Société Française de Biométrie

Journée des Jeunes Chercheurs

Paris, 29 mai 2018

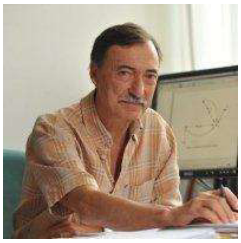
La Société Française de Biométrie (SFB)

Présentation

La SFB est la société savante dédiée aux chercheurs statisticiens intéressés par les applications dans les domaines biomédical, environnemental, agronome, etc. La SFB est la Région Française de l'International Biometric Society (IBS). Elle fait partie du "Channel Network" de l'IBS qui regroupe les régions belge, britannique, irlandaise, néerlandaise et française. Elle a pour objectif de promouvoir la biométrie en France en organisant des évènements scientifiques, en favorisant les interactions au sein du réseau des biométriciens de France et en relayant les informations et événements internationaux.

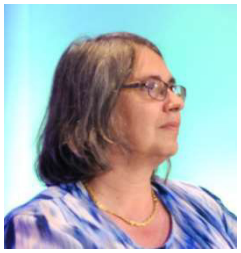
Membres du conseil

Daniel Commenges
Président



Robert Faivre
Membre

Pascale Tubert-Bitter
Secrétaire



Mounia N. Hocine
Membre

Pascal Wild
Trésorier

David Causeur
Membre

Hélène Jacqmin-Gadda
Membre

Cécile Proust-Lima
Membre

Informations pratiques

Lieu de l'événement

Conservatoire national des arts et métiers (Cnam)
Amphithéâtre Jean Prouvé
292, rue Saint-Martin
Paris 75003

Accès métro

Ligne 4, station Réaumur Sébastopol
Ligne 3 ou 11, station Arts et Métiers

Programme de la journée

10h	**Ouverture de la journée**
10h10-11h00	Geert Molenberghs (Conférencier invité)
	Hierarchical models with normal and conjugate random effects
11h00-11h15	Sara Si-moussi
	A data-driven ecological interaction patterns discovery: case study of the soil fauna
11h15-11h30	Félix Cheysson
	Estimation of exposure-attributable fractions from time series
11h30-11h45	Shaima Bel Hechmi
	Développement de modèles de survie pronostiques de grande dimension en considérant des groupes de biomarqueurs : application en génétique humaine
12h00-12h15	Loïc Ferrer
	Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment
12h15-12h30	Alessandra Meddis
	Semiparametric approach for covariate-specific time dependent ROC curves for correlated survival data
12h30-14h15	**Pause déjeuner**
14h15-15h00	Agnieszka Kroll (Lauréate du prix Daniel Schwartz)
	Prise en compte d'événements multiples pour analyser et prédire l'évolution d'un cancer
15h00-15h15	Florian Hébert
	Combinaison de tests dépendants en études d'association pangénomiques
15h15-15h30	Canelle Poirier
	Les données massives hospitalières pour la surveillance des syndromes grippaux en temps réel
15h30-15h45	**Pause café**
15h45-16h00	Bachirou O. Taddé
	Dynamic modeling of latent processes and their causal relationships: application to Alzheimer's disease
16h00-16h15	Laura Villain
	Adaptive protocols based on predictions from a mechanistic model of the effect of IL7 on CD4 counts
16h15-16h30	Emeline Courtois
	Propensity score-based approaches in high dimension for pharmacovigilance signal detection: an empirical comparison on the French spontaneous reporting database
16h30	**Clôture de la journée**

Conférencier Invité (Key-note speaker)

Geert Molenberghs

Interuniversity Institute for Biostatistics and statistical Bioinformatics

(1) I-BioStat, Hasselt University, Diepenbeek, Belgium
(2) I-BioStat, KU Leuven, Belgium

HIERARCHICAL MODELS WITH NORMAL AND CONJUGATE RANDOM EFFECTS

Molenberghs, Verbeke, and Demétrio (2007) and Molenberghs et al. (2010) proposed a general framework to model hierarchical data subject to within-unit correlation and/or overdispersion. The framework extends classical overdispersion models as well as generalized linear mixed models. Subsequent work has examined various aspects and lead to the formulation of several extensions. A unified treatment of the model framework and key extensions is provided. Particular extensions discussed are: explicit calculation of correlation and other moment-based functions, joint modeling of several hierarchical sequences, versions with direct marginally interpretable parameters, zero-inflation in the count case, and influence diagnostics. The basic models and several extensions are illustrated using a set of key examples, one per data type (count, binary, multinomial, ordinal, and time-to-event).

References

Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2017). ierarchical models with normal and conjugate random effects: A review. SORT, to appear.

