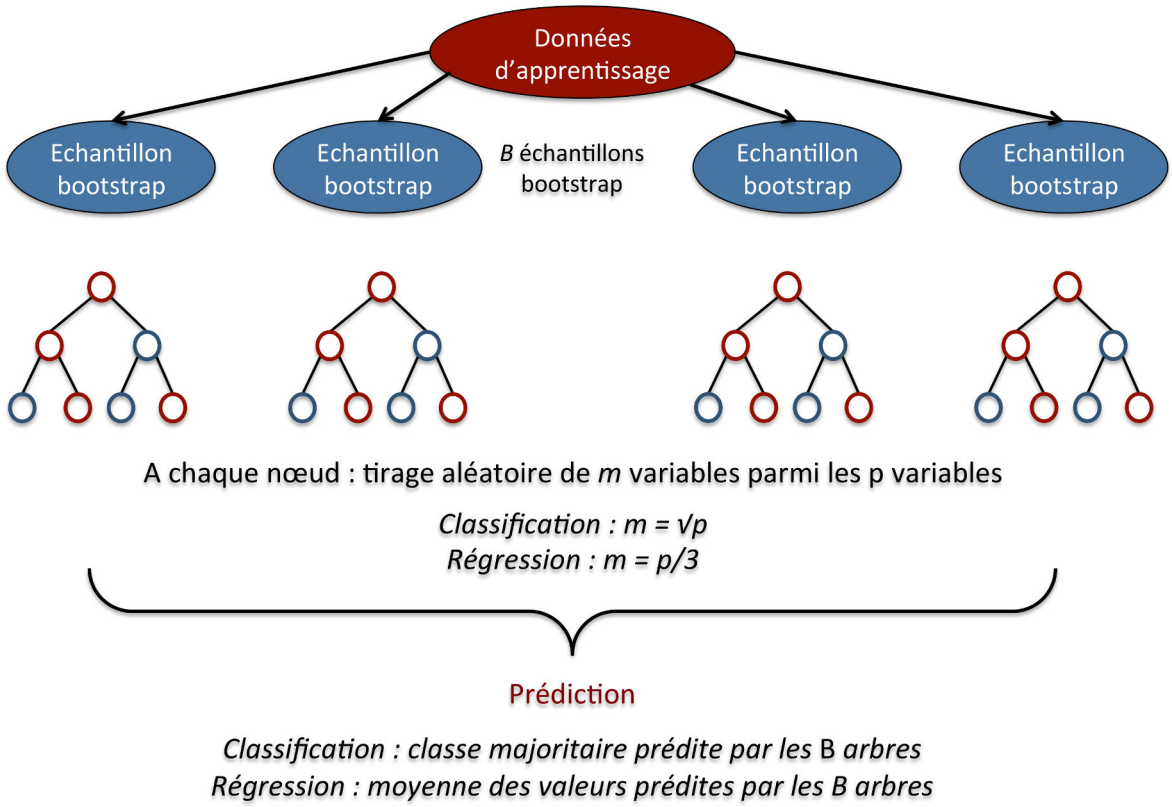


FIGURE 4.6 : Construction d'une forêt aléatoire

forêts aléatoires sur des problématiques de classification de données *omiques* [175], ces dernières ont pour avantage que l'optimisation de paramètres (*e.g.* le nombre d'arbres ou la proportion de variables m testées à chaque nœud [76]) n'est pas une condition *sine qua non* pour obtenir un modèle de qualité. Les forêts aléatoires fournissent également une estimation numérique des variables importantes pour la classification ou la régression, ceci étant d'un intérêt majeur pour identifier des associations. Cette caractéristique des forêts aléatoires est détaillée dans le chapitre 4.4.8.

4.4.6 Boosting

Contrairement au *bagging* et aux forêts aléatoires, le *boosting* ne repose pas sur un processus de moyennage des résultats d'un ensemble de modèles construits indépendamment les uns des autres, il s'agit ici d'un processus itératif [136]. La philosophie du *boosting* est de combiner les prédictions de modèles très simples, appelés "apprenants faibles" (*weak learners*) pour obtenir au final un "apprenant fort" (*strong learner*). Un *weak learner* correspond à un modèle dont les prédictions sont à peine meilleures que celles qu'on pourrait obtenir en prédisant de manière aléatoire [158]. L'algorithme de *boosting* le plus populaire est l'algorithme *AdaBoost* [68]. Les deux principales variantes de l'algorithme *AdaBoost* sont *AdaBoost.M1*, conçu pour des problématiques de classification et *AdaBoost.R* utilisé pour des problèmes de régression. Nous présentons uniquement la version *AdaBoost.M1*, étant

principalement concernés par la prédiction d'une variable dichotomique de type "infecté/non infecté". Deux extensions du *boosting*, le *gradient boosting* et le *stochastic gradient boosting* sont également discutées.

AdaBoost.M1

Le processus itératif de *boosting* par *AdaBoost.M1* est résumé sur la figure 4.7. L'idée est d'appliquer séquentiellement des arbres de classifications sur des versions pondérées des données d'apprentissage, produisant ainsi une séquence de modèles de classification $T_b(x), b = 1, 2..., B$, avec B le nombre d'itérations de *boosting*. Les prédictions Y ($Y \epsilon \{-1; 1\}$ ici) de l'ensemble des arbres de classification T_b sont combinées et pondérées pour obtenir la prédiction finale du modèle (voir la formule 4.10).

$$T(x) = sign(\sum_{b=1}^{B} \alpha_b T_b(x)) \tag{4.10}$$

Les termes α_b sont des poids estimés par l'algorithme et sont d'autant plus importants que les performances prédictives (taux de bonne classification) de chacun des arbres $T_b(x)$ sont bonnes [82]. A chaque étape de l'algorithme, les données sont également pondérées (à la première itération leur poids est identique), de telle sorte que les individus mal classés soient sur-pondérés, tandis qu'un poids moins important est appliqué sur les individus correctement classifiés. Ainsi, à l'itération suivante, l'arbre construit se concentrera majoritairement sur les individus mal classés par l'arbre construit à l'étape précédente.

Des extensions du principe algorithmique de *boosting* ont été proposées, permettant l'utilisation d'arbres de régression pour prédire des variables réponses binomiales, de Poisson ou encore des données de survie. Nous présentons l'algorithme de *gradient boosting* permettant l'application du *boosting* à ces différentes distributions.

Gradient boosting

Le principe du *gradient boosting* diffère de celle d'un algorithme de *boosting* comme *AdaBoost.M1*, notamment par le type d'arbres utilisés, s'agissant ici d'arbres de régression. L'idée est ici d'utiliser de manière itérative un algorithme de gradient (algorithme de la plus forte pente, ou *steepest descent*) pour réduire une fonction de perte (*loss function*). Cette fonction représente une perte d'un point de vue de la performance prédictive en raison d'un modèle non optimal [70].

Le *boosting* est ici la technique d'optimisation numérique permettant de minimiser la fonction de perte en ajoutant itérativement des arbres de régression réduisant au maximum (principe de la "plus forte pente") l'erreur du modèle. La fonction de perte est généralement un terme de déviance, relatif

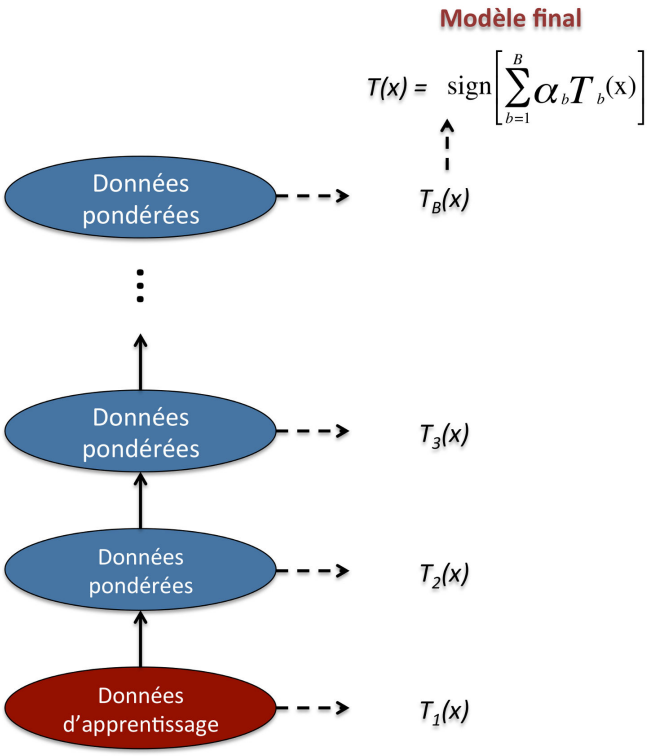


FIGURE 4.7 : Représentation schématique de l'algorithme AdaBoost.M1 (adapté de Hastie *et al.* 2009 [82])

à la log-vraisemblance du modèle [136] (voir la formule 4.11 pour le calcul de la déviance avec une variable réponse suivant une distribution binomiale [155]). Le gradient négatif de la fonction de perte dans le cas d'une variable réponse binomiale correspond simplement aux résidus du modèle (*i.e.* la différence entre les observations et les prédictions du modèle) (formule 4.12) [155].

$$deviance = -2\sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i) \tag{4.11}$$

$$z_i = y_i - p_i = y_i - \frac{1}{1 + exp(-f(x_i))} \tag{4.12}$$

Á la première étape de l'algorithme un arbre de régression est construit et des résidus sont calculés. Le second arbre est construit sur les résidus z du premier arbre, cet arbre pouvant contenir différentes variables et différents nœuds. Les résidus du modèle contenant les deux arbres sont calculés, un nouvel arbre est construit, et ainsi de suite. La modèle final obtenu est un combinaison linéaire d'arbres de régression (au minimum plusieurs centaines). Les arbres de régression constituant le modèle final sont généralement des modèles très simples, constitués d'un nombre limité de nœuds terminaux. La contribution de chaque arbre au modèle final est fixée via un paramètre appelé "taux d'apprentissage" (*learning rate*).

Si les arbres de régression "*boostés*" (*i.e.* construits par *gradient boosting*) semblent proches conceptuellement des arbres reposant sur l'algorithme *AdaBoost.M1*, il a été observé que l'utilisation de la fonction de perte binomiale est plus robuste qu'une fonction d'erreur reposant sur une proportion de mauvaise classification, et que ces derniers sont ainsi moins performants en cas de données d'apprentissage contenant des individus mal classifiés (faux positifs ou faux négatifs) [82].

Stochastic gradient boosting

Une extension du *gradient boosting* a été proposée en 2002, le *stochastic gradient boosting* [71]. Cette modification consiste à construire chaque arbre à partir d'un ré-échantillonnage sans remise du jeu de données initial (dont la taille est généralement égale à 50% du jeu de données d'apprentissage). Il a été observé que ces ré-échantillonnages successifs permettaient d'améliorer grandement les performances prédictives des modèles construits [71]. De plus, les observations non incluses dans la construction des arbres permettent d'obtenir une estimation du nombre d'itérations optimal. En effet, à chaque itération il est possible d'estimer l'erreur du modèle en prédisant la variable réponse chez les individus non inclus dans le jeu d'apprentissage. Cet échantillon est donc en quelque sorte assimilable à l'échantillon OOB rencontré dans les forêts aléatoires. La construction du modèle se poursuit alors jusqu'à ce que l'amélioration des performances prédictives soit considérée comme marginale [155].

De manière analogue aux forêts aléatoires, les modèles d'arbres de régression *boostés* construits par le *gradient boosting* ou le *stochastic gradient boosting* fournissent une estimation numérique de l'importance relative de chacune des variables incluses pour prédire la variable réponse. Cette propriété est détaillée dans le chapitre 4.4.8.

4.4.7 Performances prédictives des méthodes d'ensemble

D'un point de vue des capacités prédictives, les méthodes d'ensemble présentées ici s'avèrent très performantes. Une étude de simulation effectuée sur 11 jeux de données différents a permis d'observer que ces méthodes étaient les plus performantes parmi un panel de 10 méthodes d'apprentissage supervisé, cette même étude suggérant toutefois qu'il n'existe pas de meilleur algorithme universel [38]. Une autre étude de simulation s'est proposée de comparer la sensibilité des performances prédictives des forêts aléatoires et des arbres de régression construits par *gradient boosting* au ratio du nombre de variables associées avec la variable réponse sur le nombre de variables non associées. On peut observer une augmentation de l'erreur de test des forêts aléatoires avec la diminution de ce rapport [82] (voir la figure 4.8). Ceci s'explique par le fait que les forêts aléatoires ne sélectionnent qu'un sous-échantillon des variables disponibles pour la division de chaque nœud, la probabilité de sélectionner une variable associée diminuant avec l'augmentation du nombre de variables non associées. Le nombre de variables

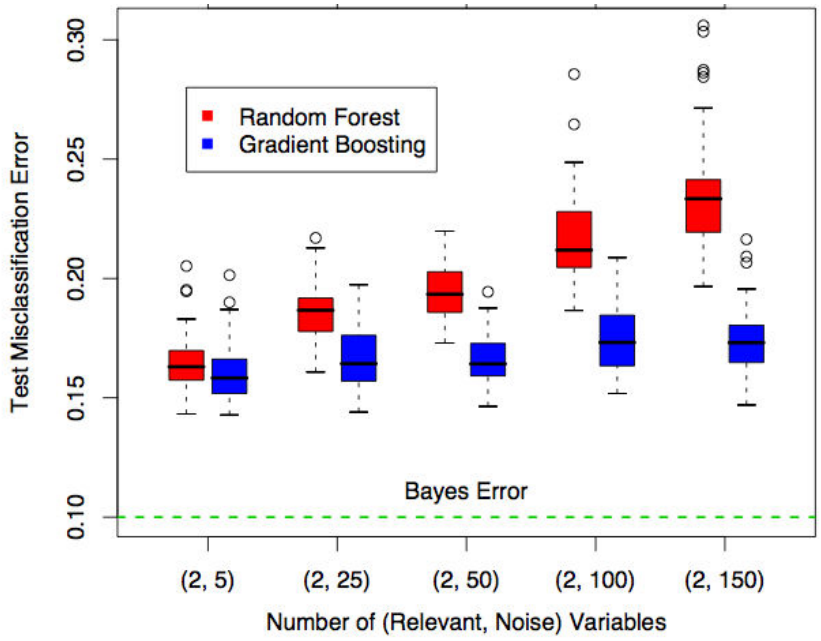


FIGURE 4.8 : Performances prédictives des forêts aléatoires et des arbres de régression construits par *gradient boosting* en fonction du nombre de variables associées et non associées avec la variable réponse. Résultats basés sur 50 simulations pour chaque cas de figure, avec des jeux d'apprentissage de taille N = 300 et des jeux test de taille N = 500 (issu de Hastie *et al.* 2009 [82])

associées avec la variable réponse est ici relativement faible. Les auteurs suggèrent toutefois qu'avec un nombre de variables associées plus important les forêts aléatoires se montrent plus robustes à l'augmentation du nombre de variables non pertinentes.

L'obtention d'un modèle de forêt aléatoire ou d'arbres de régression *boostés* avec des performances prédictives optimales peut nécessiter l'optimisation de certains paramètres. Dans le cas des forêts aléatoires, le choix du nombre de prédicteurs à considérer pour la division de chacun des nœuds des arbres ou le nombre d'arbres inclus dans le modèle pourront être considérés, le critère pour choisir un jeu de paramètres optimaux étant généralement l'erreur moyenne du modèle sur les échantillons OOB [76]. Pour les modèles d'arbres de régression *boostés*, en plus du nombre d'itérations (*i.e* le nombre d'arbres dans le modèle) abordé précédemment, l'optimisation du taux d'apprentissage (étroitement lié au nombre d'arbres) et du nombre de nœuds terminaux peut être envisagée [62].

4.4.8 Comment extraire de la connaissance des modèles data mining ?

Si les méthodes de *data mining* se distinguent par leur flexibilité et leurs bonnes performances prédictives, elles peuvent apparaitre comme des "boites noires" pour lesquelles l'interprétation des résultats, et de fait l'extraction de connaissance, ne sont pas triviales.

Une méthode de sélection récursive des covariables par des SVM a été proposée, classant les covariables par un critère relatif au poids de chaque prédicteur dans la construction de la règle de décision

(*i.e.* l'hyperplan optimal) utilisée pour classifier les individus [80]. L'idée étant de construire itérativement des SVM et de retirer à chaque fois les covariables avec un critère de poids parmi les plus faibles. L'interprétation de cette mesure d'importance n'est toutefois pas intuitive. De plus, des expériences de simulation ont montré que cette méthode était susceptible d'assigner des poids plus importants à des variables non associées avec la variable réponse qu'à des variables associées [174].

Les arbres de classification et de régression, s'ils sont construits avec un nombre relativement limité de variables, sont des modèles facilement interprétables. En revanche, l'agrégation de milliers d'arbres, comme dans le cas des forêts aléatoires, ou la combinaison linéaire d'autant de modèles, dans le cas du *gradient boosting* ou du *stochastic gradient boosting*, induit inévitablement un manque d'interprétabilité des résultats finaux. Cependant ces modèles ont la capacité de fournir des estimations numériques de l'importance des variables dans la classification des individus ou la prédiction d'une variable réponse numérique, permettant ainsi l'identification des prédicteurs potentiellement importants.

Importance des variables dans les forêts aléatoires

Deux mesures d'importance des variables sont estimées par l'algorithme des forêts aléatoires. La première d'entre elles utilise le critère d'impureté de Gini ou le terme d'erreur quadratique moyenne, selon le type d'arbre considéré, arbre de classification ou de régression (voir les formules 4.7 et 4.9). La mesure d'importance d'une variable X_j est simplement estimée en calculant la diminution moyenne sur l'ensemble des arbres du critère d'impureté lorsque la variable X_j est utilisée pour diviser un nœud [180].

La seconde mesure d'importance est calculée en utilisant les observations constituant les échantillons *OOB* (*i.e.* les observations non utilisées pour construire un arbre donné) [179].

Pour chaque arbre T_b parmi B arbres :

1. Calcul de l'erreur de l'arbre Err_{OOB} (proportion de mauvaise classification ou moyenne des résidus au carré) sur les individus de l'échantillon OOB ;
2. Pour chaque prédicteur $X_j, j = 1, ..., p$;
   (a) Permutation des valeurs de la variable X_j dans l'échantillon *OOB*
   (b) Calcul de l'erreur de l'arbre $Err_{OOB_{permuté}}$ avec les valeurs permutées ;
   (c) Calcul de la différence d'erreur de l'arbre T_b entre l'échantillon *OOB* permuté et l'échantillon *OOB* original pour la covariable X_j : $D_j = Err_{OOB_{permuté}} - Err_{OOB}$

La différence d'erreur D_j entre l'échantillon *OOB* permuté et l'échantillon *OOB* original pour la covariable X_j est ensuite moyennée sur les B arbres de la forêt aléatoire pour obtenir la valeur de la

seconde mesure d'importance, que nous appelons par la suite mesure d'importance de permutation.

L'idée de permuter chacune des variables incluses dans le modèle est de briser les associations éventuelles avec la variable réponse étudiée, conceptuellement cela équivaut à fixer à 0 un coefficient dans un modèle de régression. Ainsi, si une variable est associée avec la variable réponse, on s'attend à observer une augmentation de l'erreur de prédiction moyenne des arbres de la forêt aléatoire en supprimant cette association. Dans le cas contraire, si une variable n'est pas importante pour prédire la variable réponse, on peut penser qu'elle ne sera jamais incluse dans le modèle (auquel cas son importance sera nulle), ou qu'elle le sera très peu de fois. Dans ce cas son importance sera très faible, voire même possiblement négative [181].

Les propriétés et le comportement des mesures d'importance dans les forêts aléatoires a fait l'objet d'une abondante littérature ces dernières années, particulièrement dans le cas des arbres de classification. Il a été observé que les deux mesures d'importances fournissent des résultats proches, d'un point de vue du classement des variables les plus importantes [201]. Si ces estimations sont maintenant couramment utilisées dans des domaines très divers pour identifier des prédicteurs importants [55, 161, 201], elles ont toutefois certaines limites. Il a en effet été observé que les deux mesures d'importance des forêts aléatoires peuvent être biaisées en faveur des covariables continues ou des covariables catégorielles avec un grand nombre de catégories [180]. Reif *et al.* ont toutefois observé par une étude de simulation que les forêts aléatoires sont capables de détecter des covariables associées à une variable réponse dans des jeux de données contenant à la fois des variables continues et des variables catégorielles [153]. La mesure d'importance de Gini, tout comme la mesure d'importance de permutation, semblent également biaisées en faveur des covariables corrélées [179], bien qu'une autre étude de simulation ait démontré leur capacité à identifier des prédicteurs associés dans de telles conditions [6].

Mesure d'importance conditionnelle

Les biais potentiels observés avec les mesures d'importance ont entrainé le développement d'approches visant à améliorer l'estimation de l'impact des variables incluses dans les forêts aléatoires. Strobl *et al.* ont proposé une version modifiée de la mesure d'importance de permutation, appelée mesure d'importance conditionnelle (*conditional variable importance*) [179].

L'hypothèse nulle sous-entendue lors du calcul de la valeur d'importance de permutation d'une covariable X_j est une hypothèse d'indépendance de la covariable X_j à la fois par rapport à la variable cible Y et par rapport à l'ensemble des autres covariables (que nous appelons Z). L'hypothèse H_0 d'absence d'association peut être exprimée comme : $H_0 : X_j \perp (Y, Z)$. Le rejet de cette hypothèse (par l'estima-

tion d'une forte valeur de mesure d'importance) induit que X_j et Y ne sont pas indépendants et/ou que X_j et Z ne sont pas indépendants. Dans une optique de détection d'associations, seule la première solution est d'intérêt. Les auteurs ont ainsi développé une mesure d'importance des variables inspirée de la mesure de la mesure d'importance de permutation, basée sur une hypothèse nulle d'indépendance entre X_j et Y, conditionnellement à Z. L'hypothèse nulle d'absence d'association s'exprime alors comme $H_0 : (X_j \perp Y)|Z$. Dans le cas particulier où X_j et Z sont indépendants, la mesure d'importance conditionnelle renvoie des résultats identiques à ceux obtenus en utilisant la mesure d'importance de permutation classique. Des études de simulation ont permis de montrer que l'estimation de l'importance conditionnelle offrait des estimations plus précises de l'impact réel des variables que les indices classiques, en particulier en présence de prédicteurs corrélés.

Importance des variables dans les arbres de régression *boostés*

De manière analogue aux forêts aléatoires, les modèles d'arbres de régression *boostés* sont très difficilement interprétables en tant que tels. Il est cependant possible de calculer un indicateur numérique d'importance des prédicteurs. L'importance d'une variable X_j dans un arbre de régression *boosté* T_b est définie dans la formule 4.13. Les nœuds terminaux ne sont pas utilisés pour calculer l'importance de la variable X_j ; ainsi si l'arbre est constitué de D divisions, seuls les nœuds allant de la racine au niveau $D-1$ de l'arbre sont considérés. Il s'agit simplement de compter le nombre de fois où la variable X_j est sélectionnée pour diviser un nœud (S étant la variable utilisée pour la division du nœud considéré), pondéré par le carré de l'amélioration de la fonction de perte (ici I) suite à la division du nœud considéré par la variable X_j [72]. L'importance globale de la variable X_j est obtenue est moyennant l'importance calculée sur chacun des B arbres du modèle (formule 4.14) [136]. Les mesures d'importance sont enfin standardisées de telle sorte que leur somme soit égale à 100, fournissant ainsi la mesure d'importance relative de Friedman [72].

$$Importance_{X_j}(T_b) = \sum_{i=1}^{D-1} {I_i}^2 1(S_i = X_j) \tag{4.13}$$

$$Importance_{X_j} = \frac{1}{B} \sum_{b=1}^{B} Importance_{X_j}(T_b) \tag{4.14}$$

En dépit de ses similarités avec les forêts aléatoires et de ses très bonnes performances prédictives, les applications des modèles d'arbres de régression *boostés* sont moins courantes et l'utilisation de leur mesure d'importance pour détecter des associations plus rare [114, 186, 204]. De plus, contrairement aux mesures d'importance des forêts aléatoires, la littérature concernant l'étude du comportement de

la mesure d'importance relative de Friedman est moins riche, bien qu'une tendance à la détection de variables corrélées ait été observée [9].

Graphiques de dépendance partielle

Le principe des graphiques de dépendance partielle est de représenter l'effet d'une covariable sur la variable réponse après prise en compte des effets moyens des autres variables du modèle (voir Hastie *et al.* 2009 [82], pour plus de détails concernant les formules permettant d'obtenir les valeurs de dépendance partielle). La figure 4.9 présente deux graphiques de dépendance partielle issus d'un modèle d'arbres de régression *boostés* construit par *stochastic gradient boosting*. On peut observer que les modalités *E* et *T* de la variable catégorielle ("*Fishing method*") sont associées avec un effet marginal sur le $logit$ de la probabilité de la présence de la variable réponse plus important, le même constat pouvant être fait avec des valeurs faibles de la variable continue ("*Maximum downstream slope (°)*"). Ces graphiques, initialement développés pour les modèles d'arbres de régression *boostés*, sont maintenant également utilisables avec des modèles de forêts aléatoires. Dans le cadre des forêts aléatoires, d'autres indices sont parfois utilisés pour une analyse plus en profondeur des résultats obtenus. On peut notamment citer la mesure de proximité, qui est une mesure de similarité entre les échantillons d'un jeu de données, elle est parfois utilisée pour détecter des "*outliers*" (*i.e.* des données aberrantes). La mesure d'importance locale, indiquant l'importance d'une covariable pour la prédiction d'un échantillon particulier, peut également être employée pour repérer des variables importantes dans des sous-groupes d'individus (voir Touw *et al.* 2013 pour plus d'information concernant ces deux indices [193]).

L'identification des variables les plus importantes dans les forêts aléatoires et les arbres de régression *boostés*, ainsi que l'utilisation des graphiques de dépendance partielle, réduisent l'aspect "boite noire" de ces méthodes complexes et permettent une extraction plus ou moins aisée de connaissance de ces multiples modèles. Après avoir décrit les principes méthodologiques d'un ensemble de méthodes de *data mining*, nous nous concentrons maintenant sur les modèles de régression pénalisée, qui proposent une adaptation des modèles classiques de régression pour l'analyse de grands jeux de données. Un des avantages de ces méthodes par rapport aux modèles de *data mining* est qu'une procédure de sélection de variables est implicitement intégrée à la modélisation, permettant ainsi de s'affranchir des variables présentant peu ou pas d'intérêt pour prédire la variable réponse.

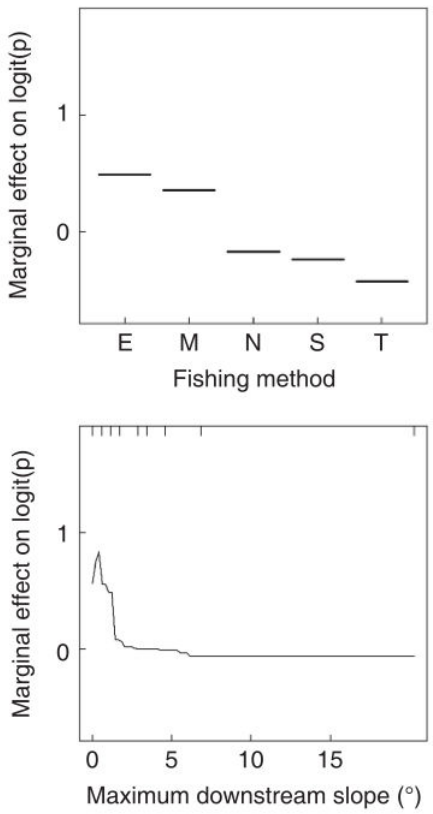


FIGURE 4.9 : Exemples de graphiques de dépendance partielle issus d'un modèle d'arbres de régression *boostés* construit par *stochastic gradient boosting*. Les axes des ordonnées sont en échelle *logit* et centrés afin d'avoir une moyenne nulle (issu de Elith *et al.* 2008 [62]).

4.5 Méthodes de régression pénalisée

Comme mentionné précédemment dans le chapitre 4.2, utiliser un grand nombre de covariables dans le modèle de régression logistique entraine des problèmes d'estimation, un possible risque de sur-ajustement du modèle aux données et rend également l'interprétation du modèle très complexe.

Un certain nombre de méthodes permettent de réduire le nombre de covariables d'un modèle de régression ; parmi les plus utilisées nous pouvons citer les méthodes de sélection *backward*, *forward* ou encore *stepwise*. Ces méthodes réduisent le nombre de variables en se basant généralement sur un critère d'ajustement du modèle aux données, tel que le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC). La sélection *backward* part d'un modèle avec toutes les covariables et les élimine successivement jusqu'à obtenir un modèle optimal. La sélection *forward* part quant à elle d'un modèle vide et ajoute des covariables les unes après les autres. La sélection *stepwise* est une méthode hybride combinant les propriétés des sélections *backward* et *forward*. Les deux procédés sont considérés à chaque étape, le meilleur des deux d'un point de vue de l'amélioration du modèle étant finalement choisi.

Á la fois très populaires et très critiquées [81, 82], ces méthodes imposent une pénalité sur le nombre de coefficients et non sur leur taille [13]. Avec un grand nombre de covariables, l'intérêt est plutôt de garder les prédicteurs avec une force d'association avec la variable réponse la plus importante et d'ignorer ceux avec une force d'association minimale, en dépit d'un possible meilleur ajustement du modèle aux données, justifiant le développement des méthodes de régression pénalisée. Ces méthodes, aussi appelées méthodes de *shrinkage*, appliquent différentes pénalités sur les coefficients du modèle de régression, ne conservant ainsi que les plus importants.

4.5.1 Régression ridge

La régression *ridge* réduit les coefficients du modèle en appliquant une pénalité sur leur taille [87]. De manière analogue à un modèle de régression logistique classique les coefficients sont estimés par maximisation de la log-vraisemblance du modèle (voir la formule 4.15).

$$logL(\hat{\beta})_{RIDGE} = \sum_{i=1}^{N}[y_i log(p_i) + (1 - y_i)log(1 - p_i)] - \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.15}$$

Le paramètre de régularisation λ ($\lambda \geqslant 0$) contrôle le niveau de contrainte appliquée sur les coefficients. Ce type de régression est adapté lorsque les prédicteurs sont fortement corrélés [59]. En revanche, elle a le défaut de conserver l'ensemble des prédicteurs dans le modèle, étant dès lors peu utile avec un grand nombre de prédicteurs [82].

4.5.2 Régression LASSO

Les coefficients de la régression *Least Absolute Shrinkage and Selection Operator* (LASSO) sont estimés par maximisation de la fonction de log-vraisemblance suivante (formule 4.16) [190] :

$$logL(\hat{\beta})_{LASSO} = \sum_{i=1}^{N}[y_i log(p_i) + (1 - y_i)log(1 - p_i)] - \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.16}$$

De manière analogue à la régression ridge, le paramètre λ contrôle le niveau de contrainte appliquée sur les coefficients. La pénalité LASSO est plus efficace que la pénalité *ridge* pour réduire des coefficients à 0. Ce phénomène s'explique notamment par le fait que pour des faibles valeurs de β, $|\beta|$ est supérieur à β^2 [210].

Bien que cette méthode soit plus efficace que la régression ridge pour réduire le nombre de covariables, elle souffre tout de même de certaines limitations. Dans le cas où $p > N$, le LASSO peut sélectionner au plus N covariables [217]. De plus en cas de présence de groupes de variables fortement corrélées, le LASSO a tendance à conserver une seule variable parmi ce groupe. Ceci peut être problématique par exemple dans des études génétiques, où l'on peut s'intéresser à des groupes de gènes ou de

variants génétiques. Dans de telles situations les performances prédictives de la régression *ridge* sont meilleures [190].

Ces dernières années ont vu le développement de nombreuses méthodes de régression pénalisée, permettant l'adaptation de la philosophie des pénalités *ridge* et LASSO à de nombreuses problématiques (voir Tibshirani 2011 [191]). On peut notamment citer la régression *group-LASSO*, dans laquelle l'ensemble des covariables sont divisées en groupes définis *a priori*, le terme de pénalisation est alors appliqué sur des groupes de variables et non sur les variables prises individuellement. Ainsi les variables sont incluses dans le modèle ou exclues par groupes [127]. Cette caractéristique peut être intéressante dans des études génétiques pour inclure des ensembles de variants génétiques dans des groupes identiques [187]. Nous présentons plus en détails une autre des évolutions du LASSO : la régression *elastic net*.

4.5.3 Régression elastic net

La régression *elastic net* permet de combiner les avantages de la régression *ridge* et du LASSO, et inclut un second terme de pénalisation dans le modèle [217]. Les coefficients s'estiment de telle façon à maximiser la fonction de log-vraisemblance suivante :

$$logL(\hat{\beta})_{ELASTICNET} = \sum_{i=1}^{N}[y_i log(p_i) + (1-y_i)log(1-p_i)] - \lambda_1 \sum_{j=1}^{p}|\beta_j| - \lambda_2 \sum_{j=1}^{p}\beta_j^2 \qquad (4.17)$$

Cette méthode, particulièrement adaptée dans le cas où $p > N$, permet comme le LASSO de réduire certains coefficients à 0. De plus cette pénalité s'avère particulièrement adaptée en présence de groupes de prédicteurs fortement corrélés, ceux-ci ayant tendance avec l'*elastic net* à être conservés dans le modèle ou à être exclus de manière groupée [217]. Ces deux propriétés en font une méthode particulièrement adaptée pour des problématiques d'analyse de données *omiques*, notamment en GWAS [202].

Caractéristiques communes aux méthodes de régression pénalisée

Nous pouvons noter que l'*intercept* β_0 est systématiquement exclu du processus de pénalisation. Dans le cas contraire le risque de base de présence de la variable réponse en l'absence de covariables aurait un impact sur la procédure de pénalisation des paramètres du modèle, ce qui n'est pas souhaité. Avant d'utiliser de telles méthodes il est également nécessaire de centrer et réduire l'ensemble des covariables incluses dans le modèle afin d'éviter tout effet d'unité de la variable sur l'estimation des coefficients.

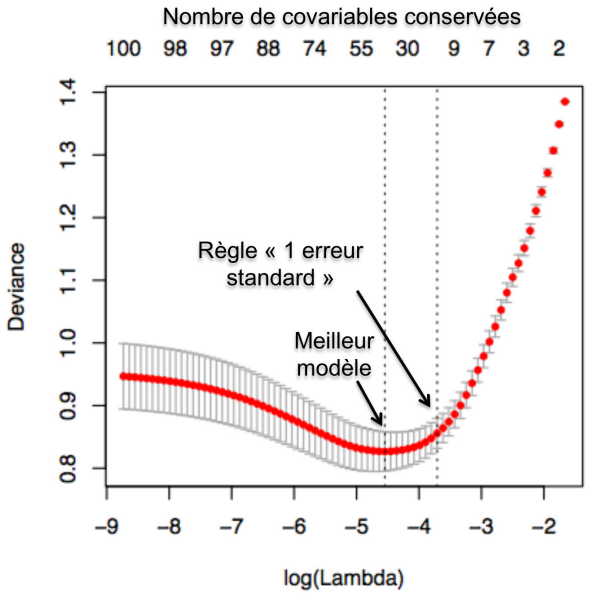


FIGURE 4.10 : Déviance moyenne en *10-fold cross-validation* d'un modèle de régression logistique LASSO en fonction du paramètre de pénalisation λ. Issu de données simulées de taille N = 1000 et p = 100 (dont 10 prédicteurs associés avec la variable réponse) (adapté de Friedman *et al.* 2010 [69]).

