

Design d'un algorithme d'IA en grande dimension pour prédire la réadmission à l'hôpital

Simon Bussy¹

simon.bussy@gmail.com

www.simonbussy.com



Étude réalisée avec

Raphaël Veil^{2,3}, Vincent Looten^{2,3}, Anita Burgun^{2,3}, Stéphane Gaïffas^{1,4},
Agathe Guilloux⁵, Brigitte Ranque^{6,7}, Anne-Sophie Jannot^{2,3}

¹ LPSM, UMR 8001, Sorbonne University, Paris

² APHP, Département d'Informatique Biomédicale et de Santé Publique, HEGP, Paris

³ INSERM, UMRS 1138, Eq22, Centre de Recherche des Cordeliers, Univ. Paris Descartes, Paris

⁴ CMAP, UMR 7641, École Polytechnique CNRS, Paris

⁵ LAMME, Univ. Evry, CNRS, Université Paris-Saclay, Paris

⁶ INSERM, UMRS 970, Univ. Paris Descartes, Paris

⁷ APHP, Département de Médecine Interne, HEGP, Paris

IA&Santé 2018

Le 03/07/2018

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Outline

IA&Santé 2018

2/10

Simon Bussy

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

- ▶ Prédiction du pronostic, médecine personnalisée

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes
 - ▶ Élaborer un algorithme qui apprend et généralise

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes
 - ▶ Élaborer un algorithme qui apprend et généralise
 - ▶ Interprétabilité des résultats!

Introduction

IA en santé

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes
 - ▶ Élaborer un algorithme qui apprend et généralise
 - ▶ Interprétabilité des résultats!

Étude de cas

- ▶ Réadmission à l'hôpital après une crise vaso-occlusive (CVO) chez des patients atteints de drépanocytose

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Introduction

IA en santé

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes
 - ▶ Élaborer un algorithme qui apprend et généralise
 - ▶ Interprétabilité des résultats!

Étude de cas

- ▶ Réadmission à l'hôpital après une crise vaso-occlusive (CVO) chez des patients atteints de drépanocytose
- ▶ But: Identifier

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Introduction

IA en santé

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes
 - ▶ Élaborer un algorithme qui apprend et généralise
 - ▶ Interprétabilité des résultats!

Étude de cas

- ▶ Réadmission à l'hôpital après une crise vaso-occlusive (CVO) chez des patients atteints de drépanocytose
- ▶ But: Identifier
 - ▶ les patients à haut risque de réadmission précoce (sous 30 jours)

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

IA en santé

- ▶ Prédiction du pronostic, médecine personnalisée
- ▶ Données massives: dossiers médicaux électroniques (EHR)
- ▶ Enjeux:
 - ▶ Modéliser ces données complexes
 - ▶ Élaborer un algorithme qui apprend et généralise
 - ▶ Interprétabilité des résultats!

Étude de cas

- ▶ Réadmission à l'hôpital après une crise vaso-occlusive (CVO) chez des patients atteints de drépanocytose
- ▶ But: Identifier
 - ▶ les patients à haut risque de réadmission précoce (sous 30 jours)
 - ▶ les facteurs prédictifs

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

- ▶ Cohorte rétrospective monocentrique sur 286 patients de HEGP

- ▶ Cohorte rétrospective monocentrique sur 286 patients de HEGP
- ▶ Entrepôt de données de santé (EDS), architecture I2B2

données de routines, numéros déidentifiés des séjours et des patients, données socio-démographiques, différents comptes-rendus textuels, diagnostics (CIM-10), procédures (CCAM), résultats biologiques (LOINC), données de prescriptions (CPOE)

- ▶ Cohorte rétrospective monocentrique sur 286 patients de HEGP
- ▶ Entrepôt de données de santé (EDS), architecture I2B2
- ▶ Critère d'inclusion: patients admis pour CVO entre le 01/01/2010 et le 31/12/2015