Molenberghs, G. and Verbeke, G. (2011). On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. Journal of Statistical Planning and Inference, 141, 861-868.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Analysis, 13, 513-531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. Statistical Science, 25, 325-347.

Vangeneugden, T., Molenberghs, G., Verbeke, G., and Demétrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. Journal of Applied Statistics, 38, 215-232.

Résumés

Consideration of multiple events for the analysis and prediction of a cancer evolution
Agnieszka Krol

Combinaison de tests dépendants en études d'association pangénomiques
Florian Hébert

A data-driven ecological interaction patterns discovery: case study of the soil fauna
Sara Si-moussi

Estimation of exposure-attributable fractions from time series
Felix Cheysson

Dynamic modeling of latent processes and their causal relationships: application to Alzheimer's disease
Bachirou O. Taddé

Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment
Loïc Ferrer

Semiparametric approach for covariate-speci_c time dependent ROC curves for correlated survival data
Alessandra Meddis

Les données massives hospitalières pour la surveillance des syndromes grippaux en temps réel
Canelle Poirier

Adaptive protocols based on predictions from a mechanistic model of the e_ect of IL7 on CD4 counts
Laura Villain

Propensity score-based approaches in high dimension for pharmacovigilance signal detection: an empirical comparison on the French spontaneous reporting database
Emeline Courtois

Développement de modèles de survie pronostiques de grande dimension en considérant des groupes de biomarqueurs : application en génétique humaine
Shaima Bel Hechmi

Consideration of multiple events for the analysis and prediction of a cancer evolution

Agnieszka Król, Innovative Medicines, Early Clinical Development, AstraZeneca, Molndal, Sweden
Stefan Michiels, Service de Biostatistique et d'Epidémiologie, Gustave Roussy, University Paris-Saclay, University Paris-Sud, CESP, INSERM U1018, Villejuif, France
Virginie Rondeau, INSERM UMR 1219, Univ. Bordeaux, Bordeaux, France

The increasing number of clinical trials for cancer treatments has led to standardization of guidelines for evaluation of tumor response. In phase III clinical trials of advanced cancer, progression-free survival is often applied as a surrogate endpoint for overall survival (OS). For solid tumors, progression is usually defined using the RECIST criteria that use information on the change of size of target lesions and progressions of non-target disease. The criteria remain the standard tool for treatment evaluation despite their limitations. In particular, repeatedly measured tumor size is used as a pointwise categorized variable to identify a patient's status. Statistical approach of joint modeling allows for more accurate analysis of the tumor response markers and survival. Moreover, joint models are useful for individual dynamic predictions of death using patient's history. In this work, we proposed to apply a trivariate joint model for a longitudinal outcome (tumor size), recurrent events (progressions of non-target disease) and survival. Using adapted measures of predictive accuracy we compared the proposed joint model with a model that considered tumor progressions defined within standard criteria and OS. For a randomized clinical trial for colorectal cancer patients, we found better predictive accuracy of the proposed joint model. In the second part, we developed freely available software for application of the proposed joint modeling and dynamic predictions approach. Finally, we extended the model to a more sophisticated analysis of tumor size evolution using a mechanistic model. An ordinary differential equation was implemented to describe the trajectory of the biomarker regarding the biological characteristics of tumor size under a treatment. This new approach contributes to clinical research on treatment evaluation in clinical trials by better understanding of the relationship between the markers of tumor response with OS.

Combinaison de tests dépendants en études d'association pangénomiques

Florian Hébert, Mathieu Emily, David Causeur
Agrocampus Ouest, IRMAR, UMR CNRS 6625, 35042 Rennes

1 Introduction

Les études d'association pangénomiques cas/témoins consistent à tester l'association entre un phénotype binaire (*e.g.* une maladie d'intérêt) et un ensemble de p marqueurs génétiques (*e.g.* des *single nucleotide polymorphisms* ou SNPs). Pour pallier les limites des analyses classiques, pour lesquelles chaque SNP est testé individuellement, il a été proposé de combiner les tests par une approche dite *SNP-set*. En considérant une région génomique d'intérêt (un gène, un bloc de déséquilibre de liaison), les méthodes *SNP-set* cherchent à tester globalement l'association de la région avec le phénotype. Statistiquement, ces méthodes permettent de combiner un ensemble d'associations en un test unique tout en intégrant la structure de dépendance entre SNPs voisins. D'un point de vue biologique, les approches *SNP-set* présentent l'avantage de tester l'association au niveau fonctionnel du génome.

Plusieurs statistiques de tests globaux ont été proposées, parmi lesquelles la statistique minP [1], ainsi que des combinaisons linéaires ou quadratiques des statistiques de test associées à chaque SNP de la région considérée [2] sont les plus utilisées. Cependant, ces méthodes sont sensibles à la structure de dépendance entre les SNPs du groupe. Nous étudions ici une méthode alternative qui s'appuie sur une première étape de décorrélation des statistiques de test, préalable au test global.