Il existe d'autres méthodes de régression permettant de gérer un grand nombre de variables, spécialement quand $p > N$ ou quand les variables sont fortement corrélées. Nous pouvons citer en particulier la régression sur composantes principales et la régression *Partial Least Squares* (PLS) [209]. Ces méthodes réalisent des combinaisons linéaires (des "composantes") des prédicteurs, et utilisent ces combinaisons linéaires pour prédire la variable réponse. Dans la régression sur composantes principales, chaque composante est une combinaison linéaire de tous les prédicteurs, cherchant à extraire le maximum de variance de ces derniers, sans tenir compte de la variable réponse [91]. La régression PLS va quant à elle construire des composantes optimisées pour être prédictives de la variable réponse [1]. S'il est possible d'estimer la proportion de variance des prédicteurs ou de la variable réponse expliquée par les composantes principales, il est difficile d'évaluer précisément l'effet des prédicteurs sur la variable réponse. De plus contrairement au LASSO ou à l'*elastic net*, ces méthodes n'ont pas de procédure explicite de réduction du nombre de covariables.

4.5.4 Estimation des paramètres de pénalisation dans la régression pénalisée

Pour les algorithmes de régression pénalisée présentés précédemment, une valeur de paramètre de pénalisation λ (deux paramètres dans le cas de l'*elastic net*) est associée avec un jeu de coefficients de régression estimés $\hat{\beta}$. Une approche classique pour la détermination des paramètres de pénalisation optimaux se base sur l'observation de l'erreur du modèle [69] (généralement la déviance ou la proportion de mauvaise classification pour une variable réponse dichotomique) avec une procédure de *cross-validation*.

La figure 4.10 représente l'erreur moyenne en *10-fold cross-validation* d'un modèle de régression logistique LASSO en fonction du paramètre de pénalisation λ. Á chaque valeur de λ est associé un nombre de covariables conservées (*i.e.* avec des coefficients de régression non nuls). Le modèle sélectionné n'est pas le "meilleur modèle" (*i.e.* le modèle avec la plus faible déviance moyenne), c'est la règle "d'une erreur standard" (*one-standard-error rule*) qui est généralement appliquée. Au lieu de conserver le modèle avec les meilleures performances, c'est généralement un modèle plus parcimonieux avec une erreur moyenne au maximum égale à l'erreur moyenne du meilleur modèle plus une erreur standard qui est considéré [69]. Cette règle permet de tenir compte du fait que les courbes de variation de la déviance en fonction des paramètres du modèle sont estimées avec une certaine erreur, et utilise ainsi une approche conservative [82].

Article : le risque d'infection grippale abordé sous l'angle data-driven

5.1 Résumé

La régression logistique est la méthode de référence utilisée en épidémiologie pour détecter des associations lorsque la variable réponse suit une distribution binomiale. Les spécificités de cette approche ainsi que ses limites ont pu être détaillées dans le chapitre précédent.

L'objet de cet article est la recherche d'associations indépendantes dans un grand jeu de données en épidémiologie, plus précisément la recherche de facteurs de risque indépendants de l'infection par la grippe pandémique H1N1pdm. Plus de 300 facteurs de risque sont étudiés ici ; l'échelle du jeu de données CoPanFlu est bien inférieure à celle de données *omiques* de type GWAS, ceci n'en reste pas moins un défi du point de vue de l'analyse statistique et nécessite de considérer des approches alternatives à la régression logistique.

Nous évaluons ici la capacité à détecter des associations indépendantes, sur les données de CoPanFlu mais aussi sur données simulées, de deux méthodes de *data mining*, les forêts aléatoires et les arbres de régression *boostés* (construits par *stochastic gradient boosting*), de la régression logistique pénalisée par le *Least Absolute Shrinkage and Selection Operator* (LASSO) et de la méthodologie couramment employée en épidémiologie en régression logistique. Celle-ci consiste en une analyse en deux étapes, premièrement les prédicteurs sont sélectionnés par une analyse univariée ; ceux ayant une *p-value* (le plus souvent issue du test de Wald) inférieure à un seuil donné (nous considérons les seuils 0.20 et 0.05) sont ensuite inclus dans un modèle multivarié, sur lequel une procédure de sélection de variables est éventuellement appliquée (nous utilisons ici une sélection *backward*).

N'ayant pas d'idée *a priori* sur les variables associées au risque d'infection grippale, la simulation de 500 jeux de données, proches en terme de dimension des données CoPanFlu, vise à avoir une idée des taux de vrais et de faux positifs associés à l'ensemble des méthodes employées.

Les forêts aléatoires, les arbres de régression *boostés* et la régression LASSO ne fournissent pas de *p-value* permettant d'évaluer la significativité statistique des associations mises en évidence. Des tests de permutation ont ainsi été implémentés afin d'interpréter les mesures d'importance des variables fournies par les méthodes de *data mining* et des coefficients estimés par le LASSO.

Selon les méthodes considérées entre trois et vingt-quatre variables ont été identifiées comme associées avec l'infection grippale, le titre d'anticorps pré-épidémique étant la seule variable pour laquelle

un consensus a été observé entre les différentes méthodes. L'analyse des 500 jeux de données simulés a permis d'observer que les taux de vrais positifs étaient les plus élevés pour les méthodes de *data mining* (80% en moyenne pour les arbres de régression *boostés*, 85% en moyenne pour les forêts aléatoires) tandis que la méthodologie d'analyse "univarié/multivarié" via des modèles logistiques était très peu performante, quel que soit le seuil de significativité utilisé pour l'analyse univariée, et indépendamment de l'utilisation d'une sélection *backward* des covariables (le taux de vrais positifs moyen était au plus de 49%). L'observation des taux de faux positifs a montré une légère augmentation du risque de première espèce (jusqu'à 9% en moyenne) pour les modèles basés sur la régression logistique, LASSO y compris, tandis que les méthodes de *data mining* ont montré un taux de faux positifs moyen de 4%.

Les méthodes de *data mining* et le LASSO doivent être considérés comme des méthodes fiables pour détecter des associations indépendantes dans des grands jeux de données en épidémiologie. La méthodologie conventionnelle employée en régression logistique n'est, à l'évidence, pas adaptée dans un tel contexte.

5.2 Article

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections

Yohann Mansiaux[1,2*] and Fabrice Carrat[1,2,3]

Abstract

Background: Big data is steadily growing in epidemiology. We explored the performances of methods dedicated to big data analysis for detecting independent associations between exposures and a health outcome.

Methods: We searched for associations between 303 covariates and influenza infection in 498 subjects (14% infected) sampled from a dedicated cohort. Independent associations were detected using two data mining methods, the Random Forests (RF) and the Boosted Regression Trees (BRT); the conventional logistic regression framework (Univariate Followed by Multivariate Logistic Regression - UFMLR) and the Least Absolute Shrinkage and Selection Operator (LASSO) with penalty in multivariate logistic regression to achieve a sparse selection of covariates. We developed permutations tests to assess the statistical significance of associations. We simulated 500 similar sized datasets to estimate the True (TPR) and False (FPR) Positive Rates associated with these methods.

Results: Between 3 and 24 covariates (1%-8%) were identified as associated with influenza infection depending on the method. The pre-seasonal haemagglutination inhibition antibody titer was the unique covariate selected with all methods while 266 (87%) covariates were not selected by any method. At 5% nominal significance level, the TPR were 85% with RF, 80% with BRT, 26% to 49% with UFMLR, 71% to 78% with LASSO. Conversely, the FPR were 4% with RF and BRT, 9% to 2% with UFMLR, and 9% to 4% with LASSO.

Conclusions: Data mining methods and LASSO should be considered as valuable methods to detect independent associations in large epidemiologic datasets.

Keywords: Data mining, Random forest, Boosted regression trees, LASSO, Logistic regression, Influenza

* Correspondence: yohann.mansiaux@iplesp.upmc.fr
[1]INSERM, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F-75013 Paris, France
[2]Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F-75013 Paris, France
Full list of author information is available at the end of the article

BioMed Central

© 2014 Mansiaux and Carrat; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

"Big data" [1] in information science refers to the collection and management of large and complex datasets. Big data is steadily growing in biomedicine with the development of electronic medical records, increased use of high-throughput technologies, and facilitated access to large environmental database [2-6]. In epidemiology, the collection of hundreds to thousands of covariates is common in large-scale cohort studies and offers new challenges for the discovery of associations between individual or collective exposures and a health outcome. The use of specific methods to explore these associations, without any pre-specified hypothesis, therefore becomes essential.

In hypothesis-driven epidemiology, the search for associations involves statistical modeling and testing of the relationships between one or several covariates and the outcome. Logistic regression is the most widely used model when the outcome follows a binomial distribution. The usual epidemiologic analytic framework consists in testing the association between each covariate and the outcome through univariate logistic models; a subset of those covariates is then selected for multivariate logistic models based on some quantile of the test statistic for the covariate coefficient under the null hypothesis, *i.e.* the Pvalue. This framework is the reference method in epidemiology for variable selection, and the use of alternative approaches remains uncommon [7,8]. With large datasets, the number of covariates selected in the univariate analyses can be high. As multivariate logistic regression can handle a limited number of covariates simultaneously [9], it might therefore be poorly adapted to large epidemiologic datasets for identifying independent associations.

"Data mining", a term which appeared in the early 1990's [10], describes data-driven analysis without any *a priori* hypothesis about the structure or the potential relationships that could exist in the data. Data mining applications are broad, ranging from consumption analysis to fraud detection in high-dimensional databases [11]. Data mining methods are non-parametric, more flexible than statistical regression methods, and are able to deal with a large number of covariates. Several studies have compared the performances of logistic regression and data mining methods for predicting a health outcome without clear conclusions about the superiority of one of these methods over the others [12-17]. Most studies explored classification and regression trees, artificial neural networks or linear discriminant analysis, but only a few focused on more recently developed "ensemble-based" methods such as random forests or boosted regression trees [13,16,17].

Shrinkage methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO) [18], have been developed to overcome the limitation of usual regression models when the number of covariates is high. However, LASSO logistic regression remains unfamiliar to epidemiologists and few applications of this method have been found [19,20].

We hereby performed a comparison of two data mining methods, random forests and boosted regression trees, with the conventional multivariate logistic regression and with the LASSO logistic regression for identifying independent associations in a large epidemiologic dataset including hundreds of covariates. Random forests and boosted regression trees were chosen among data mining methods for their ability to provide quantitative information about the strength of association between covariates and the outcome. The methods were used to detect covariates associated with H1N1 pandemic (pdm) influenza infections. We also assessed the performance of these methods to detect associations through simulations.

Methods

Data source

We used data from the CoPanFlu France cohort whose aim was to study the risk of influenza infection. Briefly, the cohort includes 601 households randomly selected between December 2009 and July 2010 and followed using an active surveillance system in order to detect influenza-like illness symptoms over two consecutive influenza seasons (2010–2011 and 2011–2012). More details about the study protocol, data collection and representativeness of households can be found elsewhere [21]. Ethics approval was given for this study by the institutional review board "Comité de Protection des Personnes Ile-de-France 1" and written informed consent was obtained for all participants.

The outcome of our study was H1N1pdm influenza infection during 2010–2011 season, defined as either a positive H1N1pdm RT-PCR [22] or a positive H1N1pdm RespiFinder assay [23] on a nasal swab collected during winter 2010–2011 or a seroconversion (4-fold increase of Haemagglutination inhibition (HAI) antibody titer [24] between post and pre-seasonal serum samples). Infection status for the 2010–2011 season was available in household members from 498 households. To neutralize within household correlation and ensure statistical independence of individuals included in the analysis, we adopted a case–control selection strategy. One case was sampled from each household where at least one influenza infection was detected and one control was selected from each household where no influenza infection occurred. Our analysis therefore focused on 498 subjects, 68 (14%) cases and 430 controls. Association with the outcome was explored for 303 covariates (a complete description can be found in Lapidus *et al.* [25]).

The mean age of study subjects was 44.3 years (SD = 21.1); 42% (208 individuals) were male. A total of 215 subjects (43%) had at least one history of chronic disease. The proportions of seasonal and pandemic vaccines recipients for 2009–2010 season were 19% and 10%, respectively. The mean number of subjects per household was 2.5 (SD = 1.3) and the number of children per household was 0.5 (SD = 0.9).

Methods for detection of independent associations

Random forests (RF)

Random Forests models were proposed by Leo Breiman [26]. RF consists of an ensemble of classification and regression trees. Each tree of the random forest is built as followed: a bootstrap sample of the original dataset is drawn with replacement. The rest of the observations compose the "out-of-bag" sample, used to assess the performances of the selected tree. At each node of the tree, a random subset of covariates is selected (usually as much as the square root of the total number of covariates). Selection of a covariate to split a "parent" node into two "child" nodes is the covariate among the subset leading to the largest decrease in the Gini impurity criterion, that is, for a binary outcome, $1 - p^2 - (1 - p)^2$ with p the proportion of individuals classified with the outcome (the influenza infection in our case). The partitioning process is iterated until the final nodes contain only individuals belonging to the same classes or until they contain only one individual. The tree is then used to classify every individual in the "out-of-bag" sample. This process is repeated until a pre-specified number of trees is reached (one thousand). Note that prediction for each individual is based on the averaged predictions over all trees. To rank potential relevant associations with covariates selected in the RF, we used the importance score, *i.e.*, the decrease in Gini impurity criterion from splitting on the covariate, averaged over all trees.

Boosted regression trees (BRT)

Boosted regression trees is another ensemble method combining regression trees with weak individual predictive performances, into a single model with high performances [27,28].

First a regression tree model is fitted to a subset of data to minimize a loss function (in our case the deviance), which quantifies the loss in predictive performance due to a suboptimal model. The "boosting" algorithm is a numerical optimization technique for minimizing the loss function by iteratively fitting new trees to the prediction residuals of the preceding tree. For example, at the second step, a tree is fitted to the residuals of the first tree, and that second tree could contain different covariates and split nodes compared with the first. The two regression trees are combined and the residuals are calculated, a new tree is fitted and so on. To improve accuracy and reduce overfitting, each regression tree is grown from a bootstrap sample (without replacement) of the original dataset (usually 50% of the original sample) [29]. The final model is built by combining weighted individual tree contributions, weights being proportional to the trees performances.

To assess and rank potential associations with covariates, we used the Friedman "relative influence" [27], *i.e.*, the number of times a covariate is selected for splitting, weighted by the squared improvement of the loss function by splitting on that covariate. One thousand trees were used for each model. To allow complex interaction detection, the "interaction.depth" parameter of the BRT models was set to a value of 3; implying models with up to 3-way interactions.

Conventional logistic regression framework

Logistic regression (LR) is a well-known Generalized Linear Model adapted to test association between a binomial outcome and covariates [30]. In order to identify independent associations, we reproduced the usual epidemiologic analytic framework – that is, Univariate Followed by Multivariate Logistic Regression (UFMLR). We explored two thresholds for the selection of covariates in the univariate step, Pvalue <0.05 (UFMLR05) and Pvalue <0.20 (UFMLR20), with coefficient tested with the Wald test. We also distinguished whether or not a backward selection of covariates was further applied in multivariate regression model.

Least Absolute Shrinkage and Selection Operator logistic regression (LASSO)

We used LASSO to fit a multivariate logistic model with penalty on the magnitude of coefficients. The LASSO method maximizes the log-likelihood of the model, while applying constraints on the sum of the absolute values of the coefficients– shrinking the less important coefficients to zero [18]. The constraint was expressed in terms of a penalty parameter; the optimal value of this parameter was determined by minimizing the deviance of the multivariate logistic model (hereafter called LASSO-max) averaged over five-fold cross-validation subsamples [31]. In addition to this optimal model, we also considered a parsimonious LASSO model with a higher penalty parameter (hereafter called LASSO-se) so that the mean deviance of the model was within one standard error of the LASSO-max average deviance [28,32]. To rank potential associations with covariates, we used the proportion of estimated non-null parameters over one hundred LASSO models.

Permutation test for statistical significance of associations

No threshold exists to assess the statistical significance of measures of association, *i.e.* the Gini impurity criterion in

RF models or the Friedman relative influence in BRT models or the proportion of estimated non-null parameters in LASSO models. We used permutation tests [33], which were applied in RF models to derive Pvalues for predictors and were suggested to be used with any method that provides a measure of covariate relevance [34]. The null hypothesis was the absence of association between a given covariate and the outcome. We computed the Pvalue by randomly permuting 999 times the values of the outcome and comparing the measures of association of the covariates in the permuted datasets with that from the original dataset [35]. Permutation of the outcome has the advantage of preserving the dependence structure between the covariates [34]. The one-sided Pvalue was computed as follows: $P_x = \frac{1+Number\ of\ permuted\ datasets\ where\ S_{per}>S_{obs}}{1+Number\ of\ permutations}$ with S_{per} and S_{obs} the measures of association for covariate x in the permuted and the real dataset, respectively. To increase consistency we also applied this method to UFMLR models, using the Wald test statistic as measure of association. We also checked whether permutation test Pvalues were similar to those obtained with the conventional Wald test (see Additional file 1).

Simulated data

To assess the performances of the different methods in detection of associations in a similar-sized dataset, we simulated 500 datasets with 500 individuals and 300 covariates, sampled from a multivariate normal distribution (the simulation process is detailed in Additional file 1). The simulated logit model involved 150 continuous and 150 binary covariates. Four covariates of each type were directly associated with the outcome. Four covariates (2 binary and 2 continuous) were involved in pairwise multiplicative interactions. Two hundred and ninety-two covariates were not associated (neither directly nor through interactions) with the outcome. Twenty covariates (the 8 associated covariates and 12 non-associated covariates - 6 continuous and 6 binary) were correlated ($\rho = 0.5$). To explore the influence of correlation among a higher number of covariates, we increased the number of correlated covariates to 58 (the 8 associated covariates and 50 non-associated covariates – see Additional file 1). For both simulations, the number of permutations of the outcome was set to 99. Performances were assessed in terms of True Positive Rate (TPR), *i.e.* the proportion of associated covariates which were detected, and False Positive Rate (FPR), *i.e.* the proportion of non-associated covariates which were detected, at 5% nominal level (see Additional file 1 for TPR and FPR at 1% nominal level). We also distinguished between covariates with and without pairwise interactions for the calculation of TPR, and between covariates with and without correlations with associated covariates for the calculation of FPR.

Fitting procedures and software

In all analyses, the covariates were centered and scaled prior to model fitting.

Statistical analyses were performed with R version 2.15.3. The following packages were used: "randomForest" [36] for RF models, "gbm" [37] for BRT models, "glmnet" [31] for LASSO-LR models.

Results

Detection of associations in the CoPanFlu dataset

With a nominal type I error of 5%, 11 independent associations were identified with RF, 12 with BRT, 24 with the LASSO-max, 8 with the LASSO-se, 3 to 5 with UFMLR05, and 9 to 10 with UFMLR20 (Table 1). Association between influenza infection and the pre-seasonal HAI titer was detected with all methods. The covariates "History of asthma", "Professional activity involves contact with ill people", "Daily frequency of hand washing (with soap or hand sanitizer) ≥ 5" and "Always or often covers mouth while coughing or sneezing", were also identified by RF, BRT, and LASSO-max. This last covariate, as well as 23 additional covariates, were not detected by any of the UFMLR methods. Three associations identified by UFMLR20, with or without backward selection, were not retrieved by any of the data mining or LASSO methods. Of note, backward selection applied in UFMLR was associated with more and sometimes different associations than UFLMR without selection. Finally, 266 (87%) covariates were not selected by any method.

Simulated data

At 5% nominal level, the True Positive Rate was 85% with RF, 80% with BRT and 71% to 78% with LASSO (Table 2). The UFMLR, with or without backward selection, was the least efficient method at detecting true associations, with a TPR ranging between 26% and 49%. All methods, except RF and BRT, exhibited higher TPR when associated covariates did not interact with other associated covariates. The TPR for continuous covariates was higher than the TPR for binary covariates with all methods. The proportion of simulated datasets in which all associated covariates were detected was 36% with RF, 17% for BRT, 10% for LASSO-max, 5% for LASSO-se, 0% for UFMLR05 with or without backward selection, 4% and 2% for UFMLR20 with and without backward selection, respectively (Figure 1). UFMLR20 without backward selection detected none of the associated covariates in 18% of the simulated datasets. Overall, the FPR was below or equal to the nominal type I error with all methods except for LASSO-max and UFMLR20 with backward selection. UFLMR05 without backward selection was the most conservative with a FPR of 2% (Table 2 and Figure 2). RF, BRT, LASSO methods and UFMLR20 with backward selection suffered from an increase of type

Table 1 Significant associations with RF, BRT, LASSO and UFMLR

	RF	BRT	LASSO-max	LASSO-se	UFMLR05	UFMLR05 backward	UFMLR20	UFMLR20 backward
Covariates								
Pre-seasonal HAI titer (log)	0.004	0.003	0.001	0.001	0.001	0.002	0.002	0.001
History of asthma[b]	0.009	0.012	0.001	0.001	0.009	0.019		0.004
Professional activity involves contact with ill people[b]	0.009	0.023	0.008				0.001	0.001
Age (years)		0.004	0.001	0.001		0.001		
Daily frequency of hand washing (with soap or hand sanitizer) ≥ 5[b]	0.014	0.023	0.019					0.008
Always or often covers mouth while coughing or sneezing[b]	0.033	0.023	0.003	0.001				
"Craftsman, shopkeeper, chief executive officer" (socio-professional group)[b]		0.036	0.006				0.047	
History of chemotherapy[b]	0.048		0.021				0.027	
Average living room temperature (°C)			0.001	0.001		0.031	0.047	0.002
Presence of a dishwasher in the kitchen[b]			0.002	0.001	0.048	0.007		0.007
Sex = male[b]	0.048							
Professional activity is primarily outdoors[b]	0.045							
Age < 15 years[b]	0.016							
Any respiratory disease[b]	0.030							
Number of children (<15 years) in the household (n)	0.004							
Body mass index (kg/m2)		0.032						
Proportion of inhabitants > 15 years without a diploma (in IRIS zone[i])		0.024						
Proportion of habitations rented by inhabitants (in IRIS zone[i])		0.023						
Proportion of habitations owned by inhabitants (in IRIS zone[i])		0.049						
Habitation = house[b]		0.016						
Duration of contacts with subjects aged between 60 and 99 years (log (min))			0.006				0.023	0.019
Number of subjects in the household (n)			0.002	0.002				
Kitchen surface area per subject (m2)			0.005	0.002				
Cardiac arrhythmia[b]			0.021					
History of radiotherapy[b]			0.016					
Daily consumption of green tea (n)			0.040					
Number of birds inside habitation (n)			0.030					
Number of rooms per subject in habitation (n)			0.009					
Number of children in the bedroom (n)			0.010					
Bedroom windows face: garden[b]			0.010					
Duration of contacts at home (log (min))			0.029					
Longitude of the habitation (degrees)			0.027					
Latitude of the habitation (degrees)			0.028					
Proportion of "farmer, primary sector" (socio-professional group) (in IRIS zone[i])			0.004					0.005
Kitchen filtration of area[b]							0.021	0.034
Tiles flooring in the kitchen[b]							0.015	0.029
Agricultural land near habitation[b]							0.048	

[b]binary covariate. [i]IRIS zones are statistical block groups of about 2000 inhabitants defined by the French Institut National de la Statistique et des Etudes Economiques (INSEE).

Table 2 Performances of RF, BRT, LASSO and UFMLR in the 500 simulated datasets

	n	RF	BRT	LASSO-max	LASSO-se	UFMLR05	UFMLR05 backward	UFMLR20	UFMLR20 backward
Type I error 5%									
True Positive Rates (TPR)	8	85% (55% - 100%)	80% (51% - 100%)	78% (52% - 100%)	71% (41% - 100%)	28% (3% - 54%)	45% (20% - 70%)	26% (0% - 65%)	49% (15% - 84%)
Covariates with pairwise interaction	4	86% (49% - 100%)	80% (41% - 100%)	77% (40% - 100%)	69% (26% - 100%)	24% (0% - 63%)	41% (0% - 83%)	24% (0% - 72%)	46% (0% - 96%)
Covariates without pairwise interaction	4	84% (46% - 100%)	79% (41% - 100%)	79% (41% - 100%)	73% (32% - 100%)	32% (0% - 74%)	50% (6% - 93%)	28% (0% - 76%)	53% (5% - 100%)
Continuous covariates	4	90% (55% - 100%)	82% (47% - 100%)	82% (49% - 100%)	77% (41% - 100%)	35% (0% - 74%)	49% (15% - 84%)	29% (0% - 76%)	55% (14% - 95%)
Binary covariates	4	80% (34% - 100%)	78% (35% - 100%)	74% (34% - 100%)	64% (20% - 100%)	22% (0% - 57%)	41% (8% - 74%)	23% (0% - 69%)	44% (0% - 90%)
False Positive Rates (FPR)	292	4% (1% - 6%)	4% (2% - 6%)	9% (0% - 17%)	4% (0% - 9%)	2% (1% - 4%)	3% (1% - 5%)	4% (0% - 12%)	9% (0% - 18%)
Covariates correlated with associated covariates	12	46% (7% - 85%)	33% (3% - 63%)	23% (0% - 48%)	18% (0% - 41%)	4% (0% - 15%)	7% (0% - 22%)	9% (0% - 39%)	19% (0% - 60%)
Covariates uncorrelated with associated covariates	280	2% (0% - 4%)	3% (1% - 5%)	8% (0% - 17%)	3% (0% - 8%)	2% (1% - 4%)	3% (1% - 4%)	4% (0% - 11%)	8% (0% - 16%)
Continuous covariates	146	3% (0% - 6%)	3% (1% - 6%)	9% (0% - 18%)	4% (0% - 9%)	2% (0% - 5%)	3% (0% - 6%)	5% (0% - 13%)	9% (0% - 18%)
Binary covariates	146	4% (0% - 8%)	5% (2% - 9%)	9% (0% - 18%)	3% (0% - 9%)	2% (0% - 5%)	3% (0% - 5%)	4% (0% - 12%)	9% (0% - 18%)

Performances are shown as: Mean (95% confidence interval).

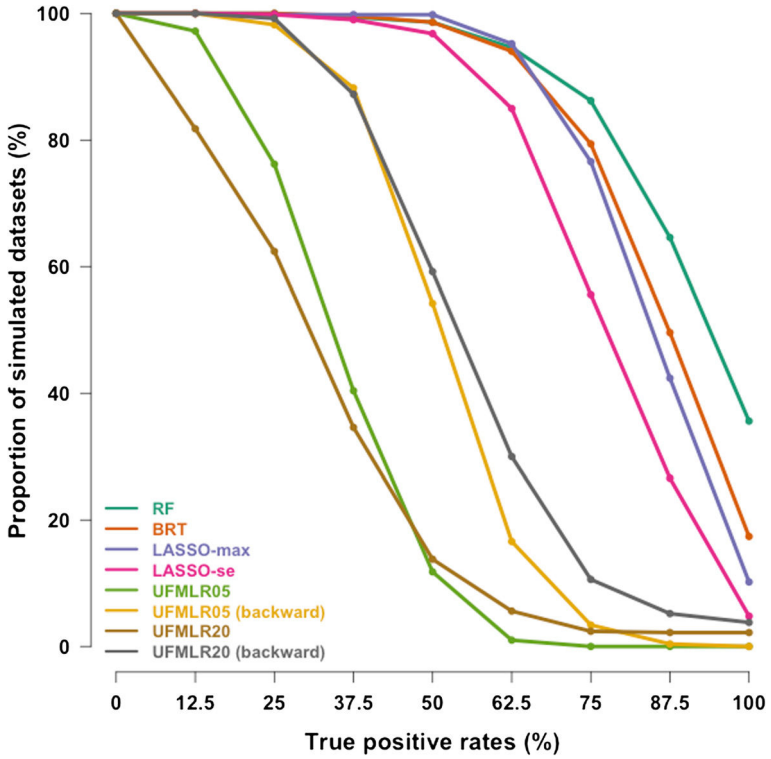


Figure 1 Cumulative distribution curves of the True Positive Rates in the 500 simulated datasets. y-axis shows the proportion of simulated datasets with True Positive Rates above or equal to the True Positive Rates on the x-axis.

I error when the non-associated covariates were correlated with associated covariates.

When increasing the number of non-associated covariates correlated with associated covariates, the TPR decreased with all methods, except for UFMLR20 with backward selection. The FPR were close to the FPR observed with a lower number of non-associated correlated covariates, with the exception of UFMLR20 with backward selection (14%, see Additional file 1).

Discussion

Without any pre-specified hypotheses, Random Forests, Boosted regression trees and LASSO models identified 8 to 24 covariates independently associated with influenza infection, among which 23 were not detected by the "univariate followed by multivariate logistic regression" framework. On the other hand, when a Pvalue threshold of 0.20 was applied to select covariates for multivariate logistic regression during univariate logistic models, a substantial number of spurious independent associations were detected which were not retrieved by any other methods. Simulations showed that RF, BRT and LASSO outperformed the conventional logistic framework to detect independent associations, while the false positive detection rates remained at the nominal significance level (RF, BRT and LASSO-se) or moderately increased above it (LASSO-max).

When covariates not associated with the outcome were correlated with covariates associated with the outcome, the false positive rate was high, particularly with RF. For this method, this finding was explained by the sensitivity of the Gini impurity criterion to between-covariates correlation [38]. More strikingly, increasing the number of correlated covariates also affected the true positive rate, which decreased with almost all methods (see Additional file 1). This finding may be attributed to a decrease of covariate strength of association due to a large number of correlated covariates and consequently, a decrease to be detected by any of the methods, as was shown in RF and LASSO [39].

In this work, we used an exploratory approach to analyze a large epidemiologic dataset, *i.e.* we aimed to detect associations between numerous covariates and an outcome, without pre-specified hypothesis. Despite the high number of covariates under study, the multiple testing issue was not considered. It is common to distinguish between two type of error rates: the comparisonwise error rate (CER) which corresponds to the probability, for an individual test, to reject the null hypothesis when it is actually true; and the experimentwise error rate (EER - also known as the familywise error rate), which corresponds to

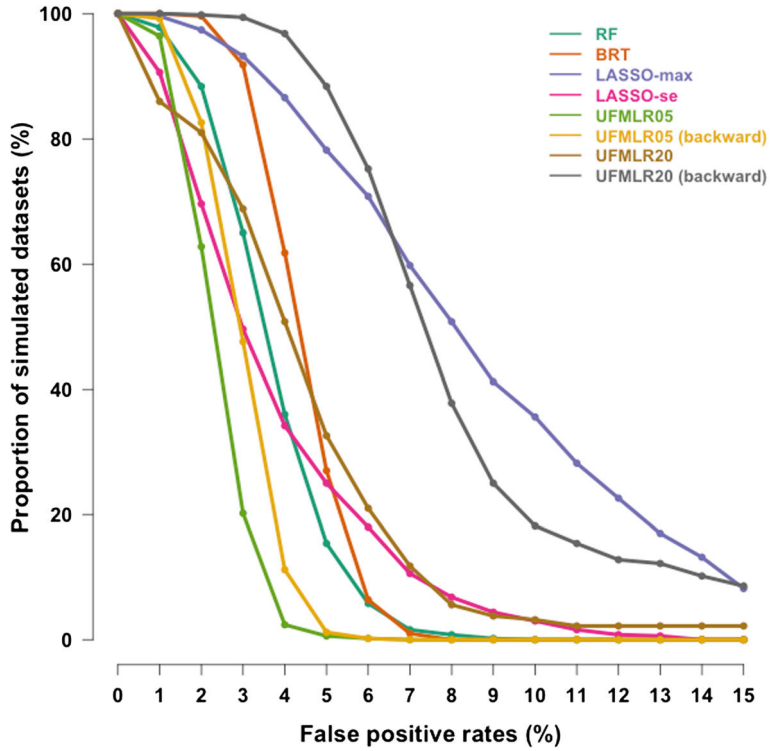


Figure 2 Cumulative distribution curves of the False Positive Rates in the 500 simulated datasets. y-axis shows the proportion of simulated datasets with False Positive Rates above or equal to the False Positive Rates on the x-axis.

the probability of rejecting at least one true null hypothesis among the multiple tests performed [40]. According to the simulations performed, we observed that the false positive rates associated to the permutation test was close to the expected CER level (5%) with almost all methods. Working at the individual covariate level, no adjustment was necessary. Adjusting Pvalues would have been required if the EER had to be controlled, *e.g.* in order to build a predictive model or to confirm the detected associations [41]. It is nevertheless essential to keep in mind that the significant results correspond to exploratory results, which require further confirmation.

To our knowledge, our study is the first to compare the performance in terms of associations detection of the random forest and boosted regression trees importance measures to the LASSO and the widely used analytic framework in the simultaneous analysis of hundreds to thousands of covariates to detect independent associations, a growing issue in epidemiology. Although such datasets offer analytic challenges, they are hardly comparable to datasets explored in omic-based approaches, in which the number of covariates (up to millions) is far higher to the number of samples, and for which the use of dedicated approaches, *e.g.* the elastic net penalty [42], would have been unavoidable.

Some associations with influenza infection detected with RF, BRT or LASSO-se methods were expected: HAI titers are well-known correlates of protection against influenza infection [43], young age is a known risk factor for H1N1pdm influenza infection [44], non-pharmaceutical preventive measures such as handwashing have been found to be determinants of H1N1pdm infection [45], and asthma was also reported as a specific risk factor [46]. Having a professional activity involving contact with ill people sounds logical as a potential risk factor, and several reports have shown that hospital staff were at increased risk of infection [47]. For other associations, *e.g.* "Always or often covers mouth while coughing or sneezing", we did not find consistent findings in the literature and it could be hard to hypothesize how the detected covariates could be linked with the risk of H1N1pdm influenza infection. However, "Professional activity involves contact with ill people" and age were correlated with this covariate ($\rho = 0.10$, Pvalue = 0.020 and $\rho = 0.29$, Pvalue < 0.001, respectively); based on our simulation findings we suspect that this association, as many others (*e.g.* "Presence of a dishwasher in the kitchen"), are likely to be false positives.

Having no prior knowledge about the covariates truly associated with influenza infections we performed a simulation study to assess the performances of the

different methods at detecting true and false associations in similar sized data, with a similar number of positive outcomes and covariates. Although we did not perform an extensive analysis exploring varying proportions of associated covariates or interactions between covariates, our simulations clearly demonstrated that UFMLR, with or without backward selection, were inefficient. We developed permutation tests to assess the significance of the covariates association with the outcome in RF, BRT and LASSO; their results with UFMLR were comparable to that of the Wald test in terms of nominal coverage (see Additional file 1). Although permutation tests exhibited slightly less power than the Wald test, this did not modify our general findings.