- ▶ Cohorte rétrospective monocentrique sur 286 patients de HEGP
- ▶ Entrepôt de données de santé (EDS), architecture I2B2
- ▶ Critère d'inclusion: patients admis pour CVO entre le 01/01/2010 et le 31/12/2015
- ▶ Sélection aléatoire d'un séjour par patient (*i.i.d.*)

Extraction

- ▶ Multiples variables hétérogènes
 - données démographiques (date de naissance, sexe, etc.)
 - horodatage de l'admission aux urgences, de l'hospitalisation et de la sortie du patient
 - horodatage de la sortie du précédent séjour pour CVO
 - horodatage de la prochaine admission aux urgences après le séjour considéré
 - résultats biologiques des bilans prélevés depuis l'admission aux urgences
 - paramètres vitaux et oxygénothérapie, relevés depuis l'admission aux urgences
 - prescriptions d'opiacés (molécule, galénique, horodatage du début et de la fin de l'administration du produit, etc.)
 - hémoglobine de base et antécédents relatifs à la maladie drépanocytaire à partir des comptes-rendus médicaux

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

- Multiples variables hétérogènes
- Outil FASTVISU

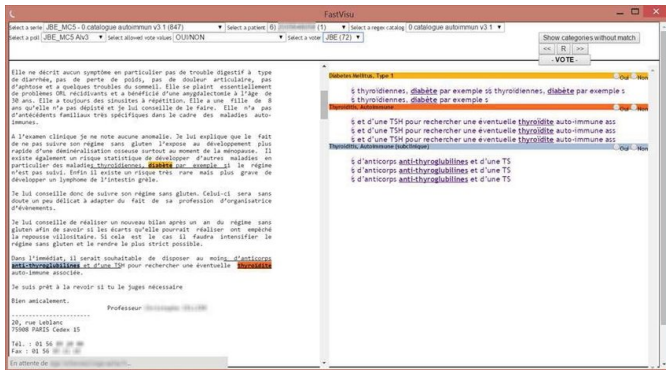


Figure 1: Visualisation de l'outil FASTVISU

Covariables

Extraction

- ▶ Multiples variables hétérogènes
- ▶ Outil FASTVISU

Création

- ▶ Nouvelles covariables
 - binarisation des covariables catégorielles
 - âge au moment de l'hospitalisation
 - durée de chaque séjour
 - présence d'une hospitalisation pour CVO dans les 18 derniers mois
 - écart entre l'hémoglobine mesurée et l'hémoglobine de base
 - délai entre l'arrêt des traitements par opiacés et la sortie
 - délai entre l'arrêt de l'oxygénothérapie et la sortie

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Extraction

- ▶ Multiples variables hétérogènes
- ▶ Outil FASTVISU

Création

- ▶ Nouvelles covariables
- ▶ Variables longitudinales: demandent une modélisation adaptée (problèmes de durées et d'alignement)

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Covariables

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
 1. grille temporelle

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Covariables

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
 1. grille temporelle
 2. nuage de point global

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Covariables

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
 1. grille temporelle
 2. nuage de point global
 3. trajectoire moyenne avec intervalle de confiance:

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Covariables

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
 1. grille temporelle
 2. nuage de point global
 3. trajectoire moyenne avec intervalle de confiance:
 - ▶ spline ajusté pour chaque trajectoire individuelle (représentation fonctionnelle)

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Covariables

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 3\sigma$)
 1. grille temporelle
 2. nuage de point global
 3. trajectoire moyenne avec intervalle de confiance:
 - ▶ spline ajusté pour chaque trajectoire individuelle (représentation fonctionnelle)
 - ▶ calcul des valeurs prises sur la grille

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Covariables

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 3\sigma$)
 1. grille temporelle
 2. nuage de point global
 3. trajectoire moyenne avec intervalle de confiance:
 - ▶ spline ajusté pour chaque trajectoire individuelle (représentation fonctionnelle)
 - ▶ calcul des valeurs prises sur la grille
 - ▶ matrice pour chaque variable: 1 ligne par patient