2 Matériel et méthodes

Nous considérons un groupe de m SNPs. Le vecteur $\mathbf{Z} = (Z_1, \ldots, Z_m)'$ est formé des m statistiques Z_i de test d'association simple-marqueur, avec $\mathbf{Z} \sim \mathcal{N}_m(\mu, \Sigma)$. Nous formalisons le test global par un problème de détection à l'échelle de la région considérée, qui correspond ainsi au test des hypothèses $H_0 : \mu = \mathbf{0}$ contre $H_1 : \mu \neq \mathbf{0}$. Nous proposons de tenir compte de la dépendance préalablement à la réalisation du test global en décorrélant le vecteur $\mathbf{Z}$, *i.e.* en calculant $\widetilde{\mathbf{Z}} = \Omega\mathbf{Z}$ avec $\Omega'\Omega = \Sigma^{-1}$ [4]. La statistique globale est ensuite calculée sur le vecteur transformé $\widetilde{\mathbf{Z}}$. Le bénéfice de ce type de transformation sur les performances des procédures statistiques a été mis en évidence dans un cadre général [3]. Dans notre contexte, pour accroître la puissance, nous proposons de plus une étape de réduction de dimension, visant à ne conserver que certaines statistiques susceptibles de comporter du signal.

A partir de simulations, nous avons ensuite comparé la puissance de notre méthode à celle de méthodes usuelles. Les méthodes comparées sont les méthodes minP [1], higher criticism (HC) [3], la norme $\mathcal{L}^2$ de $\mathbf{Z}$ et le test de Hotelling [2]. Ce test peut être vu comme étant basé sur une décorrélation de $\mathbf{Z}$, mais sans la réduction de dimension proposée ici. Pour réaliser les comparaisons dans un contexte réaliste, un bloc de SNPs est simulé selon une structure de dépendance et des distributions marginales observées. Le phénotype est ensuite simulé conditionnellement aux SNPs selon un modèle logistique dépendant d'un paramètre d'intensité de liaison noté ξ. La puissance de détection de chaque statistique est estimée en simulant 1000 phénotypes pour différentes valeurs de ξ.

3 Résultats

Le tableau 1 donne la puissance estimée de chaque méthode pour quelques valeurs de ξ sur un modèle de simulation. La transformation proposée accroît la puissance de détection. En particulier, la puissance de l'approche proposée est supérieure à celle du test de Hotelling, ce qui souligne l'importance de l'étape de réduction de dimension.

	Méthode				
ξ	minP	HC	Norme $\mathcal{L}^2$	Hotelling	Méthode proposée
0.06	0.125	0.146	0.132	0.093	0.172
0.12	0.355	0.402	0.401	0.209	0.551
0.18	0.842	0.859	0.863	0.623	0.933

Table 1: Puissance de chaque méthode pour quelques valeurs de ξ

4 Conclusion

La prise en compte de la dépendance *a priori* telle qu'introduite dans notre étude permet d'accroître la puissance de détection. Toutefois, cet accroissement de la puissance est aussi dû à la réduction de dimension, qui semble primordiale pour garantir une bonne puissance si la décorrélation est utilisée dans le contexte considéré ici.

Références

[1] Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. The American Journal of Human Genetics, 81(6) , 1158–1168.
[2] Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants : a review and some new results. Statistical Science, 29(2), 302–321.
[3] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. The Annals of Statistics, 38(3), 1686–1732.
[4] Kessy, A., Lewin, A., and Strimmer, K. (2017). Optimal whitening and decorrelation. The American Statistician, (just-accepted).

A data-driven ecological interaction patterns discovery: case study of the soil fauna

Sara Si-moussi, PhD Student at INRA, Montpellier SupAgro, INRIA.
Mickael Hedde - UMR 210 Écologie fonctionnelle et biogéochimie des sols et agroécosystèmes, INRA, Montpellier, France.
Esther Galbrun - EPI Orpailleur - Représentation de connaissances et raisonnements, INRIA, Nancy Grand Est, France.
Wilfrid Thuiller - UMR 5553 Evolution, Modeling and Analysis of BIOdiversity, CNRS.

1 Introduction

The Earth is undergoing many changes at an unprecedented pace. A major challenge for ecologists is to understand and predict their consequences on the biological interaction networks. This requires compiling generations of findings and massive knowledge that is disseminated in structurally and semantically heterogeneous formats: scientific papers, images and datasheets. We suggest in our work to employ descriptive data mining techniques to support this research at three levels: knowledge **discovery**, **modeling** and **organization**.