Conclusions

The conventional multivariate logistic regression framework is obviously not adapted for exploratory analysis of large epidemiologic datasets in view of detecting independent associations without any pre-specified hypothesis. In this respect, data mining methods and LASSO should be considered as credible alternatives to multivariate logistic regression.

Additional file

Additional file 1: Simulation process description and additional simulation results. Description of the simulation process; TPR and FPR at 1% nominal significance level; TPR and FPR with an increased number of correlated covariates; and comparisons of UFMLR TPR and FPR with Wald test and permutation test.

Abbreviations

LASSO: Least absolute shrinkage and selection operator; H1N1pdm influenza: H1N1 pandemic influenza; HAI: Haemagglutination inhibition; RF: Random forests; BRT: Boosted regression trees; LR: Logistic regression; UFMLR: Univariate followed by multivariate logistic regression; TPR: True positive rate; FPR: False positive rate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YM and FC conceived and designed the experiments. YM performed the experiments and analyzed the data. YM and FC wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank Rosemary M. Delabre for critical review of the manuscript. This work was supported by the Ministry of Higher Education and Research of France.

Author details

[1]INSERM, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F-75013 Paris, France. [2]Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F-75013 Paris, France. [3]Public Health Unit, Saint-Antoine Hospital, 75012 Paris, France.

Received: 24 April 2014 Accepted: 14 August 2014
Published: 26 August 2014

References

1. Trelles O, Prins P, Snir M, Jansen RC: **Big data, but are we ready?** *Nat Rev Genet* 2011, **12:**224.
2. Fontana JM, Alexander E, Salvatore M: **Translational research in infectious disease: current paradigms and challenges ahead.** *Transl Res* 2012, **159:**430–453.
3. Shah NH, Tenenbaum JD: **The coming age of data-driven medicine: translational bioinformatics' next frontier.** *J Am Med Informatics Assoc* 2012, **19:**e2–e4.
4. Bougnères P, Valleron A-J: **Causes of early-onset type 1 diabetes: toward data-driven environmental approaches.** *J Exp Med* 2008, **205:**2953–2957.
5. Choi H, Pavelka N: **When one and one gives more than two: challenges and opportunities of integrative omics.** *Front Genet* 2011, **2:**105.
6. Murdoch TB, Detsky AS: **The inevitable application of big data to health care.** *JAMA* 2013, **309:**1351–1352.
7. Liao H, Lynn HS: **A survey of variable selection methods in two Chinese epidemiology journals.** *BMC Med Res Methodol* 2010, **10:**87.
8. Walter S, Tiemeier H: **Variable selection: current practice in epidemiological studies.** *Eur J Epidemiol* 2009, **24:**733–736.
9. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: **A simulation study of the number of events per variable in logistic regression analysis.** *J Clin Epidemiol* 1996, **49:**1373–1379.
10. Smyth P: **Data mining: data analysis on a grand scale?** *Stat Methods Med Res* 2000, **9:**309–327.
11. Maimon O, Rokach L (Eds): *Data Mining and Knowledge Discovery Handbook.* New York: Springer; 2010.
12. Austin PC: **A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality.** *Stat Med* 2007, **26:**2937–2957.
13. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, DE Mendonca A: **Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests.** *BMC Res Notes* 2011, **4:**299.
14. Green M, Björk J, Forberg J, Ekelund U, Edenbrandt L, Ohlsson M: **Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room.** *Artif Intell Med* 2006, **38:**305–318.
15. Regnier-Coudert O, McCall J, Lothian R, Lam T, McClinton S, N'dow J: **Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers.** *Artif Intell Med* 2012, **55:**25–35.
16. Austin PC, Lee DS, Steyerberg EW, Tu JV: **Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods?** *Biometrical J* 2012, **54:**657–673.
17. Austin PC, Tu JV, Ho JE, Levy D, Lee DS: **Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes.** *J Clin Epidemiol* 2013, **66:**398–407.
18. Tibshirani R: **Regression shrinkage and selection via the Lasso.** *J R Stat Soc Ser B* 1996, **58:**267–288.
19. Xu C-J, van der Schaaf A, Schilstra C, Langendijk JA, van't Veld AA: **Impact of statistical learning methods on the predictive power of multivariate normal tissue complication probability models.** *Int J Radiat Oncol Biol Phys* 2012, **82:**e677–e684.
20. Avalos M, Adroher ND, Lagarde E, Thiessard F, Grandvalet Y, Contrand B, Orriols L: **Prescription-drug-related risk in driving: comparing conventional and lasso shrinkage logistic regressions.** *Epidemiology* 2012, **23:**706–712.
21. Lapidus N, De Lamballerie X, Salez N, Setbon M, Ferrari P, Delabre RM, Gougeon M-L, Vely F, Leruez-Ville M, Andreoletti L, Cauchemez S, Boëlle P-Y, Vivier E, Abel L, Schwarzinger M, Legeas M, Le Cann P, Flahault A, Carrat F: **Integrative study of pandemic A/H1N1 influenza infections: design and methods of the CoPanFlu-France cohort.** *BMC Public Health* 2012, **12:**417.
22. **CDC protocol of realtime RTPCR for influenza A (H1N1).** [http://www.who.int/csr/resources/publications/swineflu/realtimeptpcr/en/]
23. Reijans M, Dingemans G, Klaassen CH, Meis JF, Keijdener J, Mulders B, Eadie K, van Leeuwen W, van Belkum A, Horrevorts AM, Simons G: **RespiFinder: a new multiparameter test to differentially identify fifteen respiratory viruses.** *J Clin Microbiol* 2008, **46:**1232–1240.

24. **European Medicines Agency - Committee for proprietary medicinal products. Note for guidance on harmonization of requirements for influenza vaccines (CPMP/BWP/214/96).** [http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003945.pdf]
25. Lapidus N, de Lamballerie X, Salez N, Setbon M, Delabre RM, Ferrari P, Moyen N, Gougeon M-L, Vely F, Leruez-Ville M, Andreoletti L, Cauchemez S, Boëlle P-Y, Vivier E, Abel L, Schwarzinger M, Legeas M, Le Cann P, Flahault A, Carrat F: **Factors associated with post-seasonal serological titer and risk factors for infection with the pandemic A/H1N1 virus in the French general population.** *PLoS One* 2013, **8:**e60127.
26. Breiman L: **Random Forests.** *Mach Learn* 2001, **45:**123–140.
27. Friedman JH: **Greedy function approximation: a gradient boosting machine.** *North* 2001, **29:**1189–1232.
28. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition. New York: Springer; 2009.
29. Friedman JH: **Stochastic gradient boosting.** *Comput Stat Data Anal* 2002, **38:**367–378.
30. McCullagh P, Nelder JA: *Generalized Linear Models.* 2nd edition. London: Chapman and Hall/CRC; 1989.
31. **glmnet: Lasso and elastic-net regularized generalized linear models.** [http://cran.r-project.org/web/packages/glmnet/]
32. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J Stat Softw* 2010, **33:**1–22.
33. Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R: **Bootstrap Methods and Permutation Tests.** In *Introd to Pract Stat. Volume 5.* Edited by Moore D, McCabe G. New York: WH Freeman & Co; 2005.
34. Altmann A, Toloşi L, Sander O, Lengauer T: **Permutation importance: a corrected feature importance measure.** *Bioinformatics* 2010, **26:**1340–1347.
35. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: detecting and evaluating dependencies between variables.** *Bioinformatics* 2002, **18**(Suppl 2):S231–S240.
36. Liaw A, Wiener M: **Classification and regression by randomForest.** *R News* 2002, **2/3:**18–22.
37. Ridgeway G: **Generalized boosted models: a guide to the gbm package.** *Compute* 2007, **1:**1–12.
38. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SAFT: **Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?** *Brief Bioinform* 2013, **14:**315–326.
39. Tolosi L, Lengauer T: **Classification with correlated features: unreliability of feature ranking and solutions.** *Bioinformatics* 2011, **27:**1986–1994.
40. Bender R, Lange S: **Adjusting for multiple testing–when and how?** *J Clin Epidemiol* 2001, **54:**343–349.
41. Bender R, Lange S: **Multiple test procedures other than Bonferroni's deserve wider use.** *BMJ* 1999, **318:**600–601.
42. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Stat Soc - Ser B Stat Methodol* 2005, **67:**301–320 [Series B (Statistical Methodology)].
43. Ng S, Fang VJ, Ip DKM, Chan K-H, Leung GM, Peiris JSM, Cowling BJ: **Estimation of the association between antibody titers and protection against confirmed influenza virus infection in children.** *J Infect Dis* 2013, **208:**1320–1324.
44. Riley S, Kwok KO, Wu KM, Ning DY, Cowling BJ, Wu JT, Ho L-M, Tsang T, Lo S-V, Chu DKW, Ma ESK, Peiris JSM: **Epidemiological characteristics of 2009 (H1N1) pandemic influenza based on paired sera from a longitudinal community cohort study.** *PLoS Med* 2011, **8:**e1000442.
45. Simmerman JM, Suntarattiwong P, Levy J, Jarman RG, Kaewchana S, Gibbons RV, Cowling BJ, Sanasuttipun W, Maloney SA, Uyeki TM, Kamimoto L, Chotipitayasunondh T: **Findings from a household randomized controlled trial of hand washing and face masks to reduce influenza transmission in Bangkok, Thailand.** *Influenza Other Respi Viruses* 2011, **5:**256–267.
46. Kloepfer KM, Olenec JP, Lee WM, Liu G, Vrtis RF, Roberg KA, Evans MD, Gangnon RE, Lemanske RF, Gern JE: **Increased H1N1 infection rate in children with asthma.** *Am J Respir Crit Care Med* 2012, **185:**1275–1279.
47. Chen MIC, Lee VJM, Barr I, Lin C, Goh R, Lee C, Singh B, Tan J, Lim WY, Cook AR, Ang B, Chow A, Tan BH, Loh J, Shaw R, Chia KS, Lin RTP, Leo YS: **Risk factors for pandemic (H1N1) 2009 virus seroconversion among hospital staff, Singapore.** *Emerg Infect Dis* 2010, **16:**1554–1561.

doi:10.1186/1471-2288-14-99
Cite this article as: Mansiaux and Carrat: Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. *BMC Medical Research Methodology* 2014 **14**:99.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Additional file 1

Simulated data

Five hundred datasets were sampled from a multivariate normal distribution with standard normal marginal (the simulation process is summarized in Table 1). We simulated a binary outcome using the following logit model:

$$\log\left(\frac{p_i}{1-p_i}\right) =$$

$$\mu + \sum_{j=1}^{C} \beta cont_j c_{i,j} + \sum_{k=1}^{B} \beta bin_k b_{i,k} + \sum_{l=1}^{C}\sum_{m=1}^{C} \eta cont_{lm} c_{i,l} c_{i,m} + \sum_{n=1}^{B}\sum_{o=1}^{B} \eta bin_{no} b_{i,n} b_{i,o} + \varepsilon$$

with C (C=150) the number of continuous covariates *c* and B (B=150) the number of binary covariates *b*. To mimic binary covariates, sampled continuous covariates were set equal to 0 or 1 depending on whether the sampled values were lower or higher than the median of the standard normal distribution. The interaction parameters $\eta cont_{kl}$ and ηbin_{kl}were set equal to 0, except for the special cases $\eta cont_{12}$= ηbin_{12}= 0.5. The sample size was 500 and the mean proportion of outcome in the 500 simulated datasets was 14% (95% CI = [11%;18%]).

Two scenarios were simulated, in scenario 1 (whose results are detailed in the manuscript) twenty covariates (the 8 associated covariates and 12 non-associated covariates - 6 continuous and 6 binary) were correlated (ρ=0.5). In scenario 2, we increased the number of correlated covariates to 58 (the 8 associated covariates and 50 non-associated covariates).

Tables

Table 1 Summary of the simulation process

Scenario	Scenario 1	Scenario 2
Number of continuous covariates (C)	150	150
Number of binary covariates (B)	150	150
Number of covariates with direct effects	8	8
Number of covariates with interaction effects	4	4
Number of non-associated covariates	292	292
Intercept	$\mu = 1.5$	$\mu = 1.5$
Direct effects	$\beta cont_i = \beta bin_i$ $= \begin{cases} 0.5 \; if \; i \in \{1:4\} \\ 0 \; otherwise \end{cases}$	$\beta cont_i = \beta bin_i$ $= \begin{cases} 0.5 \; if \; i \in \{1:4\} \\ 0 \; otherwise \end{cases}$
Gaussian noise	$\varepsilon \sim \mathcal{N}(0,1)$	$\varepsilon \sim \mathcal{N}(0,1)$
Covariances	$Cov(x_{i,j}, x_{i,k}) =$ $\begin{cases} 1 \; if \; i = j \\ 0.5 \; if \; i,j \in \{1:10\} \\ 0 \; otherwise \end{cases}$ with x either a binary or a continuous covariate	$Cov(x_{i,j}, x_{i,k}) =$ $\begin{cases} 1 \; if \; i = j \\ 0.5 \; if \; i,j \in \{1:29\} \\ 0 \; otherwise \end{cases}$ with x either a binary or a continuous covariate

Table 2. Performances of RF, BRT, LASSO and UFMLR in the 500 simulated datasets (Type I error 1%) – Scenario 1. Performances are shown as: Mean (95% confidence interval)

	n	RF	BRT	LASSO-max	LASSO-se	UFMLR05	UFMLR05 backward	UFMLR20	UFMLR20 backward
Type I error 1%									
True Positive Rates (TPR)	8	58% (17% - 99%)	53% (20% - 86%)	72% (44% - 100%)	71% (41% - 100%)	11% (0% - 31%)	29% (6% - 53%)	11% (0% - 44%)	38% (1% - 75%)
Covariates with pairwise interaction	4	60% (9% - 100%)	54% (7% - 100%)	71% (30% - 100%)	69% (25% - 100%)	9% (0% - 36%)	25% (0% - 61%)	10% (0% - 46%)	35% (0% - 84%)
Covariates without pairwise interaction	4	56% (8% - 100%)	53% (9% - 96%)	74% (34% - 100%)	73% (32% - 100%)	13% (0% - 44%)	34% (0% - 75%)	12% (0% - 52%)	42% (0% - 90%)
Continuous covariates	4	71% (14% - 100%)	66% (24% - 100%)	79% (44% - 100%)	77% (41% - 100%)	15% (0% - 46%)	34% (0% - 69%)	12% (0% - 51%)	43% (0% - 86%)
Binary covariates	4	45% (0% - 96%)	40% (0% - 89%)	66% (23% - 100%)	64% (20% - 100%)	7% (0% - 31%)	25% (0% - 59%)	9% (0% - 46%)	33% (0% - 80%)
False Positive Rates (FPR)	292	1% (0% - 2%)	1% (0% - 2%)	4% (0% - 9%)	3% (0% - 8%)	1% (0% - 2%)	1% (0% - 3%)	2% (0% - 9%)	5% (0% - 15%)
Covariates correlated with associated covariates	12	18% (0% - 45%)	12% (0% - 31%)	20% (0% - 43%)	18% (0% - 41%)	1% (0% - 5%)	3% (0% - 13%)	3% (0% - 31%)	13% (0% - 52%)
Covariates uncorrelated with associated covariates	280	0% (0% - 1%)	1% (0% - 2%)	3% (0% - 8%)	3% (0% - 8%)	1% (0% - 2%)	1% (0% - 2%)	1% (0% - 8%)	5% (0% - 14%)
Continuous covariates	146	1% (0% - 3%)	1% (0% - 2%)	4% (0% - 10%)	4% (0% - 9%)	1% (0% - 2%)	1% (0% - 3%)	2% (0% - 9%)	5% (0% - 16%)
Binary covariates	146	1% (0% - 3%)	1% (0% - 4%)	4% (0% - 9%)	3% (0% - 8%)	1% (0% - 2%)	1% (0% - 3%)	1% (0% - 9%)	5% (0% - 16%)

Table 3 Performances of RF, BRT, LASSO and UFMLR in the 500 simulated datasets (Type I error 5%) – Scenario 2. Performances are shown as: Mean (95% confidence interval)

	n	RF	BRT	LASSO-max	LASSO-se	UFMLR05	UFMLR05 backward	UFMLR20	UFMLR20 backward
Type I error 5%									
True Positive Rates (TPR)	8	68% (30% - 100%)	72% (40% - 100%)	70% (41% - 98%)	63% (33% - 93%)	27% (0% - 56%)	40% (12% - 67%)	16% (0% - 70%)	56% (10% - 100%)
Covariates with pairwise interaction	4	60% (12% - 100%)	67% (22% - 100%)	67% (25% - 100%)	60% (13% - 100%)	23% (0% - 64%)	35% (0% - 77%)	15% (0% - 71%)	54% (0% - 100%)
Covariates without pairwise interaction	4	77% (31% - 100%)	77% (35% - 100%)	72% (32% - 100%)	66% (26% - 100%)	31% (0% - 74%)	44% (1% - 87%)	18% (0% - 75%)	58% (4% - 100%)
Continuous covariates	4	63% (13% - 100%)	70% (29% - 100%)	73% (36% - 100%)	68% (29% - 100%)	33% (0% - 75%)	45% (6% - 83%)	18% (0% - 76%)	60% (8% - 100%)
Binary covariates	4	74% (23% - 100%)	74% (27% - 100%)	66% (24% - 100%)	58% (15% - 100%)	20% (0% - 58%)	34% (0% - 72%)	14% (0% - 68%)	52% (0% - 100%)
False Positive Rates (FPR)	292	5% (1% - 10%)	6% (4% - 9%)	9% (1% - 17%)	4% (0% - 9%)	2% (0% - 5%)	3% (1% - 5%)	4% (0% - 20%)	14% (0% - 31%)
Covariates correlated with associated covariates	50	26% (1% - 50%)	25% (10% - 41%)	16% (5% - 26%)	12% (3% - 21%)	4% (0% - 11%)	6% (0% - 13%)	9% (0% - 56%)	33% (0% - 84%)
Covariates uncorrelated with associated covariates	242	1% (0% - 2%)	3% (1% - 5%)	8% (0% - 16%)	3% (0% - 8%)	2% (0% - 4%)	3% (1% - 5%)	3% (0% - 13%)	11% (1% - 20%)
Continuous covariates	146	4% (0% - 9%)	5% (2% - 8%)	9% (1% - 18%)	5% (0% - 10%)	3% (0% - 6%)	4% (1% - 7%)	4% (0% - 21%)	15% (0% - 32%)
Binary covariates	146	6% (0% - 12%)	8% (3% - 13%)	9% (1% - 17%)	4% (0% - 9%)	2% (0% - 5%)	3% (0% - 6%)	4% (0% - 19%)	14% (0% - 30%)

Table 4. Performances of RF, BRT, LASSO and UFMLR in the 500 simulated datasets (Type I error 1%) – Scenario 2. Performances are shown as: Mean (95% confidence interval)

	n	RF	BRT	LASSO-max	LASSO-se	UFMLR05	UFMLR05 backward	UFMLR20	UFMLR20 backward
Type I error 1%									
True Positive Rates (TPR)	8	40% (2% - 78%)	53% (20% - 87%)	65% (35% - 94%)	62% (32% - 92%)	10% (0% - 32%)	27% (2% - 51%)	10% (0% - 60%)	44% (0% - 99%)
Covariates with pairwise interaction	4	28% (0% - 73%)	44% (0% - 93%)	60% (15% - 100%)	59% (12% - 100%)	7% (0% - 32%)	22% (0% - 59%)	9% (0% - 60%)	42% (0% - 100%)
Covariates without pairwise interaction	4	52% (0% - 100%)	63% (17% - 100%)	69% (29% - 100%)	66% (25% - 100%)	14% (0% - 47%)	31% (0% - 70%)	11% (0% - 62%)	47% (0% - 100%)
Continuous covariates	4	36% (0% - 81%)	55% (13% - 97%)	71% (33% - 100%)	68% (29% - 100%)	14% (0% - 48%)	30% (0% - 65%)	11% (0% - 63%)	48% (0% - 100%)
Binary covariates	4	44% (0% - 100%)	51% (1% - 100%)	58% (14% - 100%)	56% (13% - 100%)	6% (0% - 30%)	23% (0% - 56%)	9% (0% - 59%)	41% (0% - 100%)
False Positive Rates (FPR)	292	2% (0% - 4%)	2% (0% - 4%)	5% (0% - 10%)	4% (0% - 9%)	1% (0% - 2%)	1% (0% - 3%)	3% (0% - 19%)	11% (0% - 30%)
Covariates correlated with associated covariates	50	11% (0% - 25%)	10% (0% - 20%)	13% (4% - 23%)	12% (3% - 21%)	1% (0% - 4%)	2% (0% - 7%)	7% (0% - 54%)	25% (0% - 80%)
Covariates uncorrelated with associated covariates	242	0% (0% - 1%)	0% (0% - 1%)	4% (0% - 9%)	3% (0% - 7%)	1% (0% - 2%)	1% (0% - 3%)	2% (0% - 12%)	8% (0% - 19%)
Continuous covariates	146	2% (0% - 5%)	2% (0% - 4%)	6% (0% - 11%)	4% (0% - 10%)	1% (0% - 2%)	2% (0% - 4%)	3% (0% - 20%)	11% (0% - 31%)
Binary covariates	146	2% (0% - 6%)	2% (0% - 5%)	5% (0% - 10%)	4% (0% - 9%)	1% (0% - 2%)	1% (0% - 3%)	3% (0% - 18%)	10% (0% - 29%)

Table 5 Performances of UFMLR in the 500 simulated datasets (Type I error 5%) – Scenario 1. [w] Wald test Pvalue. [p] permutation test Pvalue. Performances are shown as: Mean (95% confidence interval)

	n	UFMLR05 [w]	UFMLR05 [p]	UFMLR05 backward [w]	UFMLR05 backward [p]	UFMLR20 [w]	UFMLR20 [p]	UFMLR20 backward [w]	UFMLR20 backward [p]
Type I error 5%									
True Positive Rates (TPR)	8	31% (5% - 58%)	28% (3% - 54%)	47% (22% - 72%)	45% (20% - 70%)	32% (0% - 74%)	26% (0% - 65%)	51% (16% - 86%)	49% (15% - 84%)
Covariates with pairwise interaction	4	27% (0% - 66%)	24% (0% - 63%)	43% (0% - 86%)	41% (0% - 83%)	30% (0% - 81%)	24% (0% - 72%)	47% (0% - 97%)	46% (0% - 96%)
Covariates without pairwise interaction	4	36% (0% - 79%)	32% (0% - 74%)	51% (8% - 94%)	50% (6% - 93%)	34% (0% - 86%)	28% (0% - 76%)	55% (7% - 100%)	53% (5% - 100%)
Continuous covariates	4	38% (0% - 80%)	35% (0% - 74%)	51% (17% - 86%)	49% (15% - 84%)	35% (0% - 87%)	29% (0% - 76%)	56% (16% - 97%)	55% (14% - 95%)
Binary covariates	4	24% (0% - 61%)	22% (0% - 57%)	43% (9% - 76%)	41% (8% - 74%)	29% (0% - 78%)	23% (0% - 69%)	46% (0% - 92%)	44% (0% - 90%)
False Positive Rates (FPR)	292	3% (1% - 4%)	2% (1% - 4%)	3% (1% - 5%)	3% (1% - 5%)	6% (0% - 14%)	4% (0% - 12%)	9% (0% - 18%)	9% (0% - 18%)
Covariates correlated with associated covariates	12	4% (0% - 16%)	4% (0% - 15%)	8% (0% - 23%)	7% (0% - 22%)	12% (0% - 44%)	9% (0% - 39%)	20% (0% - 61%)	19% (0% - 60%)
Covariates uncorrelated with associated covariates	280	2% (1% - 4%)	2% (1% - 4%)	3% (1% - 5%)	3% (1% - 4%)	6% (0% - 13%)	4% (0% - 11%)	9% (1% - 16%)	8% (0% - 16%)
Continuous covariates	146	3% (0% - 5%)	2% (0% - 5%)	3% (0% - 6%)	3% (0% - 6%)	6% (0% - 15%)	5% (0% - 13%)	9% (0% - 19%)	9% (0% - 18%)
Binary covariates	146	2% (0% - 5%)	2% (0% - 5%)	3% (0% - 6%)	3% (0% - 5%)	6% (0% - 15%)	4% (0% - 12%)	9% (0% - 19%)	9% (0% - 18%)

Table 6 Performances of UFMLR in the 500 simulated datasets (Type I error 1%) – Scenario 1. [w] Wald test Pvalue. [p] permutation test Pvalue. Performances are shown as: Mean (95% confidence interval)

	n	UFMLR05 [w]	UFMLR05 [p]	UFMLR05 backward [w]	UFMLR05 backward [p]	UFMLR20 [w]	UFMLR20 [p]	UFMLR20 backward [w]	UFMLR20 backward [p]
Type I error 1%									
True Positive Rates (TPR)	8	14% (0% - 36%)	11% (0% - 31%)	34% (10% - 58%)	29% (6% - 53%)	17% (0% - 54%)	11% (0% - 44%)	43% (7% - 78%)	38% (1% - 75%)
Covariates with pairwise interaction	4	11% (0% - 40%)	9% (0% - 36%)	29% (0% - 66%)	25% (0% - 61%)	16% (0% - 59%)	10% (0% - 46%)	39% (0% - 89%)	35% (0% - 84%)
Covariates without pairwise interaction	4	17% (0% - 50%)	13% (0% - 44%)	39% (0% - 81%)	34% (0% - 75%)	19% (0% - 63%)	12% (0% - 52%)	46% (0% - 95%)	42% (0% - 90%)
Continuous covariates	4	18% (0% - 53%)	15% (0% - 46%)	39% (5% - 73%)	34% (0% - 69%)	19% (0% - 63%)	12% (0% - 51%)	48% (8% - 88%)	43% (0% - 86%)
Binary covariates	4	9% (0% - 37%)	7% (0% - 31%)	29% (0% - 61%)	25% (0% - 59%)	15% (0% - 57%)	9% (0% - 46%)	37% (0% - 84%)	33% (0% - 80%)
False Positive Rates (FPR)	292	1% (0% - 2%)	1% (0% - 2%)	1% (0% - 3%)	1% (0% - 3%)	3% (0% - 10%)	2% (0% - 9%)	6% (0% - 16%)	5% (0% - 15%)
Covariates correlated with associated covariates	12	1% (0% - 6%)	1% (0% - 5%)	4% (0% - 14%)	3% (0% - 13%)	5% (0% - 33%)	3% (0% - 31%)	15% (0% - 55%)	13% (0% - 52%)
Covariates uncorrelated with associated covariates	280	1% (0% - 2%)	1% (0% - 2%)	1% (0% - 3%)	1% (0% - 2%)	3% (0% - 9%)	1% (0% - 8%)	6% (0% - 15%)	5% (0% - 14%)
Continuous covariates	146	1% (0% - 3%)	1% (0% - 2%)	2% (0% - 4%)	1% (0% - 3%)	3% (0% - 11%)	2% (0% - 9%)	6% (0% - 17%)	5% (0% - 16%)
Binary covariates	146	1% (0% - 2%)	1% (0% - 2%)	1% (0% - 3%)	1% (0% - 3%)	3% (0% - 10%)	1% (0% - 9%)	6% (0% - 17%)	5% (0% - 16%)

Table 7 Performances of UFMLR in the 500 simulated datasets (Type I error 5%) – Scenario 2. [w] Wald test Pvalue. [p] permutation test Pvalue. Performances are shown as: Mean (95% confidence interval)

	n	UFMLR05 [w]	UFMLR05 [p]	UFMLR05 backward [w]	UFMLR05 backward [p]	UFMLR20 [w]	UFMLR20 [p]	UFMLR20 backward [w]	UFMLR20 backward [p]
Type I error 5%									
True Positive Rates (TPR)	8	33% (3% - 63%)	27% (0% - 56%)	41% (14% - 69%)	40% (12% - 67%)	20% (0% - 79%)	16% (0% - 70%)	57% (12% - 100%)	56% (10% - 100%)
Covariates with pairwise interaction	4	30% (0% - 73%)	23% (0% - 64%)	38% (0% - 81%)	35% (0% - 77%)	19% (0% - 80%)	15% (0% - 71%)	55% (0% - 100%)	54% (0% - 100%)
Covariates without pairwise interaction	4	36% (0% - 80%)	31% (0% - 74%)	45% (2% - 89%)	44% (1% - 87%)	22% (0% - 85%)	18% (0% - 75%)	59% (6% - 100%)	58% (4% - 100%)
Continuous covariates	4	41% (0% - 85%)	33% (0% - 75%)	47% (8% - 86%)	45% (6% - 83%)	23% (0% - 87%)	18% (0% - 76%)	61% (9% - 100%)	60% (8% - 100%)
Binary covariates	4	25% (0% - 66%)	20% (0% - 58%)	36% (0% - 74%)	34% (0% - 72%)	18% (0% - 76%)	14% (0% - 68%)	53% (0% - 100%)	52% (0% - 100%)
False Positive Rates (FPR)	292	3% (1% - 6%)	2% (0% - 5%)	4% (1% - 6%)	3% (1% - 5%)	5% (0% - 22%)	4% (0% - 20%)	15% (0% - 31%)	14% (0% - 31%)
Covariates correlated with associated covariates	50	6% (0% - 15%)	4% (0% - 11%)	7% (0% - 14%)	6% (0% - 13%)	12% (0% - 59%)	9% (0% - 56%)	34% (0% - 85%)	33% (0% - 84%)
Covariates uncorrelated with associated covariates	242	3% (1% - 5%)	2% (0% - 4%)	3% (1% - 5%)	3% (1% - 5%)	4% (0% - 15%)	3% (0% - 13%)	11% (2% - 20%)	11% (1% - 20%)
Continuous covariates	146	4% (0% - 7%)	3% (0% - 6%)	4% (1% - 7%)	4% (1% - 7%)	6% (0% - 23%)	4% (0% - 21%)	16% (0% - 32%)	15% (0% - 32%)
Binary covariates	146	3% (0% - 6%)	2% (0% - 5%)	3% (0% - 6%)	3% (0% - 6%)	5% (0% - 21%)	4% (0% - 19%)	14% (0% - 30%)	14% (0% - 30%)

Table 8 Performances of UFMLR in the 500 simulated datasets (Type I error 1%) – Scenario 2. [w] Wald test Pvalue. [p] permutation test Pvalue. Performances are shown as: Mean (95% confidence interval)

	n	UFMLR05 [w]	UFMLR05 [p]	UFMLR05 backward [w]	UFMLR05 backward [p]	UFMLR20 [w]	UFMLR20 [p]	UFMLR20 backward [w]	UFMLR20 backward [p]
Type I error 1%									
True Positive Rates (TPR)	8	16% (0% - 41%)	10% (0% - 32%)	30% (6% - 55%)	27% (2% - 51%)	14% (0% - 66%)	10% (0% - 60%)	51% (1% - 100%)	44% (0% - 99%)
Covariates with pairwise interaction	4	12% (0% - 44%)	7% (0% - 32%)	26% (0% - 64%)	22% (0% - 59%)	13% (0% - 67%)	9% (0% - 60%)	49% (0% - 100%)	42% (0% - 100%)
Covariates without pairwise interaction	4	20% (0% - 59%)	14% (0% - 47%)	35% (0% - 75%)	31% (0% - 70%)	15% (0% - 70%)	11% (0% - 62%)	53% (0% - 100%)	47% (0% - 100%)
Continuous covariates	4	22% (0% - 60%)	14% (0% - 48%)	35% (0% - 71%)	30% (0% - 65%)	16% (0% - 72%)	11% (0% - 63%)	55% (0% - 100%)	48% (0% - 100%)
Binary covariates	4	11% (0% - 41%)	6% (0% - 30%)	25% (0% - 61%)	23% (0% - 56%)	12% (0% - 64%)	9% (0% - 59%)	47% (0% - 100%)	41% (0% - 100%)
False Positive Rates (FPR)	292	1% (0% - 3%)	1% (0% - 2%)	2% (0% - 3%)	1% (0% - 3%)	3% (0% - 19%)	3% (0% - 19%)	13% (0% - 31%)	11% (0% - 30%)
Covariates correlated with associated covariates	50	1% (0% - 6%)	1% (0% - 4%)	3% (0% - 7%)	2% (0% - 7%)	8% (0% - 54%)	7% (0% - 54%)	29% (0% - 83%)	25% (0% - 80%)
Covariates uncorrelated with associated covariates	242	1% (0% - 3%)	1% (0% - 2%)	1% (0% - 3%)	1% (0% - 3%)	2% (0% - 12%)	2% (0% - 12%)	9% (0% - 20%)	8% (0% - 19%)
Continuous covariates	146	1% (0% - 4%)	1% (0% - 2%)	2% (0% - 4%)	2% (0% - 4%)	4% (0% - 20%)	3% (0% - 20%)	13% (0% - 32%)	11% (0% - 31%)
Binary covariates	146	1% (0% - 3%)	1% (0% - 2%)	2% (0% - 4%)	1% (0% - 3%)	3% (0% - 19%)	3% (0% - 18%)	12% (0% - 30%)	10% (0% - 29%)

5.3 Discussion

Cet article présente une analyse originale de comparaison de méthodes de *data mining* et de méthodes de régression, pénalisée ou conventionnelle, d'un point de vue de la détection d'associations indépendantes, une problématique de premier ordre en épidémiologie. Nous insistons sur la notion d'associations indépendantes, expliquant ainsi que les résultats des modèles de régression univariée ne sont pas présentés dans le manuscrit. Ce travail se démarque de certaines comparaisons "*data mining* contre régression logistique" déjà réalisées, dans le sens où celles-ci se basaient sur l'évaluation de la performance prédictive et non sur la capacité à détecter des associations [10–12, 38, 121, 211].

5.3.1 Détection d'associations indépendantes

La simulation nous a permis d'observer les qualités des méthodes de *data mining* pour la détection d'associations, tandis que la régression logistique conventionnelle a clairement montré ses limites. Les résultats de simulations nous permettent d'aborder ceux obtenus sur les données CoPanFlu avec un œil différent, celles-ci ayant été effectuées dans une configuration proche de celle de CoPanFlu d'un point de vue dimensionnel. Nous pouvons être relativement confiants dans les associations détectées par les méthodes de *data mining* et la régression LASSO (particulièrement pour les modèles LASSO-se), la plupart d'entre elles ayant déjà été rapportées dans la littérature (*e.g* l'âge ou le titre d'anticorps pré-épidémique). La détection de variables corrélées et *a priori* non associées semble également toucher les données CoPanFlu, suggérant une analyse approfondie de ces résultats.

Les mauvaises performances de la régression logistique en simulation, pour la détection de vrais positifs mais également de faux positifs, nous laisse imaginer que le même phénomène est susceptible d'être observé sur les données CoPanFlu. On peut penser que des vraies associations n'ont pas été détectées, 23 covariables détectées soit par le *data mining* soit par le LASSO, ne l'ont pas été par les modèles logistiques classiques. De plus l'inflation du risque de première espèce observée avec le modèle multivarié construit en utilisant un risque α de 0.20 pour les analyses univariées suggère qu'un certain nombre de fausses associations ont pu être capturées par ce modèle.