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Covariables

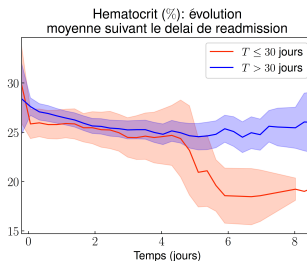
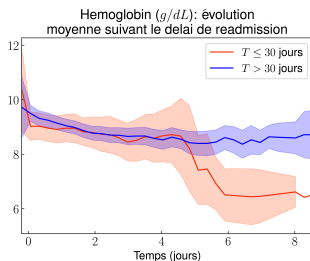
Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 3\sigma$)
 1. grille temporelle
 2. nuage de point global
 3. trajectoire moyenne avec intervalle de confiance:
 - ▶ spline ajusté pour chaque trajectoire individuelle (représentation fonctionnelle)
 - ▶ calcul des valeurs prises sur la grille
 - ▶ matrice pour chaque variable: 1 ligne par patient
 - ▶ calcul de la valeur moyenne avec IC gaussien à 95% pour chaque temps de la grille (colonne de la matrice)

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
- ▶ Résultats: <https://github.com/SimonBussy/redcvo>

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
- ▶ Résultats: <https://github.com/SimonBussy/redcvo>

Création de covariables dérivées

- ▶ Dernière valeur disponible avant la sortie

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
- ▶ Résultats: <https://github.com/SimonBussy/redcvo>

Création de covariables dérivées

- ▶ Dernière valeur disponible avant la sortie
- ▶ Pente de la régression linéaire ajustée aux données

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
- ▶ Résultats: <https://github.com/SimonBussy/redcvo>

Création de covariables dérivées

- ▶ Dernière valeur disponible avant la sortie
- ▶ Pente de la régression linéaire ajustée aux données
- ▶ Paramètres d'un processus gaussien (GP) ajusté aux données

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Visualisation des variables longitudinales

- ▶ Trajectoires moyennes stratifiées ($\pm 30j$)
- ▶ Résultats: <https://github.com/SimonBussy/redcvo>

Création de covariables dérivées

- ▶ Dernière valeur disponible avant la sortie
- ▶ Pente de la régression linéaire ajustée aux données
- ▶ Paramètres d'un processus gaussien (GP) ajusté aux données

174 covariables finales

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Modèles

- ▶ Cadre de prédiction binaire (30 jours)

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)
 - ▶ Gradient Boosting (GB)

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)
 - ▶ Gradient Boosting (GB)
 - ▶ Réseaux de neurones (NN)

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)
 - ▶ Gradient Boosting (GB)
 - ▶ Réseaux de neurones (NN)
- ▶ Cadre d'analyse de survie

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)
 - ▶ Gradient Boosting (GB)
 - ▶ Réseaux de neurones (NN)
- ▶ Cadre d'analyse de survie
 - ▶ Cox PH

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)
 - ▶ Gradient Boosting (GB)
 - ▶ Réseaux de neurones (NN)
- ▶ Cadre d'analyse de survie
 - ▶ Cox PH
 - ▶ CURE

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Modèles

- ▶ Cadre de prédiction binaire (30 jours)
 - ▶ Régression Logistique (LR)
 - ▶ SVM avec noyau linéaire
 - ▶ Forêts Aléatoires (RF)
 - ▶ Gradient Boosting (GB)
 - ▶ Réseaux de neurones (NN)
- ▶ Cadre d'analyse de survie
 - ▶ Cox PH
 - ▶ CURE
 - ▶ C-mix

[Introduction](#)[Les données](#)[Cohorte](#)[Covariables](#)[Algorithme](#)[Modèles](#)[Évaluation](#)[Résultats](#)[Conclusion](#)

Régularisation

- ▶ Elastic-Net: sélection de variables

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Régularisation

- ▶ Elastic-Net: sélection de variables
- ▶ Jeu d'apprentissage (70%) / jeu de validation (30%)