As part of the first tier of this thesis, we propose to study ecological interactions within communities of soil invertebrates. Our goal is to extract distribution and interaction patterns between these species at the population level then upwards to the community scale.

2 Methodology/Results/Conclusions

The novelty of our work lies in the combined mining of biodiversity (species occurrence, biological traits), soil measures and climate data in a systematic approach that allows to identify simultaneously distribution and interaction patterns as well as the contextual/environmental factors that shape them. To this end, we use association rules and redescription mining along with graph modeling and analysis algorithms. We implement our solution using Python programming language, particularly the graphtool and Scikit-learn libraries.

We conducted a first experiment on Earthworms distribution using the 'Lombriciens de France' dataset (Bouché, 1972). We studied in particular co-occurrence patterns on a set of 85 species of earthworms in 1500+ land parcels in France. We managed to extract 73 co-occurrence associations with a meaningful support size within few seconds. These findings are currently under expert analysis, in the meanwhile we are introducing biological traits from BETSI[1] and environmental descriptors (soil and climate measures) to explain the discovered associations. This should adduce more interesting insights.

Therewith descriptive data mining methods proved to be effective to uncover hidden knowledge in massive data. This study, though primitive and preliminary, is a building block of our research that will hopefully expand to scientific papers and images yet in the process of ecological knowledge discovery.

[1] BETSI : Biological and Ecological Traits of Soil Invertebrates database

Estimation of exposure-attributable fractions from time series

Felix Cheysson, Marie-Anne Vibet, Didier Guillemot, Laurence Watier.
Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), Inserm, UVSQ, Institut Pasteur, Université Paris-Saclay, Paris, France.

1 Introduction

Burden analysis in public health often involves the estimation of exposure-attributable fractions defined as the proportional reduction in the outcome that would occur if the exposure was removed from observed time series. When the entire population is exposed, the association between the exposure and outcome must be carefully modelled before the attributable fractions can be estimated.

2 Methods

Our work establishes asymptotic convergences of the estimates of attributable fractions for commonly used time series models (ARMAX, Poisson, negative binomial and Serfling), using for the most part the Delta method. For the Poisson regression, we detail an innovative approach for which we estimate not the expected baseline, but the conditional expectation and prediction of the baseline if the exposure was removed. This is achieved by a Monte Carlo algorithm, which takes into account both the estimation error and the prediction error.

3 Results

We then carried out a simulation study which compared these estimates in the case of an epidemic exposure and highlighted the importance of thorough analysis of the data: when the outcome was generated under an additive model, the additive models were satisfactory and the multiplicative models were poor, and vice versa, with the Serfling model performing poorly in all cases.

We illustrate the asymptotic convergences with an application to the fraction of French outpatient antibiotic use attributable to influenza between 2003 and 2010.

4 Conclusion

This work suggests that the Serfling model should be avoided when estimating attributable fractions while the model of choice should be selected after careful investigation of the association between the exposure and outcome. The asymptotic distributions are convenient tools for the statistician interested in burden analysis.

Dynamic modeling of latent processes and their causal relationships: application to Alzheimer's disease

Bachirou O. Taddé, Hélène Jaqmin-Gadda, Daniel Commenges, Jean Françcois Dartigues, Cécile Proust-Lima.
INSERM UMR 1219, Univ. Bordeaux, ISPED, F-33000 Bordeaux, France. e-mail: oladedji-bachirou.tadde@u-bordeaux.fr

1 Introduction

In some chronic diseases such as Alzheimer's disease (AD), several stochastic processes are involved in the disease progression and one challenge is to investigate the temporal relationships between these processes in a causal way. Alzheimer's disease is a neurodegenerative disease that gradually affects several dimensions including cerebral anatomy with brain atrophies, cognitive functioning with a decline in various functions and functional dependency with impairments in the daily living activities. These dimensions are interrelated and evolve over time with important heterogeneity among individuals which can lead to consider them as stochastic processes. Recently, some schemes have been proposed to describe the hypothetical sequence of impairments in AD; they highlighted the dynamic, multidimensional and interconnection aspects of AD. However, because of their complexity, they have not been translated yet into satisfying statistical models. We propose a new dynamic model that simultaneously models the multiple dimensions involved in AD and their causal relationships using a latent process approach.

2 Method

Our model defines dimensions as latent processes and combines a multivariate linear mixed model and a system of difference equations to model trajectories of the latent processes and their causal relationships in finely discrete time. The linear mixed model describes the intrinsic trajectory of each process whereas the system of difference equations models the causal links between processes. We distinguish the structural model for the latent processes and their causal relationships from the observation model which links the repeated markers (Gaussian or non-Gaussian) to their underlying latent process.

Parameters are estimated in the maximum likelihood framework enjoying a closed form for the likelihood. The estimation procedure and the impact of the time discretization on the causal interpretations are evaluated in simulations.