Nous avons remarqué que l'utilisation d'une procédure de sélection de variables *backward* était associée à la fois avec une augmentation du taux de vrais positifs mais également du taux de faux positifs. Ceci s'explique par le fait que la sélection univariée a conservé dans une majorité des 500 jeux de données un grand nombre de covariables, conduisant ainsi à des modèles multivariés contenant de nombreuses covariables simultanément. Un grand nombre de prédicteurs, couplé à des problèmes potentiels de multicolinéarité,web a conduit dans beaucoup de modèles sans sélection *backward* à ce qu'aucun ou peu de coefficients ne soient significatifs, expliquant les faibles taux de vrais et de faux

positifs dans les modèles sans sélection. *A contrario*, dans les modèles avec sélection *backward*, l'ensemble des covariables avaient des coefficients significatifs, des covariables associées avec la variable réponse et des covariables non associées ont alors pu être détectées.

Les résultats de simulation présentés ne sont le reflet que d'un cas particulier. Nous n'avons pas effectué d'analyse intensive de l'influence de paramètres tels que la force d'association entre les prédicteurs et la variable réponse, la proportion de variables continues/dichotomiques ou encore l'impact d'une optimisation des paramètres des forêts aléatoires ou des arbres de régression *boostés* (*e.g* le nombre d'arbres). Il est ainsi délicat d'affirmer que les méthodes de *data mining* et le LASSO sont universellement performantes. En effet, si l'augmentation du nombre de covariables corrélées n'a pas modifié les conclusions générales de l'article (résultats détaillés présentés en annexe de l'article), elle a toutefois permis d'observer une baisse généralisée des performances du *data mining* et du LASSO. Ces résultats invitent néanmoins à s'interroger sur la pertinence du paradigme d'analyse largement admis en épidémiologie et à avoir un esprit critique sur les associations identifiées (et non identifiées) par les modèles logistiques.

5.3.2 Utilisation des mesures d'importance des variables

Deux mesures d'évaluation de l'importance des variables dans les modèles de forêts aléatoires sont disponibles. Ces deux mesures ont chacune leurs défauts, notamment la détection de variables corrélées non associées avec la variable cible. Nous avons ici considéré la mesure d'importance de Gini, et non la mesure d'importance de permutation. Le jeu de données CoPanFlu peut être qualifié de déséquilibré. Ce déséquilibre concerne la proportion d'individus classés parmi les deux modalités possibles de la variable cible (infecté/non-infecté), en effet seuls 14% des individus étaient considérés comme infectés. Rappelons que la mesure d'importance de permutation va comparer pour chaque prédicteur la proportion de mauvaises classifications avec ou sans permutation des valeurs de cette variable. Il a été montré que l'utilisation d'un tel critère est non adaptée dans le cas d'un jeu de données déséquilibré [42, 89], justifiant ainsi l'utilisation de la mesure d'importance de Gini.

Le mesure d'importance conditionnelle, abordée dans le chapitre précédent, est attendue comme étant une solution aux biais associés aux mesures d'importance des forêts aléatoires. Cependant cet indice alternatif n'a pas pu être utilisé dans notre cas, la complexité des calculs requis était incompatible avec la dimension des jeux de données utilisés [138]. La taille des données était pourtant relativement modeste, en tout cas bien loin des jeux de type GWAS. L'applicabilité de cette méthode non-biaisée reste donc pour le moment relativement limitée, en dépit de son potentiel.

Concernant la régression LASSO, la question de l'évaluation de l'importance des variables incluses dans le modèle s'est posée. Sur un seul modèle LASSO, on pourrait considérer comme importantes toutes les variables ayant un coefficient de régression non nul. Ceci sous-entend d'utiliser un seul ré-échantillonnage par *cross-validation*. Les paramètres estimés par le modèle sont dès lors potentiellement très dépendants de l'échantillonnage choisi. Nous avons considéré que répéter cette procédure 100 fois permettait d'obtenir une estimation des coefficients moins dépendante des échantillons de *cross-validation* fixés. Sur les 100 modèles construits par *cross-validation* nous avons pu noter que certaines variables avait des coefficients nuls dans certains cas et pas dans d'autres, justifiant l'intérêt d'itérer un grand nombre de fois ce procédé. Nous avons ainsi considéré une mesure d'importance non rapportée jusqu'ici dans la littérature correspondant au nombre de fois où une variable avait un coefficient estimé non nul parmi 100 modèles.

Un des principaux avantages des mesures d'importance utilisées pour le *data mining* et le LASSO, comparativement à des méthodes de sélection univariée, est qu'elles n'ignorent pas les autres prédicteurs du modèle, permettant ainsi d'estimer des effets indépendants et de prendre en compte de potentielles interactions multivariées [181].

5.3.3 Tests de permutation pour la détection d'associations

Les mesures d'importances des variables des forêts aléatoires et des arbres de régression *boostés* ne fournissent pas de *p-value* permettant d'évaluer la significativité statistique de ces indices numériques. Les coefficients des modèles LASSO ne sont quant à eux pas accompagnés d'erreurs standards qui pourraient permettre le calcul d'une statistique de test, et l'élaboration de méthodes de calcul de *p-value* dans ces modèles est toujours au stade du développement [110].

Le recours à la permutation pour déterminer les prédicteurs les plus importants dans des modèles de forêts aléatoires a déjà été observé [2,4,157]. En revanche, l'utilisation de tels tests pour effectuer de la sélection de variables dans des modèles d'arbres de régression *boostés* ou de régression pénalisée n'a, à notre connaissance, jamais été observée. Les résultats de simulation ont permis d'observer l'efficacité des tests de permutation pour la détection de covariables ; si ces tests sont toutefois légèrement moins puissants qu'un test de référence comme le test de Wald, le risque de première espèce associé à cette procédure est quant à lui plus faible, sinon identique (voir les annexes de l'article).

L'une des limites de cette méthode trouve son origine dans la relation directe entre le niveau de précision des *p-values* obtenues et le nombre de permutations à réaliser. Obtenir des *p-values* précises requiert de répéter un grand nombre de fois les procédures d'estimation des modèles. Si les temps de calculs requis pour un seul jeu de données sont raisonnables, ceux-ci deviennent plus problématiques

avec un grand nombre de jeux de données. Ceci explique le nombre relativement limité de permutations réalisées pour les jeux de données simulés. La justesse des résultats obtenus en simulation, en particulier avec un seuil de significativité à 1%, n'est, dès lors, probablement pas optimale. Le niveau de précision limité des *p-values* aurait également pu être problématique si un ajustement des *p-values*, de telle sorte à limiter l'*experimentwise error rate* (*i.e.* la probabilité de rejeter à tort au moins une hypothèse nulle parmi les multiples tests réalisés) avait été envisagé. Comme discuté dans l'article, si un tel ajustement aurait été indispensable dans l'optique de construire un modèle prédictif ou de valider des associations, nous avons considéré que dans un but exploratoire l'utilisation d'une procédure d'ajustement n'était pas requis [21].

5.3.4 Limites du plan d'analyse classique en régression logistique

La méthodologie usuellement employée en régression logistique, basée sur une sélection univariée pour construire un modèle multivarié avec un nombre réduit de prédicteurs puis sur une procédure de sélection de variables, n'est clairement pas adaptée pour la recherche d'associations dans un grand jeu de données. Les critiques, tant sur la philosophie d'une analyse en deux étapes que sur l'utilisation de procédure de sélection *backward*, *forward* ou *stepwise* sont nombreuses [132, 184, 206, 207]. Cette méthodologie de référence est pourtant couramment utilisée en épidémiologie, et les recours aux méthodes de pénalisation peu nombreux [108, 203]. S'il apparait évident que sur un jeu de données *omiques* contenant plusieurs centaines de milliers de variables l'évaluation de cette procédure aurait été non pertinent, sur un jeu de taille plus modeste comme le notre cette question méritait d'être considérée.

Dans le cas où un modèle logistique est choisi, l'utilisation d'une méthode de régression pénalisée avec l'ensemble des prédicteurs doit être privilégiée. Nous avons ici considéré deux types de modèles, l'un avec des performances optimales d'un point de vue de la déviance, l'autre moins performant mais plus parcimonieux. Si dans une perspective de prédiction le modèle parcimonieux apparait comme la solution à privilégier, dans une optique de détection d'associations le modèle optimal semble être le plus à même de détecter les covariables associées, le prix à payer en retour étant une augmentation modérée du risque de première espèce (9% en moyenne dans nos simulations avec un seuil de significativité fixé à 5%).

Ces deux derniers chapitres ont été consacrés à l'analyse *data-driven* avec pour application la recherche de facteurs de risque de l'infection grippale dans les données CoPanFlu. Nous revenons maintenant à la philosophie *hypothesis-driven* et présentons une méthodologie d'analyse axée cette fois-ci non pas sur l'analyse d'une grande quantité de données mais l'intégration de données décrivant des

dimensions très diverses, les modèles d'équations structurelles.

Les modèles d'équations structurelles, une alternative hypothesis-driven aux méthodes de régression

6.1 Introduction

Les modèles d'équations structurelles (*structural equation modeling*, SEM) sont une méthode statistique permettant de tester et d'estimer des relations causales entre une ou plusieurs variables indépendantes et une ou plusieurs variables dépendantes.

Ces méthodes *hypothesis-driven*, développées au début du XXème siècle avec l'apparition de l'analyse factorielle et de l'analyse de chemin (concepts développés par la suite) ont vu leur utilisation fortement augmenter dans les années 1970 [26], et sont maintenant très fréquemment utilisées dans les sciences sociales et comportementales. Elles s'utilisent principalement dans un but confirmatoire, il s'agit de poser des hypothèses quant à un ensemble de relations causales existant au sein des données, et de vérifier si ces hypothèses sont supportées par les données. Les modèles SEM combinent l'utilisation de méthodes graphiques comme les graphiques acycliques orientés (*Directed Acyclic Graphs*, ou *DAG*), permettant une représentation simple des relations à tester, et d'un ensemble d'équations mathématiques [195]. Ces équations correspondent à des modèles de régression représentant le processus causal étudié et sont utilisées pour tester la significativité statistique des relations supposées.

Nous présentons maintenant les concepts généraux des modèles SEM, leurs similitudes avec les modèles de régression conventionnels mais aussi la plus-value apportée par ces méthodes.

6.2 Concepts généraux

6.2.1 Variables observées et variables latentes

Les méthodes d'analyse statistique usuelles se basent sur l'utilisation de variables observées uniquement. Les modèles SEM utilisent également un autre type de variables, appelées variables latentes. Les variables latentes sont des variables non observées [25]. L'introduction de variables latentes offre une extension de la technique d'analyse de chemin (*path analysis*), proposant comme les modèles SEM l'exploration de relations causales multiples, ces relations étant toutefois limitées à l'utilisation de variables observées. Le concept de variables latentes trouve son sens en sciences sociales et comportementales où l'on cherche à mesurer des phénomènes non directement observables, tels que la

perception d'un phénomène donné ou encore l'aptitude dans un domaine particulier, expliquant ainsi la popularité des modèles SEM dans ces disciplines.

Les variables latentes étant des variables non observées, elles sont estimées à partir d'un ensemble de variables observées, appelées indicateurs. Le choix minutieux des indicateurs est essentiel, ces variables étant supposées être le reflet d'un phénomène latent non observable [33]. Ceci sous-entend que les indicateurs d'une variable latente ne sont pas indépendants et qu'un certain niveau de corrélation existe entre eux.

6.2.2 Variables exogènes et variables endogènes

Un modèle de régression classique implique généralement l'étude de la relation entre une variable cible dépendante et une ou plus variables explicatives indépendantes. Dans un modèle SEM, on peut avoir un grand nombre d'équations, impliquant chacune un nombre multiple de variables. Ainsi, une variable donnée peut être une variable indépendante dans une équation tout en étant une variable dépendante dans une autre équation.

Ce concept de variable dépendante/indépendante étant plus flou dans le cadre des modèles SEM, les termes de variables endogènes et exogènes sont employés, pour les variables observées comme pour les variables latentes. Les variables exogènes ne sont influencées par aucune autre variable dans le modèle tandis que les variables endogènes sont l'équivalent d'une variable dépendante dans au moins une des équations structurelles [5, 73].

6.2.3 L'analyse factorielle

L'analyse factorielle (*factor analysis*) décrit un ensemble de procédures utilisées pour étudier les relations entre des variables observées et des variables latentes (généralement appelées facteurs dans le contexte de l'analyse factorielle). On distingue deux types d'analyses factorielles, l'analyse factorielle exploratoire (*exploratory factor analysis*, EFA) et l'analyse factorielle confirmatoire (*confirmatory factor analysis*, CFA).

L'analyse factorielle exploratoire

L'analyse factorielle exploratoire est utilisée pour mettre en évidence une éventuelle structure sous-jacente parmi un certain nombre de variables observées. Il s'agit de déterminer le nombre idéal de facteurs nécessaires pour expliquer les corrélations parmi les variables observées. Cette analyse est exploratoire dans le sens où aucune hypothèse n'est posée sur les relations entre les variables latentes et les variables observées.

Cette technique est principalement employée pour le développement d'échelles de mesure, l'analyse factorielle exploratoire permet alors d'identifier le nombre de facteurs optimal en fonction des données mesurées [63]. Prenons l'exemple d'une échelle supposée mesurer les aptitudes intellectuelles d'un individu en explorant des dimensions telles que la mémoire, les aptitudes verbales, les aptitudes numériques ou encore la rapidité de raisonnement ; ce test consiste donc en quatre questionnaires, composés d'un nombre variables d'items (*i.e.* des variables observées). On s'attend si le test est correctement construit à identifier quatre facteurs, censés représenter les dimensions étudiées par chacun des questionnaires.

La capacité de ce test à capturer le phénomène étudié (les aptitudes intellectuelles de l'individu) peut être évaluée par une EFA. L'analyse factorielle exploratoire nous indiquera si le nombre de facteurs optimal est bien quatre facteurs, si certains items sont des variables indicatrices de faible qualité, si certains facteurs sont mal "mesurés" par leurs indicateurs ou encore si certains indicateurs peuvent être associés à plus d'un facteur. La relation entre une variable observée et une variable latente est quantifiée par un *factor loading*, assimilable la plupart du temps à un coefficient de corrélation [24]. On s'attend ainsi à ce que par exemple les *factor loadings* des items censés mesurer la rapidité de raisonnement aient des valeurs élevées pour ce facteur et des valeurs faibles ou négatives pour les autres facteurs.

Si cette méthode est appropriée dans une optique de création d'une échelle de mesure, elle n'est pas adaptée pour tester des hypothèses et est dès lors d'un intérêt moindre dans une problématique de compréhension d'un mécanisme tel que celui de l'infection grippale. Nous ne détaillons donc pas les procédures d'estimation des paramètres d'une EFA et de sélection du nombre optimal de facteurs et nous concentrons sur l'analyse factorielle confirmatoire, qui est d'un intérêt supérieur dans notre cas.

L'analyse factorielle confirmatoire

Contrairement à l'analyse factorielle exploratoire, la CFA se base sur une connaissance *a priori* de la structure des variables latentes, cette connaissance étant basée sur la théorie ou sur des recherches empiriques [31]. Il s'agit ici de tester si les hypothèses de relations entre les variables latentes et leurs indicateurs respectifs sont validées d'un point de vue statistique. La CFA permet également d'évaluer si un modèle de facteurs construit sur une population donnée est également valide sur un échantillon indépendant. Reprenons l'exemple de l'échelle de mesure des aptitudes intellectuelles. Dans une CFA, la présence des quatre variables latentes est déterminée *a priori*, tout comme le sont les relations entre les indicateurs et leurs variables latentes respectives. Par exemple, un indicateur lié à la variable latente "mémoire" ne sera pas utilisé pour estimer la variable latente "aptitudes numériques" (concrètement on fixera les valeurs des *factor loadings* de ces items vers cette variable latente à 0), ce qui aurait été possible dans une EFA. La pertinence du modèle construit dans la CFA est évaluée par différents

indices d'adéquation du modèle aux données ou d'amélioration relative par rapport à un modèle de base prédéfini [31] (les procédures d'évaluation des modèles SEM sont détaillées dans le chapitre 6.4.4).

6.2.4 Le modèle d'équations structurelles complet

Nous avons pu observer que les deux types d'analyses factorielles, exploratoire et confirmatoire, se concentrent uniquement sur les relations entre les variables latentes et leurs indicateurs. Aucune de ces méthodes n'étudie de relations éventuelles entre variables latentes. C'est l'intérêt du modèle SEM complet, qui autorise la spécification de relations de régression entre variables latentes.

Le modèle SEM complet est composé de deux compartiments, le modèle de mesure (*measurement model*) et le modèle structurel (*structural model*) [134]. Le modèle de mesure décrit les relations entre les variables latentes et leurs indicateurs respectifs, il est donc assimilable à un modèle CFA. Le modèle structurel décrit quant à lui les relations entre variables latentes [5]. La combinaison du modèle de mesure, du modèle structurel et d'éventuelles covariables observées endogènes ou exogènes non indicatrices de variables latentes constitue le modèle SEM complet.

6.3 Le diagramme de chemin et les équations structurelles

En fonction de la complexité du phénomène à modéliser, les modèles SEM peuvent rapidement contenir un grand nombre de variables et de relations, et par conséquent reposer sur un grand nombre d'équations. Une façon simple de représenter les hypothèses posées est d'utiliser un diagramme de chemin (*path diagram*) [152].

La figure 6.1 représente un exemple un diagramme de chemin d'un modèle SEM. Un certain nombre de conventions régissent la représentation des modèles SEM. Les cercles (ou ellipses) représentent généralement les variables latentes tandis que les rectangles (ou carrés) sont le plus souvent utilisés pour les variables observées. Les flèches simples représentent le sens de la relation entre deux variables, alors que les doubles flèches (non utilisées ici) sont parfois employées pour signifier des corrélations entre des paires de variables [5]. La convention des SEM implique généralement que les variables observées indicatrices sont des variables dépendantes de leur facteur, *i.e.* que les flèches vont de la variable latente vers les indicateurs (d'autres types de facteurs existent, voir Borsboom *et al.* 2003 [27] pour davantage d'informations sur les différents types de variables latentes).

A partir de cette représentation graphique, nous pouvons facilement reconstituer l'ensemble des équations structurelles associées au modèle SEM de la figure 6.1. Les équations correspondant au modèle de mesure peuvent être représentées ainsi :

$$\begin{cases} x_1 = \lambda_1\xi_1 + \delta_1 & y_1 = \lambda_4\eta_1 + \epsilon_1 & y_4 = \lambda_7\eta_2 + \epsilon_4 \\ x_2 = \lambda_2\xi_1 + \delta_2 & y_2 = \lambda_5\eta_1 + \epsilon_2 & y_5 = \lambda_8\eta_2 + \epsilon_5 \\ x_3 = \lambda_3\xi_1 + \delta_3 & y_3 = \lambda_6\eta_1 + \epsilon_3 & y_6 = \lambda_9\eta_2 + \epsilon_6 \end{cases}$$

Les x et les y correspondent aux indicateurs des variables latentes ξ_1, η_1 et η_2 (on remarque une notation différente en fonction du caractère endogène ou exogène de la variable latente, voir les tableaux 6.1 et 6.2). Les *factor loadings* sont représentés par les termes λ et quantifient les relations entre les variables latentes et leurs indicateurs ; en d'autres termes ils représentent le changement attendu dans les variables indicatrices en cas de variation de la variable latente correspondante. Les termes δ et ϵ sont des "erreurs de mesure" (*measurement error*), correspondant à la variance des variables indicatrices non capturée par la variable latente [129]. Les erreurs de mesure reflètent l'adéquation des indicateurs à correctement mesurer le phénomène représenté par la variable latente, pour des indicateurs correctement choisis on s'attend à observer des erreurs de mesure plutôt faibles [33]. L'erreur de mesure dérive de deux sources, une source d'erreur aléatoire non systématique (par exemple si un sujet comprend mal la question correspondant à l'indicateur) et une source d'erreur systématique non aléatoire (causée par exemple par un indicateur mal mesuré). Par défaut les erreurs de mesures des différents indicateurs d'une même variable latente sont supposées être indépendantes les unes des autres, il est toutefois possible d'ajouter dans le modèle des termes de corrélation entre les erreurs de mesure [26] (voir le tableau 6.2) .

De manière identique au modèle de mesure on peut représenter les équations du modèle structurel (en remarquant une nouvelle fois une notation variable en fonction du caractère endogène/exogène de la variable latente) :

$$\begin{cases} \eta_1 = \gamma_{11}\xi_1 + \zeta_1 \\ \eta_2 = \beta_{21}\eta_1 + \zeta_2 \end{cases}$$

Les termes γ et β représentent des coefficients de régression et indiquent l'association entre deux variables latentes. On note une nouvelle fois la présence de termes d'erreur ζ, appelés "erreurs résiduelles" (*residual error*), analogues à des résidus dans un modèle de régression classique. Les différents termes d'erreurs, erreur de mesure et erreur résiduelle, sont considérés comme des termes latents et sont donc représentés par des cercles [33].

Après avoir introduit les concepts généraux et présenté les différentes composantes d'un modèle SEM, nous présentons maintenant les cinq principales étapes de la construction d'un tel modèle : la spécification, l'identification, l'estimation, l'évaluation et enfin la manipulation du modèle.

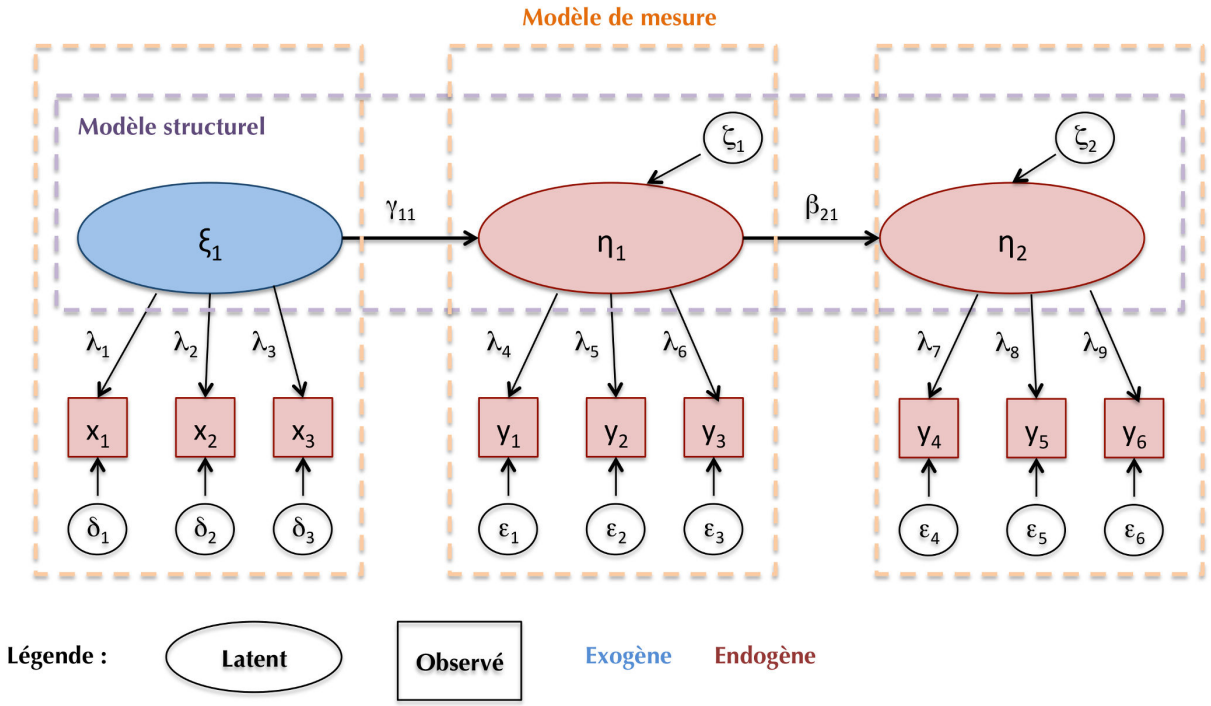


FIGURE 6.1 : Exemple de diagramme de chemin d'un modèle SEM

Symbole	Définition
η	Variable latente endogène
ξ	Variable latente exogène
y	Indicateur d'une variable latente endogène
x	Indicateur d'une variable latente exogène
ϵ	Erreur de mesure d'un indicateur d'une variable latente endogène
δ	Erreur de mesure d'un indicateur d'une variable latente exogène
ζ	Erreur résiduelle d'une variable latente endogène

TABLE 6.1 : Types de variables d'un modèle SEM

Symbole	Définition
λ	Coefficient liant une variable latente endogène/exogène à un indicateur
β	Coefficient liant deux variable latentes endogènes
γ	Coefficient liant une variable latente exogène à une variable latente endogène
ϕ	Corrélations entre les variables latentes exogènes
ψ	Corrélations entre les erreurs résiduelles ζ
Θ_ϵ	Corrélations entre les erreurs de mesure ϵ
Θ_δ	Corrélations entre les erreurs de mesure δ

TABLE 6.2 : Types de paramètres dans un modèle SEM

6.4 Construction d'un modèle d'équations structurelles

6.4.1 Spécification du modèle

Il s'agit ici de définir formellement le modèle qui va être testé, *i.e.* de construire le diagramme de chemin du modèle. C'est à cette étape qu'il est nécessaire de définir les variables latentes et leurs indicateurs, les paramètres "fixés" et les paramètres "libres". Les paramètres fixés ne seront pas estimés à partir des données, et seront contraints à une valeur numérique choisie, généralement 0, signifiant dans ce cas l'absence de relation. Les paramètres libres seront quant à eux estimés à partir des données, sous-entendant que les relations concernées sont non-nulles. Le sens des relations testées doit également être précisé.

6.4.2 Identifiabilité du modèle

La construction du diagramme de chemin et la spécification du caractère fixé/libre de l'ensemble des paramètres du modèle SEM va soulever la question de l'identifiabilité de ce dernier. Un modèle SEM peut être identifié s'il existe une unique valeur que l'on peut estimer à partir des données pour chacun des paramètres libres dans le modèle [26]. Un modèle identifié peut être correctement estimé et évalué alors que pour un modèle non-identifié, une infinité de valeurs peuvent être estimées pour un paramètre donné.

Pour qu'un modèle SEM soit identifié, il faut que le nombre de données soit supérieur au nombre de paramètres à estimer. Dans un modèle SEM, les données utilisées pour l'estimation correspondent à la matrice de variance/covariance du jeu de données. Le nombre d'éléments constituant la matrice de variance/covariance d'un jeu de données correspond simplement à $p(p+1)/2$, p étant le nombre de variables observées [94]. Les modèles peuvent ainsi être qualifiés de "juste-identifiés", "sous-identifiés" ou "sur-identifiés". Dans le premier cas le nombre de données et le nombre de paramètres à estimer sont identiques. Bien qu'une seule solution soit estimable pour l'ensemble des paramètres, le nombre de degrés de liberté de ce modèle est nul, il est dès lors impossible de le rejeter par des tests statistiques. Dans un modèle sous-identifié, le nombre de paramètres à estimer est supérieur au nombre de données, les paramètres du modèle ne peuvent avoir de solution unique. Enfin on distingue le modèle sur-identifié où le nombre de paramètres à estimer est inférieur au nombre de données. Un jeu unique de paramètres est estimable et ce modèle peut être testé.

Dimensionnement des variables latentes

Avoir un modèle sur-identifié' est une condition *sine qua non* dans la construction d'un modèle SEM. Seulement, la résolution du problème d'identifiabilité des modèles nécessite aussi que la dimension de chacune des variables latentes soit déterminée. Les variables latentes sont souvent utilisées pour représenter des concepts abstraits, par conséquent elles sont non mesurées et n'ont pas d'échelle.

L'approche la plus courante pour fixer la dimension d'une variable latente est celle de la "variable de référence", qui consiste à assigner une valeur constante (en général 1.0) à un *path loading* d'un des indicateurs de la variable latente. Sur l'exemple de la figure 6.1, la dimension des variables latentes ξ_1, η_1 et η_2 pourrait être fixée en choissant par exemple λ_1, λ_4 et λ_7 égaux à 1. Le choix de l'indicateur de référence peut être fait de manière arbitraire, ou en fonction de la variable jugée comme étant la plus représentative du phénomène latent. Une seconde approche est de fixer la variance de la variable latente à une valeur donnée (généralement 1.0), cette technique ayant l'avantage que l'ensemble des *path loadings* peuvent être librement estimés [31].

6.4.3 Estimation du modèle

Considérons S la matrice de covariance des données, θ un vecteur contenant l'ensemble des paramètres du modèle et $\Sigma(\theta)$ la matrice de covariance des variables associée aux paramètres θ estimés par le modèle. Le but de cette étape est d'estimer le vecteur θ permettant de minimiser au maximum la différence entre la matrice S et la matrice $\Sigma(\theta)$ estimée par le modèle [56].

Il existe un certain nombre de méthodes d'estimation, différant en fonction de la taille de l'échantillon et des distributions des données. L'estimateur du maximum de vraisemblance (*maximum likelihood*, ML) est couramment utilisé, et cherche à maximiser la vraisemblance des estimations par rapport aux données. On peut également citer l'estimateur des moindres carrés (*least squares*, LS) qui tend à minimiser la somme des carrés des résidus du modèle [24]. D'autres adaptations de ces estimateurs ont été proposées (tel que l'estimateur robuste du maximum de vraisemblance), autorisant une violation de certaines hypothèses relative à la distribution des données ou l'utilisation de variables dépendantes catégorielles, de survie ou encore des variables suivant des distributions de Poisson [135].

L'intérêt de la construction d'un modèle SEM est de savoir dans quelle mesure les hypothèses construites sont concordantes avec les données. Une fois les paramètres estimés l'étape suivante est l'évaluation du modèle. Celle ci se fait premièrement sous un angle global, offrant un aperçu général de la qualité d'ajustement du modèle, puis de manière plus précise, où chacune des estimations est évaluée.

6.4.4 Evaluation globale du modèle

De manière analogue aux modèles de régression classiques, des indices d'information tels que le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC) sont parfois utilisés dans le cadre des SEM. Ces deux indices, dont le calcul se base sur la vraisemblance du modèle et sur le nombre de paramètres, sont utiles pour sélectionner un modèle parmi un sous-ensemble de modèles construits. En revanche, ils ne permettent pas d'évaluer l'ajustement d'un modèle pris individuellement. Nous présentons maintenant un ensemble de statistiques alternatives, permettant une évaluation globale de la qualité d'ajustement d'un modèle SEM.

Tests du χ^2

Le test du χ^2 est utilisé pour évaluer dans quelle mesure la matrice $\Sigma(\theta)$ est éloignée de la matrice de covariance des données S, l'hypothèse nulle testée étant que la différence entre les deux matrices est nulle. La statistique de test calculée nous informe sur la vraisemblance des estimations par rapport aux données. Une adéquation maximale du modèle aux données étant espérée, le non-rejet de l'hypothèse nulle est souhaité.

Un autre test du χ^2 est généralement réalisé, comparant cette fois le modèle estimé à un modèle représentant une indépendance complète de toutes les variables (*independence model* ou encore *baseline model*, *i.e.* l'ensemble des variables du modèle sont non associées). Contrairement au premier test du χ^2 on attend ici que le modèle construit soit supérieur au modèle d'indépendance, c'est à dire à ce que l'hypothèse nulle du test soit rejetée.

L'utilisation de statistiques du χ^2 pour évaluer les modèles SEM est toutefois critiquée, celles-ci étant notamment sensibles à la taille de l'échantillon [94], suggérant l'utilisation de statistiques alternatives.

Indices comparatifs

Les indices d'évaluation de l'ajustement d'un modèle SEM dits "comparatifs" (ou incrémentaux) mesurent l'amélioration proportionnelle apportée par le modèle construit par rapport à un modèle de base (généralement le modèle d'indépendance). Deux indices incrémentaux sont généralement utilisés : le *Comparative Fit Index* (CFI) [23] et le *Tucker Lewis Index* (TLI) [196]. Le CFI est un indice normé variant entre 0 et 1, des valeurs supérieures à 0.90, voire 0.95, sont généralement considérées comme satisfaisantes [33]. Le TLI induit lui une pénalité sur le nombre de paramètres inclus dans le modèle et va défavoriser les modèles excessivement complexes en pénalisant l'inclusion de paramètres n'apportant qu'une contribution marginale à l'amélioration de la qualité d'ajustement du modèle. Le TLI est un indice non normé, toutefois comme pour le CFI des valeurs proches de 1 sont recommandées (voir

Byrne 2012 [33] pour plus d'informations sur le calcul de ces indices).

Indices absolus

Contrairement au CFI et au TLI qui comparent le modèle construit à un modèle de base, les indices absolus se basent seulement sur une évaluation de l'adéquation du modèle aux données.

L'indice absolu le plus couramment utilisé est le *root mean square error of approximation* (RMSEA) [176]. Le RMSEA est une fonction du χ^2 indiquant à quel point le modèle est proche des données, tout en prenant en compte le nombre de paramètres du modèle. Des valeurs inférieures à 0.05 sont généralement associées avec un modèle avec une adéquation aux données satisfaisante [33]. L'utilisation du RMSEA est recommandée, notamment en raison de sa sensibilité à de mauvaises spécifications du modèle et de la possibilité de construire des intervalles des confiance autour des valeurs estimées [115].

Nous pouvons également citer le *standardized root mean square residual* (SRMR), qui correspond à la moyenne des résidus (*i.e.* la différence entre la matrice de covariance estimée par le modèle et celle correspondant aux données) standardisés [31]. Des valeurs inférieures à 0.08 sont généralement recommandées pour cet indice, qui est toutefois peu adapté avec des données catégorielles. Cet indice doit être privilégié par rapport au *root mean square residual* (RMR), qui ne considère pas les résidus dans leur version standardisée, rendant dès lors l'interprétation de cet indice plus complexe.

6.4.5 Évaluation des paramètres estimés

La première question à traiter est celle de la viabilité des paramètres et de leur cohérence, d'un point de vue du signe et de l'ordre de grandeur, par rapport aux hypothèses établies *a priori*. De mauvaises estimations doivent interroger soit sur la pertinence du modèle construit soit sur la qualité et/ou la quantité de données disponibles pour estimer les relations. La procédure d'estimation peut parfois conduire à l'observation de *Heywood cases*, se manifestant par l'estimation de paramètres incohérents comme des variances négatives ou des corrélations supérieures à 1 [96].

Il est également nécessaire d'observer les erreurs standards des estimations. Des erreurs standards anormalement petites ou élevées peuvent être problématiques. Dans un premier cas, si l'erreur standard d'un paramètre est proche de 0, la statistique de test permettant d'évaluer la significativé du paramètre ne pourra être calculée. Dans le second cas, une erreur standard élevée est symptomatique d'un paramètre qui ne peut pas être correctement estimé.