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Régularisation

- ▶ Elastic-Net: sélection de variables
- ▶ Jeu d'apprentissage (70%) / jeu de validation (30%)
- ▶ Cross-validation des hyper-paramètres avec recherche aléatoire

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

Régularisation

- ▶ Elastic-Net: sélection de variables
- ▶ Jeu d'apprentissage (70%) / jeu de validation (30%)
- ▶ Cross-validation des hyper-paramètres avec recherche aléatoire

Métrique

- ▶ Modèles de survie: prédictions binaires à 30 jours à partir de la fonction de survie estimée

Régularisation

- ▶ Elastic-Net: sélection de variables
- ▶ Jeu d'apprentissage (70%) / jeu de validation (30%)
- ▶ Cross-validation des hyper-paramètres avec recherche aléatoire

Métrique

- ▶ Modèles de survie: prédictions binaires à 30 jours à partir de la fonction de survie estimée
- ▶ Score AUC pour évaluer les prédictions

Table 1: Comparaison des performances de prédiction (AUC)

Method	SVM	GB	LR	NN	RF	CURE	Cox	C-mix
AUC	0.524	0.561	0.616	0.707	0.738	0.831	0.855	0.940

Table 1: Comparaison des performances de prédiction (AUC)

Method	SVM	GB	LR	NN	RF	CURE	Cox	C-mix
AUC	0.524	0.561	0.616	0.707	0.738	0.831	0.855	0.940

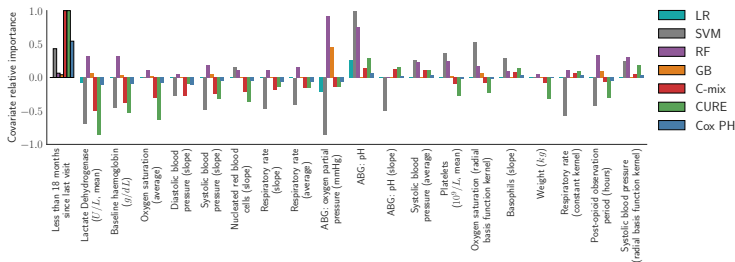


Figure 1: Comparaison de l'importance des 20 premières covariables

Introduction

Les données

Cohorte

Covariables

Algorithme

Modèles

Évaluation

Résultats

Conclusion

- Construire de nombreuses covariables sans *a priori*

- ▶ Construire de nombreuses covariables sans *a priori*
- ▶ Modéliser l'information longitudinale (9/20)

- ▶ Construire de nombreuses covariables sans *a priori*
- ▶ Modéliser l'information longitudinale (9/20)
- ▶ Considérer un grand nombre d'algorithmes sans *a priori*

- ▶ Construire de nombreuses covariables sans *a priori*
- ▶ Modéliser l'information longitudinale (9/20)
- ▶ Considérer un grand nombre d'algorithmes sans *a priori*
- ▶ Les variables sélectionnées sont cliniquement pertinentes

- ▶ Construire de nombreuses covariables sans *a priori*
- ▶ Modéliser l'information longitudinale (9/20)
- ▶ Considérer un grand nombre d'algorithmes sans *a priori*
- ▶ Les variables sélectionnées sont cliniquement pertinentes
- ▶ Différences de sélection: hypothèses sous-jacentes!
nécessité d'utiliser plusieurs méthodes pour avoir une interprétation des covariables

- ▶ Construire de nombreuses covariables sans *a priori*
- ▶ Modéliser l'information longitudinale (9/20)
- ▶ Considérer un grand nombre d'algorithmes sans *a priori*
- ▶ Les variables sélectionnées sont cliniquement pertinentes
- ▶ Différences de sélection: hypothèses sous-jacentes!
nécessité d'utiliser plusieurs méthodes pour avoir une interprétation des covariables
- ▶ Aggrégation d'experts