3 Result

The model is applied to the data of the Alzheimer's Disease Neuroimaging Initiative (ADNI). We considered three dimensions (cerebral atrophy, cognition and functional dependency) and contrasted their causal structure at three preclinical stages of AD defined at inclusion: healthy elderly subjects, elderly subjects with a mild cognitive impairment and subjects diagnosed with AD. We found that cerebral atrophy impacted cognitive and functional abilities at each stage and that cognitive and functional abilities were inter-related. In addition, with the development of the disease, we found some interesting reverse impact of cognitive ability on the cerebral anatomy.

4 Conclusion

This application gave a first insight on interconnections between dimensions and their trajectories at different preclinical stages of AD. More generally, the statistical model we proposed provides a promising statistical tool to better understand the natural history of diseases and their underlying mechanisms, which is essential for the development of novel treatments. Keywords: causality, mixed models, difference equations, latent process, longitudinal data

Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment

Loïc Ferrer, loic.ferrer@curie.fr, Institut Curie U900, Saint-Cloud, France,
Hein Putter, Leiden University Medical Center, Leiden, the Netherlands,
and Cécile Proust-Lima, INSERM UMR 1219, Univ. Bordeaux, Bordeaux, France

Abstract. After the diagnosis of a disease, one major objective is to predict cumulative probabilities of events such as clinical relapse or death from the individual information collected up to a prediction time, including usually biomarker repeated measurements. Several competing estimators have been proposed to calculate these individual dynamic predictions, mainly from two approaches: joint modelling and landmarking. These approaches differ by the information used, the model assumptions and the complexity of the computational procedures. It is essential to properly validate the estimators derived from joint models and landmark models, quantify their variability and compare them in order to provide key elements for the development and use of individual dynamic predictions in clinical follow-up of patients. Motivated by the prediction of two competing causes of progression of prostate cancer from the history of prostate-specific antigen, we conducted an in-depth simulation study to validate and compare the dynamic predictions derived from these two methods. Specifically, we formally defined the quantity to estimate and its estimators, proposed techniques to assess the uncertainty around predictions and validated them. We also compared the individual dynamic predictions derived from joint models and landmark models in terms of prediction error, discriminatory power, efficiency and robustness to model assumptions. We show that these prediction tools should be handled with care, in particular by properly specifying models and estimators.

Semiparametric approach for covariate-specific time dependent ROC curves for correlated survival data

Alessandra Meddis, Paul Blanche, Aurélien Latouche

1 Introduction

Considerable research has focused on the development of new biomarkers. The first step in developing a clinically useful biomarker is to identify its ability in discriminating patients at high risk of dying within the next t-years (e.g. 5-years) from those who will not. The standard methodology to quantify the discrimination performance of a biomarker, with right censored data, is to estimate time dependent ROC curves, ROC(t). In presence of clustered failure times, the common strategy is to ignore heterogeneity in the phase of evaluation of the performance of a candidate biomarker, but to confirm its discriminatory capacity, it is important to account for heterogeneity while adjusting for clinical covariates. The usefulness of our approach is illustrated on our motivating example, which consists in the first meta-analysis on individual data of more than 2000 patients from 15 centers with non metastatic breast cancer. Its objective was to quantify the clinical usefulness of circulating tumor cells (CTCs) count as a prognostic marker of survival.

2 Methodology

ROC(t) allows to study the capacity of a biomarker Y to discriminate between patients who experience event prior time t (cumulative cases) from those who do not up to time t (dynamic controls). The current methodology does not account for heterogeneity while estimating ROC(t). In this work, we fill this gap by proposing an extension to clustered data of the Song & Zhou method (Statistica Sinica, 2008). To estimate the covariate-specific time dependent ROC curve we consider a joint model: (i) shared frailty model which links the covariates and the biomarker to the time-to-event, (ii) location scale model to link the covariates to the biomarker. We evaluate the performance of the proposed method in a simulation study. We demonstrate an application of the estimator to data derived from a meta-analysis on individual patient data with non metastatic breast cancer where the goal is to understand the clinical usefulness of CTCs count for this scenario. In particular, we estimate the covariate-specific ROC curves that quantify the discrimination performance of CTCs count within subgroups of patients having the same tumor stage at time of diagnosis, since subjects with inflammatory tumor show a higher number of CTCs and a poorer prognosis. A bootstrap method is proposed for calculating confidence intervals.

3 Results

The estimator is computationally simple and the simulation results highlighted the robustness of the method at varying of censoring with negligible bias ($\approx 10^{-3}$). Moreover, we provide the results for the motivating example with the time dependent ROC curves and respective AUCs for different tumor stage. The wide confidence intervals highlighted that having inflammatory tumor does not influence the discrimination of the CTCs count.