Enfin, la significativité statistique des estimations est le dernier critère employé pour évaluer les paramètres individuels du modèle. La statistique utilisée correspond simplement au rapport entre l'estimation de la valeur d'un paramètre et de son erreur standard. Au seuil 5% si la valeur absolue de ce

ratio est supérieure à 1.96 on conclut que ce paramètre est significativement différent de 0. Des paramètres non significatifs peuvent être le reflet soit d'une relation non nécessaire ou encore d'un manque de puissance pour mettre en évidence ces relations.

6.4.6 Manipulation du modèle

Différentes manipulations peuvent être réalisées pour améliorer la qualité d'ajustement du modèle. Premièrement l'observation de la matrice de covariance résiduelle, qui reflète la différence entre les matrices S et $\Sigma(\theta)$ peut être informative sur des relations qui seraient sur-estimées ou sous-estimées, suggérant par exemple l'ajout de relations supplémentaires ou l'incorporation de variables supplémentaires susceptibles de modifier la relation supposée et ainsi de diminuer les résidus [24, 31].

Un mauvais indice d'adéquation du modèle aux données peut également être le reflet d'un paramètre fixé à 0 à tort. Les *modifications indices* (MI) permettent d'identifier des contraintes exercées sur des paramètres qui nuisent à la qualité du modèle [173]. Chacun des paramètres fixés à une valeur nulle est évalué et une éventuelle diminution significative du χ^2 du modèle suite au changement de statut (de fixé à libre) de ce paramètre en fait un candidat pour l'ajout d'une nouvelle relation. Le choix de l'inclusion ou non d'une relation dans le modèle suite à l'utilisation des MI dépend de sa pertinence d'un point de vue théorique, et de la cohérence de la valeur attendue de ce paramètre (*expected parameter change*, EPC). De plus l'ajout excessif de paramètres dans le modèle soulève la question du sur-ajustement du modèle aux données, renforçant la nécessité d'utiliser les MI avec parcimonie.

La dernière étape de manipulation du modèle SEM consiste à supprimer les relations pour lesquelles les coefficients estimés sont non significatifs. Le modèle final attendu est un modèle offrant un compromis entre un modèle consistant avec les données et un modèle parcimonieux [56].

6.4.7 Modèle final

Dès lors qu'un modèle avec des indices de qualité d'ajustement satisfaisants et des paramètres estimés significatifs est obtenu, les paramètres du modèle peuvent être standardisés, permettant ainsi une comparaison des forces d'associations relatives entre variables observées et/ou latentes. Les *path loadings* standardisés sont directement interprétables en terme de corrélation entre les indicateurs et les variables latentes correspondantes. Notons que le carré d'un *path loading* standardisé correspond à la proportion de variance (R^2) de l'indicateur expliquée par la variable latente [94]. Un bémol à l'utilisation de ces indices est que l'obtention d'erreurs standards autour des paramètres standardisés n'est pas toujours possible, notamment lorsque l'estimateur "*weighted least squares with mean and variance adjustment*" (particulièrement adapté en présence de variables dépendantes catégorielles comme le statut "infection grippale" étudié ici) est utilisé en présence de covariables observées non indicatrices de

variables latentes.

6.5 Avantages et limites des modèles d'équations structurelles

Nous avons pu voir les différentes spécificités des modèles d'équations structurelles. La construction d'un tel modèle n'est pas triviale, spécialement si l'on compare avec l'utilisation d'un modèle de régression multivariée classique.

Contrairement aux modèles de régression plus communs, les modèles SEM permettent d'utiliser des variables observées ou latentes, mais aussi de modéliser des relations directes ou indirectes, avec des variables pouvant être tantôt des variables à expliquer tantôt des variables explicatives. Ces méthodes sont ainsi considérées comme plus performantes que la régression multiple, qui ne considère que des associations indépendantes entre les prédicteurs et la variable réponse, sans modèle causal sous-jacent [7]. Le prix à payer en retour est la multiplication des hypothèses [57] ; dans les modèles SEM, des hypothèses sont faites sur la forme de la relation entre des prédicteurs et une variable dépendante pour chacune des équations structurelles, et non pour une seule relation comme dans un modèle de régression. De plus l'absence de relation dans le modèle SEM (*i.e.* l'absence de flèche entre deux variables sur le diagramme de chemin) doit être considérée comme une hypothèse en elle-même.

Une autre différence notable entre les modèles de régression et les modèles SEM est la prise en compte de termes d'erreurs multiples. Les modèles SEM permettent de modéliser une erreur de mesure sur les prédicteurs observés tandis que dans un modèle de régression classique, seul un terme d'erreur résiduelle sur la prédiction de la variable cible est estimé.

L'interprétation des modèles SEM doit être faite avec précaution. En effet, ceux-ci ne peuvent être utilisés pour affirmer que les mécanismes traduits sous forme d'équations structurelles sont vrais. Il est nécessaire de garder à l'esprit que les indices numériques utilisés pour évaluer les modèles SEM traduisent l'adéquation du modèle aux données et son amélioration par rapport à un modèle d'indépendance. Ainsi, un certain nombre de modèles alternatifs construits à partir de jeux de variables différents sont susceptibles de fournir des indices de qualité d'ajustement aussi bons que le modèle considéré comme final.

Un modèle causal dans le cadre des équations structurelles doit être interprété comme un modèle dont les hypothèses sont cohérentes avec les données, non comme la preuve de la véracité des mécanismes testés. De plus, comme dans toute analyse statistique, il convient de distinguer la significativité statistique de la pertinence scientifique des résultats obtenus, la cohérence des résultats doit prévaloir. Une sur-interprétation des modèles SEM peut conduire à des conclusions erronées, dans la mesure du possible un modèle SEM devrait être validé avec des données indépendantes des données d'apprentis-

sage.

Nous présentons maintenant une étude des facteurs de risque de l'infection grippale H1N1pdm par un modèle d'équations structurelles. La philosophie de ce travail est à l'opposé de celle de l'étude exploratoire présentée précédemment dans le chapitre 5. Nous allons évaluer l'apport d'une approche *hypothesis-driven* dans le cadre de l'analyse de CoPanFlu après avoir travaillé jusqu'alors avec des méthodes *data-driven*.

Article : Le risque d'infection grippale abordé sous l'angle hypothesis-driven

7.1 Résumé

Le précédent article s'est proposé d'étudier les données CoPanFlu d'une manière globale, toutes les variables étant mises sur le même niveau. Nous proposons une étude causale du risque d'infection, par le biais d'un modèle d'équations structurelles. L'intérêt est ici de tirer parti de la diversité des variables récoltées dans CoPanFlu (figure 3.1 page 19) pour modéliser différents niveaux d'information.

L'hypothèse fondamentale faite en amont de ce travail est que l'infection par la grippe d'un individu est la résultante de deux conditions indispensables, premièrement cet individu doit être exposé au virus (le terme exposition est ici très large), deuxièmement il doit être susceptible (d'un point de vue biologique) de contracter une infection. Partant de cette hypothèse, nous avons exploité la spécificité des modèles d'équations structurelles de modéliser des variables latentes pour étudier des facteurs très divers (biologiques, socio-démographiques, comportementaux et environnementaux), régulant l'exposition au virus ou la susceptibilité, et estimer leur association avec le risque d'infection ainsi que leurs relations éventuelles.

Sept variables latentes ont été utilisées pour représenter l'exposition d'un sujet au virus, traitant de dimensions aussi diverses que la perception du risque d'infection, celle des mesures préventives, l'adoption de comportements préventifs, l'environnement intérieur et extérieur, l'environnement socio-économique ou encore l'intensité des contacts sociaux. Une seule variable a été utilisée pour modéliser la susceptibilité individuelle à l'infection : le titre d'anticorps pré-épidémique. Nous avons supposé l'existence de relations directes entre les variables latentes construites et le risque d'infection, à l'exception de la perception du risque d'infection et de celle des mesures préventives, considérant comme peu probable que ces facteurs comportementaux soient associés de manière directe avec le risque d'infection.

Les effets des variables latentes et observées sont représentés par des coefficients standardisés β_{std} permettant une comparaison des forces d'associations relatives. Parmi l'ensemble des relations testées, seules les variables latentes décrivant la susceptibilité de l'hôte et l'adoption de comportements préventifs étaient directement associées avec le risque d'infection, l'une positivement l'autre négativement ($\beta_{std} = 0.31$ et $\beta_{std} = -0.21$, respectivement). Les facteurs comportementaux décrivant la perception du risque et des mesures préventives ont montré une influence positive sur l'adoption des comportements préventifs ($\beta_{std} = 0.14$ et $\beta_{std} = 0.27$, respectivement).

Aucune association significative n'a pu être montrée entre le nombre et la durée des contacts sociaux et le risque d'infection. Une analyse approfondie de l'exposition par les contacts en fonction de l'âge et du statut infectieux n'a pas permis de mettre en évidence une exposition supérieure chez les infectés par rapport aux non-infectés, et ce quelle que soit la classe d'âge observée, alors qu'une forte association entre les contacts et l'âge a pu être observée, les individus les plus jeunes ayant le plus de contacts, principalement avec des sujets d'un âge similaire.

L'utilisation des modèles d'équations structurelles nous a permis d'obtenir une représentation synthétique des déterminants de l'infection grippale et de classer ces facteurs par ordre de force d'association avec l'infection, autorisant ainsi l'identification de cibles prioritaires potentielles pour des interventions de santé publique. Nos résultats suggèrent que la vaccination anti-grippale et que des campagnes de communication promouvant les comportements préventifs pourraient être les mesures de santé publique à prioriser lors des futures épidémies de grippe.

7.2 Article

Manuscript title: Causal analysis of H1N1pdm09 influenza infection risk in a household cohort

Authors: Yohann Mansiaux [a,b*], Nicolas Salez [c], Nathanael Lapidus [a,b,d], Michel Setbon [e], Laurent Andreoletti [f,g], Marianne Leruez-Ville [h], Simon Cauchemez [i], Marie-Lise Gougeon [j], Frédéric Vély [k,l,m,n], Michael Schwarzinger [o,p,q], Laurent Abel [r,s], Rosemary Markovic Delabre [a,b], Antoine Flahault [t,u], Xavier de Lamballerie [c], Fabrice Carrat [a,b,d]

[a] INSERM, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F-75013, Paris, France

[b] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Institut Pierre Louis d'Epidémiologie et de Santé Publique, F-75013, Paris, France

[c] Aix Marseille Univ, IRD French Institute of Research for Development, EHESP French School of Public Health, UMR_D 190 "Emergence des Pathologies Virales", 13005, Marseille, France

[d] Public Health Unit, Saint-Antoine Hospital, 75012 Paris, France

[e] CNRS-Aix Marseille Université, IRD French Institute of Research for Development, EHESP French School of Public Health, EPV UMR_D 190 "Emergence des Pathologies Virales", 13385, Marseille, France

[f] Laboratoire de Virologie médicale et moléculaire Hôpital Robert Debré, CHU Reims

[g] EA 4684, Faculté de Médecine, Reims, France

[h] Laboratory of Virology. Hospital Necker-.Enfants-malades, Assistance Publique-Hôpitaux de Paris APHP - University Paris Descartes, Sorbonne Paris Cité

[I] Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, Paris, France

[j] Institut Pasteur, Antiviral Immunity, Biotherapy and Vaccine Unit, Paris, France

[k] Centre d'Immunologie de Marseille-Luminy, INSERM, U1104, F-13009 Marseille, France

[l] CNRS, UMR7280, F-13009 Marseille, France

[m] Aix Marseille Université, UM2, F-13009 Marseille, France

[n] Service d'Immunologie, Assistance Publique – Hôpitaux de Marseille, Hôpital de la Conception, Marseille, France

[o] INSERM, IAME, UMR 1137, F-75018 Paris, France

[p] Univ Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, F-75018 Paris, France

[q] Translational Health Economics Network, F-75010 Paris, France

[r] Laboratory of Human Genetics of Infectious Diseases, Necker Branch, Institut National de la Santé et de la Recherche Médicale U1163, 75015 Paris, France

[s] Paris Descartes University, Sorbonne Paris Cité, Imagine Institute, 75015 Paris, France

[t] Centre Virchow-Villermé, Descartes, Université Sorbonne Paris Cité, France

[u] Global Health Institute, University of Geneva, Switzerland

[*]**Corresponding author**

Mailing address: Equipe 2, UMRS-1136 , Faculté de Médecine Saint Antoine, 27 Rue Chaligny, 75571 Paris, Cedex 12, France.

Phone: +33 1 44 73 84 32.

E-mail: yohann.mansiaux@upmc.fr

Keywords

Statistical Model; Influenza, Human; Epidemiology

Word count

Abstract: 204 words.

Main text: 2934 words.

Abstract

Background: Obtaining a comprehensive quantitative figure of the determinants of influenza infection will help identify priority targets for future influenza mitigation interventions. We developed an original causal model integrating highly diverse factors and their dependencies, to identify the most critical determinants of pandemic influenza infection (H1N1pdm09) during the 2010-11 influenza season.

Methods: We used data from 601 households (1450 subjects) included in a dedicated cohort. Structural equations were used to model direct and indirect relationships between infection and risk perception, compliance with preventive behaviors, social contacts, indoor and outdoor environment, socio-demographic factors and pre-epidemic host susceptibility. Standardized estimates (β_{std}) were used to assess the strength of associations (ranging from -1 for a completely negative association to 1 for a completely positive association).

Results: Host susceptibility to H1N1pdm09 and compliance with preventive behaviors were the only two factors directly associated with the infection risk (β_{std}=0.31 and β_{std}=-0.21). Compliance with preventive behaviors was influenced by risk perception and preventive measures perception (β_{std}=0.14 and β_{std}=0.27). Number and duration of social contacts were not associated with H1N1pdm09 infection.

Conclusions: Our findings suggest that influenza vaccination in addition to public health communication campaigns focusing on personal preventive measures should be prioritized as potentially efficient interventions to mitigate influenza epidemics.

INTRODUCTION

Infection with the influenza virus results from a complex interplay between biological characteristics of the virus and host, individual and collective behaviors, social interactions, and environmental factors. Since 2009, numerous studies have investigated risk factors for pandemic influenza infection (H1N1pdm09); factors as diverse as individual characteristics (e.g. young age,[1] female gender,[1] chronic comorbidity,[2] and low pre-epidemic antibody titer [3]), indoor and outdoor environments (*e.g.* number of subjects in the household,[4] and residence in an urban area [3]), contacts with infected individuals,[1,5] and use of preventive measures such as hand washing,[6] have been found to be determinants of H1N1pdm09 infection. However, no study has simultaneously explored these factors to provide a comprehensive figure of the determinants of influenza infection. The transmission mechanisms of the virus are therefore unclear and a question remains of how this information can be used by policy makers to effectively plan future mitigation strategies.

Conceptually, the determinants condition two unobserved components causally linked to infection, namely exposure to the virus and host susceptibility. For example, the number of contacts or the use of preventive measures are factors related to exposure while haemagglutination antibody titer (HAI) titers, whether acquired naturally or after influenza vaccination, are factors related to host susceptibility. Other factors such as risk perception, beliefs and behaviors may act on both exposure and susceptibility, *e.g.* by increasing the social distance or use of vaccine.

In this study, a causal model with structural equations was used to explicitly model direct and indirect relationships between H1N1pdm09 infection and perception of infection risk, compliance with preventive behaviors, social contacts, indoor and outdoor environment, socio-demographic factors and pre-epidemic host susceptibility. Through the integration of

various risk factors known to play a role in H1N1pdm09 infection, we aimed to identify the most critical determinants, and thus, help prioritize clear targets for future public health interventions.

METHODS

Data source

We used data from the CoPanFlu-France cohort, created to study the risk of H1N1pdm09 infection. The cohort comprises 601 households (1450 subjects) randomly selected between December 2009 and July 2010 and actively monitored for influenza-like illness (ILI) during two consecutive influenza seasons (2010-11 and 2011-12). When ILI was reported (fever ≥ 37.8°C and cough and/or sore throat, with no other known cause [7]), nasal swabs were collected from every household member by a study nurse, and screened for respiratory viruses. Serum samples were collected for pre/post–season HAI titration. Questionnaires exploring medical history, risk perception, use of preventive measures, and duration and location of social contacts were administered at inclusion. Details on the study protocol, data collection and representativeness of the study households can be found elsewhere.[8] The study protocol of the study was approved by the research ethics committee "Comité de Protection des Personnes Ile-de-France 1". All participants gave their written informed consent. H1N1pdm09 infection was defined by either positive H1N1pdm09 RT-PCR or positive H1N1pdm09 RespiFinder assay on a nasal swab, or seroconversion (4-fold increase in the HAI titer). Among the 1450 individuals included in the study, 132 withdrew participation before the end of the 2010-11 season (17 for familial and/or health reasons and 64 due to study burden; 47 participants were lost to follow-up, and 4 died). This analysis therefore focused on 1318 subjects (559 households). Infection status was unavailable for 197

individuals (15%) because at least one serological sample was missing. These subjects were nevertheless included in the analysis of other covariates.

The hypothesized structural model

We postulated that influenza infection results from two unobserved phenomena, namely virus exposure and host susceptibility (Figure 1). Their simultaneous impact on the H1N1pdm09 infection risk was modeled using a structural equation model (SEM, see Supplementary File for an introduction to causal modeling with SEM). SEM allows estimating direct and indirect relationships between observed or latent variables. Latent variables are unobserved variables, estimated from several observed variables, called indicators.

We modeled exposure to the H1N1pdm09 virus with seven latent variables and one observed variable: the cumulative incidence of ILI in the relevant French administrative regions. Host susceptibility to H1N1pdm09 infection was modeled with a single indicator: the pre-epidemic HAI titer (see Supplementary File and Lapidus *et al.,*[9] for additional information on covariates).

Contact network

Seven continuous indicators were used for this latent variable: the daily number and duration of contacts with individuals younger than 15 years, between 15 and 50 years, and over 50 years, and the daily time spent in public transport. A contact was a verbal exchange (at least 3 words) between the subject and another individual or a physical contact. The duration and location of contacts, as well as the age of the individuals, were collected annually. For each subject we used all the contact data collected as close to and/or during the 2010-11 epidemic season (20 December 2010-20 February 2011) (*e.g.* if a subject had multiple pre- or post-epidemic contact data, we considered the information collected as close to the start and the end of the influenza season, respectively). The duration and number of individual contacts

were averaged over the repeated contact data and summed according to the contacts' age and were log-transformed (a value of 0.01 was imputed for subjects reporting a zero summed duration or number of contacts). The reported durations of a given contact were additional, *i.e.* a 20-minute contact with three individuals corresponded to a summed contact duration of 60 minutes.

Perception of the risk of infection and of preventive measures

Risk perception and opinions on preventive measures have been shown to impact individual preventive behaviors.[10,11] Two latent variables were used to model the subjects' perception and beliefs about H1N1pdm09 influenza based on responses to a dedicated questionnaire (see Figure 1 for covariates used). Specifically, all household subjects over 15 years indicated whether they "totally agree" / "partly agree" / "partly disagree"/ "totally disagree" to statements such as "prevention highly depends on behaviors" and "H1N1pdm influenza is most often fatal". For this analysis, the answers were dichotomized ("agree"/"disagree"). For household subjects <15 years, we imputed for each missing covariate the answer given most frequently by adults in the same household.

Compliance with preventive behaviors

Compliance with preventive measures has been shown to reduce the risk of H1N1pdm09 infection.[6,12] Three binary indicators were used: "always/often washes hands after coughing/sneezing", "always/often covers mouth while coughing or sneezing" and "daily frequency of handwashing (with soap or hand sanitizer) $\geq$ 5", dichotomized around the median daily frequency of handwashing.

Indoor characteristics

Six categorical indicators were used to describe living room and bedroom heating (electric/gas/other), average temperatures (dichotomized around their medians) and the presence of air humidifiers. A possible impact of relative humidity on influenza aerosol transmission has been reported, and high indoor heating levels, especially during winter months, may lead to continuous circulation of dry air, creating an environment favorable for the persistence of infectious viral particles.[13]

Neighborhood socioeconomic status

Geocoding of the subjects' addresses allowed us to assign each subject a geographic unit defined by INSEE (Institut National de la Statistique et des Etudes Economiques), IRIS zones corresponding to statistical block groups of about 2000 inhabitants.[14] Three continuous indicators, provided by INSEE, were used to describe neighborhood socioeconomic status: "employment rate (%) among 15-65 years", "proportion (%) of inhabitants >15 years without a diploma" and "mean annual income (k€)" (log-transformed).

Urban environment

Three binary indicators were used for this latent variable: "lives in a urban area" (defined by a 6-item variable provided by INSEE), "presence of agricultural land near habitation" and "presence of livestock near habitation".

Host susceptibility to H1N1pdm09 infection

We used the log-transformed pre-epidemic HAI titer to measure host susceptibility to H1N1pdm09 infection. The HAI titer is used by the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA) to evaluate vaccine protection.[15,16]

Additional observed covariates

We postulated that the subjects' age class (<15 years, 15-50 years, >50 years) would influence the contact network, compliance with preventive behaviors, and differential susceptibility. Factors associated with high pre-epidemic HAI titers, such as influenza vaccination and history of ILI during the previous season, were also investigated. Potential associations between age class, gender and vaccination were tested (relationships between observed covariates and observed/latent variables are shown on Supplementary File).

The ILI incidence in the relevant French administrative regions was used as a proxy of exposure. Using data from the French General Practitioner Sentinel Network we estimated for each French administrative region the cumulative weekly incidence of ILI (defined as abrupt onset of fever above 39°C with myalgia and respiratory signs) cases per 100,000 population for the 2009-10 (7 September 2009–27 December 2009) and 2010-11 (20 December 2010-20 February 2011) seasons.[17] While the cumulative incidence of ILI per region in 2010-11 was expected to be associated with the 2010-11 risk of infection, we assumed that the 2009-10 cumulative incidence of ILI would affect host susceptibility to H1N1pdm09.

Model estimation

Mplus 7 software was used to estimate the parameters of the structural model while accounting for within-household correlations. For easier interpretation of the coefficients, we report standardized estimates (β_{std}) with pvalues below 0.05. Standardized parameter estimates (ranging from -1 for a completely negative association to 1 for a completely positive association) can be interpreted with reference to other parameters in the model and relative strength of associations can be compared. Goodness of fit was assessed with two criteria: the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). The CFI ranges from 0 to 1, with values above 0.90 corresponding to an acceptable fit, while an RMSEA value below 0.08 is recommended.[18] To improve goodness of fit, new

relationships can be added following "modification indices" (approximations of the improvement in model fit when new structural equations are added) or non-significant and inconsistent relationships can be removed.

RESULTS

Data

Infection status could be determined for 1121 subjects (89 infections, 8%), which were used to estimate relationships with H1N1pdm09 infection risk. The proportion of males was 47% (612 subjects). Two hundred fifty-seven subjects were younger than 15 years (19%; 24 infections; 11%), 545 were aged 15-50 years (41%; 36 infected; 7%) and 516 were older than 50 years (60%; 29 infected; 7%). One hundred twenty-seven subjects received the 2010-11 seasonal influenza vaccine (10%), compared to 180 subjects vaccinated against H1N1pdm09 for the 2009-10 season (14%). Ninety-six subjects (7%) had a history of ILI during the 2009-10 season.

Structural-equation model of H1N1 infection

The final SEM model is shown in Figure 2 (coefficients of the relationships between latent variables and their indicators as well as a summary of the added and removed relationships are provided on Supplementary File). Host susceptibility to H1N1pdm09 and compliance with preventive behaviors were the only two factors directly associated with the risk of H1N1pdm09 infection (β_{std}=0.31, pvalue<0.001 and β_{std}=-0.21, pvalue=0.011). Compliance with preventive behaviors was positively influenced by the latent variables "risk perception of H1N1pdm09 infection" and "perception of preventive measures" (β_{std}=0.14, pvalue=0.029 and β_{std}=0.27, pvalue=0.001). Subjects <15 years and those aged 15-50 years were more exposed through contacts than were subjects over 50 years (β_{std}=0.82 and β_{std}=0.75,

pvalues<0.001). These two younger populations were less frequently vaccinated against H1N1pdm09 for the 2010-11 season (β_{std}=-0.26, pvalue=0.018 and β_{std}=-0.29, pvalue<0.001) and were more likely to have had ILI during the previous season (β_{std}=0.10, pvalue=0.022 and β_{std}=0.11, pvalue=0.017). Subjects <15 years were less compliant with preventive behaviors than the older age groups (β_{std}=-0.63, pvalue<0.001), and males were less compliant than females (β_{std}=-0.26, pvalue<0.001). Subjects <15 years were the least susceptible to H1N1pdm09 (β_{std}=-0.17, pvalue<0.001). Vaccination during both the 2009-10 and 2010-11 seasons, a history of ILI in the 2009-10 season, and a high cumulative incidence of ILI in the subjects' administrative regions (2009-10 season) were associated with lower host susceptibility to H1N1pdm09 infection (β_{std}=-0.29, pvalue=0.007; β_{std}=-0.19, pvalue<0.001; β_{std}=-0.07, pvalue=0.033 and β_{std}=-0.13, pvalue=0.018 respectively). Individuals vaccinated during the 2009-10 season were more likely to be vaccinated during the following season (β_{std}=0.34, pvalue=0.008). Individuals living in an urban area with "high" neighborhood socioeconomic status (*e.g.* a high mean income in the IRIS zone) were more likely to be vaccinated against pandemic influenza for the 2009-10 season (β_{std}=0.32, pvalue=0.006 and β_{std}=0.23, pvalue=0.010). The goodness-of-fit indices of the model were satisfactory (RMSEA=0.023, CFI=0.943).

DISCUSSION

Using structural equation modeling we were able to obtain a coherent quantitative picture of the complex mechanisms determining the risk of H1N1pdm09 infection during the 2010-11 influenza season. To our knowledge, our study is the first to simultaneously integrate factors as diverse as contacts, risk perception, preventive behaviors, and environmental and socioeconomic factors. Studies of risk factors reported since 2009 have mainly been based on regression analysis, which may fail to capture the complex mechanisms underlying infection

by estimating only direct, independent associations between covariates and H1N1pdm09 infection. Use of a structural-equation model allowed us to explicitly model direct and indirect relationships between highly diverse factors and their respective impacts on the infection risk. Our results are particularly relevant to informing public health strategies; they can be used to determine which dimensions should be priority targets of public health communication campaigns.

Host susceptibility, measured here with the pre-epidemic HAI titer, was the major factor explaining infection.[3,19] Influenza vaccination and history of ILI during the first H1N1pdm09 season, or level of community incidence, were factors associated with host susceptibility.[20]

Our results also suggest that compliance with preventive behaviors played an important role in influenza infection.[12] Compliance with preventive behaviors was related to positive perception of the impact of such measures and by risk perception of infection.[21] Lower preventive behaviors compliance among men compared with women has previously been observed.[11] The urban area of residence and neighborhood socioeconomic status had no direct influence on the infection risk, but both factors positively influenced the probability of vaccination for the 2009-10 season. Urban residence has been reported to have a positive influence on seasonal influenza vaccination.[22] No relationship between vaccination and high neighborhood socioeconomic status has previously been reported, but a high level of education has been linked to pandemic vaccine uptake,[23] as has living in a household where the head of family has high professional status.[24]

The infection rate during the second season was 8% in the CoPanFlu cohort, a figure consistent with estimated seroconversion rates of 6.2% and 6.8% reported by Chen *et al.* during the second and third H1N1pdm09 waves in Singapore.[25] Younger subjects (<15

years) were the least susceptible in our study (see Supplementary File). This may be explained by a higher rate of H1N1pdm09 infection during the 2009-10 season; 10% of young subjects in the CoPanFlu cohort reported ILI during this period. A similar relationship between a high rate of infection among children during the 2009-10 season and lower susceptibility in 2010-11 has been reported.[20]

No relationship was found between exposure through the contact network and H1N1pdm09 infection. Several studies have nevertheless linked school closures or holiday periods to a reduced infection rate,[26] implying a major influence of contacts with children on the infection risk. However, deeper analyses (see Supplementary File) did not suggest a higher level of exposure among infected subjects. There are two possible explanations of this finding. First, contact information was not necessarily collected at the time when influenza was present in the community – especially during winter months. We used a methodology similar to those of the POLYMOD study and obtained highly consistent findings regarding intensity of contacts according to age.[27] Contacts were also stable across the calendar seasons (see Supplementary File). However we cannot formally exclude that the pattern of contacts observed during the study might not be representative of contacts with infected subjects; contacts patterns/behaviors of susceptible individuals with infectious subjects may be modified in this context. Second, the daily number and duration of contacts were high in all age groups. Consequently the likelihood of contacts with one or several infectious subjects over the two-month influenza season might be so high that transmissibility would not be influenced by the number of occasions, but rather by the host characteristics.

Another limitation of our study related to the modeling of the host susceptibility, which was only based on the pre-epidemic HAI titer. Integration of alternative correlates of protection such as T-cell responses,[28] or of genetic factors,[29] may provide a more robust picture of host susceptibility.

Compared with previous studies, our work has two major strengths. First, households were randomly sampled through a probabilistic procedure and the selected households were shown to be representative of the general population (see Lapidus *et al.,*[8] for more details). Second, participants were prospectively followed through an active weekly surveillance system over the two consecutive influenza periods. Both limit the risk of selection and information bias.

To our knowledge, our study is also the first to simultaneously integrate such a variety of influenza risk factors through a causal approach. Structural equation models, however, cannot unequivocally prove causal relationships: acceptable indices of goodness-of-fit merely indicate that postulated relationships are supported by the data. Other models may provide equivalent or better performance, and covariates potentially involved in the phenomenon under study might have been omitted.[30]

Conclusion

Structural equation modeling should be considered as a valuable approach to steer public health policies. While our results rely on data collected from the 2010-11 influenza season, we feel the risk factors identified would be relevant in future epidemic and pandemic contexts. The nature of the influenza virus requires production of a new annual vaccine to lower individual susceptibility. Risk perception and compliance with preventive behaviors, though modifiable, are unlikely to change over time without targeted intervention. We suggest that communication campaigns focus on infection risk and preventive measures efficacy, specifically targeting men and children. Additionally, we find that measures based on social distancing should not be a primary public health mitigation strategy, but warrant further investigation.

What is already known on this subject?

- Since 2009, numerous studies have investigated risk factors for pandemic influenza infection, reporting factors as diverse as individual characteristics, indoor and outdoor environments, contacts with infected individuals and use of preventive measures.
- Questions remain about the respective roles of the reported determinants in the mechanism of influenza virus transmission.
- Knowledge on how this information can be used to guide policy makers to implement effective mitigation strategies is lacking.

What this study adds?

- Using a novel methodological approach we were able to rank the relative impacts of all the determinants of infection, thereby determining priority targets for public health communication campaigns. Pre-epidemic antibody titers, compliance with preventive behaviors, as well as risk perception, were the only factors directly or indirectly associated with the infection risk.
- Social behavior does not seem to be a major determinant of infection.
- Based on our findings, it is likely that public health communication campaigns stressing the efficacy of preventive measures and the risks associated with pandemic influenza infection would be effective to mitigate influenza burden.

Acknowledgements: We thank David Young for critical review of the manuscript.

Conflicts of interest: Pr Carrat reported not having shares or paid employment with pharmaceutical companies; received honoraria from Novartis, GlaxoSmithKline, Astra Zeneca and Boiron and received travel support to attend scientific meetings from Novartis. Other authors have no competing interest to declare.

Funding: This study was supported by the Institut de Microbiologie et Maladies Infectieuses (IMMI-AVIESAN), the French Ministry of Health and Assistance Publique Hôpitaux de Paris - PHRC 2010 #AOM10199, the French Ministry of Research and the Institut de Recherche en Sante Publique (IReSP – TGIR 2009). The funders had no role in study design, in collection, analysis and interpretation of data, or writing of the report.

REFERENCES

1 France AM, Jackson M, Schrag S, *et al.* Household transmission of 2009 influenza A (H1N1) virus after a school-based outbreak in New York City, April-May 2009. *J Infect Dis* 2010;**201**:984–92. doi:10.1086/651145

2 Sugimoto JD, Borse NN, Ta ML, *et al.* The effect of age on transmission of 2009 pandemic influenza A (H1N1) in a camp and associated households. *Epidemiology* 2011;**22**:180–7. doi:10.1097/EDE.0b013e3182060ca5

3 Chao D-Y, Cheng K-F, Li T-C, *et al.* Factors associated with infection by 2009 pandemic H1N1 influenza virus during different phases of the epidemic. *Int J Infect Dis* 2011;**15**:e695–701. doi:10.1016/j.ijid.2011.05.012

4 Cauchemez S, Donnelly CA, Reed C, *et al.* Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *N Engl J Med* 2009;**361**:2619–27.

5 Lau LLH, Nishiura H, Kelly H, *et al.* Household transmission of 2009 pandemic influenza A (H1N1): a systematic review and meta-analysis. *Epidemiology* 2012;**23**:531–42. doi:10.1097/EDE.0b013e31825588b8

6 Simmerman JM, Suntarattiwong P, Levy J, *et al.* Findings from a household randomized controlled trial of hand washing and face masks to reduce influenza transmission in Bangkok, Thailand. *Influenza Other Respi Viruses* 2011;**5**:256–67. doi:10.1111/j.1750-2659.2011.00205.x

7 CDC. Flu activity and surveillance: reports and surveillance methods in the United States. http://www.cdc.gov/flu/weekly/fluactivitysurv.htm (accessed 31 Mar2014).

8 Lapidus N, De Lamballerie X, Salez N, *et al.* Integrative study of pandemic A/H1N1 influenza infections: design and methods of the CoPanFlu-France cohort. *BMC Public Health* 2012;**12**:417. doi:10.1186/1471-2458-12-417

9 Lapidus N, de Lamballerie X, Salez N, *et al.* Factors associated with post-seasonal serological titer and risk factors for infection with the pandemic A/H1N1 virus in the French general population. *PLoS One* 2013;**8**:e60127. doi:10.1371/journal.pone.0060127

10 Liao Q, Cowling BJ, Lam WWT, *et al.* Factors affecting intention to receive and self-reported receipt of 2009 pandemic (H1N1) vaccine in Hong Kong: a longitudinal study. *PLoS One* 2011;**6**:e17713. doi:10.1371/journal.pone.0017713

11 Lin Y, Huang L, Nie S, *et al.* Knowledge, attitudes and practices (KAP) related to the pandemic (H1N1) 2009 among Chinese general population: a telephone survey. *BMC Infect Dis* 2011;**11**:128. doi:10.1186/1471-2334-11-128

12 Gordon SM. Update on 2009 pandemic influenza A (H1N1) virus. *Cleve Clin J Med* 2009;**76**:577–82.

13 Lofgren E, Fefferman NH, Naumov YN, *et al.* Influenza seasonality: underlying causes and modeling theories. *J Virol* 2007;**81**:5429–36. doi:10.1128/JVI.01680-06

14 INSEE. Découpage infra-communal en IRIS. http://www.insee.fr/fr/methodes/default.asp?page=zonages/iris.htm (accessed 11 Nov2013).

15 EMA. European Medicines Agency - Committee for proprietary medicinal products. Note for guidance on harmonization of requirements for influenza vaccines (CPMP/BWP/214/96). 1997.http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003945.pdf (accessed 4 Oct2012).