4 Conclusions

In presence of clustered failure times it is important to take into account heterogeneity. In fact, the introduction of a random effect (frailty) is needed to estimate the performance of the biomarker in the general population. In this scenario, the covariate-specific time dependent ROC curve can be easily estimated with the proposed approach.

Les données massives hospitalières pour la surveillance des syndromes grippaux en temps réel

Canelle Poirier[1,2,*], Audrey Lavenu[3,4], Valérie Bertaud[1,2], Boris Campillo-Gimenez[1,5], Emmanuel Chazard[6], Marc Cuggia[1,2,7] and Guillaume Bouzillé[1,2,7]

[1] INSERM, U1099, Rennes, F-35000, France

[2] Université de Rennes 1, LTSI, Rennes, F-35000, France

[3] INSERM CIC 1414, Université de Rennes 1, Rennes, F-35000, France

[4] Université de Rennes 1, Rennes, F-35000, France

[5] Comprehensive Cancer Regional Center, Eugene Marquis, Rennes, F-35000, France

[6] Département de Santé Publique, Université de Lille EA 2694, CHU Lille, F-59000 Lille, France

[7] CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France

[*] Laboratoire D'informatique Médicale, 2 Rue Henri Le Guilloux, F-35033 Rennes, France, canelle.poirier@outlook.fr, 02.99.28.42.15

1 Introduction

La grippe est un enjeu de santé publique majeur, responsable de 5 millions de cas graves chaque année dans le monde, elle perturbe les systèmes de santé au moment des épidémies. Afin de réduire son impact, les réseaux de surveillance traditionnels produisent des estimations hebdomadaires des taux d'incidence mais avec un délai de une à trois semaines. Plusieurs alternatives ont alors été proposées, notamment basées sur les données massives et en particulier les données du web. Cependant, avec l'adoption du dossier patient électronique, les hôpitaux sont également devenus des producteurs de gros volumes de données.

L'objectif de ce travail est d' évaluer l'apport des données du web et des données massives hospitali ères ainsi que diff érents modèles statistiques pour la surveillance des épidémies de grippe en temps réel.

2 Matériel et Méthodes

Afin d'évaluer les données du web, nous avons pu récupérer grâce au service Google Correlate, les 100 requêtes de Google les plus corrélées à notre signal épidémique. Pour les données hospitali ères, nous avons utilis é l'entrepôt de données eHOP, incluant tous les dossiers patients électroniques du CHU de Rennes. Nous avons alors effectué une trentaine de requêtes sur cette base de données hospitalières. Nous avons comparé trois modèles statistiques : les forêts aléatoires, Elastic Net et les machines à vecteurs supports (SVM). Cette comparaison a été faite à l'aide de différents indicateurs dont le coefficient de corrélation de Pearson et l'erreur quadratique moyenne.

3 Résultats

Au niveau national, le meilleur coefficient de corrélation de Pearson (PCC), obtenu entre nos prévisions et les estimations effectuées par le r éseau traditionnel, est égal à 0.98 avec une erreur quadratique moyenne (MSE) égale à 866. Ce résultat a été obtenu grâce aux données hospitalières et au modèle SVM. Pour les données du web, le meilleur PCC est égal à 0.97 et le meilleur MSE est égal à 968, obtenu également avec le modèle SVM. Au niveau régional (Bretagne), le meilleur PCC est égal à 0.923 et le MSE égal à 2364, obtenu également grâce aux données hospitalières et au modèle SVM. Pour les données du web, le meilleur PCC est égal à 0.912 et le MSE est égal à 2736 obtenu grâce à un modèle de forêts aléatoires.

4 Conclusion

Nous avons montré que les données hospitalières combinées aux données historiques des r éseaux traditionnels de surveillance, permettaient d'estimer efficacement en temps réel les taux d'incidence des syndromes grippaux. Que ce soit au niveau national ou au niveau régional, et quelque soit le modèle statistique utilisé, les résultats sont meilleurs avec les données hospitali ères qu'avec les données du web. Concernant les modèles statistiques, le modèle SVM est celui qui nous permet d'obtenir les meilleurs résultats.

Adaptive protocols based on predictions from a mechanistic model of the effect of IL7 on CD4 counts

Laura Villain, Daniel Commenges, Mélanie Prague, Chloé Pasin, Rodolphe Thiébaut

1 Introduction

In HIV infected patients, antiretroviral therapy suppresses the viral replication which is followed in most patients by a restoration of the CD4+T cells pool. For patients who fail to do so, repeated injections of exogenous IL-7 is considered, as IL-7 is a cytokine involved in the T cell homeostasis. The INSPIRE 2 and 3 evaluates a first cycle of IL7 injections followed by a new cycle each time the patient is under 550 CD4 cells/μL, with a visit every 3 months. Restauration of the CD4 levels has been demonstrated by the phase I/II INSPIRE clinical trials, but as patients have a different CD4 dynamics, this fixed criterion is not always appropriate and the long-term best adaptive protocol is yet to determine.

2 Method

Ordinary Differential Equation models of the evolution of CD4 after IL7 injections, which include random effects, have been developed (Thiebaut et al., 2014, Plos Comp. Biol). Based on this model, we use a Bayesian approach to forecast the dynamic of CD4 of new patients. We propose realistic protocols which reduce the time spent under the limit of 500 CD4 cells/μL and limit the number of IL7 injections. Using the estimation made with the INSPIRE study as a prior, a Metropolis within Gibbs is used to estimate the posterior of the random effect of a new patient, in order to predict the evolution of their levels of CD4. Two approaches are compared: adapting the criterion for a new cycle based on the risk of falling under 500 cells/μL before the next visit, and adapting the times of control visits. A total of 150 pseudo-patients are simulated to compare the proposed adaptive protocols to the original INSPIRE protocol.

3 Results

We show that our model has a good predictive ability on real data. On the simulated pseudo-patients, we show that our protocols significantly reduce the time spent under 500 CD4 over a period of two years, without increasing the number of injections. Such protocols have the potentiality of increasing the efficiency of this therapy.

Propensity score-based approaches in high dimension for pharmacovigilance signal detection: an empirical comparison on the French spontaneous reporting database

Emeline Courtois[(1,*)], Antoine Pariente[(2)], Francisco Salvo[(2)], Etienne Volatier[(1)], Pascale Tubert-Bitter[(1,**)] and Ismaïl Ahmed[(1,**)]

[(1)] Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases (B2PHI), INSERM, UVSQ (Université Paris-Saclay), Institut Pasteur, Villejuif, France

[(2)] Bordeaux Population Health Research Center, Pharmacoepidemiology team (UMR 1219), INSERM, University of Bordeaux, Bordeaux, France

[(**)] Joint last authors

1 Introduction

Classical methods used for signal detection in pharmacovigilance rely on disproportionality analysis of counts aggregating spontaneous reports of a given adverse drug reaction. In recent years, alternative methods have been proposed to analyze individual spontaneous reports such as penalized multiple logistic regression approaches. These approaches address some well-known biases resulting from disproportionality methods. However, while penalization accounts for computational constraints due to high-dimensional data, it raises the issue of determining the regularization parameter and eventually that of an error-controlling decision rule.

We present a new automated signal detection strategy for pharmacovigilance systems, based on propensity scores (PS) in high dimension. PSs are increasingly used to assess a given association with high-dimensional observational healthcare databases in accounting for confusion bias. Our main aim was to develop a method having the same advantages as multiple regression approaches in dealing with bias, while relying on the statistical multiple comparison framework as regards decision thresholds, by considering false discovery rate (FDR)-based decision rules.

2 Methods

We investigate four PS estimation methods in high dimension: a gradient tree boosting algorithm from machine-learning and three variable selection algorithms. For each (drug, adverse event) pair, the PS is then applied as adjustment covariate or by using two kinds of weighting: inverse proportional treatment weighting and matching weights. The different versions of the new approach were compared to a univariate approach, which can be assimilated to a disproportionality method, and to two penalized multiple logistic regression approaches, directly applied on spontaneous reporting data.

3 Material

Performance was assessed through an empirical comparative study conducted in the French national pharmacovigilance database (2000-2016). It was based on a large drug reference set pertaining to drug-induced liver injury adverse reaction.

4 Results

Multiple regression approaches performed better in detecting true positives and false positives. Nonetheless, the performances of the PS-based methods using matching weights was very similar to that of multiple regression and better than with the univariate approach. Versions of the PS-based approach show a similar behaviour within the same PS strategy (adjustment, weightings), regardless of the PS estimation method.

5 Discussion

The results suggest that the proposed PS-based methodology which relies on matching weights is an interesting complement to other existing methods. In a sense, it combines the main strengths of both univariate and multiple regression approaches: it makes it possible to account for co-reported drugs while using multiple hypothesis testing theory as regards the detection threshold.

Développement de modèles de survie pronostiques de grande dimension en considérant des groupes de biomarqueurs : application en génétique humaine

Shaima Bel Hechmi[*,1,2], Riccardo De Bin[3], Anne-Laure Boulesteix[4], Stefan Michiels[1,2] & Federico Rotolo[1,2]

[1] Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM U1018 Oncostat, F-94805, Villejuif, France.