16 FDA. U.S. Food and Frug Administration - Guidance for Industry: Clinical Data Needed to Support the Licensure of Seasonal Inactivated Influenza Vaccines. 2009.http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/Vaccines/ucm074794.htm (accessed 4 Oct2012).

17 Carrat F, Flahault A, Boussard E, *et al.* Surveillance of influenza-like illness in France. The example of the 1995/1996 epidemic. *J Epidemiol Community Health* 1998;**52 Suppl 1**:32S–38S.

18 Byrne BM. *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge Academic New York 2011.

19 Kieffer A, Paboriboune P, Crépey P, *et al.* 2009 A(H1N1) seroconversion rates and risk factors among the general population in Vientiane Capital, Laos. *PLoS One* 2013;**8**:e61909. doi:10.1371/journal.pone.0061909

20 Hoschler K, Thompson C, Andrews N, *et al.* Seroprevalence of influenza A(H1N1)pdm09 virus antibody, England, 2010 and 2011. *Emerg Infect Dis* 2012;**18**:1894–7. doi:10.3201/eid1811.120720

21 Rubin GJ, Amlôt R, Page L, *et al.* Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: cross sectional telephone survey. *BMJ* 2009;**339**:b2651.

22 Andrew MK, McNeil S, Merry H, *et al.* Rates of influenza vaccination in older adults and factors associated with vaccine use: a secondary analysis of the Canadian Study of Health and Aging. *BMC Public Health* 2004;**4**:36. doi:10.1186/1471-2458-4-36

23 Schwarzinger M, Flicoteaux R, Cortarenoda S, *et al.* Low acceptability of A/H1N1 pandemic vaccination in French adult population: did public health policy fuel public dissonance? *PLoS One* 2010;**5**:e10199. doi:10.1371/journal.pone.0010199

24 Vaux S, Van Cauteren D, Guthmann J-P, *et al.* Influenza vaccination coverage against seasonal and pandemic influenza and their determinants in France: a cross-sectional survey. *BMC Public Health* 2011;**11**:30. doi:10.1186/1471-2458-11-30

25 Chen MIC, Cook AR, Lim WY, *et al.* Factors influencing infection by pandemic influenza A(H1N1)pdm09 over three epidemic waves in Singapore. *Influenza Other Respi Viruses* Published Online First: 5 July 2013. doi:10.1111/irv.12129

26 Cauchemez S, Valleron A-J, Boëlle P-Y, *et al.* Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature* 2008;**452**:750–4. doi:10.1038/nature06732

27 Mossong J, Hens N, Jit M, *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* 2008;**5**:e74. doi:10.1371/journal.pmed.0050074

28 McElhaney JE, Xie D, Hager WD, *et al.* T cell responses are better correlates of vaccine protection in the elderly. *J Immunol* 2006;**176**:6333–9. doi:176/10/6333 [pii]

29 Franco LM, Bucasas KL, Wells JM, *et al.* Integrative genomic analysis of the human immune response to influenza vaccination. *Elife* 2013;**2**:e00299–e00299. doi:10.7554/eLife.00299

30 DiLalla LF. A structural equation modeling overview for medical researchers. *J Dev Behav Pediatr* 2008;**29**:51–4. doi:10.1097/DBP.0b013e31815f250c

FIGURE LEGENDS

Figure 1 Hypothesized structural equation model. Ellipses: latent variables; boxes: observed variables. Dotted background: latent variables related to exposure to H1N1pdm09. Striped background: latent variable related to susceptibility to H1N1pdm09 infection. For clarity, relationships involving the additional observed variables are not shown

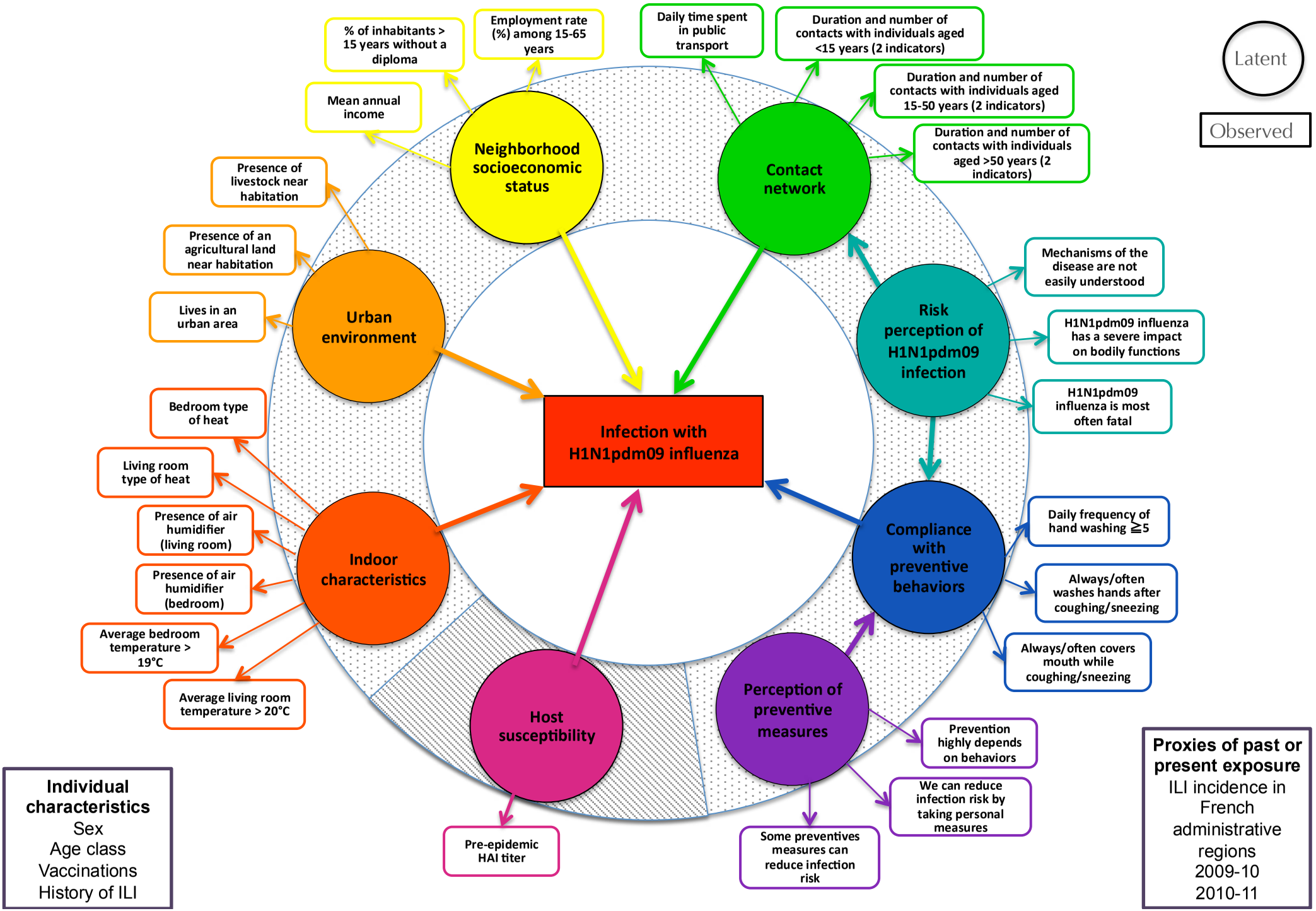
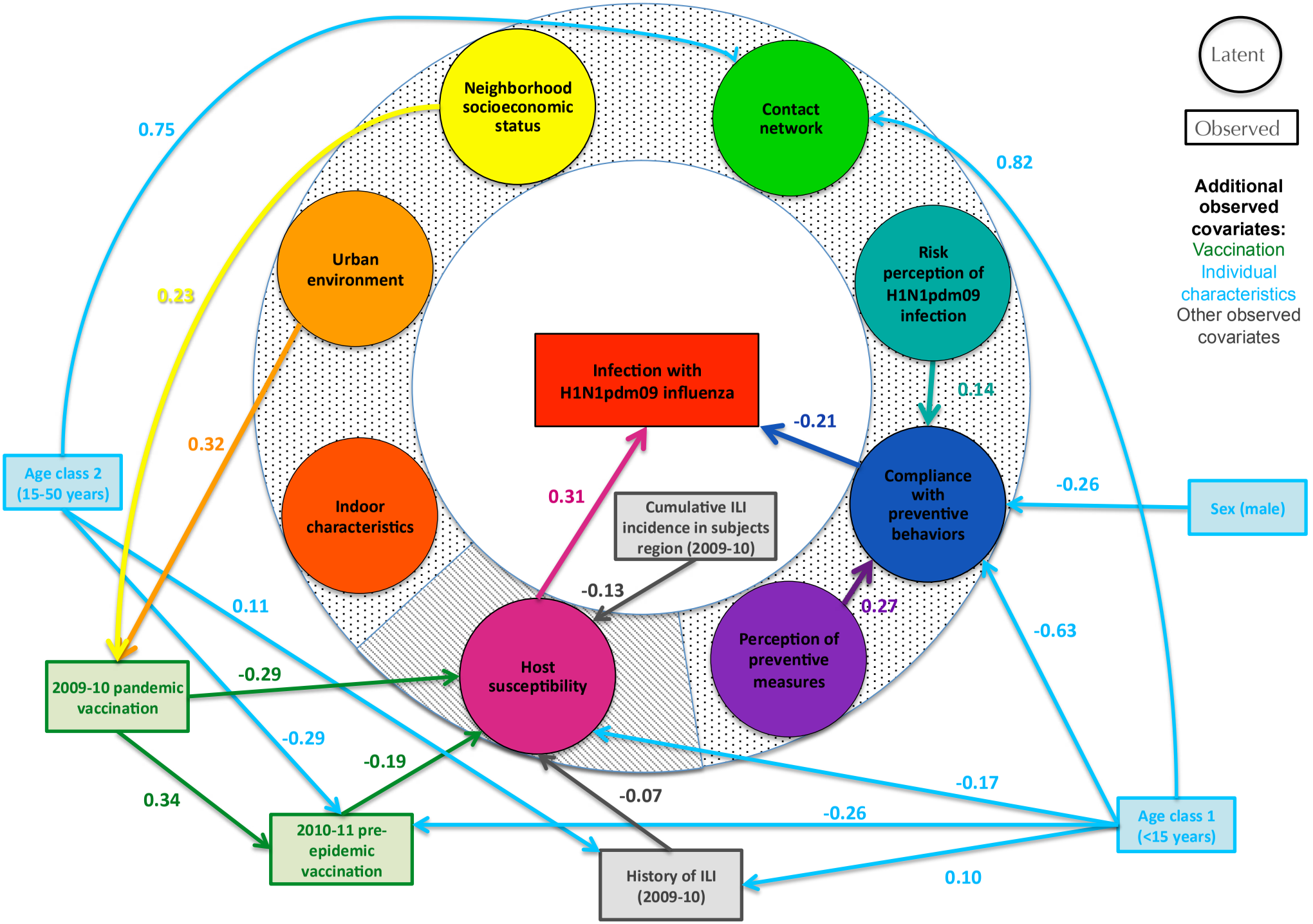

Latent
Observed
Neighborhood socioeconomic status
% of inhabitants > 15 years without a diploma
Employment rate (%) among 15-65 years
Mean annual income
Contact network
Daily time spent in public transport
Duration and number of contacts with individuals aged <15 years (2 indicators)
Duration and number of contacts with individuals aged 15-50 years (2 indicators)
Duration and number of contacts with individuals aged >50 years (2 indicators)
Urban environment
Presence of livestock near habitation
Presence of an agricultural land near habitation
Lives in an urban area
Risk perception of H1N1pdm09 infection
Mechanisms of the disease are not easily understood
H1N1pdm09 influenza has a severe impact on bodily functions
H1N1pdm09 influenza is most often fatal
Infection with H1N1pdm09 influenza
Indoor characteristics
Bedroom type of heat
Living room type of heat
Presence of air humidifier (living room)
Presence of air humidifier (bedroom)
Average bedroom temperature > 19°C
Average living room temperature > 20°C
Compliance with preventive behaviors
Daily frequency of hand washing ≧5
Always/often washes hands after coughing/sneezing
Always/often covers mouth while coughing/sneezing
Host susceptibility
Pre-epidemic HAI titer
Perception of preventive measures
Prevention highly depends on behaviors
We can reduce infection risk by taking personal measures
Some preventives measures can reduce infection risk
Individual characteristics
Sex
Age class
Vaccinations
History of ILI
Proxies of past or present exposure
ILI incidence in French administrative regions
2009-10
2010-11

Figure 2 Final structural equation model. All coefficients have pvalues<0.05. Ellipses: latent variables; boxes: observed variables. Dotted background: latent variables related to exposure to H1N1pdm09 virus. Striped background: latent variable related to susceptibility to H1N1pdm09 infection


Latent
Observed
Additional observed covariates:
Vaccination
Individual characteristics
Other observed covariates
Neighborhood socioeconomic status
Contact network
Urban environment
Risk perception of H1N1pdm09 infection
Infection with H1N1pdm09 influenza
Indoor characteristics
Compliance with preventive behaviors
Cumulative ILI incidence in subjects region (2009-10)
Host susceptibility
Perception of preventive measures
2009-10 pandemic vaccination
2010-11 pre-epidemic vaccination
History of ILI (2009-10)
Age class 2 (15-50 years)
Age class 1 (<15 years)
Sex (male)
0.75
0.82
0.23
0.32
0.14
-0.21
0.31
-0.26
0.11
-0.13
0.27
-0.63
-0.29
-0.29
-0.19
-0.17
0.34
-0.07
-0.26
0.10

Causal analysis of H1N1pdm09 influenza infection risk in a household cohort (Supplementary File)

Causal modeling with structural equations

Structural equation modeling (SEM) is a multivariate statistical method frequently used in social sciences [1,2]. This hypothesis-driven approach can be used to test a series of structural equations (*i.e.* relationships between variables) representing a theoretical causal process. Two types of variables, observed and latent, are included in SEM models. Latent variables are unobserved variables that cannot be measured directly [3]. They are estimated from several observed variables, called indicators, presumed to represent an underlying unobserved phenomenon [4]. Latent and observed variables can be classified as exogenous or endogenous. The former are independent variables not influenced by the other variables of the model, while the latter are dependent variables [5,6]. The set of hypotheses of a given SEM can be summarized with a path diagram (Figure 1). The SEM model is composed of two compartments: the measurement model and the structural model [7]. The first shows the relations between the latent variables and their indicators; arrows going from the latent variable to its respective indicators are called factor loadings and represent correlation coefficients [8]. The structural model represents the relations between the latent variables. The SEM model is a combination of the measurement model, the structural model, and potential endogenous or exogenous observed variables.

Additional information on covariates used in the SEM model

A summary of the covariates included in the SEM model is provided on table 1. Table 2 shows the correlations between the covariates included in the measurement model.

HAI titers of the subjects included in the study

Pre-epidemic HAI titers depending on subjects' infection status and age class are shown on Figure 2. For each age class, we observed that the non-infected subjects showed higher log pre-epidemic HAI titers than the infected individuals (Wilcoxon rank-sum test $P<0.001$).

Model estimation

As we included categorical dependent variables, we used the weighted least squares with mean and variance adjustment (WLSMV) estimator; standard errors for the standardized path coefficients are not provided with this estimator. Missing data were not imputed. All available information was used: a subject with a missing value for a given covariate was omitted when

estimating relationships involving this covariate, but available information for the same subject was used to estimate the other relationships.

Modifications of the hypothesized structural-equation model

Prior to the addition of the H1N1pdm09 infection outcome and of additional observed covariates, the first required step was the measurement model validation (Figure 1 in the manuscript). The measurement model had good fit indices (RMSEA=0.024, CFI=0.962). However, the factor loading of the indicator "time spent in public transport", as well as those of the average living room and bedroom temperatures, were not significant, and these indicators were therefore removed from the model (see Table 3 for the coefficients of the measurement model).

There were several differences between the final structural equation model (Figure 2 in the manuscript, RMSEA=0.023, CFI=0.943) and our hypothesized model (Figure 3, RMSEA=0.034, CFI=0.868). Using modifications indices, we added 2 regression paths: the regression of the latent variable of the compliance with preventive behaviors on the observed covariate "sex", and the regression of the 2010-11 pre-epidemic vaccination status on the 2009-10 pandemic vaccination status. We also added correlations between measurement error terms (measurement errors refer to the indicator's adequacy in measuring their underlying latent variable) for the daily duration and number of contacts for the three age classes of contacts considered (<15 years, 15-50 years and > 50 years).

Regarding the estimated coefficients, non-significant relations were removed. No association was observed between the age class and the pandemic vaccination status for 2009-10. Sex was not associated with vaccination for both the 2009-2010 and 2010-2011 seasons. Risk perception of H1N1pdm09 infection was not associated with the contact network latent variable. Latent variables describing the urban area and the neighborhood socio-economic status were not associated with 2010-2011 pandemic vaccination. Finally, four latent variables ("Indoor characteristics", "Urban environment", "Neighborhood socioeconomic status" and "Contact network") were not associated with the infection outcome, nor was the cumulative ILI incidence for 2010- 2011 season.

Additional analysis about duration and number of contacts

We further investigated the absence of a significant relationship between exposure through contact network and H1N1pdm09 infection (P=0.340). Exploration of the daily number and duration of contacts depending on subjects' age class and infection status (Figure 4) showed that subjects mainly reported contacts with individuals belonging to their age class. However, no significant differences (with Wilcoxon sum-rank tests) were observed between infected and non-infected subjects for such contacts. Non-infected subjects younger than 15 years of age had a higher daily number and duration of contacts with subjects 15 to 50 years old (P <0.05 and P <0.05) and with individuals over 50 years old (P <0.001 and P <0.001) than infected subjects. The same observation was made for the subjects aged between 15 and 50 years concerning the daily number and duration of contacts with individuals aged younger than 15 years (P <0.05 for the number of contacts and P <0.05 for the duration of contacts). We also observed that contact patterns were stable across the calendar seasons (Kruskal-Wallis rank sum test P for the daily duration of contacts=0.670 and Kruskal-Wallis rank sum test P for the daily number of contacts=0.39 – see Figure 5).

References

1 González J, De Boeck P, Tuerlinckx F. A double-structure structural equation model for three-mode data. *Psychol Methods* 2008;**13**:337–53. doi:10.1037/a0013269

2 Bollen KA, Noble MD. Structural equation models and the quantification of behavior. *Proc Natl Acad Sci U S A* 2011;**108 Suppl**:15639–46. doi:10.1073/pnas.1010661108

3 Kline RB. *Principles and practice of structural equation modeling*. Third. The Guilford Press 2010. doi:10.1038/156278a0

4 Byrne BM. *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge Academic New York 2011.

5 Geiser C. *Data Analysis with Mplus*. Guilford Press 2013.

6 Amorim LDAF, Fiaccone RL, Santos CAST, *et al.* Structural equation modeling in epidemiology. *Cad saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Esc Nac Saúde Pública* 2010;**26**:2251–62.

7 Stein CM, Morris NJ, Nock NL. Structural equation modeling. *Methods Mol Biol* 2012;**850**:495–512. doi:10.1007/978-1-61779-555-8_27

8 Beran TN, Violato C. Structural equation modeling in medical research: a primer. *BMC Res Notes* 2010;**3**:267. doi:10.1186/1756-0500-3-267

Tables

Table 1 Description of the observed covariates included in the structural equation model. All covariates are categorical except quantitative (q) and log-transformed quantitative covariates (l). (i) refers to covariates describing subjects' IRIS zone. * : for categorical covariates the number of classes is shown, for quantitative covariates mean and standard deviation range are shown.

Observed variable	Number of classes / Mean(SD)*	Missing values : N (%)
Prevention highly depends on behaviors	2 (agree/disagree)	168 (13%)
Some preventive measures can reduce infection risk	2 (agree/disagree)	157 (12%)
We can reduce infection risk by taking personal measures	2 (agree/disagree)	126 (9%)
H1N1pdm09 influenza is most often fatal	2 (agree/disagree)	165 (12%)
Mechanisms of the disease are not easily understood	2 (agree/disagree)	222 (17%)
H1N1pdm09 influenza has a severe impact on bodily functions	2 (agree/disagree)	288 (22%)
Pre-epidemic HAI titer (q) (l)	3.65 (0.66)	30 (2%)
Duration of contacts with individuals aged <15 years (q) (l) (min)	3.24 (4.70)	0 (0%)
Duration of contacts with individuals aged 15-50 years (q) (l) (min)	6.07 (1.61)	0 (0%)
Duration of contacts with individuals aged >50 years (q) (l) (min)	4.74 (2.89)	0 (0%)
Daily time spent in public transport (q) (l) (min)	-2.39 (4.08)	0 (0%)
Number of contacts with individuals aged <15 years (q) (l)	-0.46 (2.52)	0 (0%)
Number of contacts with individuals aged 15-50 years (q) (l)	1.38 (1.02)	0 (0%)

Number of contacts with individuals aged >50 years (q) (l)	0.39 (1.59)	0 (0%)
Always/often washes hands after coughing/sneezing	2 (yes/no)	23 (2%)
Daily frequency of hand washing ≧5	2 (yes/no)	37 (3%)
Always/often covers mouth while coughing/sneezing	2 (yes/no)	16 (1%)
Employment rate (%) among 15-65 years (q) (i)	63.63 (7.50)	0 (0%)
% of inhabitants >15 years without a diploma (q) (i)	12.87 (5.90)	0 (0%)
Mean annual income (q) (l) (i) (k€)	10.06 (0.28)	0 (0%)
Living room type of heat	3 (electric/gas/other)	145 (11%)
Bedroom type of heat	3 (electric/gas/other)	185 (14%)
Presence of air humidifier (bedroom)	2 (yes/no)	0 (0%)
Presence of air humidifier (living room)	2 (yes/no)	1 (1%)
Average bedroom temperature >19°C	2 (yes/no)	120 (9%)
Average living room temperature >20°C	2 (yes/no)	117 (9%)
Age class	3 (<15 years, 15-50 years, >50 years)	0 (0%)
Sex (male)	2 (yes/no)	0 (0%)
2009-10 pandemic vaccination	2 (yes/no)	2 (1%)
2010-11 pre-epidemic vaccination	2 (yes/no)	15 (1%)
History of ILI (2009-10)	2 (yes/no)	48 (4%)
Cumulative ILI incidence in subjects region (2009-10) (q)	5165 (1531)	0 (0%)
Cumulative ILI incidence in subjects region (2010-11) (q)	3391 (1152)	0 (0%)
Infection with H1N1pdm09 influenza	2 (yes/no)	197 (15%)

Table 2. Correlations between the covariates included in the measurement model. All covariates are categorical except quantitative (q) and log-transformed quantitative covariates (l). (i) refers to covariates describing subjects' IRIS zone. Matrix diagonal shows variances of quantitative covariates and proportions of ‘‘ones’’ observed for dichotomous categorical covariates. For categorical covariates with more than 2 modalities (covariates 24-25), the proportion of the first modality is shown.

Covariates	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Prevention highly depends on behaviors 1	83.68																												
Some preventive measures can reduce infection risk 2	0.41	94.61																											
We can reduce the infection risk by taking personal measures 3	0.48	0.71	95.68																										
H1N1pdm09 influenza is most often fatal 4	-0.05	0.08	0.1	10.27																									
Mechanisms of the disease are not easily understood 5	0.06	-0.09	-0.09	0.22	67.12																								
H1N1pdm09 influenza has a severe impact on bodily functions 6	0.1	0.22	0.19	0.56	0.24	53.41																							
Pre-epidemic HAI titer (q) (l) 7	0.01	0	0.03	-0.03	0.06	0	0.45																						
Duration of contacts with indiv. aged <15 years (q) (l) 8	0.05	0.08	-0.02	0.03	-0.14	-0.05	0.04	22.09																					
Duration of contacts with indiv. aged 15-50 years (q) (l) 9	0.06	-0.13	-0.08	-0.08	-0.15	-0.08	-0.04	0.3	2.62																				
Duration of contacts with indiv. aged >50 years (q) (l) 10	0.01	0.08	-0.06	-0.02	0.04	-0.04	0.01	-0.11	-0.09	8.36																			
Daily time spent in public transport (q) (l) 11	-0.03	0	-0.16	0	0.05	-0.08	0.1	0.05	0.07	0.03	16.67																		
Number of contacts with indiv. aged <15 years (q) (l) 12	0.08	0.08	0	0.02	-0.14	-0.05	0.06	0.98	0.29	-0.11	0.05	6.4																	
Number of contacts with indiv. aged 15-50 years (q) (l) 13	0.08	-0.06	-0.01	-0.1	-0.16	-0.08	-0.03	0.19	0.85	-0.05	0.09	0.19	1.05																
Number of contacts with indiv. aged >50 years (q) (l) 14	0.02	0.09	-0.01	-0.04	0.03	-0.03	-0.01	-0.12	-0.11	0.95	0.03	-0.12	-0.02	2.54															
Always/often washes hands after coughing/sneezing 15	0.07	0.13	0.15	0.22	0.05	0.26	-0.07	-0.11	0.01	0.02	-0.02	-0.14	0	0.04	32.2														
Daily frequency of hand washing >=5 16	0.03	0.2	0.1	0.02	0.15	0.07	-0.08	-0.19	-0.14	0.12	-0.01	-0.21	-0.04	0.15	0.42	63.15													
Always/often covers mouth while coughing/sneezing 17	0.03	0.11	0.05	0.08	0.08	0.02	-0.07	-0.3	-0.1	0.16	0.12	-0.32	0.05	0.2	0.4	0.41	84.18												
Lives in a urban area 18	-0.01	0.13	-0.13	0.03	0.02	0.01	0.12	-0.05	0	-0.02	0.01	-0.05	0.01	-0.07	-0.06	0.06	0	78.22											
Presence of an agricultural land near habitation 19	0.08	0.04	0.07	-0.22	-0.06	0.03	-0.06	0.07	0.02	0.06	-0.03	0.08	0.04	0.1	0.04	0.03	0.03	-0.62	50.99										
Presence of livestock near habitation 20	0.06	-0.2	0.19	0.13	-0.04	0.18	-0.05	0.07	-0.02	0.04	-0.02	0.07	-0.02	0.05	0.02	-0.05	0.03	-0.59	0.62	5.24									
% of employment among 15-65 years (q) (i) 21	0.06	-0.02	0.07	-0.02	-0.03	-0.08	-0.06	0.03	0.02	0.14	0.03	0.03	0.04	0.14	0.02	0.01	0.07	-0.48	0.31	0.37	56.25								
% of inhabitants > 15 years without a diploma (q) (i) 22	-0.12	0.07	0	0.13	0.12	0.05	0	0.01	-0.03	-0.16	-0.06	0.01	-0.07	-0.17	0.02	0.06	-0.09	0.23	-0.14	-0.16	-0.55	34.91							
Mean annual income (q)(l)(i) (k€) 23	0.16	0.22	0.19	0	-0.06	-0.05	-0.03	0.03	0.01	0.07	0.04	0.04	0.05	0.07	0	0	0.11	-0.04	-0.02	0.09	0.39	-0.43	0.08						
Living room type of heat 24	-0.14	-0.06	0.05	-0.11	0.04	0	-0.04	0.01	-0.03	0.07	0.01	0.01	-0.01	0.07	0.01	0.01	0.13	-0.14	0.13	0.07	0.02	0.04	-0.05	30.26					
Bedroom type of heat 25	-0.1	0.08	0	-0.09	0.04	0.01	0.02	0.01	-0.02	0.03	-0.01	0.02	0.01	0.05	-0.02	0.02	0.08	-0.11	0.16	0.14	-0.01	0.04	-0.07	0.9	32.74				
Presence of air humidifier (bedroom) 26	0.15	-0.08	0.04	-0.17	0.09	0.13	0.05	-0.07	-0.07	-0.08	-0.15	-0.07	-0.09	-0.07	0.1	0.03	-0.02	0.03	-0.07	-0.3	0.06	0.03	-0.05	0.1	0.16	6.83			
Presence of air humidifier (living room) 27	0.31	0.07	0.15	-0.2	0.21	0.34	-0.05	0.06	-0.01	-0.2	-0.14	0.07	0.04	-0.18	0.12	-0.04	-0.01	0.07	-0.09	-0.02	-0.07	0.19	-0.24	-0.01	0.11	0.76	4.7		
Average bedroom temperature >19°C 28	-0.07	0.04	-0.07	0.18	-0.03	0.09	-0.02	0.01	-0.04	-0.19	-0.03	0.01	-0.09	-0.2	-0.01	-0.08	-0.11	0.29	-0.35	-0.11	-0.14	0.13	0.12	-0.09	-0.08	0.08	-0.03	27.73	
Average living room temperature >20°C 29	-0.17	0.04	0.1	0.12	0.04	0.18	-0.01	0	-0.12	-0.11	-0.03	-0.03	-0.17	-0.12	0.04	0.08	0.04	-0.04	-0.08	-0.05	-0.03	0.13	-0.03	-0.03	-0.11	0.12	-0.09	0.48	25.63

Table 3. Coefficients of the SEM measurement model.

Latent variable	Indicator	Unstandardized coefficient	Standard error	pvalue	Standardized coefficient
Perception of preventive measures	Prevention highly depends on behaviors	0.606	0.115	<0.001	0.545
	Some preventive measures can reduce infection risk	0.878	0.203	<0.001	0.79
	We can reduce the infection risk by taking personal measures	1	NA	NA	0.899
Risk perception of H1N1pdm09 infection	H1N1pdm09 influenza is most often fatal	1	NA	NA	0.598
	Mechanisms of the disease are not easily understood	0.436	0.129	0.001	0.261
	H1N1pdm09 influenza has a severe impact on bodily functions	1.591	0.545	0.003	0.951
Host susceptibility	Pre-epidemic HAI titer	-1	NA	NA	-1
Contact network	Duration of contacts with indiv. aged <15 years	1	NA	NA	0.481
	Duration of contacts with indiv. aged 15-50 years	0.417	0.049	<0.001	0.547
	Duration of contacts with indiv. aged >50 years	-0.336	0.099	0.001	-0.259
	Number of contacts with indiv. aged <15 years	0.532	0.026	<0.001	0.484
	Number of contacts with indiv. aged 15-50 years	0.201	0.026	<0.001	0.427
	Number of contacts with indiv. aged >50 years	-0.224	0.05	<0.001	-0.314
Compliance with preventive	Always/often washes hands after	0.822	0.097	<0.001	0.583

behaviors	coughing/sneezing				
	Daily frequency of hand washing >=5	0.962	0.122	<0.001	0.663
	Always/often covers mouth while coughing/sneezing	1	NA	NA	0.683
Urban environment	Lives in a urban area	1	NA	NA	0.889
	Presence of an agricultural land near habitation	-0.789	0.114	<0.001	-0.701
	Presence of livestock near habitation	-0.848	0.134	<0.001	-0.754
Neighborhood socioeconomic status	% of employment among 15-65 years	1	NA	NA	0.786
	% of inhabitants > 15 years without a diploma	-0.753	0.057	<0.001	-0.742
	Mean annual income	0.024	0.003	<0.001	0.489
Indoor characteristics	Bedroom type of heat	1	NA	NA	0.98
	Living room type of heat	0.938	0.087	<0.001	0.919
	Presence of air humidifier (bedroom)	0.369	0.09	<0.001	0.362
	Presence of air humidifier (living room)	0.415	0.103	<0.001	0.406

Figures

Figure 1 Example of a path diagram. Ellipses: latent variables; boxes: observed variables.

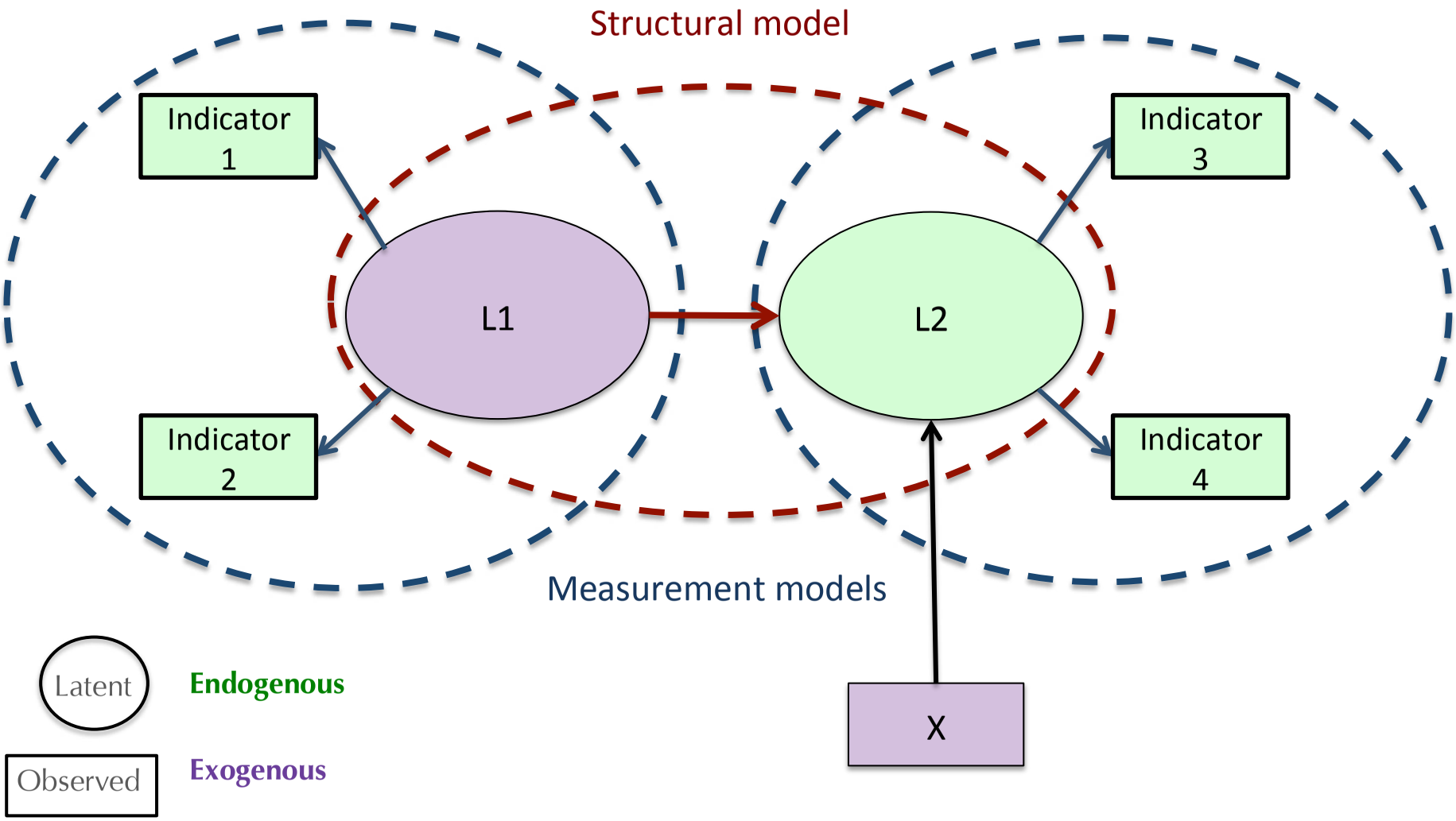

Structural model
Indicator 1
Indicator 2
L1
L2
Indicator 3
Indicator 4
Measurement models
X
Latent
Endogenous
Observed
Exogenous

Figure 2 Pre-epidemic immunity depending on subjects' infection status and age class. y-axis shows the proportion of subjects with pre-epidemic HAI titers above or equal to the HAI titers on the x-axis.