[2] Service de Biostatistique et d'Epidémiologie, Institut Gustave Roussy, F-94805, 114 Rue Edouard Vaillant, 94800 Villejuif, France.

[3] Department of Mathematics, University of Oslo, Oslo, Norway.

[4] Biometry and Epidemiology, Institute for Medical Information Processing, University of Munich, D-81377, Munich, Germany.

[*] Shayma.BEL-HECHMI@gustaveroussy.fr

Résumé. La médecine stratifiée ou de précision vise à sélectionner, en fonction d'un ensemble de biomarqueurs, les patients susceptibles de bénéficier d'un traitement. En oncologie, le critère de jugement final est souvent un critère de survie et le modèle de Cox est fréquemment utilisé pour évaluer l'efficacité des traitements dans les essais de phase III. La pénalisation lasso et ses extensions sont couramment utilisées pour sélectionner efficacement les biomarqueurs dans le contexte de données de grande dimension [1]. Bien que la plupart de ces méthodes de sélection considèrent un ensemble homogène de biomarqueurs, les données génomiques peuvent être regroupées en fonction de leur voie biologique ou de leur nature.

Nous présentons différentes pénalisations pour le modèle de Cox avec des biomarqueurs groupés par voie afin de favoriser la sélection de biomarqueurs qui, en plus d'avoir un effet individuel important, appartiennent à un groupe ayant un fort effet global. Nous considérons le cas de groupes pré-spécifiés et disjoints. Nous nous intéressons à deux familles de méthodes, l'une basée sur le Sparse-Group lasso (SG) [2] et l'autre sur le lasso adaptatif ou le lasso intégratif avec des facteurs de pénalisation [3]. Dans le sparse-group lasso, nous étudions différents poids pour la pénalisation de biomarqueurs individuels par rapport à la pénalisation des groupes. Pour l'approche du lasso adaptatif, nous considérons différentes stratégies de pondération comportant l'inverse de la moyenne par groupe des coefficients univariés, l'inverse du coefficient univarié de la première composante principale (supervisée) de chaque groupe, éventuellement après une sélection préliminaire avec le lasso groupé. Pour toutes ces méthodes, le paramètre de pénalisation est choisi par validation croisée.

Nous illustrons ces méthodes en utilisant les données d'expression génétique de 614 patientes atteints de cancer du sein traités avec une chimiothérapie adjuvante.

Mots-clés. Médecine stratifiée, données de grande dimension, régression pénalisée, biomarqueurs pronostiques, données génomiques, voies biologiques.

Bibliographie

[1] Hastie, Tibshirani, Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. 2015 Chapman and Hall/CRC.
[2] Simon, Friedman, Hastie, Tibshirani, A sparse-group lasso. J Comp Graph Stat. 2013.
[3] Boulesteix, De Bin, Jiang, Fuchs. IPF-LASSO: Integrative -Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. Comp Math Meth Med. 2017.

Liste des participants

Prénom	Nom	Email
Sophie	Ancelet	sophie.ancelet@irsn.fr
Vincent	Audigier	vincent.audigier@cnam.fr
Nadim	Ballout	nadim-ballout-r@hotmail.com
Shayma	Bel-Hechmi	shayma.bel-hechmi@gustaveroussy.fr
David	Causeur	david.causeur@agrocampus-ouest.fr
Félix	Cheysson	felix.cheysson@pasteur.fr
Daniel	Commenges	d.commenges@gmail.com
Emeline	Courtois	emeline.courtois@inserm.fr
Damien	Drubay	damien.drubay@gustaveroussy.fr
Robert	Faivre	robert.faivre@inra.fr
Loïc	Ferrer	loic.ferrer@u-bordeaux.fr
Florian	Hebert	florian.hebert@agrocampus-ouest.fr
Mounia N	Hocine	nacima.hocine@lecnam.net
Hélène	Jacqmin-Gadda	Helene.Jacqmin-Gadda@bordeaux.inserm.fr
Agnieszka	Krol	agnieszka.listwon@u-bordeaux.fr
Alicia	Larive	larive.alicia@gmail.com
Alessandra	Meddis	alessandra.meddis@curie.fr
Geert	Molenberghs	geert.molenberghs@uhasselt.be
Canelle	Poirier	canelle.poirier@univ-rennes1.fr
Cécile	Proust-Lima	cecile.proust-lima@u-bordeaux.fr
Sarah-Laure	Rincourt	sarah-laure.rincourt@gustaveroussy.fr
Pascale	Tubert-Bitter	pascale.tubert@inserm.fr
Sara	Si-Moussi	cs_si-moussi@esi.dz
Bachirou	Tadde	oladedji-bachirou.tadde@u-bordeaux.fr
Laura	Villain	laura.villain@u-bordeaux.fr
Pascal	Wild	Pascal.wild@inrs.fr