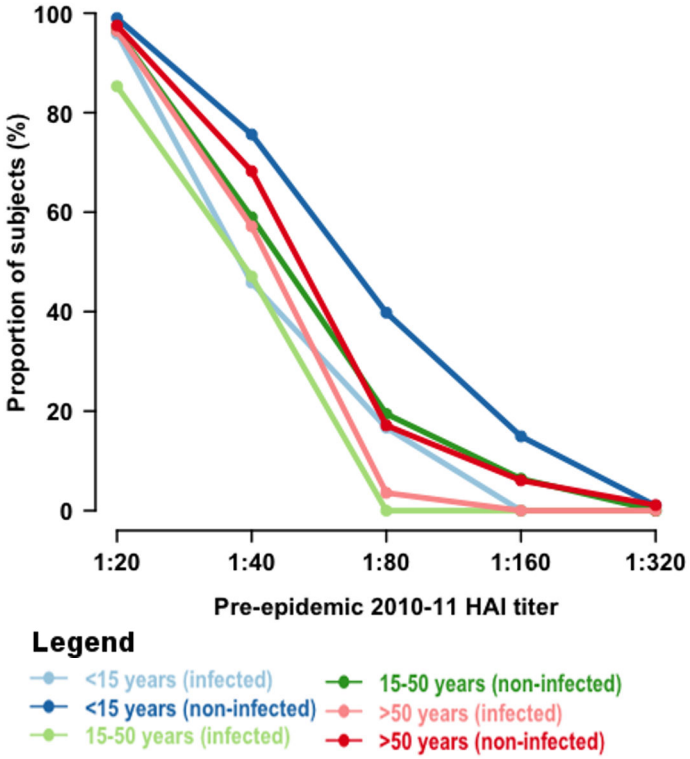

100
80
60
40
20
0
Proportion of subjects (%)
1:20
1:40
1:80
1:160
1:320
Pre-epidemic 2010-11 HAI titer
Legend
<15 years (infected)
<15 years (non-infected)
15-50 years (infected)
15-50 years (non-infected)
>50 years (infected)
>50 years (non-infected)

Figure 3 Hypothesized relationships between latent variables and exogenous/endogenous observed variables. Ellipses: latent variables; boxes: observed variables. Dotted background: latent variables related to exposure to H1N1pdm09. Striped background: latent variable related to susceptibility to H1N1pdm09 infection. For clarity, latent variables indicators are not shown.

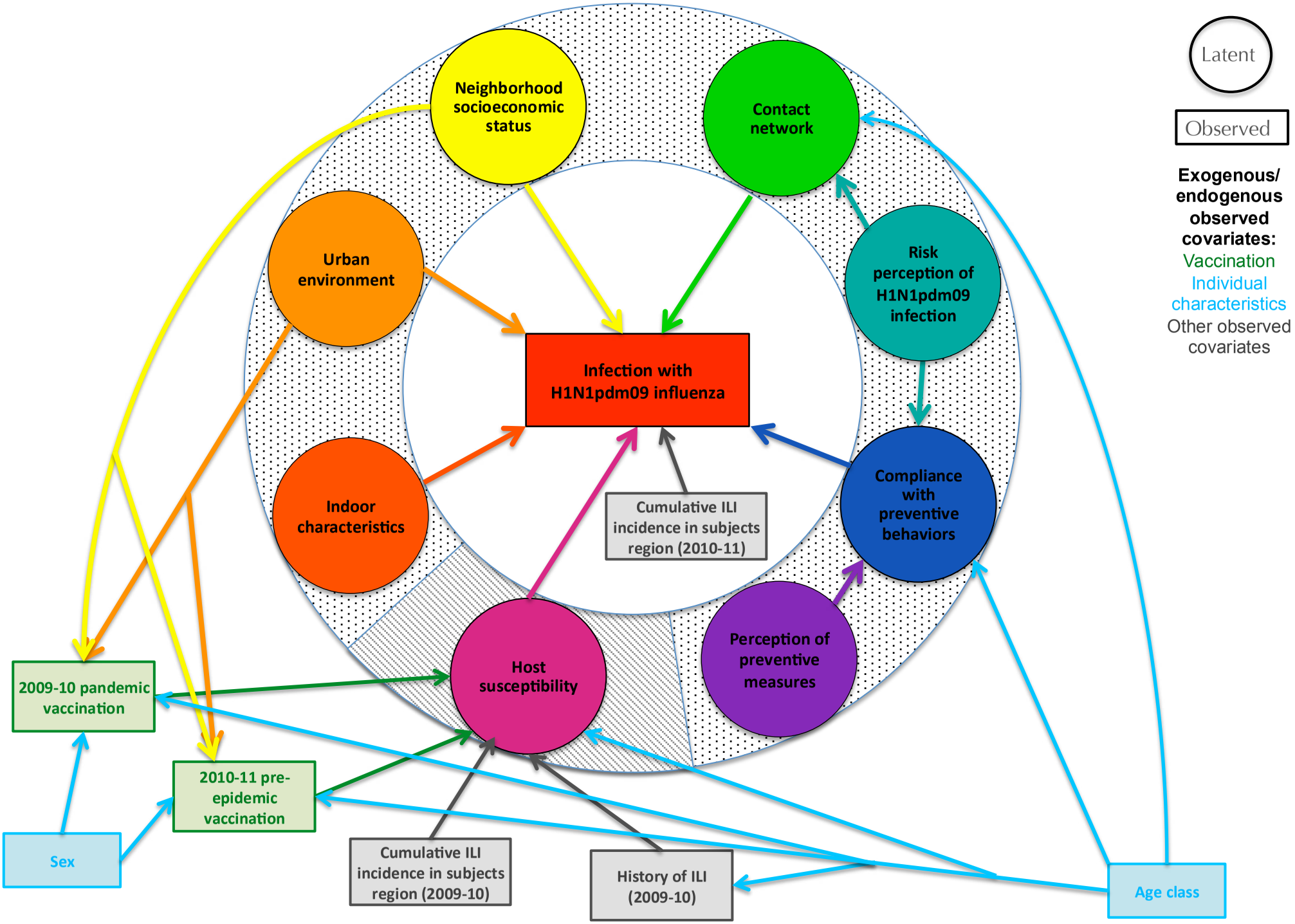

Latent
Observed
Exogenous/ endogenous observed covariates:
Vaccination
Individual characteristics
Other observed covariates
Neighborhood socioeconomic status
Contact network
Urban environment
Risk perception of H1N1pdm09 infection
Infection with H1N1pdm09 influenza
Indoor characteristics
Cumulative ILI incidence in subjects region (2010-11)
Compliance with preventive behaviors
Host susceptibility
Perception of preventive measures
2009-10 pandemic vaccination
2010-11 pre-epidemic vaccination
Sex
Cumulative ILI incidence in subjects region (2009-10)
History of ILI (2009-10)
Age class

Figure 4 Contact patterns depending on subjects' infection status and age class. y-axis shows the proportion of subjects with daily number/duration of contacts above or equal to the daily number/duration of contacts on the x-axis.

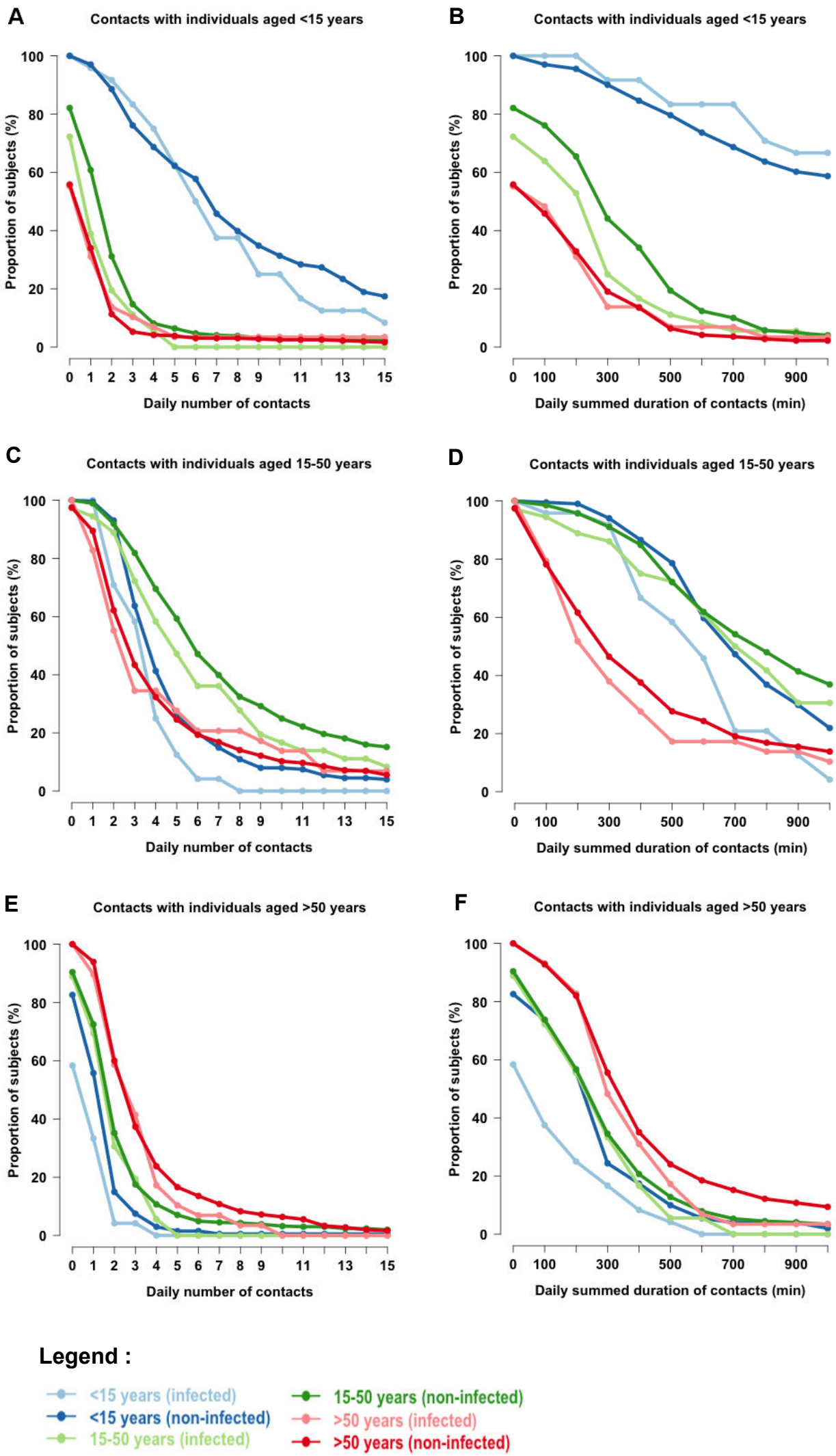
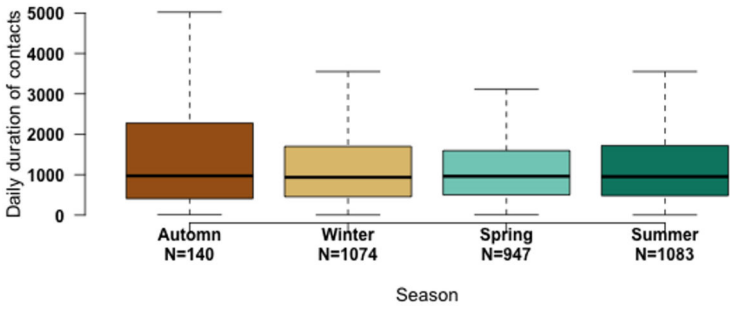
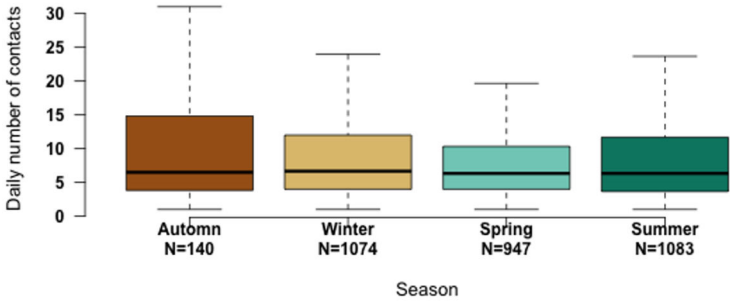

A
Contacts with individuals aged <15 years
Proportion of subjects (%)
Daily number of contacts
B
Contacts with individuals aged <15 years
Proportion of subjects (%)
Daily summed duration of contacts (min)
C
Contacts with individuals aged 15-50 years
Proportion of subjects (%)
Daily number of contacts
D
Contacts with individuals aged 15-50 years
Proportion of subjects (%)
Daily summed duration of contacts (min)
E
Contacts with individuals aged >50 years
Proportion of subjects (%)
Daily number of contacts
F
Contacts with individuals aged >50 years
Proportion of subjects (%)
Daily summed duration of contacts (min)
Legend :
<15 years (infected)
<15 years (non-infected)
15-50 years (infected)
15-50 years (non-infected)
>50 years (infected)
>50 years (non-infected)

Figure 5 Boxplots of the daily duration and number of contacts depending on calendar seasons. N corresponds to the number of subjects reporting contacts for the considered calendar season.


Daily duration of contacts
5000
4000
3000
2000
1000
0
Autumn
N=140
Winter
N=1074
Spring
N=947
Summer
N=1083
Season


Daily number of contacts
30
25
20
15
10
5
0
Autumn
N=140
Winter
N=1074
Spring
N=947
Summer
N=1083
Season

7.3 Discussion

Depuis l'émergence du virus pandémique H1N1pdm en 2009 un grand nombre de facteurs de risque de l'infection ont été recensés. Comme observé précédemment dans le chapitre 3, ces facteurs concernent des dimensions très diverses, cette diversité sous-entend que l'infection par la grippe résulte de l'interaction complexe de mécanismes hétérogènes. Les études menées jusqu'alors ont généralement consisté en une analyse de régression conventionnelle. Ces analyses ont le défaut d'étudier uniquement des associations directes entre les expositions et l'infection. Si l'existence de mécanismes sous-jacents est abordée dans l'interprétation des résultats de ces études, elle est ignorée au moment de l'analyse statistique et les paramètres estimés n'en tiennent par conséquent pas compte.

Les variables décrivant les croyances et la perception du risque d'infection ont jusqu'ici été ignorées dans les études basées sur les données CoPanFlu (c'est notamment le cas de l'analyse exploratoire présentée dans le chapitre 5), au motif qu'il apparait délicat de faire l'hypothèse qu'une telle dimension puisse être directement associée au risque d'infection. Le modèle d'équations structurelles nous a permis d'inclure ces variables, comme un déterminant associé de manière indirecte à l'infection, régulant ici l'adoption de comportements préventifs. Un tel résultat souligne la pertinence d'avoir employé cette méthodologie pour gérer l'intégration des divers niveaux d'information recueillis dans CoPanFlu.

7.3.1 L'absence d'association entre l'infection et les contacts sociaux

Il est toujours délicat de discuter une absence d'association, celle-ci pouvant être due à un manque de puissance pour la mettre en évidence. L'analyse approfondie des profils de contacts n'a pas permis de montrer de différence entre les sujets infectés et les non infectés. S'il est possible de discuter la période ou la manière dont l'exposition est mesurée ou de considérer que les durées et nombres de contacts ne sont pas des bons indicateurs des contacts susceptibles de propager une infection, nous pouvons cependant penser que les profils de contacts observés paraissent cohérents. On pouvait en effet s'attendre *a priori* à une exposition supérieure chez les sujets jeunes, avec des individus du même âge ou avec des adultes de l'âge de leurs parents, et une exposition moindre chez les sujets âgés [104]. Des voies de transmission alternatives, comme les surfaces contaminées, ou d'autres mesures d'exposition par les contacts sociaux, notamment par le biais de capteurs mériteraient d'être approfondies [79, 163]. Le recueil de telles données parait néanmoins complexe à l'échelle d'une étude de cohorte.

La grippe est une maladie infectieuse avec des taux d'incidences non négligeables, on peut ainsi imaginer qu'un grand nombre ou une grande durée de contacts ne sont probablement pas des conditions indispensables pour être infecté. Les résultats présentés ici nous laissent penser qu'à un niveau d'exposition équivalent, c'est la susceptibilité de l'individu qui va déterminer s'il sera infecté ou non.

Notons toutefois qu'à notre connaissance aucune étude n'a démontré de relation entre de telles mesures d'exposition et l'infection, les associations rapportées relatives à des contacts concernaient généralement des relations avec un cas index infecté au sein d'un environnement restreint comme le foyer. Une publication récente, basée sur l'utilisation de données de contacts de même nature que celles employées dans notre étude, a notamment affirmé que le risque d'infection individuel était influencé par l'intensité moyenne des contacts collectifs chez les individus d'un âge proche de celui du sujet considéré plutôt que par l'intensité des contacts individuels rapportés par ce sujet [102]. Ceci sous-entend que des mesures de contacts telles que celles mesurées dans CoPanFlu peuvent être utiles pour observer des comportements moyens au sein de classes d'âge données, ces profils moyens pouvant ensuite être employés pour prédire un risque individuel.

Il est néanmoins nécessaire de rappeler pour l'interprétation des résultats de ce travail que notre étude a porté sur la saison grippale 2010-2011, correspondant à la deuxième année de circulation du virus H1N1pdm en France métropolitaine. L'analyse des profils sérologiques des individus a permis d'observer une forte proportion de titres d'anticorps élevés chez les jeunes comparativement aux autres classes d'âge, laissant penser à des taux d'infection plus élevés chez ces individus l'année précédente. Il est ainsi regrettable de n'avoir pu effectuer la même analyse sur une population naïve d'un point immunologique face au virus H1N1pdm pour obtenir des estimations plus proches de celles attendues en cas d'émergence d'un nouveau virus.

7.3.2 Modélisation de la susceptibilité individuelle

Un seul indicateur a été utilisé pour modéliser la susceptibilité de l'hôte à contracter un infection grippale, le titre d'anticorps inhibant l'hémagglutination (titre HAI). Il s'agit d'une mesure de référence pour évaluer le niveau de protection individuelle contre le virus de la grippe.

La robustesse de la variable latente représentant la susceptibilité aurait pu être renforcée par l'inclusion d'autres marqueurs corrélats du niveau de protection individuel. On peut penser à d'autres marqueurs de la réponse humorale comme les anticorps IgA ou IgG [20,61] ou encore à des marqueurs immunocellulaires, notamment les lymphocytes T cytotoxiques [66,125]. Si des analyses de marqueurs immunocellulaires ont été prévues dans le cadre du suivi de l'étude CoPanFlu, ces données n'étaient pas disponibles au moment de l'analyse.

7.3.3 L'infection vue comme une variable latente

L'infection par la grippe pandémique H1N1pdm a ici été considérée comme une variable observée, basée sur une définition de cas, virologique et sérologique. Nous avions envisagé de traiter l'infection comme un phénomène latent, pour lequel il serait possible d'observer un certain nombre de manifesta-

tions biologiques ou cliniques, *e.g.* un titre d'anticorps post-épidémique élevé, une détection de virus positive par RT-PCR ou des symptômes constitutifs d'un syndrome grippal.

Cependant la construction d'une variable latente "infection" avec pour indicateurs les trois définitions de cas utilisées n'a pas permis d'obtenir une variable robuste. Ceci peut s'expliquer par le manque de concordance entre les définitions de cas, en effet parmi les 49 sujets présentant un critère virologique positif, seuls huit étaient également positifs pour le critère de séroconversion. Si ce dernier critère est largement utilisé pour définir les infections, son manque de sensibilité a déjà été observé [40]. Le manque de concordance entre le critère de séroconversion et le critère virologique pourrait également s'expliquer par un délai trop important entre la date d'infection potentielle d'un individu et la date du prélèvement sérologique réalisé. Une décroissance au fil du temps des titres d'anticorps HAI acquis suite à une infection a pu être observée [205]. On peut dès lors suspecter que si les prélèvements avaient été réalisés juste après l'épidémie 2010-2011, le taux de séroconversion aurait été probablement été plus important et de fait la concordance entre les critères virologiques et sérologiques aurait été meilleure.

En dépit du manque de sensibilité du critère de séroconversion, nous avons néanmoins considéré que l'utilisation simultanée de critères virologique et sérologique permettait d'identifier un maximum de cas et de limiter ainsi des problèmes de mauvaise classification des sujets (principalement un risque de faux négatifs). Rappelons que dans le protocole de l'étude CoPanFlu-France, la récolte de prélèvements destinés à la recherche de virus par RT-PCR ou puce RespiFinder était conditionnée par la déclaration d'un syndrome grippal au sein du foyer. Les prélèvements sérologiques ayant été réalisés chez une grande majorité des individus, l'utilisation du critère de séroconversion a ainsi permis de détecter des sujets potentiellement infectés mais non symptomatiques, la proportion estimée d'infections grippales non symptomatiques étant en effet de l'ordre de 34% [36].

Durant la saison grippale 2010-2011, en plus du virus H1N1pdm, la circulation d'autres souches de virus grippaux, principalement des virus de type B et dans une moindre mesure des virus de type A/H3N2, a pu être observée en France métropolitaine [17]. L'utilisation des techniques RT-PCR et RespiFinder sur les prélèvements virologiques réalisés chez les sujets de l'étude a notamment mis en évidence 37 infections par la grippe B (dont 10 co-infections avec le virus H1N1pdm). Il apparait donc qu'un certain nombre de sujets infectés par une grippe saisonnière durant la saison 2010-2011 ont été considérés comme non-infectés dans notre étude. Rappelons toutefois que celle-ci était concentrée sur les facteurs de risque de l'infection par le virus H1N1pdm uniquement. La réalisation d'une étude complémentaire sur les déterminants de l'infection grippale, sans distinction de souche particulière, nécessiterait de considérer ces sujets en tant qu'individus infectés. L'absence de sérologies disponibles pour les virus saisonniers est cependant une limite à la réalisation d'une telle analyse.

Les modèles d'équations structurelles, méthode de référence en sciences sociales plus rarement utilisée en épidémiologie, ont permis d'intégrer avec succès les différents niveaux d'information contenue dans les données CoPanFlu. Les résultats obtenus sont crédibles, et pour la plupart déjà rapportés dans la littérature. Contrairement à l'article précédent où les hypothèses portaient exclusivement sur le choix de la modélisation et ignoraient totalement la structure des données, des choix ont été faits ici sur les relations à inclure mais aussi sur les relations à exclure. Le modèle présenté ici est satisfaisant d'un point de vue des estimations et des indices numériques d'évaluation globale du modèle. Il est néanmoins primordial de garder à l'esprit que des dizaines de modèles auraient pu fournir des indices de qualité d'ajustement identiques, sinon meilleurs, et que les associations montrées ici ne peuvent être considérées comme étant le reflet exact des mécanismes causaux impliqués dans l'infection par la grippe pandémique H1N1pdm.

Conclusions et perspectives

Ce document a abordé la question de l'augmentation de la taille des jeux de données en épidémiologie et de la méthodologie statistique à employer pour en extraire de l'information, en particulier pour rechercher des associations. Nous avons décrit des approches variées, empruntées à des disciplines comme la sociologie ou la génomique, et les avons confrontées à l'un des paradigmes d'analyse statistique de référence en épidémiologie, le modèle de régression.

Les facteurs à explorer pour expliquer l'étiologie des maladies sont maintenant en grand nombre et sont extrêmement variés. L'épidémiologie évolue, les principes méthodologiques d'analyse statistique des études épidémiologiques doivent évoluer en parallèle.

La question de l'exploration des grands jeux de données a été abordée sous deux angles. C'est premièrement la philosophie *data-driven* qui a été employée, en s'inspirant des méthodes utilisées dans les sciences *omiques* et de leur caractère "pêche à la ligne" pour détecter des associations avec le risque d'infection grippale. Puis nous nous sommes rapprochés de la méthode scientifique conventionnelle *hypothesis-driven*, basée l'élaboration puis le test d'une série d'hypothèses pour proposer une approche plus mécanistique de l'infection par le virus de la grippe.

Que retenir de l'analyse *data-driven* ?

L'exploration *data-driven* des données CoPanFlu par des méthodes *data mining* ou par la régression LASSO n'a pas débouché sur la découverte de facteurs de risque "inattendus" associés à l'infection grippale mais a en revanche permis de retrouver un certain nombre d'associations déjà connues. Si ce travail a apporté relativement peu de nouveautés d'un point de vue épidémiologique, il a en revanche soulevé des questions méthodologiques. Nous avons pu observer que la méthodologie très largement utilisée en épidémiologie pour détecter des associations était très peu performante, sa puissance est faible et la probabilité de détecter des faux positifs est supérieure à celle attendue *a priori*. L'analyse présentée est limitée à un nombre réduit de cas et il est bien sur impossible de prétendre à l'universalité de ces conclusions. Cette approche exploratoire doit toutefois inviter à la réflexion quant à la pertinence des résultats généralement obtenus par le biais des modèles construits selon la procédure "régressions logistiques univariées/régression logistique multivariée".

Le point faible de ces méthodes, comparativement à la régression logistique, réside dans l'extraction d'information des modèles construits. Nous avons pu, par le biais de tests de permutation, calculer des *p-values*, qui bien que critiquables, constituent une mesure numérique appréciée des épidémiogistes. En revanche, il est difficile d'obtenir des indices de force d'association via les modèles de *data mining*, ceci nécessite obligatoirement une analyse en profondeur des variables détectées alors que la régression

logistique fournit des *odds ratios*, mesure intuitive et facilement interprétable. Enfin les temps de calculs nécessaires, notamment pour obtenir des *p-values* de tests de permutation ne sont pas négligeables, tandis que l'estimation des coefficients d'un modèle de régression logistique ne requiert que quelques secondes.

Ces méthodes d'exploration sont toutefois relativement récentes, on peut penser que leur potentiel reste encore partiellement inexploité et que des développements de ces approches permettant une extraction facilitée de connaissance peuvent être attendus dans les prochaines années. On peut notamment penser au développement, actuellement en cours, de mesure de significativité des estimations des modèles de régression LASSO. Ces méthodes ont pour grande valeur ajoutée de fournir des coefficients de régression auxquels les épidémiologistes sont familiers, tout en contournant de nombreuses limites des modèles conventionnels.

Pour conclure sur le versant exploratoire de l'analyse, nous pouvons également attendre beaucoup du développement de méthodes non paramétriques de détection d'associations complexes telles que le *maximum information coefficient* (MIC) [154]. Le MIC est une mesure de dépendance entre des paires de covariables, capable de détecter des associations plus ou moins complexe (linéaire, cubique, périodique, sinusoïdale, exponentielle ...) avec une grande efficacité. Sa capacité à capturer des relations hautement non linéaire, à les évaluer numériquement et à évaluer leur significativité statistique font du MIC une méthode extrêmement prometteuse. L'utilisation de cet indice est toutefois limitée à des paires de covariables, et l'utilisation de variables catégorielles n'est pour le moment pas possible.

Que retenir de l'analyse *hypothesis-driven* ?

Les modèles d'équations structurelles sont apparus comme étant une alternative de qualité à la régression multivariée conventionnelle. Contrairement à ce modèle et aux autres méthodes employées dans le travail précédent, les équations structurelles offrent l'opportunité de placer les covariables à des niveaux différents, ne supposant pas nécessairement que les prédicteurs sont associés de manière directe avec la variable réponse. Nous avons ainsi pu tirer pleinement parti du caractère intégratif de l'étude CoPanFlu-France. Ces modèles nous ont notamment permis d'inclure des facteurs relatifs à la perception du risque de l'infection grippale et à la perception des mesures préventives, ce qui n'avait pas été fait dans l'analyse exploratoire.

Les modèles d'équations structurelles, bien que fortement contraints par les hypothèses faites au moment de la construction du diagramme de chemin, constituent une méthode prometteuse. Les modèles utilisant des variables latentes sont relativement rares en épidémiologie ; les perspectives sont pourtant larges, non seulement d'un point de vue de la description de mécanismes "causaux" mais aussi en

termes d'impact potentiel sur les politiques de santé publique, comme souligné dans le dernier travail présenté dans ce document.

***Data-driven* ou *hypothesis-driven* ?**

Les deux travaux de modélisation statistique présentés dans ce document ne permettent pas d'affirmer de manière péremptoire qu'une approche est supérieure à une autre pour explorer un jeu de données en épidémiologie. Ceux-ci ont tous deux fourni de l'information pertinente d'un point de vue épidémiologique, l'analyse *data-driven* ayant également soulevé de nombreuses questions méthodologiques. On peut penser que le choix de l'approche à employer, exploratoire ou confirmatoire, est très dépendant de la problématique. Á l'ère des données *omiques*, il apparait délicat de se passer des approches *data-driven* et d'occulter leur caractère générateur d'hypothèses. L'une des limites de l'épidémiologie *hypothesis-driven* conventionnelle est que les données sont collectées en fonction d'hypothèses de travail pré-établies, souvent issues de la littérature, limitant potentiellement la génération de nouvelles connaissances. Dans un contexte comme celui de l'analyse *omique*, pour lequel construire des hypothèses apparait *a priori* complexe, l'utilisation d'une approche moins biaisée, car "aveugle", semble difficilement contournable.

Plutôt que d'opposer ces deux philosophies, il convient de les considérer comme complémentaires. Le *data-driven* apparait adapté pour "fouiller" dans les données, trouver des associations inattendues, au risque d'extraire des signaux non pertinents, la probabilité de trouver des faux positifs augmentant avec le nombre de variables analysées. Au contraire, la méthodologie *hypothesis-driven* parait vouée à être utilisée pour comprendre des signaux détectés ou des mécanismes. Elle est utile pour explorer plus en profondeur des associations déjà connues, comme dans le cas du modèle d'équations structurelles construit, ou des facteurs identifiés par une analyse *data-driven*, comme montré par les nombreuses études gènes-candidats dérivant d'hypothèses issues d'études GWAS.

Les grands jeux de données en épidémiologie : "*better or just bigger*" ?

L'analyse approfondie des données CoPanFlu, par des approches très diverses, d'un point de vue des algorithmes comme de la philosophie sous-jacente, amène à s'interroger sur la plus-value apportée par la récolte d'un tel nombre de données pour étudier les déterminants du risque d'infection grippale. En d'autres termes on peut se demander si l'augmentation de la taille des données s'est faite au bénéfice de l'information.

Les résultats de l'analyse exploratoire, confirmant principalement des associations déjà connues, ne sont *a priori* pas en faveur de la récolte massive de variables. Cette interprétation est toutefois à nuancer, certaines collections biologiques réalisées dans le cadre de CoPanFlu n'ayant pas encore été analysées.

L'addition d'autres facteurs, génomiques, transcriptomiques ou encore environnementaux (notamment climatiques) pourrait permettre de détecter des facteurs de risque inattendus. De plus, la question de la puissance statistique pour mettre en évidence des associations mérite également d'être soulignée. Le nombre de cas (*i.e.* le nombre d'individus infectés) était relativement faible dans les données analysées, il est possible de penser que ce nombre était potentiellement insuffisant pour mettre en évidence des facteurs de risque potentiels supplémentaires. La relation entre la taille de l'échantillon étudié et la puissance statistique a déjà été abordée dans le cadre des GWAS [83, 192], la notion de puissance reste une limite de ces études, notamment pour mettre en évidence des associations avec des tailles d'effets faibles (en particulier des *odds ratios* < 1.2) [101]. S'il apparait complexe de fournir en l'état une règle définissant le nombre d'individus à inclure dans une étude épidémiologique exploratoire comme celle présentée ici, cette question cruciale mérite d'être abordée en amont de la récolte des facteurs de risque à étudier (notamment par des études de simulation). La récolte de nombreuses covariables semble vouée à apporter peu d'information pertinente sans une puissance statistique suffisante. Le recours à des projets de très grandes cohortes [74, 216], notamment via Internet [85], pourrait alors être une solution pour l'étude de nombreuses pathologies.

Comme souligné au début de ce document avec l'exemple des analyses de données *omiques*, le recul sur l'études des grands jeux de données en biologie et en épidémiologie est encore faible, la plus value apportée par le *big data* dans ce contexte apparait dès lors difficilement estimable. Le risque de la multiplication des grands jeux de données et de l'utilisation systématique d'approches "aveugles" serait d'obtenir uniquement des catalogues d'associations sans être capables d'en comprendre le sens et d'oublier le principe fondamental que corrélation statistique et causalité sont deux notions distinctes. Si l'utilisation du *big data* en sciences de la vie apparait maintenant incontournable il est prématuré d'affirmer que la science basée sur les hypothèses est obsolète ; sans théorie, conceptuelle ou mathématique, ces données, aussi riches et variées soient-elles, ne resteront rien d'autre que du bruit.

Références

[1] H. Abdi. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews : Computational Statistics*, 2(1) :97–106, Jan. 2010.

[2] A. Acharjee, B. Kloosterman, R. C. H. de Vos, J. S. Werij, C. W. B. Bachem, R. G. F. Visser, and C. Maliepaard. Data integration and network reconstruction with omics data using Random Forest regression in potato. *Analytica chimica acta*, 705(1-2) :56–63, Oct. 2011.

[3] J. S. Aguilar-Ruiz, J. H. Moore, and M. D. Ritchie. Filling the gap between biology and computer science. *BioData mining*, 1(1) :1, Jan. 2008.

[4] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance : a corrected feature importance measure. *Bioinformatics (Oxford, England)*, 26(10) :1340–7, May 2010.

[5] L. D. A. F. Amorim, R. L. Fiaccone, C. A. S. T. Santos, T. N. dos Santos, L. T. L. P. de Moraes, N. F. Oliveira, S. O. Barbosa, D. N. dos Santos, L. M. dos Santos, S. M. A. Matos, and M. L. Barreto. Structural equation modeling in epidemiology. *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública*, 26(12) :2251–62, Dec. 2010.

[6] K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4) :2249–2260, Jan. 2008.

[7] A. Arlinghaus, D. A. Lombardi, J. L. Willetts, S. Folkard, and D. C. Christiani. A structural equation modeling approach to fatigue-related risk factors for occupational injury. *American journal of epidemiology*, 176(7) :597–607, Oct. 2012.

[8] C. Auffray, T. Caulfield, M. J. Khoury, J. R. Lupski, M. Schwab, and T. Veenstra. 2012 highlights in translational 'omics. *Genome medicine*, 5(1) :10, Jan. 2013.

[9] L. Auret and C. Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105(2) :157–170, 2011.

[10] P. C. Austin, D. S. Lee, E. W. Steyerberg, and J. V. Tu. Regression trees for predicting mortality in patients with cardiovascular disease : what improvement is achieved by using ensemble-based methods ? *Biometrical journal*, 54(5) :657–73, Sept. 2012.

[11] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee. Using methods from the data-mining and machine-learning literature for disease classification and prediction : a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, 66(4) :398–407, Apr. 2013.

[12] P. C. Austin, J. V. Tu, and D. S. Lee. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *Journal of clinical epidemiology*, 63(10) :1145–55, Oct. 2010.

[13] K. L. Ayers and H. J. Cordell. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic epidemiology*, 34(8) :879–91, Dec. 2010.

[14] M. Baker. Big biology : The 'omes puzzle. *Nature*, 494(7438) :416–419, Feb. 2013.

[15] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 7(10) :781–91, Oct. 2006.

[16] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms : Bagging, boosting, and variants. *Machine learning*, 36(1-2) :105–139, 1999.

[17] E. Belchior. Surveillance épidémiologique et virologique de la grippe en France, saison 2010-2011. *Bulletin épidémiologique hebdomadaire*, (37-38) :394–398, 2011.

[18] R. Bellazzi and B. Zupan. Towards knowledge-based gene expression data mining. *Journal of biomedical informatics*, 40(6) :787–802, Dec. 2007.

[19] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine : current issues and guidelines. *International journal of medical informatics*, 77(2) :81–97, Feb. 2008.

[20] R. B. Belshe, W. C. Gruber, P. M. Mendelman, H. B. Mehta, K. Mahmood, K. Reisinger, J. Treanor, K. Zangwill, F. G. Hayden, D. I. Bernstein, K. Kotloff, J. King, P. A. Piedra, S. L. Block, L. Yan, and M. Wolff. Correlates of immune protection induced by live, attenuated, cold-adapted, trivalent, intranasal influenza virus vaccine. *The Journal of infectious diseases*, 181(3) :1133–7, Mar. 2000.

[21] R. Bender and S. Lange. Adjusting for multiple testing–when and how ? *Journal of clinical epidemiology*, 54(4) :343–9, Apr. 2001.

[22] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[23] P. M. Bentler. Comparative fit indexes in structural models. *Psychological bulletin*, 107(2) :238–46, Mar. 1990.

[24] T. N. Beran and C. Violato. Structural equation modeling in medical research : a primer. *BMC research notes*, 3 :267, Jan. 2010.

[25] K. A. Bollen. Latent variables in psychology and the social sciences. *Annual review of psychology*, 53 :605–634, 2002.

[26] K. A. Bollen and M. D. Noble. Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl :15639–46, Sept. 2011.

[27] D. Borsboom, G. J. Mellenbergh, and J. van Heerden. The theoretical status of latent variables. *Psychological review*, 110(2) :203–219, 2003.

[28] P. Bougnères and A.-J. Valleron. Causes of early-onset type 1 diabetes : toward data-driven environmental approaches. *The Journal of experimental medicine*, 205(13) :2953–7, Dec. 2008.

[29] L. Breiman. Bagging predictors. *Machine learning*, 24(2) :123–140, 1996.

[30] L. Breiman. Random Forests. *Machine Learning*, 45 :123–140, 2001.

[31] T. A. Brown. *Confirmatory factor analysis for applied research*. Guilford Press, 2006.

[32] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. Identifying SNPs predictive of phenotype using random forests. *Genetic epidemiology*, 28(2) :171–82, Feb. 2005.

[33] B. M. Byrne. *Structural equation modeling with Mplus : basic concepts, applications, and programming*. Routledge, 2012.

[34] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS results : A review of statistical methods and recommendations for their application. *American journal of human genetics*, 86(1) :6–22, Jan. 2010.

[35] D. Carcione, C. M. Giele, L. S. Goggin, K. S. Kwan, D. W. Smith, G. K. Dowse, D. B. Mak, and P. Effler. Secondary attack rate of pandemic influenza A(H1N1) 2009 in Western Australian households, 29 May-7 August 2009. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, 16(3), Jan. 2011.

[36] F. Carrat, E. Vergu, N. M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach, and A.-J. Valleron. Time lines of infection and disease in human influenza : a review of volunteer challenge studies. *American journal of epidemiology*, 167(7) :775–85, Apr. 2008.

[37] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning (2008)*, pages 96–103, 2008.

[38] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, volume C of *ICML '06*, pages 161–168. ACM, ACM, 2006.

[39] S. Cauchemez, C. A. Donnelly, C. Reed, A. C. Ghani, C. Fraser, C. K. Kent, L. Finelli, and N. M. Ferguson. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *The New England Journal of Medicine*, 361(27) :2619–2627, 2009.

[40] S. Cauchemez, P. Horby, A. Fox, L. Q. Mai, L. T. Thanh, P. Q. Thai, L. N. M. Hoa, N. T. Hien, and N. M. Ferguson. Influenza infection rates, measurement errors and the interpretation of paired serology. *PLoS pathogens*, 8(12) :e1003061, 2012.

[41] D.-Y. Chao, K.-F. Cheng, T.-C. Li, T.-N. Wu, C.-Y. Chen, C.-A. Tsai, J.-H. Chen, H.-T. Chiu, J.-J. Lu, M.-C. Su, Y.-H. Liao, W.-C. Chan, and Y.-H. Hsieh. Factors associated with infection by 2009 pandemic H1N1 influenza virus during different phases of the epidemic. *International journal of infectious diseases*, 15(10) :e695–701, Oct. 2011.

[42] N. V. Chawla. Data Mining for Imbalanced Datasets : An Overview. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer US, 2010.

[43] M. I. C. Chen, V. J. M. Lee, I. Barr, C. Lin, R. Goh, C. Lee, B. Singh, J. Tan, W. Y. Lim, A. R. Cook, B. Ang, A. Chow, B. H. Tan, J. Loh, R. Shaw, K. S. Chia, R. T. P. Lin, and Y. S. Leo. Risk factors for pandemic (H1N1) 2009 virus seroconversion among hospital staff, Singapore. *Emerging infectious diseases*, 16(10) :1554–61, Oct. 2010.

[44] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. K. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, Y. Cheng, M. J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J. T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. A. Blasco, P. L. Greenberg, P. Snyder, T. E. Klein, R. B. Altman, A. J. Butte, E. A. Ashley, M. Gerstein, K. C. Nadeau, H. Tang, and M. Snyder. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6) :1293–307, Mar. 2012.

[45] R. Chen and M. Snyder. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews. Systems biology and medicine.*, Nov. 2012.

[46] X. Chen, M. Wang, and H. Zhang. The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(1) :55–63, Jan. 2011.

[47] H. Choi and N. Pavelka. When one and one gives more than two : challenges and opportunities of integrative omics. *Frontiers in genetics*, 2 :105, Jan. 2011.

[48] H.-Y. Chuang, M. Hofree, and T. Ideker. A decade of systems biology. *Annual review of cell and developmental biology*, 26 :721–44, Jan. 2010.

[49] F. Collins. Has the revolution arrived ? *Nature*, 464(7289) :674–5, Apr. 2010.

[50] J. Couzin-Frankel. Major heart disease genes prove elusive. *Science*, 328(5983) :1220–1, June 2010.

[51] Cutler, T. Edwards, K. Beard, A. Cutler, and K. Hess. Random forests for classification in ecology. *Ecology*, 88(11) :2783–2792, 2007.

[52] A. Cutler, D. R. Cutler, and J. R. Stevens. Tree-based methods. In *High-Dimensional Data Analysis in Cancer Research*, pages 1–19. Springer, 2009.

[53] T. H. Davenport and D. J. Patil. Data Scientist : The Sexiest Job of the 21st Century. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/.

[54] F. S. Dawood, S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom, R. J. Garten, L. V. Gubareva, X. Xu, C. B. Bridges, and T. M. Uyeki. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *The New England journal of medicine*, 360(25) :2605–15, June 2009.

[55] R. Díaz-Uriarte and S. Alvarez De Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1) :3, 2006.

[56] L. F. DiLalla. A structural equation modeling overview for medical researchers. *Journal of developmental and behavioral pediatrics : JDBP*, 29(1) :51–4, Feb. 2008.

[57] O. Dumas, V. Siroux, N. Le Moual, and R. Varraso. Approches d'analyse causale en épidémiologie. *Revue d'epidemiologie et de sante publique*, Dec. 2013.

[58] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293) :52–64, 1961.

[59] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2) :407–499, 2004.

[60] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1993.

[61] M. Eichelberger, H. Golding, M. Hess, J. Weir, K. Subbarao, C. J. Luke, M. Friede, and D. Wood. FDA/NIH/WHO public workshop on immune correlates of protection against influenza A viruses in support of pandemic vaccine development, Bethesda, Maryland, US, December 10-11, 2007. *Vaccine*, 26(34) :4299–303, Aug. 2008.

[62] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4) :802–813, 2008.

[63] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3) :272–299, 1999.

[64] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3) :37, 1996.

[65] J. M. Fontana, E. Alexander, and M. Salvatore. Translational research in infectious disease : current paradigms and challenges ahead. *Translational Research*, 159(6) :430–53, Jan. 2012.

[66] B. D. Forrest, M. W. Pride, A. J. Dunning, M. R. Z. Capeding, T. Chotpitayasunondh, J. S. Tam, R. Rappaport, J. H. Eldridge, and W. C. Gruber. Correlation of cellular immune responses with protection against culture-confirmed influenza virus in young children. *Clinical and vaccine immunology*, 15(7) :1042–1053, 2008.

[67] A. M. France, M. Jackson, S. Schrag, M. Lynch, C. Zimmerman, M. Biggerstaff, and J. Hadler. Household transmission of 2009 influenza A (H1N1) virus after a school-based outbreak in New York City, April-May 2009. *The Journal of infectious diseases*, 201(7) :984–92, Apr. 2010.

[68] Y. Freund and R. R. E. Schapire. Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[69] J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1) :1–22, Jan. 2010.

[70] J. H. Friedman. Greedy function approximation : a gradient boosting machine. *North*, 29(2) :1189–1232, 2001.

[71] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4) :367–378, 2002.

[72] J. H. Friedman and J. J. Meulman. Multiple additive regression trees with application in epidemiology. *Statistics in medicine*, 22(9) :1365–81, May 2003.

[73] C. Geiser. *Data Analysis with Mplus*. Guilford Press, 2013.

[74] German National Cohort (GNC) Consortium. The German National Cohort : aims, study design and organization. *European journal of epidemiology*, 29(5) :371–82, May 2014.

[75] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232) :1012–4, Feb. 2009.

[76] B. A. Goldstein, A. E. Hubbard, A. Cutler, and L. F. Barcellos. An application of Random Forests to a genome-wide association dataset : methodological considerations & new findings. *BMC genetics*, 11 :49, Jan. 2010.

[77] B. A. Goldstein, E. C. Polley, and F. B. S. Briggs. Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, 10(1) :32, Jan. 2011.

[78] T. Golub. Counterpoint : Data first. *Nature*, 464(7289) :679, Apr. 2010.

[79] J. S. Greatorex, P. Digard, M. D. Curran, R. Moynihan, H. Wensley, T. Wreghitt, H. Varsani, F. Garcia, J. Enstone, and J. S. Nguyen-Van-Tam. Survival of influenza A(H1N1) on materials found in households : implications for infection control. *PloS one*, 6(11) :e27932, Jan. 2011.

[80] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3) :389–422, 2002.

[81] F. E. Harrell. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.

[82] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, second edition, 2009.

[83] A. T. Hattersley and M. I. McCarthy. What makes a good genetic association study ? *Lancet*, 366(9493) :1315–23, Oct. 2005.

[84] K. Heikamp and J. Bajorath. Support vector machines for drug discovery. *Expert opinion on drug discovery*, Dec. 2013.

[85] S. Hercberg, K. Castetbon, S. Czernichow, A. Malon, C. Mejean, E. Kesse, M. Touvier, and P. Galan. The Nutrinet-Santé Study : a web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. *BMC public health*, 10 :242, Jan. 2010.

[86] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23) :9362–7, June 2009.

[87] A. Hoerl and R. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.

[88] S. Issenberg. How President Obama's campaign used big data to rally individual voters. http://www.technologyreview.com/featuredstory/508836/how-obama-used-big-data-to-rally-voters-part-1/.

[89] S. Janitza, C. Strobl, and A.-L. Boulesteix. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14(1) :119, Jan. 2013.

[90] P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records : towards better research applications and clinical care. *Nature reviews. Genetics*, 13(6) :395–405, June 2012.

[91] I. T. Jolliffe. Principal Components in Regression Analysis. *Principal Component Analysis*, pages 167–198, 2002.

[92] C. Kingsford and S. L. Salzberg. What are decision trees ? *Nature biotechnology*, 26(9) :1011–3, Sept. 2008.

[93] H. Kitano. Systems biology : a brief overview. *Science*, 295(5560) :1662–4, Mar. 2002.

[94] R. B. Kline. *Principles and practice of structural equation modeling*, volume 156. The Guilford Press, third edition, 2010.

[95] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1143. Citeseer, 1995.

[96] S. Kolenikov and K. A. Bollen. Testing Negative Error Variances : Is a Heywood Case a Symptom of Misspecification ? *Sociological Methods & Research*, 41(1) :124–167, 2012.

[97] N. Komiya, Y. Gu, H. Kamiya, Y. Yahata, Y. Yasui, K. Taniguchi, and N. Okabe. Household transmission of pandemic 2009 influenza A (H1N1) virus in Osaka, Japan in May 2009. *The Journal of infection*, 61(4) :284–8, Oct. 2010.

[98] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh. A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. *BioMed research international*, 2013 :432375, Jan. 2013.

[99] S. Koscielny. Why most gene expression signatures of tumors have not been useful in the clinic. *Science translational medicine*, 2(14) :1–3, Jan. 2010.

[100] J. Kruppa, A. Ziegler, and I. R. König. Risk estimation and risk prediction using machine-learning methods. *Human genetics*, 131(10) :1639–54, Oct. 2012.

[101] C. S. Ku, E. Y. Loy, Y. Pawitan, and K. S. Chia. The pursuit of genome-wide association studies : where are we now ? *Journal of human genetics*, 55(4) :195–206, Apr. 2010.

[102] A. J. Kucharski, K. O. Kwok, V. W. I. Wei, B. J. Cowling, J. M. Read, J. Lessler, D. A. Cummings, and S. Riley. The contribution of social behaviour to the transmission of influenza a in a human population. *PLoS pathogens*, 10(6) :e1004206, June 2014.

[103] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333) :187–97, Mar. 2011.

[104] N. Lapidus, X. de Lamballerie, N. Salez, M. Setbon, R. M. Delabre, P. Ferrari, N. Moyen, M.-L. Gougeon, F. Vely, M. Leruez-Ville, L. Andreoletti, S. Cauchemez, P.-Y. Boëlle, E. Vivier, L. Abel, M. Schwarzinger, M. Legeas, P. Le Cann, A. Flahault, and F. Carrat. Factors associated with post-seasonal serological titer and risk factors for infection with the pandemic A/H1N1 virus in the French general population. *PloS one*, 8(4) :e60127, Jan. 2013.

[105] N. Lapidus, X. De Lamballerie, N. Salez, M. Setbon, P. Ferrari, R. M. Delabre, M.-L. Gougeon, F. Vely, M. Leruez-Ville, L. Andreoletti, S. Cauchemez, P.-Y. Boëlle, E. Vivier, L. Abel, M. Schwarzinger, M. Legeas, P. Le Cann, A. Flahault, and F. Carrat. Integrative study of pandemic A/H1N1 influenza infections : design and methods of the CoPanFlu-France cohort. *BMC Public Health*, 12(1) :417, 2012.

[106] D. Lazer, R. Kennedy, G. King, and A. Vespignani. Big data. The parable of Google Flu : traps in big data analysis. *Science*, 343(6176) :1203–5, Mar. 2014.

[107] J. Leber. In a Data Deluge, Companies Seek to Fill a New Role. `http://m.technologyreview.com/news/513866/in-a-data-deluge-companies-seek-to-fill-a-new-role/`.

[108] H. Liao and H. S. Lynn. A survey of variable selection methods in two Chinese epidemiology journals. *BMC medical research methodology*, 10 :87, Jan. 2010.

[109] P. M. Lind, U. Risérus, S. Salihovic, B. van Bavel, and L. Lind. An environmental wide association study (EWAS) approach to the metabolic syndrome. *Environment international*, 55C :1–8, Feb. 2013.

[110] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso, 2013. `http://arxiv.org/abs/1301.7161`.

[111] W.-Y. Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(1) :14–23, 2011.

[112] C. Looker, K. Carville, K. Grant, and H. Kelly. Influenza A (H1N1) in Victoria, Australia : a community case series and analysis of household transmission. *PloS one*, 5(10) :e13702, Jan. 2010.

[113] J. Loscalzo and A.-L. Barabasi. Systems biology and the future of medicine. *Wiley interdisciplinary reviews. Systems biology and medicine*, 3(6) :619–27, 2011.

[114] G. Lubke, C. Laurin, R. Walters, N. Eriksson, P. Hysi, T. Spector, G. Montgomery, N. Martin, S. Medland, and D. Boomsma. Gradient Boosting as a SNP Filter : an Evaluation Using Simulated and Hair Morphology Data. *Journal of data mining in genomics & proteomics*, 4, Oct. 2013.

[115] R. C. MacCallum and J. T. Austin. Applications of structural equation modeling in psychological research. *Annual review of psychology*, 51 :201–226, 2000.

[116] O. Maimon and L. Rokach, editors. *Data mining and knowledge discovery handbook*. Springer edition, 2010.

[117] B. Manenti. Comment le big data va révolutionner 2013. http://obsession.nouvelobs.com/high-tech/20130103.OBS4353/comment-le-big-data-va-revolutionner-2013.html.

[118] T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine*, 363(2) :166–76, July 2010.

[119] T. A. Manolio. Bringing genome-wide association findings into clinical use. *Nature reviews. Genetics*, 14(8) :549–58, Aug. 2013.

[120] T. L. Marchbanks, A. Bhattarai, R. P. Fagan, S. Ostroff, S. V. Sodha, M. E. Moll, B. Y. Lee, C.-C. H. Chang, B. Ennis, P. Britz, A. Fiore, M. Nguyen, R. Palekar, W. R. Archer, T. L. Gift, R. Leap, B. L. Nygren, S. Cauchemez, F. J. Angulo, and D. Swerdlow. An outbreak of 2009 pandemic influenza A (H1N1) virus infection in an elementary school in Pennsylvania. *Clinical infectious diseases*, 52 Suppl 1 :S154–60, Jan. 2011.

[121] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. D. Mendonca. Data mining methods in the prediction of Dementia : A real-data comparison of the accuracy, sensitivity and specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests. *BMC research notes*, 4(1) :299, Aug. 2011.

[122] L. J. Martin, J. G. Woo, C. L. Avery, H.-S. Chen, K. E. North, K. Au, P. Broët, C. Dalmasso, M. Guedj, P. Holmans, B. Huang, P.-H. Kuo, A. C. Lam, H. Li, A. Manning, I. Nikolov, R. Sinha, J. Shi, K. Song, M. Tabangin, R. Tang, and R. Yamada. Multiple testing in the genomics era : findings from Genetic Analysis Workshop 15, Group 15. *Genetic epidemiology*, 31 Suppl 1 :S124–31, Jan. 2007.

[123] P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC, 1989.

[124] J. E. McDermott, J. Wang, H. Mitchell, B.-J. Webb-Robertson, R. Hafen, J. Ramey, and K. D. Rodland. Challenges in Biomarker Discovery : Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert opinion on medical diagnostics*, 7(1) :37–51, Jan. 2013.

[125] J. E. McElhaney, D. Xie, W. D. Hager, M. B. Barry, Y. Wang, A. Kleppinger, C. Ewen, K. P. Kane, and R. C. Bleackley. T cell responses are better correlates of vaccine protection in the elderly. *Journal of immunology*, 176(10) :6333–6339, 2006.

[126] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore. Machine learning for detecting gene-gene interactions : a review. *Applied bioinformatics*, 5(2) :77–88, Jan. 2006.

[127] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(1) :53–71, Jan. 2008.

[128] D. Meyer, F. Leisch, and K. Hornik. The support vector machine under test. *Neurocomputing*, 55(1) :169–186, 2003.

[129] L. Meyers, G. Gamst, and A. Guarino. *Applied Multivariate Research : Design and Interpretation*, volume 45. Second edition, 2006.

[130] J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics (Oxford, England)*, 26(4) :445–55, Feb. 2010.

[131] M. Mosca. Les 10 ans de Facebook en 10 chiffres. http://www.challenges.fr/high-tech/20140203.CHA9959/les-10-ans-de-facebook-en-10-chiffres.html/.

[132] R. Mundry and C. L. Nunn. Stepwise model fitting and statistical inference : turning noise into signal pollution. *The American naturalist*, 173(1) :119–23, Jan. 2009.

[133] T. B. Murdoch and A. S. Detsky. The inevitable application of big data to health care. *JAMA : the journal of the American Medical Association*, 309(13) :1351–2, Apr. 2013.

[134] B. Muthén. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1) :115–132, 1984.

[135] L. Muthén and B. Muthén. Mplus User's Guide. Seventh Edition. (1998-2012). Los Angeles, CA : Muthén & Muthén.

[136] A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7 :21, Jan. 2013.

[137] National Human Genome Research Institute. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/.

[138] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, 11 :110, Jan. 2010.

[139] H. Nishiura and H. Oshitani. Household transmission of influenza (H1N1-2009) in Japan : age-specificity and reduction of household transmission risk by zanamivir treatment. *The Journal of international medical research*, 39(2) :619–28, Jan. 2011.

[140] W. S. Noble. What is a support vector machine ? *Nature biotechnology*, 24(12) :1565–7, Dec. 2006.

[141] F. Odaira, H. Takahashi, T. Toyokawa, Y. Tsuchihashi, T. Kodama, Y. Yahata, T. Sunagawa, K. Taniguchi, and N. Okabe. Assessment of secondary attack rate and effectiveness of antiviral prophylaxis among household contacts in an influenza A(H1N1)v outbreak in Kobe, Japan, May-June 2009. *Eurosurveillance*, 14(35), Jan. 2009.

[142] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza : a comparative epidemiological study at three geographic scales. *PLoS computational biology*, 9(10) :e1003256, Jan. 2013.

[143] D. Opitz and R. Maclin. Popular ensemble methods : An empirical study. *Journal of Artificial Intelligence Research*, (11) :169–198, 1999.

[144] C. J. Patel, J. Bhattacharya, and A. J. Butte. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PloS one*, 5(5) :e10746, Jan. 2010.

[145] R. G. Pebody, R. Harris, G. Kafatos, M. Chamberland, C. Campbell, J. S. Nguyen-Van-Tam, E. McLean, N. Andrews, P. J. White, E. Wynne-Evans, J. Green, J. Ellis, T. Wreghitt, S. Bracebridge, C. Ihekweazu, I. Oliver, G. Smith, C. Hawkins, R. Salmon, B. Smyth, J. McMenamin, M. Zambon, N. Phin, and J. M. Watson. Use of antiviral drugs to reduce household transmission of pandemic (H1N1) 2009, United Kingdom. *Emerging infectious diseases*, 17(6) :990–9, June 2011.

[146] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12) :1373–9, Dec. 1996.

[147] D. A. Pendlebury. White paper using bibliometrics in evaluating research. http://wokinfo.com/media/mtrp/UsingBibliometricsinEval_WP.pdf.

[148] R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3) :21–45, 2006.

[149] V. M. Prieto, S. Matos, M. Alvarez, F. Cacheda, and J. L. Oliveira. Twitter : a good place to detect health conditions. *PloS one*, 9(1) :e86191, Jan. 2014.

[150] S. M. Rappaport. Implications of the exposome for exposure science. *Journal of exposure science & environmental epidemiology*, 21(1) :5–9, Jan. 2011.

[151] S. M. Rappaport and M. T. Smith. Environment and disease risks. *Science*, 330(6003) :460–1, Oct. 2010.

[152] T. Raykov and G. A. Marcoulides. *A First Course in Structural Equation Modeling*, volume 13. Lawrence Erlbaum Associates Publishers, 2006.

[153] D. M. Reif, A. A. Motsinger, B. A. McKinney, J. E. Crowe, and J. H. Moore. Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types. *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006.

[154] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062) :1518–1524, Dec. 2011.

[155] G. Ridgeway. Generalized Boosted Models : A guide to the gbm package. *Compute*, 1(4) :1–12, 2007.

[156] S. Riley, K. O. Kwok, K. M. Wu, D. Y. Ning, B. J. Cowling, J. T. Wu, L.-M. Ho, T. Tsang, S.-V. Lo, D. K. W. Chu, E. S. K. Ma, and J. S. M. Peiris. Epidemiological characteristics of 2009 (H1N1) pandemic influenza based on paired sera from a longitudinal community cohort study. *PLoS medicine*, 8(6) :e1000442, June 2011.

[157] W. Rodenburg, A. G. Heidema, J. M. A. Boer, I. M. J. Bovee-Oudenhoven, E. J. M. Feskens, E. C. M. Mariman, and J. Keijer. A framework to identify physiological responses in microarray-based gene expression studies : selection and interpretation of biologically relevant genes. *Physiological genomics*, 33(1) :78–90, Mar. 2008.

[158] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2) :1–39, 2009.

[159] H. Ruet. La victoire d'Obama : cas d'étude concret d'utilisation des Big Data. http://www.journaldunet.com/ebusiness/expert/53760/la-victoire-d-obama---cas-d-etude-concret-d-utilisation-des-big-data.shtml.

[160] A. Sadilek, S. Brennan, H. Kautz, and V. Silenzio. nEmesis : Which Restaurants Should You Avoid Today ? In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[161] Y. Saeys, I. n. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19) :2507–17, Oct. 2007.

[162] M. Salathé, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, and A. Vespignani. Digital epidemiology. *PLoS computational biology*, 8(7) :e1002616, Jan. 2012.

[163] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences of the United States of America*, 107(51) :22020–5, Dec. 2010.

[164] G. Saporta. Epidémiologie et data mining ou fouille de données. *L'épidémiologie humaine : conditions de son développement en France et rôle des mathématiques*, pages 137–142, 2006.

[165] J. Shen, S. I. Chang, E. S. Lee, Y. Deng, and S. J. Brown. Determination of cluster number in clustering microarray data. *Applied Mathematics and Computation*, 169(2) :1172–1185, 2005.

[166] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS one*, 6(5) :e19467, Jan. 2011.

[167] C. Sikora, S. Fan, R. Golonka, D. Sturtevant, J. Gratrix, B. E. Lee, J. Jaipaul, and M. Johnson. Transmission of pandemic influenza A (H1N1) 2009 within households : Edmonton, Canada. *Journal of clinical virology*, 49(2) :90–3, Oct. 2010.

[168] S. D. Silvey. Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 539–552, 1969.

[169] J. M. Simmerman, P. Suntarattiwong, J. Levy, R. G. Jarman, S. Kaewchana, R. V. Gibbons, B. J. Cowling, W. Sanasuttipun, S. A. Maloney, T. M. Uyeki, L. Kamimoto, and T. Chotipitayasunondh. Findings from a household randomized controlled trial of hand washing and face masks to reduce influenza transmission in Bangkok, Thailand. *Influenza and other respiratory viruses*, 5(4) :256–67, July 2011.

[170] A. Sinha, G. Hripcsak, and M. Markatou. Large datasets in biomedicine : a discussion of salient analytic issues. *Journal of the American Medical Informatics Association*, 16(6) :759–67, 2009.

[171] P. Smyth. Data mining : data analysis on a grand scale ? *Statistical methods in medical research*, 9(4) :309–27, Aug. 2000.

[172] W. W. Soon, M. Hariharan, and M. P. Snyder. High-throughput sequencing for biology and medicine. *Molecular systems biology*, 9 :640, Jan. 2013.

[173] D. Sörbom. Model modification. *Psychometrika*, 54(3) :371–384, 1989.

[174] A. Statnikov, D. Hardin, and C. F. Aliferis. Using SVM weight-based methods to identify causally relevant and non-causally relevant variables. In *Proceedings of Neural Information Processing Systems (NIPS) Workshop on Causality and Feature Selection*, Cambridge, MA, USA, 2006.

[175] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9 :319, Jan. 2008.

[176] J. H. Steiger and J. C. Lind. Statistically based tests for the number of common factors. *Annual Meeting of the Psychometric Society*, 1980.

[177] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :479–498, 2002.

[178] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16) :9440–5, Aug. 2003.

[179] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9 :307, Jan. 2008.

[180] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures : illustrations, sources and a solution. *BMC bioinformatics*, 8(1) :25, Jan. 2007.

[181] C. Strobl, J. Malley, and G. Tutz. An introduction to recursive partitioning : rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4) :323–48, Dec. 2009.

[182] A. Sturn, J. Quackenbush, and Z. Trajanoski. Genesis : cluster analysis of microarray data. *Bioinformatics*, 18(1) :207–208, Jan. 2002.

[183] J. D. Sugimoto, N. N. Borse, M. L. Ta, L. J. Stockman, G. E. Fischer, Y. Yang, M. E. Halloran, I. M. Longini, and J. S. Duchin. The effect of age on transmission of 2009 pandemic influenza A (H1N1) in a camp and associated households. *Epidemiology*, 22(2) :180–7, Mar. 2011.

[184] G. W. Sun, T. L. Shook, and G. L. Kay. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of clinical epidemiology*, 49(8) :907–16, Aug. 1996.

[185] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest : a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6) :1947–58, 2003.

[186] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, and Q. Song. Boosting : an ensemble learning tool for compound classification and QSAR modeling. *Journal of chemical information and modeling*, 45(3) :786–99, 2005.

[187] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, and Y. V. Sun. Machine learning in genome-wide association studies. *Genetic epidemiology*, 33(Supplement 1) :S51–7, Jan. 2009.

[188] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici. Machine Learning and Its Applications to Biology. *PLoS Computational Biology*, 3(6) :11, 2007.

[189] R. Thiébaut, B. Hejblum, and L. Richert. The analysis of "Big Data" in clinical research. *Revue d'epidemiologie et de sante publique*, Jan. 2014.

[190] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1) :267–288, 1996.

[191] R. Tibshirani. Regression shrinkage and selection via the lasso : a retrospective. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(3) :273–282, June 2011.

[192] T. Tong and H. Zhao. Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments. *Statistics in medicine*, 27(11) :1960–72, May 2008.

[193] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. van Hijum. Data mining in the Life Sciences with Random Forest : a walk in the park or lost in the jungle ? *Briefings in bioinformatics*, 14(3) :315–26, May 2013.

[194] O. Trelles, P. Prins, M. Snir, and R. C. Jansen. Big data, but are we ready ? *Nature reviews. Genetics*, 12(3) :224, Mar. 2011.

[195] Y.-K. Tu. Commentary : Is structural equation modelling a step forward for epidemiologists ? *International journal of epidemiology*, 38(2) :549–51, Apr. 2009.

[196] L. R. Tucker and C. Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1) :1–10, Mar. 1973.

[197] C. van Gemert, M. Hellard, E. S. McBryde, J. Fielding, T. Spelman, N. Higgins, R. Lester, H. Vally, and I. Bergeri. Intrahousehold transmission of pandemic (H1N1) 2009 virus, Victoria, Australia. *Emerging infectious diseases*, 17(9) :1599–607, Sept. 2011.

[198] P. van Helden. Data-driven hypotheses. *EMBO reports*, Dec. 2012.

[199] L. van Kerkhoff. Integrated research : concepts of connection in environmental science and policy. *Environmental Science & Policy*, 8(5) :452–463, Oct. 2005.

[200] P. Velardi, G. Stilo, A. E. Tozzi, and F. Gesualdo. Twitter mining for fine-grained syndromic surveillance. *Artificial Intelligence in Medicine*, Jan. 2014.

[201] A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests : A survey and results of new tests. *Pattern Recognition*, 44(2) :330–349, 2011.

[202] P. Waldmann, G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4 :270, Jan. 2013.

[203] S. Walter and H. Tiemeier. Variable selection : current practice in epidemiological studies. *European journal of epidemiology*, 24(12) :733–6, Jan. 2009.

[204] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, and W. Yu. MegaSNPHunter : a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC bioinformatics*, 10 :13, Jan. 2009.

[205] M. Wang, J. Yuan, T. Li, Y. Liu, J. Wu, B. Di, X. Chen, X. Xu, E. Lu, K. Li, Y. Liu, Y. Wu, X. Chen, P. He, Y. Wang, and J. Liu. Antibody dynamics of 2009 influenza A (H1N1) virus in infected patients and vaccinated people in China. *PloS one*, 6(2) :e16809, Jan. 2011.

[206] M. J. Whittingham, P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. Why do we still use stepwise modelling in ecology and behaviour ? *The Journal of animal ecology*, 75(5) :1182–9, Sept. 2006.

[207] R. E. Wiegand. Performance of using multiple stepwise algorithms for variable selection. *Statistics in medicine*, 29(15) :1647–59, July 2010.

[208] C. P. Wild. Complementing the genome with an "exposome" : the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer epidemiology, biomarkers & prevention*, 14(8) :1847–50, Aug. 2005.

[209] H. Wold. Partial Least Squares. *Medicinal Chemistry*, 147(2) :581–591, 1985.

[210] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)*, 25(6) :714–21, Mar. 2009.

[211] C.-J. Xu, A. van der Schaaf, C. Schilstra, J. A. Langendijk, and A. A. van't Veld. Impact of statistical learning methods on the predictive power of multivariate normal tissue complication probability models. *International journal of radiation oncology, biology, physics*, 82(4) :e677–84, Mar. 2012.

[212] Q. Yang, J. Cui, I. Chazaro, L. A. Cupples, and S. Demissie. Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC genetics*, 6 Suppl 1(Suppl 1) :S134, Jan. 2005.

[213] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. Data mining in healthcare and biomedicine : a survey of the literature. *Journal of medical systems*, 36(4) :2431–48, Aug. 2012.

[214] A. Ziegler, A. L. DeStefano, I. R. König, C. Bardel, D. Brinza, S. Bull, Z. Cai, B. Glaser, W. Jiang, K. E. Lee, C. X. Li, J. Li, X. Li, P. Majoram, Y. Meng, K. K. Nicodemus, A. Platt, D. F. Schwarz, W. Shi, Y. Y. Shugart, H. H. Stassen, Y. V. Sun, S. Won, W. Wang, G. Wahba, U. A. Zagaar, and Z. Zhao. Data mining, neural nets, trees–problems 2 and 3 of Genetic Analysis Workshop 15. *Genetic epidemiology*, 31 Suppl 1 :S51–60, Jan. 2007.

[215] A. Ziegler, I. R. König, and J. R. Thompson. Biostatistical aspects of genome-wide association studies. *Biometrical journal. Biometrische Zeitschrift*, 50(1) :8–28, Feb. 2008.

[216] M. Zins, S. Bonenfant, M. Carton, M. Coeuret-Pellicer, A. Guéguen, J. Gourmelen, M. Nachtigal, A. Ozguler, A. Quesnot, C. Ribet, G. Rodrigues, A. Serrano, R. Sitta, A. Brigand, J. Henny, and M. Goldberg. The CONSTANCES cohort : an open epidemiological laboratory. *BMC public health*, 10 :479, Jan. 2010.

[217] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society - Series B : Statistical Methodology*, 67(2) :301–320, 2005.